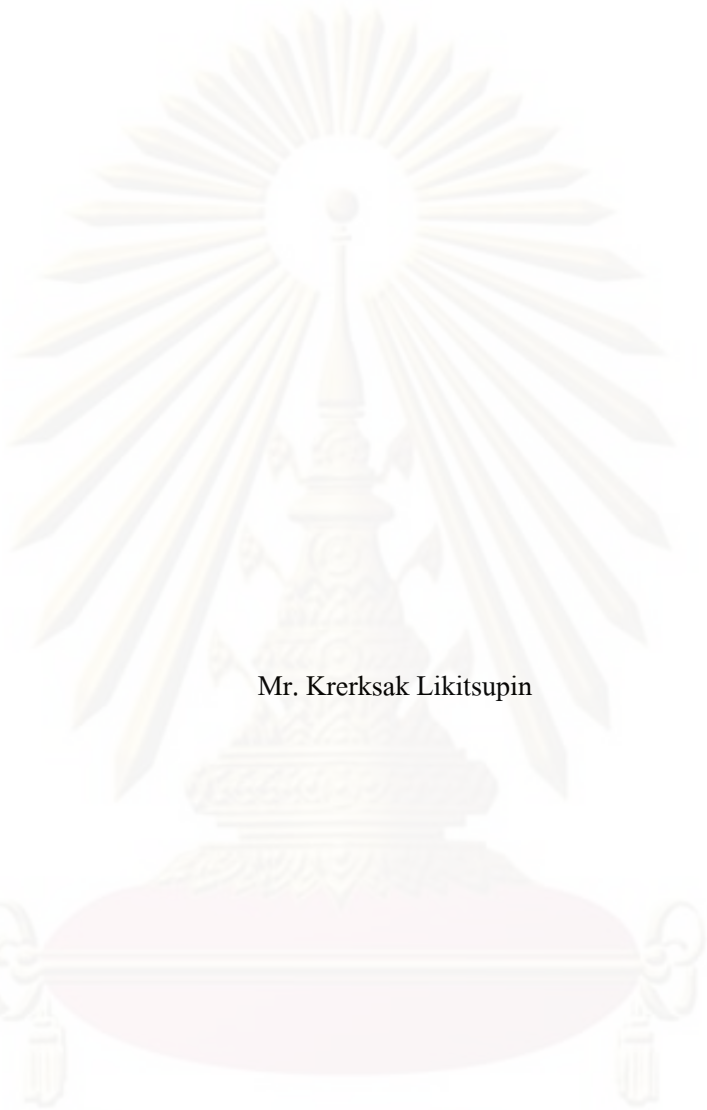
วิธีการทางสวนสัทศาสตร์สำหรับการปรับปรุงการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

นายเกริกศักดิ์ ลิขิตสุภิณ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2552

ACOUSTIC-PHONETIC APPROACHES TO IMPROVING SEGMENT-BASED SPEECH RECOGNITION

Mr. Krerksak Likitsupin



A Dissertation Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy Program in Computer Engineering

Department of Computer Engineering

Chulalongkorn University

Academic Year 2009

หัวข้อวิทยานิพนธ์	วิธีการทางสวนสัทศาสตร์สำหรับการปรับปรุงการรู้จำเสียงพูดแบบอาศัยเซกเมนต์
โดย	นายเกริกศักดิ์ ลิขิตสุภิณ
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.อดิวงศ์ สุชาโต
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ ดร. ชัย วุฒิวิวัฒน์ชัย

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาดุษฎีบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.อดิวงศ์ สุชาโต)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ดร. ชัย วุฒิวิวัฒน์ชัย)

..... กรรมการ
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... กรรมการภายนอกมหาวิทยาลัย
(ดร. อนันต์ลดา โชติมงคล)

เกริกศักดิ์ ลิขิตสุภิณ : วิธีการทางสวนสัทศาสตร์สำหรับการปรับปรุงการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ (ACOUSTIC-PHONETIC APPROACHES TO IMPROVING SEGMENT-BASED SPEECH RECOGNITION). อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ. ดร. อติวงศ์ สุชาโต, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ผศ. ดร.โปรดปราน บุณยพุกกณะ, ดร. ชัย วุฒิวิวัฒน์ชัย, 98 หน้า

ปัจจุบันนี้มีหลายวิธีที่ใช้ในการรู้จำเสียงพูด ซึ่งวิธีที่ได้รับความนิยมมากที่สุดคือวิธีที่มีการดึงเอาเวกเตอร์คุณสมบัติออกจากกรอบเวลาที่แน่นอน เช่น การรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ (HMM) ซึ่งได้มีการพิสูจน์แล้วว่าการที่จะเพิ่มความรู้ด้านสวนสัทศาสตร์ลงไปในการรู้จำเสียงพูดแบบอาศับแบบจำลองฮิดเดนมาร์คอฟนั้นเป็นไปได้ยาก ดังนั้นจึงได้มีการนำเสนอวิธีการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ขึ้นมา ซึ่งวิธีการนี้จะมีการดึงเอาเวกเตอร์คุณสมบัติออกจากเซกเมนต์ที่มีขนาดแตกต่างกันไปแทนที่จะดึงออกจากกรอบเวลาที่มีขนาดเท่าๆกัน วิทยานิพนธ์นี้แสดงให้เห็นว่าการรู้จำเสียงพูดแบบอาศัยเซกเมนต์มีความแม่นยำสูงกว่าการรู้จำเสียงพูดแบบอาศัยกรอบเวลา ในการทดลองรู้จำเสียงพูดภาษาไทยในระดับหน่วยเสียง อย่างไรก็ตาม การรู้จำเสียงพูดแบบอาศัยเซกเมนต์จะมีการค้นหาคำตอบที่อยู่ในกราฟของเซกเมนต์ ดังนั้น ความแม่นยำในการรู้จำเสียงพูดของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์จึงขึ้นอยู่กับคุณภาพของกราฟของเซกเมนต์ หากต้องการเพิ่มความแม่นยำในการรู้จำเสียงพูดของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ จึงต้องมีการปรับปรุงคุณภาพกราฟของเซกเมนต์ โดยเพิ่มจำนวนเซกเมนต์ที่ถูกต้องลงในกราฟของเซกเมนต์ ดังนั้น วิทยานิพนธ์นี้มุ่งเน้นในปรับปรุงคุณภาพกราฟของเซกเมนต์ โดยการแก้ไขความผิดพลาดในกราฟของเซกเมนต์ ซึ่งเกิดจากการที่มีขอบเขตของหน่วยเสียงแทรกมา และเกิดจากการตัดออกของขอบเขตของหน่วยเสียง ที่เกิดจากขั้นตอนการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น นอกจากนี้ เพื่อเพิ่มความแม่นยำของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ วิทยานิพนธ์นี้ยังมีการนำคะแนนที่เกิดจากความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างมาใช้ในขั้นตอนการให้คะแนน และค้นหาคำตอบของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ จากผลการทดลองแสดงให้เห็นว่าการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ผ่านกระบวนการที่นำเสนอในวิทยานิพนธ์นี้สามารถรู้จำเสียงพูดในระดับหน่วยเสียงได้แม่นยำถึง 58.26% ขณะที่การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ไม่มีการผ่านกระบวนการที่นำเสนอ และการรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟมีความแม่นยำในการรู้จำเสียงพูดในระดับหน่วยเสียงน้อยกว่า 50%

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2552 ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

4871801921 : MAJOR COMPUTER ENGINEERING

KEYWORDS : SEGMENT-BASED SPEECH RECOGNITION / SPEECH SEGMENTATION / DISTINCTIVE FEATURE / MANNERS OF ARTICULATION

KRERKSAK LIKITSUPIN : ACOUSTIC-PHONETIC APPROACHES TO IMPROVING SEGMENT-BASED SPEECH RECOGNITION. THESIS ADVISOR : ASST. PROF. ATIWONG SUCHATO, Ph.D., THESIS CO-ADVISOR : ASST. PROF. PROADPRAN PUNYABUKKANA, Ph.D., CHAI WUTIWIWATCHAI, Ph.D., 98 pp.

Today, there are many approaches to automatic speech recognition. However, most of them represent an observation space based on a temporal sequence of measurements extracted from fixed-length frames, such as the Hidden Markov Model (HMM)-based speech recognition. Incorporating acoustic-phonetic knowledge into those HMM-based approaches are proved to be difficult. A segment-based approach, in which acoustic feature vectors represent underlying speech segments instead of speech frames, was introduced. Segment-based approaches have been shown as competitive alternatives to HMM-based techniques. In this dissertation, we show that using a segment-based approach can yield better accuracies than the HMM-based ones in Thai phoneme recognition tasks. Still, its accuracies rely heavily on the quality of the segment graph from which the recognizer searches for the most likely recognition hypotheses. In order to increase the inclusion rate of actual segments in the graph, we recover possible missing segments due to boundary insertion and deletion errors based on acoustic discontinuities together with manner distinctive features from segment graphs provided by a typical frame-based segmentation. Scores based on how likely a segment belongs to some phoneme broad classes are also incorporated to the probabilistic framework used for scoring segments. The best phoneme recognition accuracy achieved is 58.26%, while they are less than 50% for the baseline HMM-based and the traditional segment-based recognizers.

Department: Computer Engineering Student's Signature Krerksak Likitsupin

Field of Study: Computer Engineering Advisor's Signature

Academic Year: 2009 Co-Advisor's Signature

Co-Advisor's Signature

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงได้ด้วยดีก็ด้วยความกรุณาจากท่านอาจารย์ที่ปรึกษา ผศ. ดร. อติวงศ์ สุชาโต, ผศ. ดร. โปรดปราน บุณยพุกกณะ และ ดร. ชัย วุฒิวิวัฒน์ชัย ที่ได้สละเวลาในการให้คำปรึกษา ชี้แนะแนวทางในงานวิจัยต่างๆ ซึ่งผู้เขียนวิทยานิพนธ์ไม่ได้มีพื้นฐานทางด้านเทคโนโลยีเสียงพูดมาก่อน ถ้าไม่ได้ความช่วยเหลือของท่านเหล่านี้ วิทยานิพนธ์นี้ก็มิอาจสำเร็จลุล่วงไปได้

นอกจากนี้ ผมต้องขอขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ ซึ่งได้แก่ ศ. ดร. ประภาส จงสถิตย์วัฒนา, ศ. ดร. บุญเสริม กิจศิริกุล และ ดร. อนันต์ลดา โชติมงคล ที่สละเวลาเป็นกรรมการสอบวิทยานิพนธ์ และให้คำแนะนำอันเป็นประโยชน์ ซึ่งช่วยเพิ่มคุณค่าให้แก่งานวิจัยนี้

ขอขอบคุณ ผศ. ดร. อติวงศ์ สุชาโต และภาควิชาวิศวกรรมคอมพิวเตอร์ที่ได้ให้โอกาสให้ผมเข้ามาศึกษาต่อในระดับดุษฎีบัณฑิต

งานวิจัยนี้ได้รับเงินทุนสนับสนุนจากโครงการทุนสถาบันบัณฑิตวิทยาศาสตร์และเทคโนโลยีไทย (TGIST) ซึ่งอยู่ภายใต้สวทช. โดยมี ดร. ชัย วุฒิวิวัฒน์ชัย เป็นผู้แนะนำและช่วยดำเนินการเรื่องการขอทุน จึงขอขอบคุณไว้ ณ ที่นี้ด้วย นอกจากนี้ผมยังขอขอบคุณพี่ๆ เจ้าหน้าที่ของ TGIST ด้วยที่ช่วยเหลือในด้านต่างๆ

ขอขอบคุณพี่ๆ เพื่อนๆ และน้องๆ ในภาควิชาวิศวกรรมคอมพิวเตอร์ โดยเฉพาะสมาชิกในห้องปฏิบัติการ SLS และ ATL ที่ให้ความช่วยเหลือ และเป็นกำลังใจที่ดีเสมอมา ขอบคุณเจ้าหน้าที่ในภาควิชาวิศวกรรมคอมพิวเตอร์ที่ช่วยอำนวยสะดวกในการดำเนินเรื่องต่างๆ

ขอบคุณ Mr. Gun Jae Lee ที่เป็นผู้ผลักดันให้ผมได้ศึกษาต่อให้ระดับดุษฎีบัณฑิตนี้ ซึ่งทำให้ผมได้รับประสบการณ์ใหม่ๆอีกมากมาย

สุดท้ายนี้ ขอขอบคุณญาติๆ และคนในครอบครัว โดยเฉพาะ คุณพ่อ คุณแม่ที่ให้การสนับสนุน และเป็นกำลังใจตลอดมา เป็นเรื่องที่น่าเสียดายที่ผมไม่สามารถจบศึกษาให้คุณพ่อได้เห็นความสำเร็จได้ทัน ขอให้คุณพ่อได้รับรู้ถึงความสำเร็จของลูกคนนี้บนสวรรค์ด้วยเถิด ขอบคุณคุณวารุณี ลิขิตสุภิณ ที่เป็นกำลังใจให้ตลอดมา ด้วยกำลังใจนี้ทำให้ผมผ่านพ้นช่วงเวลาที่ยากลำบากที่สุดเวลาหนึ่งในชีวิตเลยก็ว่าได้

สารบัญ

หน้า

หน้า

หน้า

หน้า

สารบัญตาราง

หน้า

หน้า

สารบัญรูปภาพ

หน้า

หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันระบบรู้จำเสียงพูดเริ่มเข้ามามีบทบาทและมีการใช้งานในชีวิตประจำวันมากยิ่งขึ้น ไม่ว่าจะเป็นระบบตอบรับโทรศัพท์อัตโนมัติ (Interactive Voice Response: IVR) ระบบเขียนตามคำบอก (Dictation System) ระบบนำทาง (Navigation System) เป็นต้น

การรู้จำเสียงในปัจจุบันนั้นมีวิธีการที่ใช้กันอยู่หลายวิธี ซึ่งวิธีที่ได้รับความนิยมสูงสุด คือ การแบ่งเสียงออกเป็นลำดับของเวกเตอร์คุณสมบัติของเสียงที่ดึงลักษณะคุณสมบัติของเสียงมาจาก “กรอบเวลา (Frame)” ที่มีขนาดเท่าๆกัน โดยวิธีดังกล่าวนี้จะเรียกว่า “การรู้จำเสียงพูดแบบอาศัยกรอบเวลา (Frame-based Speech Recognition)” ปกติแล้ววิธีการนี้จะใช้โมเดลสถานะจำกัดที่เรียกกันว่า แบบจำลองฮิดเดนมาร์คอฟ (HMM) [1] ในการโมเดลลำดับของลักษณะเฉพาะให้อยู่ในรูปแบบของหน่วยเสียง ถึงแม้ว่าวิธีการรู้จำเสียงพูดแบบอาศัยกรอบเวลานั้นจะประสบความสำเร็จเพียงใด แต่วิธีการนี้ก็ยังมีข้อด้อยต่างๆอยู่ ข้อด้อยสำคัญของวิธีการรู้จำเสียงพูดแบบอาศัยกรอบเวลา คือ การตั้งสมมติฐานไว้ว่าแต่ละกรอบเวลานั้นมีความเป็นอิสระต่อกัน ซึ่งในความเป็นจริงแล้วแต่ละกรอบเวลาที่อยู่ภายในหน่วยเสียงเดียวกันนั้นมีความสัมพันธ์กัน เนื่องจากสัญญาณเสียงนั้นจะมีความต่อเนื่องกันอยู่ ด้วยข้อจำกัดของระบบรู้จำเสียงพูดแบบอาศัยกรอบเวลานั้น จึงทำให้เกิดการคิดค้นวิธีการขึ้นมาใหม่ โดยแบ่งเสียงออกเป็นลำดับของเวกเตอร์คุณสมบัติของเสียงที่ดึงลักษณะคุณสมบัติของเสียงมาจาก “เซกเมนต์ (Segment)” ที่มีขนาดแตกต่างกันไปแทนที่จะเป็นกรอบเวลาที่มีขนาดเท่าๆกัน ซึ่งวิธีการดังกล่าวเรียกว่า “การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ (Segment-Based Speech Recognition)” [2] วิธีการรู้จำเสียงพูดแบบอาศัยเซกเมนต์จะมีการตั้งสมมติฐานความเป็นอิสระต่อกันน้อยกว่าการรู้จำเสียงพูดแบบอาศัยกรอบเวลา ยังสามารถรองรับการใช้เวกเตอร์คุณสมบัติ และตัวจำแนกที่แตกต่างกันได้ [3, 4] นอกจากนี้การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ยังสามารถนำคุณสมบัติของเสียงในส่วนของขอบเขตของหน่วยเสียง (Boundary) มาใช้เป็นข้อมูลเพิ่มเติมเพื่อเพิ่มประสิทธิภาพของการรู้จำเสียงพูดด้วย ซึ่งเป็นที่เชื่อกันในเหล่านักวิจัยว่าในสัญญาณเสียงที่ขอบเขตของหน่วยเสียงนั้นจะมีคุณสมบัติต่างๆที่ใช้สามารถแบ่งแยกความแตกต่างของเสียงต่างๆได้ดี ตัวอย่างเช่น ตำแหน่งของการเปิดหรือปิดปาก การยืดหรือหดตัวของอวัยวะที่ใช้ในการเปล่งเสียง สำหรับการรู้จำเสียงพูดภาษาอังกฤษ ระบบ SUMMIT ของ MIT [2] เป็นการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ประสบความสำเร็จเป็นอย่างดี โดยผลการทดลองรู้จำเสียงพูดใน

ระดับหน่วยเสียงกับฐานข้อมูลเสียง TIMIT [5] ระบบ SUMMIT มีความผิดพลาดเพียง 24.4% และในการทดลองรู้จำเสียงพูดในระดับคำในงานด้านการสอบถามสภาพอากาศ SUMMIT มีความผิดพลาดเพียง 6.1% เท่านั้น

โดยปกติแล้วการรู้จำเสียงพูดแบบอาศัยเซกเมนต์จะดำเนินการ 2 ขั้นตอนหลักๆด้วยกัน คือ ขั้นตอนแรกเป็นการแบ่งเสียงพูดเป็นเซกเมนต์ (Segmentation) ในขั้นตอนนี้จะสร้างกราฟของเซกเมนต์ (Segment Graph) จากการนำเซกเมนต์ที่เป็นไปได้มาต่อๆกันโดยขนาดของเซกเมนต์ก็จะขึ้นอยู่กับหน่วยเสียง ขั้นตอนต่อมาจะเป็นขั้นตอนการรู้จำ ซึ่งในขั้นตอนนี้จะมีการให้คะแนนแต่ละเส้นทาง และค้นหาคำตอบจากเส้นทางที่มีคะแนนมากที่สุดในกราฟที่อยู่ในรูปแบบของตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก (Weighted Finite State Transducer) [6] ที่เกิดจากการประกอบกัน (Composite) ของกราฟของเซกเมนต์ และกราฟการออกเสียงที่ได้มาจากไวยากรณ์ (Pronunciation Graph) ที่ขึ้นอยู่กับงานที่ต้องการให้รู้จำเสียงพูด ซึ่งจะมีหลายอัลกอริทึมที่ใช้ในการค้นหาเส้นทางที่มีคะแนนมากที่สุด การค้นหาของไวเทอร์บิ (Viterbi Search) ก็เป็นวิธีการหนึ่งที่นิยมใช้กันอย่างแพร่หลาย

เมื่อคำตอบของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์มีการค้นหาคำตอบจากกราฟที่เกิดจากการประกอบกันของกราฟการออกเสียง และกราฟของเซกเมนต์แล้ว ดังนั้น เราจะพบว่าความแม่นยำของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์นั้นจะขึ้นอยู่กับจำนวนเซกเมนต์ที่มีการแบ่งอย่างถูกต้องในกราฟของเซกเมนต์ ถ้ากราฟของเซกเมนต์มีเซกเมนต์ที่ถูกต้องอยู่มาก ความแม่นยำของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ก็จะดีขึ้นเช่นกัน แต่อย่างไรก็ตามสำหรับภาษาไทยการรู้จำเสียงพูดแบบอาศัยเซกเมนต์มีความแม่นยำในการรู้จำเสียงพูดในระดับหน่วยเสียงได้เพียงประมาณ 50% [7] ซึ่งเกิดจากมีทรัพยากรไม่เพียงพอที่จะทำให้ระบบรู้จำเสียงพูดแบบอาศัยกรอบเวลาซึ่งใช้ในการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น (Probabilistic Segmentation) [8-10] ซึ่งเป็นวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ที่ระบบ SUMMIT ใช้อยู่ในปัจจุบัน มีความสมบูรณ์ถูกต้องจึงทำเกิดความผิดพลาดอยู่ในกราฟของเซกเมนต์ และส่งผลกระทบต่อกระบวนการรู้จำเสียงพูดของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

วิทยานิพนธ์นี้มีวัตถุประสงค์ เพื่อที่จะศึกษา วิจัย และพัฒนาการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ สำหรับภาษาไทยโดยจะมุ่งเน้นในการปรับปรุงประสิทธิภาพของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ 2 ส่วนด้วยกัน คือ 1) การเพิ่มจำนวนเซกเมนต์ที่ถูกต้องในกราฟของเซกเมนต์ โดยการแก้ไขความผิดพลาดที่เกิดจากการแบ่งเสียงพูดเป็นเซกเมนต์ ซึ่งนำเอา สมบัติลักษณะเฉพาะ (Distinctive Features) และความไม่ต่อเนื่องของเสียง (Acoustic Discontinuities) มาใช้ในการเพิ่มจำนวนเซกเมนต์ที่ถูกต้องในกราฟของเซกเมนต์ 2) ปรับปรุงในส่วนของการให้คะแนนของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์โดยการจะมีการนำคะแนนซึ่งเป็นความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้าง (Phoneme Broad Class) ใดๆมาใช้เป็นคะแนนเพิ่มเติม

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีวัตถุประสงค์ในการพัฒนาระบบรู้จำเสียงพูดต่อเนื่องภาษาไทย โดยใช้วิธีการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ และเสนอวิธีปรับปรุง แก้ไข ข้อผิดพลาดที่ปรากฏในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์แบบเดิม เพื่อเป็นการเพิ่มความแม่นยำให้แก่การรู้จำเสียงพูดแบบอาศัยเซกเมนต์

1.3 ขอบเขตของการวิจัย

งานวิจัยนี้นำเสนอวิธีการเพิ่มความแม่นยำของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์จากการเพิ่มจำนวนเซกเมนต์ที่ถูกต้องในกราฟของเซกเมนต์โดยมีการแก้ไขความผิดพลาดที่เกิดจากการแทรก และการตัดออกของขอบเขตของหน่วยเสียง ซึ่งเป็นผลมาจากการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น และการนำความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างมาใช้ในขั้นตอนการให้คะแนนและค้นหาคำตอบของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ ซึ่งขอบเขตของวิทยานิพนธ์นี้จะมี ดังนี้

1. พัฒนาระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์สำหรับภาษาไทย
2. นำเสนอวิธีการเพิ่มความแม่นยำของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์โดยการเพิ่มจำนวนเซกเมนต์ที่ถูกต้องในกราฟของเซกเมนต์โดยมีการแก้ไขความผิดพลาดที่เกิดจากการแทรก และการตัดออกของขอบเขตของหน่วยเสียง ซึ่งเป็นผลมาจากขั้นตอนการแบ่งเสียงพูดเป็นเซกเมนต์ ที่ใช้วิธีการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น
3. นำเสนอวิธีการเพิ่มความแม่นยำของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์โดยการเพิ่มความน่าจะเป็นที่เซกเมนต์นั้นถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างใดๆ เข้าไปในกระบวนการให้คะแนนของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์
4. เปรียบเทียบผลการรู้จำเสียงพูดภาษาไทยของวิธีต่างๆ ได้แก่ การรู้จำเสียงพูดแบบอาศัยกรอบเวลา การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ และการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ผ่านกระบวนการที่นำเสนอ
5. ในการทดสอบผลการรู้จำเสียงพูดภาษาไทยนั้น จะเป็นการเปรียบเทียบผลการรู้จำเสียงพูดในระดับหน่วยเสียง และดำเนินการบนฐานข้อมูลเสียงพูดต่อเนื่องภาษาไทยจากฐานข้อมูลเสียงโลตัส (Large Vocabulary Thai Continuous Speech Recognition Corpus : LOTUS) [11-13]

1.4 ขั้นตอนและวิธีดำเนินการวิจัย

1. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องกับวิทยานิพนธ์นี้ ทั้งในส่วนของการแบ่งเสียงพูดเป็นเซกเมนต์ การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ สมบัติลักษณะเฉพาะ และความไม่ต่อเนื่องของเสียง
2. พัฒนาระบบการรู้จำเสียงพูดแบบอาศัยเซกเมนต์
3. ทดลองและปรับปรุงการรู้จำเสียงพูดแบบอาศัยเซกเมนต์กับฐานข้อมูลเสียงภาษาไทย
4. เปรียบเทียบ และวิเคราะห์ผลลัพธ์ที่ได้จากการปรับปรุงการรู้จำเสียงพูดแบบอาศัยเซกเมนต์
5. เขียนรายงานและจัดทำวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

สิ่งที่ได้รับจากงานวิจัยนี้ คือ ระบบการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่สามารถใช้งานได้กับภาษาไทย ที่ได้รับการปรับปรุงความแม่นยำโดยวิธีที่นำเสนอในวิทยานิพนธ์นี้ ซึ่งระบบการรู้จำเสียงพูดแบบอาศัยเซกเมนต์นี้จะสามารถทำให้เกิดงานวิจัยต่อยอดไปได้อีก นอกจากนี้ ในขั้นตอนในการปรับปรุงความแม่นยำต่างๆ ยังได้มีศึกษาองค์ความรู้ในด้านของสมบัติลักษณะเฉพาะ และมีการสร้างตัวจำแนกสมบัติลักษณะเฉพาะที่ยังสามารถนำไปประยุกต์ใช้งานในการพัฒนาระบบการรู้จำเสียงพูดแบบอื่นๆได้อีก

1.6 เนื้อหาในวิทยานิพนธ์

รายละเอียดต่างๆในวิทยานิพนธ์ฉบับนี้จะนำเสนอเป็นลำดับดังต่อไปนี้

บทที่ 2 กล่าวถึงทฤษฎี และงานวิจัยที่เกี่ยวข้อง ซึ่งได้แก่ ทฤษฎีเกี่ยวกับการรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ งานวิจัยที่เกี่ยวกับการแบ่งเสียงพูดเป็นเซกเมนต์ และงานวิจัยที่ได้มีการนำเอาสวนสัทศาสตร์ เช่น สมบัติลักษณะเฉพาะมาใช้ในการรู้จำเสียงพูด เป็นต้น

บทที่ 3 เป็นการทดลองเบื้องต้นเกี่ยวกับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ และมีการแสดงถึงความผิดพลาดในกราฟของเซกเมนต์ที่เกิดจากการแบ่งเสียงพูดเป็นเซกเมนต์ ซึ่งเป็นปัจจัยสำคัญที่มีผลกระทบต่อความแม่นยำในการรู้จำเสียงพูดของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ ซึ่งแรงจูงใจของงานวิจัยนี้

บทที่ 4 อธิบายถึงวิธีการที่วิทยานิพนธ์นี้นำเสนอในการปรับปรุงความแม่นยำของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ โดยการแก้ไขความผิดพลาดที่เกิดจากขั้นตอนการแบ่งเสียงพูดเป็นเซกเมนต์ รวมทั้งการนำค่าความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้าง

มาใช้เป็นคะแนนในขั้นตอนการให้คะแนนและค้นหาคำตอบของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

บทที่ 5 อธิบายรายละเอียดเกี่ยวกับการทดลองต่างๆในงานวิจัยนี้ และในบทที่ 6 จะกล่าวถึงผลการทดลอง รวมทั้งการอธิปรายผลการทดลอง

บทที่ 6 จะเป็นการสรุปผลงานวิจัยของวิทยานิพนธ์ฉบับนี้ รวมทั้งข้อเสนอแนะ

1.7 งานตีพิมพ์

ในระหว่างการศึกษาได้มีการตีพิมพ์ผลงานวิจัยดังนี้

Proceedings

- Krerksak Likitsupin, Atiwong Suchato, Proadpran Punyabukkana, and Chai Wutiwiwatchai, “Improving Segment-based Speech Recognition by Recovering Missing Segments in Segment Graphs – A Thai Case Study”, In Proceedings of The International Symposium on Communications and Information Technologies (ISCIT 2008), Vientiane, Lao PDR, October 21-23, 2008.
- Krerksak Likitsupin, Sirinart Tangruamsub, Atiwong Suchato, and Proadpran Punyabukkana, “Phoneme Recognition from Thai Continuous Speech using a Segment-based Approach”, In Proceedings of The 11th National Computer Science and Engineering Conference (NCSEC2007), Bangkok, Thailand, November 19-21, 2007.

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 โครงสร้างพยางค์ในภาษาไทย

พยางค์ (Syllable) หมายถึง หน่วยหนึ่งขององค์ประกอบเสียงที่ใช้ในการสื่อสารด้วยคำพูด ตามฐานข้อมูลเสียงโลตัสพยางค์ จะประกอบด้วย แกนพยางค์ (Syllable Nucleus) ซึ่งเป็นเสียงสระ (Vowel: V) จำนวน 24 หน่วยเสียง เป็นสระเดี่ยว (Monophthongs) 18 หน่วยเสียงและสระผสม (Diphthongs) 6 หน่วยเสียง ส่วนเสียงพยัญชนะ (Consonant) ที่เกิดขึ้นในส่วนต้นพยางค์ เรียกว่า พยัญชนะต้น (Initial Consonant: Ci) ซึ่งมีจำนวน 38 หน่วยเสียง หรือท้ายพยางค์ เรียกว่าตัวสะกด (Final Consonant: Cf) จำนวน 12 หน่วยเสียง ซึ่งโครงสร้างพยางค์มีรูปแบบเป็น Ci V (Cf) โดยบ้างพยางค์ไม่จำเป็นต้องมีตัวสะกด

2.2 ระบบเสียงในภาษาไทย [14]

เสียงในภาษาไทย [15] ประกอบด้วยหน่วยเสียงจากตัวอักษรภาษาไทยมีทั้งหมด 44 หน่วยเสียงได้แก่ หน่วยเสียงพยัญชนะ 38 หน่วยเสียง และหน่วยเสียงสระ 24 หน่วยเสียง

2.2.1 เสียงพยัญชนะ

ในงานวิจัยนี้จะอ้างอิงหน่วยเสียงจากฐานข้อมูลเสียงโลตัส [11-13] ซึ่งมีการกำหนดให้หน่วยเสียงที่เป็นพยัญชนะต้นมีทั้งหมด 38 หน่วยเสียง แบ่งเป็น เสียงพยัญชนะต้น 21 หน่วยเสียง เสียงพยัญชนะต้นควบกล้ำในภาษาไทยแท้ 11 หน่วยเสียง เสียงพยัญชนะต้นควบกล้ำในภาษาไทยใช้ในการทับศัพท์ภาษาอังกฤษ 6 หน่วยเสียง ดังที่แสดงในตารางที่ 2.1

ตารางที่ 2.1 หน่วยเสียงพยัญชนะต้นภาษาไทยในฐานข้อมูลเสียงโลตัส [15]

พยัญชนะต้น (Ci)			
เดี่ยว	**ตัวอย่าง**	**ผสม**	**ตัวอย่าง**
p	ปาก	pr	ประสาน
t	เต้น, กุฏิ	phr	พราน
c	จะ	tr	เตรียม
k	ก่อน	kr	กราบ
z	อาน	khr	คร่า
ph	พบ, ภัย, ผ่าน	pl	ปลา
th	ทิ้ง, ธง, เฒ่า, ฐาน, มณโฑ	phl	พลาด
ch	ชอบ, เฌอ	thr	จันทรา
kh	คน, เงิน, ฆ่า	kl	เกลอ
b	บอก	khl	เคลื่อน
d	ด้าน, ชฎา	kw	กวาง
m	ไม่	khw	ขวา
n	นาน, เณร		
ng	เงิน	เสียงทับศัพท์	
l	เล่น, กีฬา	br	เบรน
r	รอ, ฤทัย	bl	บลู
f	ฝน, ฟัน	fr	ฟราย
s	สาย, ศิลา, รักษา, ซ่อน	fl	เฟลม
h	โหน, เฮฮา	dr	ดรากอน
w	ว่า	17 หน่วย	
j	ย้อน, หญิง		
21 หน่วย			

ในส่วนของหน่วยเสียงพยัญชนะที่สามารถเป็นตัวสะกดจะมีทั้งหมด 12 หน่วยเสียง โดยจะมี 4 หน่วยเสียงที่เป็นเสียงพยัญชนะที่เป็นตัวสะกดควบกล้ำในภาษาไทยใช้ในการทับศัพท์ภาษาอังกฤษ ดังที่แสดงในตารางที่ 2.2

ตารางที่ 2.2 หน่วยเสียงตัวสะกดภาษาไทยในฐานข้อมูลเสียงโลตัส [15]

ตัวสะกด (Cf)	
เดี่ยว	**ตัวอย่าง**
p^	พบ
t^	เกร็ด
k^	ปาก
n^	หาร
m^	ลม
ng^	ฟาง
j^	ยาย
w^	กาว
เสียงทับศัพท์	
f^	กราฟ
l^	แอล
s^	เอส
ch^	คลัช
12 หน่วย	

คุณสมบัติที่ใช้จำแนกความแตกต่างของเสียงพยัญชนะภาษาไทย [16] มีดังนี้

1) คุณสมบัติความก้องของเสียง ความก้องของเสียงเป็นคุณสมบัติที่ใช้ในการแบ่งแยกเสียงพยัญชนะออกได้เป็นสองชนิด คือ
   - เสียงพยัญชนะโฆษะ (Voiced) หรือเสียงก้อง เป็นเสียงพยัญชนะที่เส้นเสียงสั่นสะเทือนขณะที่เปล่งเสียง
   - เสียงพยัญชนะอโฆษะ (Voiceless) หรือเสียงไม่ก้อง เป็นเสียงพยัญชนะที่เส้นเสียงไม่สั่นสะเทือนขณะที่เปล่งเสียง
2) ลักษณะของลมที่ผ่านเส้นเสียง เสียงพยัญชนะสามารถแบ่งตามลักษณะลมที่ผ่านเส้นเสียงออกมาได้ดังนี้
   - เสียงหยุด (Stop) สามารถแบ่งได้เป็น 2 กลุ่มย่อยคือ เสียงพยัญชนะระเบิด (Plosive Stop) และเสียงพยัญชนะกัก (Unreleased Stop) เสียงพยัญชนะระเบิดเกิดจากการที่มีการปิดกั้นลมไว้ภายในปาก แล้วเปิดจุดที่กักลมออก ทำให้เกิดลมพุ่งออกมา

ทันที โดยเสียงพยัญชนะระเบิดแบ่งได้เป็น เสียงพยัญชนะระเบิดธนิต (Aspirated Plosive) ซึ่งจะมีกลุ่มลมพ่นออกมา หลังเปล่งเสียง และเสียงพยัญชนะระเบิดสิถิล (Unaspirated Plosive) ซึ่งจะไม่มีกลุ่มลมพ่นออกมา ส่วนเสียงพยัญชนะกักเกิดจากลมที่เปล่งออกถูกกักไว้ ณ ที่ใดที่หนึ่งในปาก และไม่ได้ถูกปล่อยออกมา ซึ่งเสียงพยัญชนะกักนี้จะเป็นเสียงตัวสะกดท้ายพยางค์

- เสียงนาสิก (Nasal) เป็นเสียงพยัญชนะที่เกิดจากมีการปิดกั้นลมไว้ในปาก และเกิดการลดระดับของเพดานอ่อนและช่องวีลิคเปิดทำให้ลมถูกส่งออกผ่านโพรงจมูก
- เสียงเสียดแทรก (Fricative) เป็นเสียงพยัญชนะที่เกิดจากการปิดกั้นลมที่ไม่สมบูรณ์ในช่องเสียงทำให้ลมที่ผ่านออกมาต้องผ่านช่องแคบเล็กๆ ณ ที่ใดที่หนึ่งในช่องปาก ทำให้ลมแทรกผ่านไปในลักษณะเสียดสี เกิดเป็นเสียงเหมือนเสียงรบกวน
- เสียงกึ่งเสียดแทรก (Affricate) เป็นเสียงพยัญชนะที่เกิดจากมีการปิดกั้นลมไว้ภายในปาก เหมือนเสียงหยุดหรือเสียงระเบิด แต่แทนที่จะปล่อยลมออกมาทันที อวัยวะที่ใช้ในการออกเสียงจะเปิดออกอย่างช้าๆ ตามด้วยการเกิดเสียงเสียดแทรกและเพดานอ่อนยกขึ้นปิดช่องวีลิค ทำให้ลมออกทางปาก
- เสียงข้างลิ้น (Lateral) เป็นเสียงพยัญชนะที่เกิดจากมีการปิดกั้นลมไว้ภายในปาก และมีจุดปิดกั้นอยู่ภายในปาก โดยใช้ลิ้นปิดบริเวณปุ่มเหงือกหรือเพดานแข็งส่วนกลางไว้ ทำให้ลมไหลผ่านออกมาท้างข้างลิ้น ซึ่งจะไหลออกมาข้างเดียวหรือสองข้างก็ได้
- เสียงรัว (Trill) เป็นเสียงพยัญชนะที่เกิดจากปลายลิ้นกระดกขึ้นไปแตะปุ่มเหงือกอย่างเร็วและแตะหลายครั้งจนได้ยินเป็นเสียงรัว
- เสียงพยัญชนะกึ่งสระ (Semi-Vowel) หรือ เสียงเปิด (Approximant) เป็นเสียงพยัญชนะที่เกิดขึ้น โดยการเปิดกว้างของช่องปาก ทำให้ลมผ่านออกมาได้โดยสะดวก โดยไม่มีการปิดกั้นของลม หรือไม่มีการบังคับให้ลมแทรกออกมาผ่านตามช่องแคบๆ ในลักษณะเสียดสี

3) ฐานที่เกิดของเสียง ไม่ว่าลมที่ใช้ในการออกเสียงพยัญชนะนั้นจะมาถูกกัก ถูกดัดแปลงจนเกิดการกัก หรือการเสียดแทรก จำเป็นต้องมีตำแหน่งที่เกิดอยู่ด้วยเสมอในช่องปาก

หน่วยเสียงพยัญชนะในภาษาไทย และคุณสมบัติที่ใช้จำแนกความแตกต่างของเสียงพยัญชนะภาษาไทย แสดงดังตารางที่ 2.3

ตารางที่ 2.3 หน่วยเสียงพยัญชนะภาษาไทย (ปรับปรุงจาก [16])

1 (*) ปรากฏท้ายพยางค์ได้

2 (**) ปรากฏท้ายพยางค์เฉพาะในคำไทยทับศัพท์ภาษาอังกฤษ

3 (/…/) ปรากฏเฉพาะในคำไทยทับศัพท์ภาษาอังกฤษ

4 [/…/] ปรากฏในคำไทยทับศัพท์ภาษาอังกฤษ หรือคำไทยที่ยืมมาจากภาษาสันสกฤต

หน่วยเสียง[1]	หน่วยเสียงควบกล้ำ	ลักษณะของลม	การพ่นลม	ความก้อง	ฐานที่เกิด	รูปพยัญชนะ
/p/ (*)	/pr/, /pl/	กัก	ไม่พ่นลม	ไม่ก้อง	ริมฝีปาก	ป
/p^h/	/p^hr/,/p^hl/	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
/t^h/	[/t^hr/]	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
/c^h/(**)		กัก	พ่นลม	ไม่ก้อง	เพดานแข็ง	ฉ ช ฌ
/k/ (*)	/kr/, /kl/, /kw/	กัก	ไม่พ่นลม	ไม่ก้อง	เพดานอ่อน	ก
/k^h/	/k^hr/, /k^hl/, /k^hw/	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)

2.2.2 เสียงสระ

เสียงสระเป็นเสียงที่เปล่งออกมาโดยไม่มีอวัยวะส่วนใดในปากที่เป็นอุปสรรคปิดกั้นทางลมไว้ ลักษณะการออกเสียงสระ เส้นเสียงจะมีการสั่นสะเทือน และส่งผลทำให้เกิดเสียงดังเป็นเสียงก้อง เสียงสระเป็นเสียงที่ทำหน้าที่เป็นใจกลาง หรือแกนของพยางค์ จากฐานข้อมูลเสียงโลตัสเสียงสระสามารถแบ่งเป็น 24 หน่วยเสียง ซึ่งในฐานข้อมูลเสียงโลตัสไม่มีการนับรวมเสียงสระเกิน (Vowel Letter) จำนวน 8 หน่วยเสียง โดยเสียงสระในฐานข้อมูลเสียงโลตัสจะแบ่งออกเป็น 2 กลุ่มด้วยกัน คือ

1) สระเดี่ยว เป็นสระเสียงแท้ ซึ่งการออกเสียงสระตั้งแต่เริ่มต้นจนสิ้นสุดไม่มีการเปลี่ยนรูปร่างของลิ้นและช่องปาก สระเดี่ยวในภาษาไทยมีทั้งสิ้น 18 หน่วยเสียง เป็นสระเสียงสั้น 9 หน่วยเสียง และสระเสียงยาว 9 หน่วยเสียง
2) สระประสม เป็นสระที่เกิดจากการออกเสียงผสมกันของสระแท้ โดยลิ้นและช่องปากจะเปลี่ยนจากรูปร่างการออกเสียงของสระหนึ่งไปยังอีกสระหนึ่งอย่างค่อนข้างกลมกลืนและรวดเร็ว สระประสมในภาษาไทยมีทั้งสิ้น 6 หน่วยเสียง เป็นสระเสียงสั้น 3 หน่วยเสียง และสระเสียงยาว 3 หน่วยเสียง

หน่วยเสียงสระภาษาไทยของฐานข้อมูลเสียงโลตัส จะแสดงในตารางที่ 2.4

อวัยวะที่สำคัญที่ทำให้เกิดเสียงสระต่างๆ คือ ลิ้น ริมฝีปาก ดังนั้น จึงสามารถแบ่งชนิดของสระได้จาก

1) ส่วนของลิ้นที่ใช้ในการเปล่งเสียง (Place of Articulation) ส่วนนี้จะเป็นการบอกถึงตำแหน่งของลิ้นที่ใช้ในการออกเสียงสระ เช่น ลิ้นส่วนหน้า ลิ้นส่วนกลาง หรือลิ้นส่วนหลัง
2) ความสูงของลิ้น หรือระยะห่างระหว่างลิ้นและเพดานปาก (Degree of Stricture) ในกรณีที่ลิ้นอยู่ต่ำ หรืออยู่ห่างจากเพดานปากมาก จะทำให้ช่องโพรงปากกว้างทำให้ลมผ่านออกมาได้ง่าย จึงเรียกเสียงสระจำพวกนี้ว่า สระเปิด (Open Vowel) แต่ถ้าลิ้นอยู่ในสูง ช่องโพรงจะแคบมาก ทำให้ลมผ่านได้น้อย จึงเรียกสระเหล่านี้ว่า สระปิด (Close Vowel) นอกจากนี้ยังความสูงของลิ้นที่อยู่ระหว่างสระเปิด และสระปิด ซึ่งเสียงสระที่มีตำแหน่งลิ้นอยู่สูงกว่าสระปิดเล็กน้อยจะเรียกว่า สระกลางปิด หรือสระกึ่งปิด (Close-mid, Half-close Vowel) แต่ถ้าความสูงของลิ้นอยู่สูงขึ้นไปอีกหน่อยจะเรียกเสียงนี้ว่าเป็น สระกลางเปิด หรือสระกึ่งเปิด (Open-mid, Half-open Vowel)
3) การห่อริมฝีปาก (Labialization) จะมีสองลักษณะคือ ห่อริมฝีปากกลม (Rounded) และไม่มีการห่อริมฝีปากกลม (Unrounded)
4) ลักษณะนาสิก (Nasalization) เป็นลักษณะที่ใช้บอกว่าเสียงสระเกิดขึ้นที่จมูกหรือไม่

5) ความยาวในการออกเสียง (Duration) จะแสดงความสั้นยาวในการออกเสียง โดยจะแบ่งออกเป็น 2 ประเภท คือ สระเสียงสั้น (รัสสระ) และสระเสียงยาว (ทีฆสระ)

หน่วยเสียงสระในภาษาไทย และคุณสมบัติที่ใช้จำแนกความแตกต่างของเสียง แสดงดังตารางที่ 2.5 และตารางที่ 2.6

ตารางที่ 2.4 หน่วยเสียงสระภาษาไทยในฐานข้อมูลเสียงโลตัส [15]

ตัวสะกด (Cf)	
เดี่ยว	**ตัวอย่าง**
p^	พบ
t^	เกร็ด
k^	ปาก
n^	หาร
m^	ลม
ng^	ฟาง
j^	ยาย
w^	กาว
เสียงทับศัพท์	
f^	กราฟ
l^	แอล
s^	เอส
ch^	คลัช
12 หน่วย	

ตารางที่ 2.5 หน่วยเสียงสระเดี่ยวภาษาไทย [16]

หน่วยเสียง[1]	ส่วนของลิ้นที่ใช้เปล่งเสียง	ความสูงของลิ้น	การห่อริมฝีปาก	ความยาวเสียง	รูปสระ
/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

[1]ใช้หน่วยเสียงตามสัทอักษรสากล (International Phonetic Alphabet – IPA)

ตารางที่ 2.6 หน่วยเสียงสระประสมภาษาไทย [16]

หน่วยเสียง[1]	ส่วนประกอบ	ความยาวเสียง	รูปสระ
/ia/	/i/ + /a/	สั้น	เอียะ
/iːa/	/iː/ + /a/	ยาว	เอีย
/i̱a/	/i̱/ + /a/	สั้น	เอือะ
/i̱ːa/	/i̱ː/ + /a/	ยาว	เอือ
/ua/	/u/ + /a/	สั้น	อัวะ
/uːa/	/uː/ + /a/	ยาว	อัว

[1]ใช้หน่วยเสียงตามสัทอักษรสากล (International Phonetic Alphabet – IPA)

2.3 สมบัติลักษณะเฉพาะ

สมบัติลักษณะเฉพาะเป็นสิ่งที่สามารถใช้แยกความแตกต่างของแต่ละกลุ่มของหน่วยเสียง ซึ่งจะมีการแบ่งแยกโดยค่า บวก (+) และ ลบ (–) ซึ่งค่าบวกจะหมายถึงมีสมบัติลักษณะเฉพาะดังกล่าว และลบหมายถึงไม่มีสมบัติลักษณะเฉพาะ สมบัติลักษณะเฉพาะสามารถแบ่งออกเป็น 3 ประเภท คือ คุณสมบัติความก้องของเสียงตามลักษณะของแหล่งกำเนิดเสียง (Source Characteristics) ลักษณะการออกเสียง (Manner of Articulation) และตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียง (Place of Articulation)

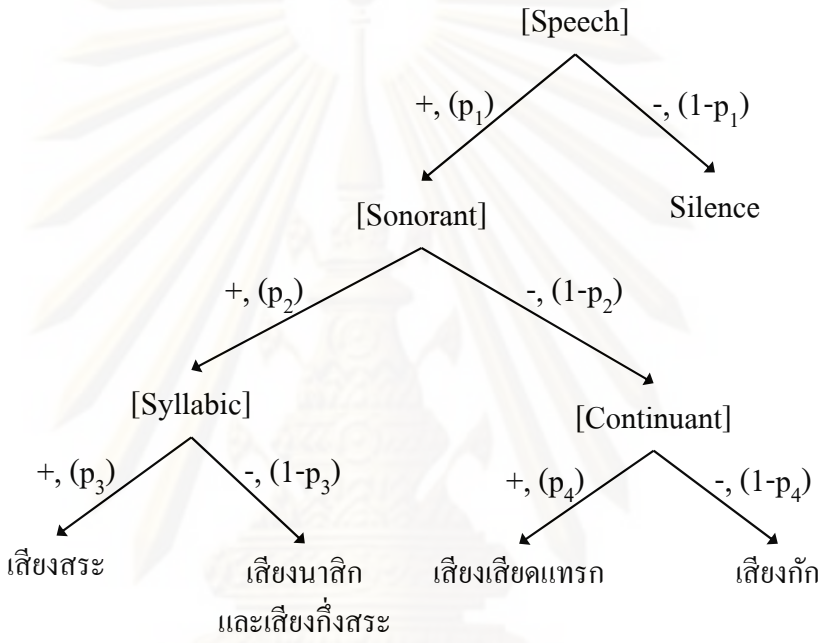
2.3.1 คุณสมบัติความก้องของเสียง

ขณะที่มีการออกเสียงอากาศจะถูกดันออกมาจากปอดผ่านเส้นเสียงและจะทำให้เส้นเสียงสั่น เมื่อเส้นเสียงมีการสั่นจะทำให้สัญญาณเสียงที่ออกมาจากแหล่งกำเนิดเสียงมีลักษณะเป็นคาบ ซึ่งคุณสมบัตินี้ก็คือ คุณสมบัติความเป็นเสียงก้อง ในกรณีนี้แทนด้วย “+” หรือ [+voiced] และในส่วนของเสียงที่ไม่มีลักษณะไม่เป็นคาบจะแสดงด้วยคุณสมบัติความเป็นเสียงไม่ก้อง ซึ่งจะแทนด้วย “-” หรือ [-voiced]

2.3.2 ลักษณะการออกเสียง

ลักษณะการออกเสียง เป็นพารามิเตอร์อธิบายเกี่ยวกับ การเปิดหรือปิดของช่องเสียง มีการหดตัวที่แข็งหรืออ่อน จนถึงกระทั่งว่ามีอากาศคลื่นผ่านไปยังช่องปากหรือโพรงจมูก ลักษณะการออกเสียงนั้นมีลักษณะต่างๆ ดังนี้ คุณสมบัติความเป็นเสียงพูด (Speech) คุณสมบัติความเป็นเสียงที่อากาศผ่านช่องปากไปโดยไม่ได้ถูกกักเอาไว้เพียงพอต่อการสร้างเสียงรบกวนหรือกักการไหลของอากาศ (Sonorant) คุณสมบัติความเป็นเสียงที่มีคุณสมบัติความเป็นเสียงที่มีการกักเสียงเอาไว้โดยที่

ช่องทางเดินเสียงอยู่ในสภาพปิดอยู่แต่ยังปิดไม่สมบูรณ์ (Continuant) และคุณสมบัติความเป็นเสียงที่เป็นศูนย์กลางของพยางค์ (Syllabic) ในการแสดงความเป็นลักษณะการออกเสียงจะแทนด้วยเครื่องหมาย บวก และ ลบ ซึ่งเครื่องหมายบวกจะแสดงความเป็นลักษณะการออกเสียงนั้นๆ ตัวอย่างเช่น [+Sonorant] คือเสียงมีการเปิดช่องทางการไหลของอากาศ เป็นต้น นอกจากนี้เมื่อนำลักษณะการออกเสียงยังสามารถทำเป็นโครงสร้างลำดับชั้น ก็จะสามารถใช้ในการระบุกลุ่มของหน่วยเสียงแบบกว้างได้ ดังรูปที่ 2.1



รูปที่ 2.1 โครงสร้างลำดับชั้นของลักษณะการออกเสียง [17, 18]

จากรูปที่ 2.1 จะเห็นว่าลักษณะการออกเสียงสามารถแบ่งแยกกลุ่มของหน่วยเสียงได้ ตัวอย่างเช่น เสียงสระมีคุณสมบัติการออกเสียงเป็น [+Speech], [+Sonorant] และ [+Syllabic] เป็นต้น

นอกจากนี้เรายังสามารถหาความน่าจะเป็นที่เสียงจะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างใดๆได้ โดยถ้าเราให้ o_t เป็นเวกเตอร์ของข้อมูลทางเสียงที่สังเกตได้ (Observation Vector) ที่กรอบเวลาที่ t เราจะสามารถคำนวณหาค่าความน่าจะเป็นที่ o_t จะจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้าง ณ ที่ตำแหน่งปลายสุดของโครงสร้างลำดับชั้นตามรูปที่ 2.1 ได้โดยนำหาความน่าจะเป็นแบบมีเงื่อนไขของสมบัติลักษณะตั้งแต่ลำดับบนจนถึงลำดับล่างสุดของโครงสร้างลำดับชั้น ซึ่งเราได้ประยุกต์สมการของ [17] เพื่อการคำนวณหาค่าความน่าจะเป็นที่ o_t จะจัดอยู่ในกลุ่มของหน่วยเสียงสระ จะแสดงดังสมการ (2.1)

$$
\begin{aligned}
P(Vowel \mid o_t) &= P(+Speech, +Sonorant, +Syllabic \mid o_t) \\
&= P(+Speech \mid o_t) P(+Sonorant \mid +Speech, o_t) \\
&\quad P(+Syllabic \mid +Speech, +Sonorant, o_t) \\
&= p_1 p_2 p_3
\end{aligned}
\tag{2.1}
$$

จากโครงสร้างลำดับชั้นตามรูปที่ 2.1 จะเห็นว่ากลุ่มของหน่วยเสียงสระ จะต้องประกอบไปด้วย [+Speech] [+Sonorant] และ [+Syllabic] และจากสมการ (2.1) จะเห็นว่าความน่าจะเป็นที่ o_t จะจัดอยู่ในกลุ่มของหน่วยเสียงสระ จะเกิดจากผลคูณของค่าความน่าจะเป็น p_1 p_2 และ p_3

$$
\begin{aligned}
P(Stop, o_t) &= P(+Speech, -Sonorant, -Continuant \mid o_t) \\
&= P(+Speech \mid o_t) P(-Sonorant \mid +Speech, o_t) \\
&\quad P(-Continuant \mid +Speech, -Sonorant, o_t) \\
&= p_1 (1 - p_2)(1 - p_4)
\end{aligned}
\tag{2.2}
$$

สมการ (2.2) แสดงถึงการคำนวณหาความน่าจะเป็นที่ o_t จะจัดอยู่ในกลุ่มของหน่วยเสียงหยุด อย่างไรก็ตาม การเปรียบเทียบความน่าจะเป็นดังกล่าวโดยตรงนั้นจะไม่มีความยุติธรรม หากมีการนำค่าความน่าจะเป็นมาเปรียบเทียบกันจะพบว่าความน่าจะเป็นที่ o_t จะจัดอยู่ในกลุ่มของหน่วยเงียบจะมีความน่าจะเป็นสูงที่สุด เนื่องจากค่าความน่าจะเป็นดังกล่าวนี้มีค่ามาจาก (1-p_1) เท่านั้น เมื่อเปรียบเทียบกับค่าความน่าจะเป็นอื่นๆ ซึ่งเกิดจากผลคูณของค่าความน่าจะเป็นตามลำดับชั้น ดังนั้นในวิทยานิพนธ์นี้เราจะมีการตรวจสอบความเป็น [Speech] เวกเตอร์ของข้อมูลทางเสียงที่สังเกตได้ก่อน (เนื่องจากตัวจำแนกของคุณสมบัติการเป็นเสียงพูดนั้นมีความถูกต้องสูง) ถ้าเวกเตอร์ของข้อมูลทางเสียงที่สังเกตได้นั้นจำแนกได้เป็น [+Speech] เราจะตั้งค่าให้ความน่าจะเป็น p_1 มีค่าเข้าใกล้ 1 และจะตั้งค่าให้ p_1 มีค่าเข้าใกล้ 0 เมื่อจำแนกเป็น [-Speech]

2.3.3 ตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียง

ตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียงเป็นพารามิเตอร์ที่ใช้อธิบายถึงลักษณะทางกายภาพในกระบวนการทำให้เกิดเสียง ตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียงสามารถใช้ในการแบ่งแยกหน่วยเสียงที่อยู่ในกลุ่มที่มีลักษณะการออกเสียงเดียวกัน

ตัวอย่างเช่น เสียงนาสิกจะสามารถระบุหน่วยเสียงโดยตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียง ดังนี้ ตำแหน่งที่ปุ่มเหงือก ริมผีปาก และเพดานอ่อน โดยที่เสียงมุทธชะจะเกิดจากเอาลิ้นแตะ หรือวางอยู่ใกล้บริเวณปุ่มเหงือกในขณะออกเสียง ซึ่งในเสียงนาสิก คือ หน่วยเสียง /น/ เสียงโอฐชะจะเกิดจากการบังคับริมฝีปากให้ปิด ซึ่งในเสียงนาสิก คือ หน่วยเสียง /ม/ สุดท้ายเสียงที่เกิดที่เพดานอ่อนจะเกิดจากการใช้เพดานอ่อน ซึ่งเป็นเนื้อเยื่อที่ใช้แยกระหว่างช่องปากและจมูก ซึ่ง

ในเสียงนาสิก คือ หน่วยเสียง /ง/ ในเสียงพยัญชนะกักก็สามารถแยกแยะหน่วยเสียงโดยใช้ตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียง เช่น ตำแหน่งที่ปุ่มเหงือก คือ หน่วยเสียง /ด/ และ /ท/ ตำแหน่งริมผีปาก คือ หน่วยเสียง /บ/ และ /พ/ และตำแหน่งเพดานอ่อน คือ หน่วยเสียง /ค/ และ /ก/

2.4 การรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ

การรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ [1] เป็นการรู้จำเสียงพูดที่ได้รับความนิยมมากวิธีหนึ่ง ด้วยเหตุผลสองประการ คือ 1) แบบจำลองฮิดเดนมาร์คอฟอาศัยโครงสร้างทางคณิตศาสตร์ และสามารถเปลี่ยนแปลงทฤษฎีพื้นฐานเพื่อประยุกต์ใช้งานได้อย่างกว้างขวาง และ 2) แบบจำลองฮิดเดนมาร์คอฟสามารถทำงานได้เป็นอย่างดีเมื่อประยุกต์ใช้อย่างเหมาะสม

แบบจำลองฮิดเดินมาร์คอฟ จะมีองค์ประกอบ ดังนี้

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

$$a_{ij}=P[q_{t+1}=S_j \mid q_t=S_i], \qquad 1\le i,\ j<N \tag{2.3}$$

ในกรณีเฉพาะที่สถานะใดๆ สามารถเข้าถึงสถานะอื่นๆได้ภายในขั้นตอนเดียว จะกำหนดให้ ส่วนในกรณีอื่นนอกเหนือจากนี้จะกำหนดให้ สำหรับ (i,j) เพียงคู่เดียวหรือมากกว่า

4) การแจกแจงของความน่าจะเป็นของสัญลักษณ์ของค่าสังเกต (Observation Symbol Probability Distribution) ใช้สัญลักษณ์แทนด้วยเมทริกซ์ (Matrix) "B" โดย $B=\{b_j(k)\}$ ในสถานะที่ j เมื่อ

$$b_j(k)=P[v_k \text{ at } t \mid q_t=S_j], \qquad \begin{matrix}1\le j\le N\\ 1\le k\le M\end{matrix} \tag{2.4}$$

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

$$\pi_i = P[q_1 = S_i], \qquad 1 \le i \le N \tag{2.5}$$

โดยการกำหนดค่าที่เหมาะสมให้กับองค์ประกอบ N, M, A, B, π ของแบบจำลองฮิดเดนมาร์คอฟซึ่งใช้ในการกำหนดลำดับค่าสังเกต

$$O = O_1 O_2 ... O_T \tag{2.6}$$

เมื่อแต่ละค่าสังเกต O_t เป็นสัญลักษณ์ที่ได้จาก V และ T เป็นจำนวนค่าสังเกตทั้งหมดที่มีในลำดับซึ่งมีขั้นตอนวิธีการดังนี้

1) เลือกสถานะเริ่มต้น $q_1 = S_i$ ที่สัมพันธ์กับการกระจายของสภาวะเริ่มต้น π
2) กำหนดให้ $t = 1$
3) เลือก $O_t = v_k$ ตามสัญลักษณ์ของค่าสังเกตในสถานะ S_i เช่น $b_i(k)$
4) เคลื่อนย้ายไปยังสถานะใหม่ $q_{t+1} = S_j$ ที่สัมพันธ์กับการกระจายความน่าจะเป็นในการเปลี่ยนแปลงสถานะสำหรับสถานะ S_i เช่น a_{ij}
5) กำหนดให้ $t = t + 1$ แล้วกลับไปทำซ้ำขั้นตอนที่ 3 ใหม่ถ้า $t < T$ นอกเหนือจากนี้ให้ยุติกระบวนการ

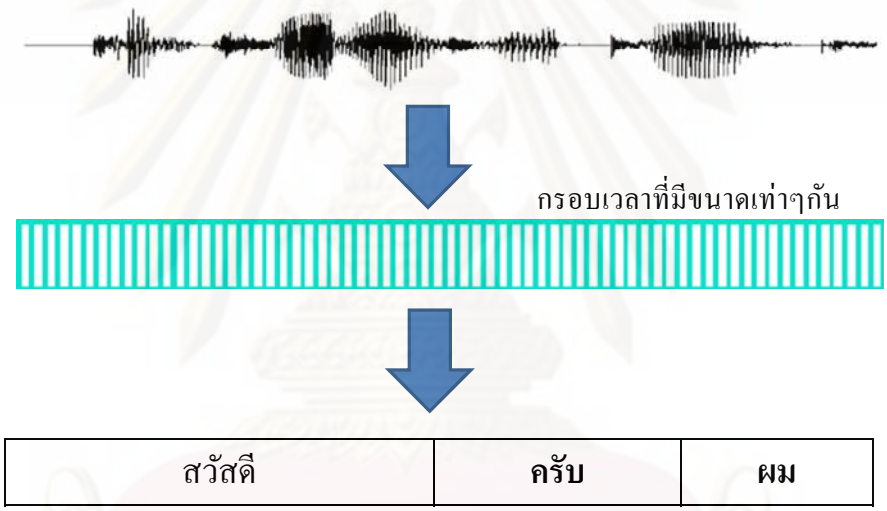
ขั้นตอนดังกล่าวนี้เป็นได้ทั้งการกำเนิดค่าสังเกต และเป็นแบบจำลองเพื่อบอกถึงความเหมาะสมในการกำเนิดลำดับค่าสังเกตด้วยแบบจำลองฮิดเดนมาร์คอฟ ดังนั้นการกำหนดคุณสมบัติเฉพาะของแบบจำลองฮิดเดนมาร์คอฟ จึงต้องการคุณสมบัติเฉพาะของพารามิเตอร์ของแบบจำลอง (นั่นคือ N และ M) คุณสมบัติเฉพาะของสัญลักษณ์ของค่าสังเกต และคุณสมบัติเฉพาะของการวัดค่าความน่าจะเป็นอันได้แก่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

$$\lambda = (A, B, \pi) \tag{2.7}$$

จากแบบจำลองฮิดเดนมาร์คอฟที่ได้กล่าวมาข้างต้น มีปัญหาพื้นฐาน 3 อย่างของแบบจำลองฮิดเดนมาร์คอฟที่จะต้องแก้ไขเพื่อที่จะนำแบบจำลองฮิดเดนมาร์คอฟมาใช้งานจริงได้แก่

1) เมื่อมีลำดับค่าสังเกต $O = O_1 O_2 ... O_T$ และแบบจำลอง $\lambda = (A, B, \pi)$ จะสามารถคำนวณหาค่าความน่าจะเป็นของลำดับค่าสังเกตเมื่อรู้แบบจำลอง $P(O|\lambda)$ ได้อย่างไร
2) เมื่อมีลำดับค่าสังเกต $O = O_1 O_2 ... O_T$ และแบบจำลอง $\lambda = (A, B, \pi)$ จะเลือกลำดับของสถานะอย่างไรให้มีความเหมาะสมที่สุด
3) จะสามารถปรับค่าพารามิเตอร์ $\lambda = (A, B, \pi)$ อย่างไร จึงจะทำให้ $P(O|\lambda)$ มีค่าสูงสุด

ในการแก้ไขปัญหาที่ 1) จะใช้อัลกอริทึมแบบเดินหน้าและย้อนกลับ (Forward-Backward Algorithm) ปัญหาที่ 2) จะแก้ไขโดยใช้อัลกอริทึมไวเทอร์บิ ส่วนปัญหาที่ 3) จะสามารถทำได้โดยใช้การกระบวนการประมาณซ้ำของ Baum-Welch



รูปที่ 2.2 การทำงานของการรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ

สำหรับการรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ ในขั้นตอนแรกก็จะมีการนำสัญญาณเสียงพูดมาแบ่งเป็นกรอบเวลาเท่าๆกัน จากนั้นก็จะมีการดึงเอาเวกเตอร์คุณสมบัติออกมาจากแต่ละกรอบเวลา เพื่อสร้างเป็นลำดับค่าสังเกต ดังรูปที่ 2.2 เวกเตอร์คุณสมบัติที่นิยมใช้การรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ คือ สัมประสิทธิ์เมลฟรีเคว็นซีเคปสตรอล (Mel-Frequency Cepstral Coefficients: MFCC) โดยปกติแล้วในแบบจำลองทางเสียงของการรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟมักจะมีการแทนแต่ละหน่วยเสียงด้วยแบบจำลองฮิดเดนมาร์คอฟหนึ่งตัว ดังรูปที่ 2.3 ซึ่งจะเป็นแบบจำลองฮิดเดนมาร์คอฟแบบซ้ายไปขวา (Left to Right) ที่มีจำนวนสถานะ 3 สถานะ (ไม่รวมสถานะเริ่มต้น และสถานะสิ้นสุด) ส่วนการค้นหาคำตอบก็จะเป็น

การค้นหาลำดับของคำ $W* = w_1,...,w_N$ ที่มีค่าความน่าจะเป็นภายหลังที่มากที่สุด (Maximum a Posteriori Probability: MAP) ของ $P(W|O)$ ดังสมการ

$$W^* = \arg\max_W P(W|O) \tag{2.8}$$

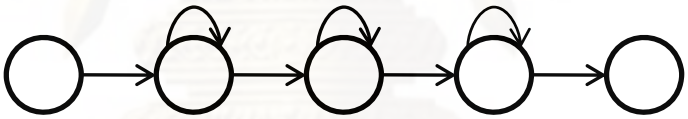
จากสมการ (2.8) เมื่อใช้กฎของเบส์ (Bayes' rule) จะได้

$$W^* = \arg\max_W \frac{P(O|W)P(W)}{P(O)} \tag{2.9}$$

เมื่อ $P(O)$ ไม่ขึ้นอยู่กับ $P(O|W)$ และ $P(W)$ จึงได้ว่า

$$W^* = \arg\max_W P(O|W)P(W) \tag{2.10}$$

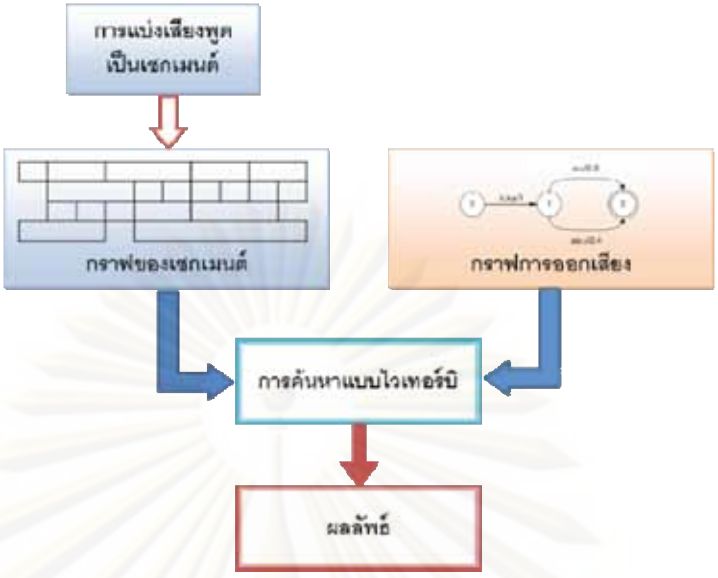
โดยที่ $P(O|W)$ เป็นแบบจำลองทางเสียง (Acoustic Model) และ $P(W)$ เป็นแบบจำลองทางภาษา (Language Model)



รูปที่ 2.3 แบบจำลองฮิดเดนมาร์คอฟ

2.5 การรู้จำเสียงพูดแบบอาศัยเซกเมนต์

การรู้จำเสียงพูดแบบอาศัยเซกเมนต์จะทำงานโดยนำเสียงพูดมาตัดแบ่งเป็นเซกเมนต์เพื่อสร้างเป็นกราฟของเซกเมนต์ หลังจากนั้นก็จะดำเนินการค้นหาคำตอบที่ให้ค่าความน่าจะเป็นสูงสุดออกมาเป็นคำตอบ โดยวิธีการที่ใช้ในการค้นหานั้นจะใช้วิธีการค้นหาแบบไวเทอร์บิ ดังที่แสดงในรูปที่ 2.4



รูปที่ 2.4 การทำงานของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

2.5.1 หลักความน่าจะเป็นที่ใช้สำหรับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

ในหัวข้อนี้จะกล่าวถึงหลักความน่าจะเป็นที่ใช้สำหรับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ [19] กล่าวโดยจุดมุ่งหมายของการรู้จำเสียงพูด คือ ค้นหาลำดับของคำ $W* = w_1,...,w_N$ ที่ทำให้ค่าความน่าจะเป็นภายหลังมากสุดของ $P(W \mid A)$ โดยที่ A เป็นเซตของค่าสังเกตของเสียง (Acoustic Observations) ที่ได้จากสัญญาณเสียงพูด

$$W^* = \arg\max_W P(W \mid A) \tag{2.11}$$

ในขั้นตอนการรู้จำของการรู้จำเสียงพูดส่วนใหญ่จะแบ่งเสียงพูด (S) จากสัญญาณเสียงพูดให้เป็นลำดับของหน่วยเสียง (U) ดังนั้นจะสามารถเขียนสมการ (2.11) ได้ใหม่เป็น

$$W^* = \arg\max_W \sum_{\forall S,U} P(S,U,W \mid A) \approx \arg\max_{S,U,W} P(S,U,W \mid A) \tag{2.12}$$

เราสามารถแปลง $P(S,U,W \mid A)$ โดยใช้กฎของเบส์ ได้ดังนี้

$$P(S,U,W \mid A) = \frac{P(A \mid S,U,W)P(S \mid U,W)P(U \mid W)P(W)}{P(A)} \tag{2.13}$$

เมื่อ $P(A)$ เป็นอิสระต่อ S, U และ W จึงไม่มีผลในการค้นหา ดังนี้เราจึงไม่สนใจความน่าจะเป็นในส่วนนี้

$P(W)$ คือ แบบจำลองทางภาษา

$P(U|W)$ จะพิจารณาเป็นแบบจำลองการออกเสียง (Pronunciation Model) ซึ่งจะเป็นตัวบอกความน่าจะเป็นของลำดับของหน่วยย่อยของคำ (Sub-word units: U) ที่สร้างขึ้นมาตามลำดับของคำ W ซึ่งโดยทั่วไปจะใช้วิธีการค้นหาจากในพจนานุกรม

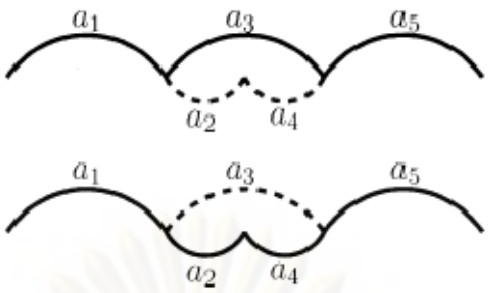
$P(S|U,W)$ จะพิจารณาเป็นความน่าจะเป็นของการแบ่งเป็นเซกเมนต์นั้นๆ ซึ่งปกติจะขึ้นอยู่กับ U นอกจากนี้ยังสามารถพิจารณาเป็นแบบจำลองระยะเวลา (Duration Model) ซึ่งจะไว้ทำนายความน่าจะเป็นของระยะเวลาของเซกเมนต์นั้นๆ

$P(A|S,U,W)$ คือ แบบจำลองเสียง ซึ่งในที่นี้ได้ตั้งสันนิษฐานว่า A ไม่ขึ้นอยู่กับ W เมื่อรู้ U อย่างมีเงื่อนไข ดังนั้น $P(A|S,U,W) = P(A|S,U)$

2.5.2 การจำลองแบบจำลองเสียงของเซกเมนต์

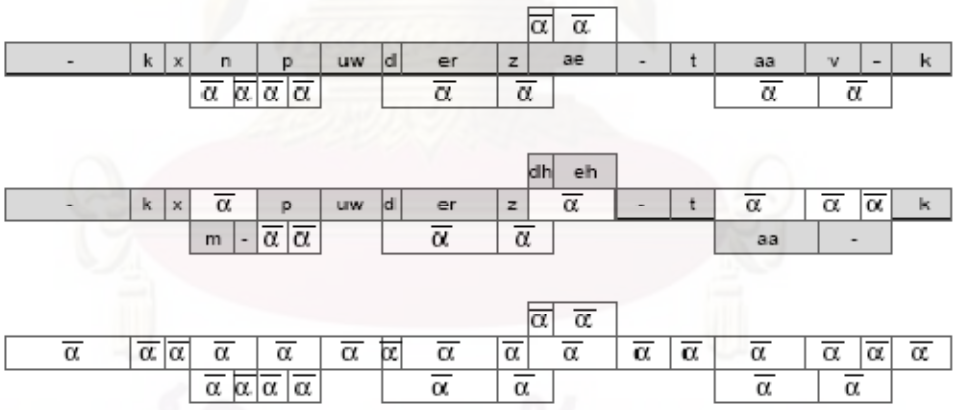
วิธีการจำลองแบบจำลองเสียงของเซกเมนต์ในการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ [19] จะแทนเซกเมนต์ที่มีความยาวไม่แน่นอนด้วย s_i และแต่ละเซกเมนต์นั้นจะมีเวกเตอร์คุณสมบัติของเสียงที่มีจำนวนมิติที่แน่นอน x_i ดังนั้น ค่าสังเกต A ประกอบไปด้วยทุกๆเวกเตอร์คุณสมบัติของเสียงในเซกเมนต์กราฟ ถ้าจำนวนมีเซกเมนต์ n จำนวน ดังนั้น $S = \{s_1, \ldots, s_n\}$ และจะมีเวกเตอร์คุณสมบัติของเสียง n จำนวนเช่นกัน ดังนั้น $X = \{x_1, \ldots, x_n\}$

ดังรูปที่ 2.5 จะแสดงถึงกราฟของเซกเมนต์ที่มีการแบ่งเสียงพูดเป็นเซกเมนต์อยู่สองแบบหรือสองเส้นทาง ในแต่ละเส้นทางของการแบ่งเสียงพูดเป็นเซกเมนต์ (S) จะประกอบไปด้วยลำดับของเซกเมนต์ที่ต่อเนื่องกันที่มีจำนวนของค่าสังเกต (X) ที่แตกต่างกัน ดังนั้นการที่จะเปรียบเทียบค่าความเป็นไปได้ $P(X|S,U)$ โดยตรงนั้นจึงไม่ถูกต้อง ในการเปรียบเทียบสองเส้นทางนั้นจะต้องมีการพิจารณาค่าสังเกตทั้งหมด ดังนั้นถ้าจะมองว่าค่าสังเกต X จะเป็นเส้นทางที่พิจารณาที่เกิดจากการแบ่งเสียงพูดเป็นเซกเมนต์ S เราจะต้องพิจารณาทุกๆค่าสังเกตในเส้นทางอื่นๆที่เรียกว่า Y ด้วย ดังนั้น $X \cap Y = \emptyset$ และ $X \cup Y = A$ ตัวอย่างเช่น ในเส้นทางบนของรูปที่ 2.5 จะมี X และ Y เป็นดังนี้ $X = \{a_1, a_3, a_5\}$ และ $Y = \{a_2, a_4\}$ ส่วนในเส้นทางล่างนั้น $X = \{a_1, a_2, a_4, a_5\}$ และ $Y = \{a_3\}$ ในทุกๆค่าสังเกตใน A จะต้องประกอบไปด้วย X และ Y ดังนั้น การถอดรหัสหาค่าความน่าจะเป็นภายหลังมากสุดจะต้องประมาณค่าความเป็นไปได้ $P(X,Y|S,U)$ และเนื่องจาก X สามารถแสดงนัยถึง S ได้ เราจึงสามารถพิจารณาได้ว่า $P(X,Y|S,U) = P(X,Y|U)$



รูปที่ 2.5 กราฟของเซกเมนต์ที่มีการแบ่งเสียงพูดเป็นเซกเมนต์อยู่สองแบบ (ในเส้นทึบ) ที่ประกอบไปด้วยค่าสังเกต 5 แบบ $\{a_1,\ldots,a_5\}$ การแบ่งเสียงพูดเป็นเซกเมนต์แบบบนจะประกอบไปด้วยค่าสังเกต $\{a_1,a_3,a_5\}$ ส่วนการแบ่งเสียงพูดเป็นเซกเมนต์แบบล่างนั้นจะประกอบไปด้วยค่าสังเกต $\{a_1,a_2,a_4,a_5\}$ [19]

ในการจำลอง *P(X,Y|S,U)* จะทำได้โดยเพิ่มหน่วยในพจนานุกรมพิเศษที่เรียกว่า "หน่วยที่ไม่ใช่หน่วยเสียง (Anti-phone: $\bar{\alpha}$)" ขึ้นมาเพื่อใช้แทนเซกเมนต์ที่ไม่ได้ปรากฏอยู่ในหน่วยเสียง *U* ซึ่งอาจจะเป็นเสียงที่สั้นเกินไป ยาวเกินไป หรือมีการซ้อนทับกัน เป็นต้น จากรูปที่ 2.6 มีสองเส้นทางที่ต้องพิจารณาเป็นคำตอบจะต้องมีการจำลองแต่ละเซกเมนต์ด้วยหน่วยเสียง *U* หรือหน่วยที่ไม่ใช่หน่วยเสียง $\bar{\alpha}$



รูปที่ 2.6 วิธีการจำลองโดยใช้หน่วยที่ไม่ใช่หน่วยเสียง [19]

จากรูปที่ 2.6 แสดงให้เห็นว่าวิธีนี้ จะให้ทุกๆเซกเมนต์ที่ไม่ได้อยู่ในเส้นทางที่กำลังพิจารณาด้วยหน่วยที่ไม่ใช่หน่วยเสียง $\bar{\alpha}$ ในกราฟของเซกเมนต์อันล่างสุดจะแสดงวิธีการที่แทนทุกเซกเมนต์ด้วยหน่วยที่ไม่ใช่หน่วยเสียง เพื่อหลีกเลี่ยงการที่จะต้องจำแนกทุกๆเซกเมนต์ในการค้นหา โดยการมีรู้จำว่าค่าความเป็นไปได้ที่ทุกๆเซกเมนต์จะไม่ใช่หน่วยเสียงในพจนานุกรม $P(X,Y|\bar{\alpha})$ จะเป็นค่าคงที่ในกราฟที่กำหนด ดังนั้นมันจะไม่มีผลในขั้นตอนการค้นหาคำตอบ ถ้าเราสมมุติว่า *X*

และ Y จะไม่ขึ้นต่อกันอย่างมีเงื่อนไขเมื่อรู้ U และ $P(Y|U)$ ในขึ้นอยู่กับ $\bar{\alpha}$ เท่านั้น เราจะสามารถเขียนสมการของ $P(X,Y|U)$ ได้เป็นดังนี้

$$P(X,Y \mid U) = P(X \mid U)P(Y \mid \bar{\alpha})\frac{P(X \mid \bar{\alpha})}{P(X \mid \bar{\alpha})} \propto \frac{P(X \mid U)}{P(X \mid \bar{\alpha})} \tag{2.14}$$

เมื่อพิจารณาการแบ่งเสียงพูดเป็นเซกเมนต์ S หนึ่งๆจะสนใจเฉพาะค่าสังเกตจำนวน n ค่าที่เกี่ยวข้องกับ S ในแต่ละเซกเมนต์ s_i ต้องสนใจสองส่วนด้วยกัน คือ ค่าความเป็นไปได้ที่จะเป็นหน่วยเสียง $P(x_i|U)$ (หรือ $P(x_i|u_i)$) และค่าความเป็นไปได้ที่เซกเมนต์จะเป็นหน่วยที่ไม่ใช่หน่วยเสียง $P(x_i|\bar{\alpha})$ ซึ่งจะได้สมการที่ต้องหาค่าสูงสุดในการค้นหาดังนี้

$$W^* = \arg\max_{S,U,W} \prod_{i=1}^{n} \frac{P(x_i \mid u_i)}{P(x_i \mid \bar{\alpha})} P(s_i \mid u_i)P(U \mid W)P(W) \tag{2.15}$$

2.5.3 การจำลองแบบจำลองเสียงของขอบเขตของหน่วยเสียง

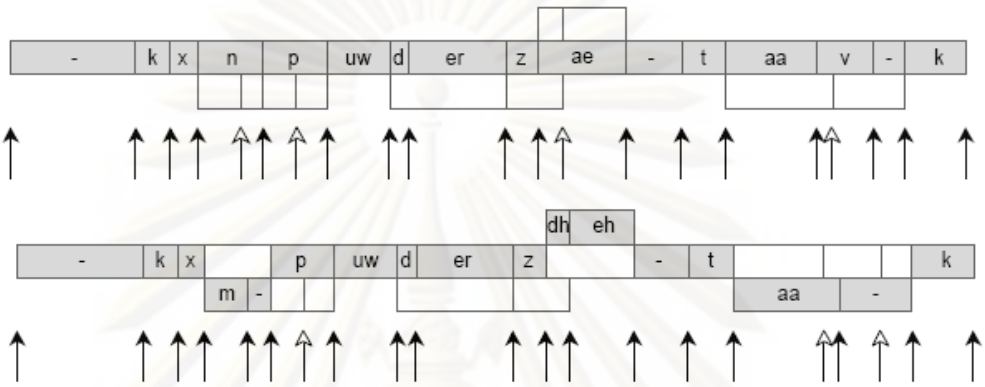
นอกจากจะมีการจำลองในส่วนของเซกเมนต์แล้ว การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ก็นิยมที่จะจำลองข้อมูลที่เกี่ยวกับขอบเขตของหน่วยเสียง หรือแลนด์มาร์ค (Landmarks) เพิ่มเติม [19] ถ้าเรากำหนดให้ค่าสังเกตของเสียงที่ดึงที่ตำแหน่งของแลนด์มาร์ด Z เราจะต้องพิจารณาค่าสังเกตทั้งหมดเป็น $A = XYZ$ ดังนั้นจะสามารถทำเป็นค่าความน่าจะเป็น $P(X, Y, Z, | S, U)$ ถ้าพิจารณาโดยสมมุติว่าค่าสังเกตของเซกเมนต์ XY และแลนด์มาร์ด Z ไม่ขึ้นต่อกัน เราจะสามารถเขียนสมการได้ดังนี้

$$P(X,Y,Z \mid S,U) = P(X,Y \mid S,U)P(Z \mid S,U) \tag{2.16}$$

ถ้ากำหนดให้ Z เป็นเซตของค่าสังเกตที่ดึงจากขอบเขตของหน่วยเสียง หรือแลนด์มาร์ด ซึ่งจะแบ่งเป็นสองส่วนด้วยกันคือ แลนด์มาร์ดที่อยู่ระหว่างหน่วยเสียงในพจนานุกรม และแลนด์มาร์ดที่อยู่ภายในหน่วยเสียง ดังที่แสดงในรูปที่ 2.7 จะเห็นว่าจำนวนแลนด์มาร์ดในแต่ละเส้นทางที่ค้นหามีจำนวนที่เท่ากัน ดังนั้นจะไม่จำเป็นจะต้องมีการทำให้เป็นบรรทัดฐานเหมือนกับการจำลองเซกเมนต์ ถ้าเราสมมุติว่าค่าสังเกตของขอบเขตของหน่วยเสียงแต่ละตัวที่ m ที่อยู่ใน Z นั้นเป็นอิสระต่อกันเมื่อรู้ U ดังนั้น จะสามารถเขียน $P(Z \mid S, U)$ ได้ดังนี้

$$P(Z \mid S,U) = \prod_{i=1}^{m} P(z_i \mid S,U) \qquad (2.17)$$

เมื่อ z_i เป็นค่าสังเกตที่ดึงมาจากแลนด์มาร์คที่ i



รูปที่ 2.7 การจำลองของเขตของหน่วยเสียง

โดยในรูปจะแสดงสองเส้นทางในกราฟของเซกเมนต์เดียวกัน ลูกศรหัวทึบจะเป็นแลนด์มาร์คที่อยู่ระหว่างหน่วยเสียงและลูกศรหัวโปร่งจะเป็นแลนด์มาร์คที่อยู่ภายในหน่วยเสียง [19]

2.6 การแบ่งเสียงพูดเป็นเซกเมนต์

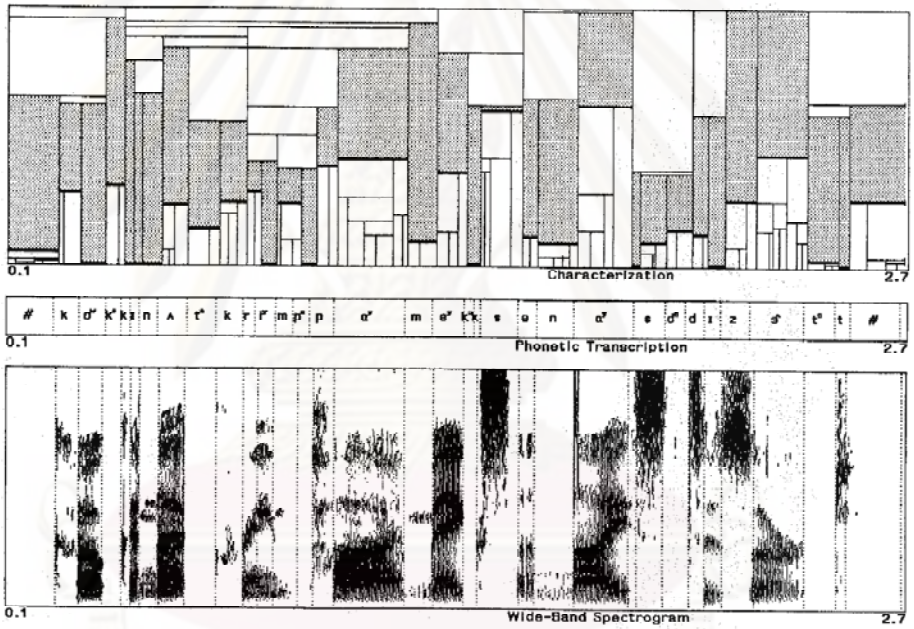
การแบ่งเสียงพูดเป็นเซกเมนต์ที่ใช้สำหรับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์นั้นแบ่งเป็นสองกลุ่ม คือ การแบ่งเสียงพูดเป็นเซกเมนต์ด้วยการค้นหาตำแหน่งที่มีการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียง (Acoustic Segmentation) และการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น

2.6.1 การแบ่งเสียงพูดเป็นเซกเมนต์ด้วยการค้นหาตำแหน่งที่มีการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียง

การแบ่งเสียงพูดเป็นเซกเมนต์ด้วยการค้นหาตำแหน่งที่มีการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียง อาศัยหลักการที่ว่าสัญญาณเสียงที่บริเวณขอบเขตของหน่วยเสียงจะมีความไม่ต่อเนื่องสูงกว่าบริเวณที่เป็นหน่วยเสียง Glass และ Zue [20, 21] ได้ศึกษาหาขอบเขตของหน่วยเสียงโดยวัดค่าระยะห่างยูคลีเดียน (Euclidean Distance) ของเวกเตอร์ทางสเปกตรัม (Spectral Vector) ระหว่างกรอบเวลาทั้งสองกรอบเวลาที่อยู่ติดกันว่ามีระยะห่างยูคลีเดียนมากหรือน้อยเพียงใด ถ้า

ระยะห่างยูคลีเดียนมีมากแสดงว่าสัญญาณเสียงมีความแตกต่างกันมาก และมีโอกาสที่จะเป็นขอบเขตของหน่วยเสียง

ขั้นตอนในการสร้างกราฟของเซกเมนต์ของ Glass และ Zue จะใช้วิธีที่เรียกว่า “การแบ่งเสียงพูดเป็นเซกเมนต์โดยอาศัยการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียงแบบหลายระดับ (Multi-level Acoustic Segmentation)” ซึ่งมีขั้นตอนดังนี้ 1) ทำให้ทุกๆขอบเขตของกรอบเวลาเป็นเสมือนขอบเขตของหน่วยเสียง 2) หาการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียงระหว่างกรอบเวลาด้านซ้าย และขวาของขอบเขตของหน่วยเสียงโดยใช้ระยะห่างยูคลีเดียน เมื่อค่าระยะห่างยูคลีเดียนมีค่าน้อยจะยุบขอบเขตนั้นลง ทำต่อไปเรื่อยๆจนไม่สามารถยุบขอบเขตลงได้ โดยในการยุบขอบเขตแต่ละครั้งก็จะเก็บค่าไว้ให้อยู่ในรูปของต้นไม้ที่เรียกว่า “เดนโดรแกรม (Dendrogram)” ดังที่แสดงในรูปที่ 2.8 กราฟของเซกเมนต์จะได้จากการเลือกระดับที่ต้องการจากเดนโดรแกรม

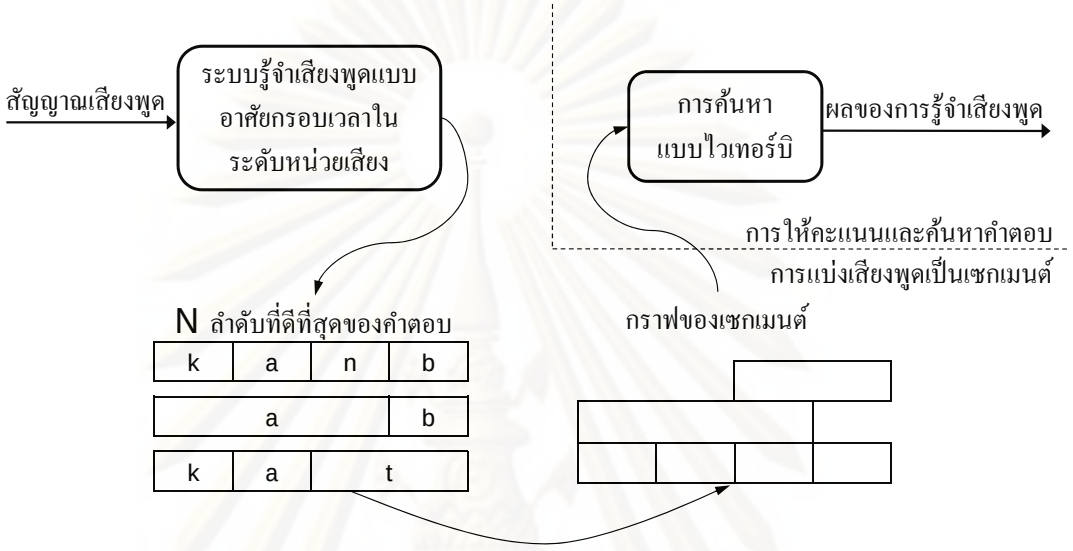


รูปที่ 2.8 การแบ่งเสียงพูดเป็นเซกเมนต์แบบหลายระดับ [20]

2.6.2 การแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น

การแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีการความน่าจะเป็นหรือเรียกอีกอย่างว่า “การแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยการรู้จำเสียง (Segmentation by Recognition)” [8-10] จะหาขอบเขตของหน่วยเสียงโดยอาศัยระบบรู้จำเสียงพูดแบบอาศัยกรอบเวลาในระดับหน่วยเสียงมารู้จำเสียงพูดออกมาเป็นลำดับของหน่วยเสียง N ลำดับที่ดีที่สุด แล้วจึงนำแต่ละเซกเมนต์ของหน่วยเสียงที่รู้จำได้มาประกอบรวมกันเป็นกราฟของเซกเมนต์ ดังที่แสดงในรูปที่ 2.9 โดยที่ N ซึ่งเป็นตัวแปรที่กำหนดจำนวนลำดับที่ดีที่สุดที่ต้องการรู้จำ ซึ่งจะเป็นตัวกำหนดขนาดของกราฟของเซกเมนต์

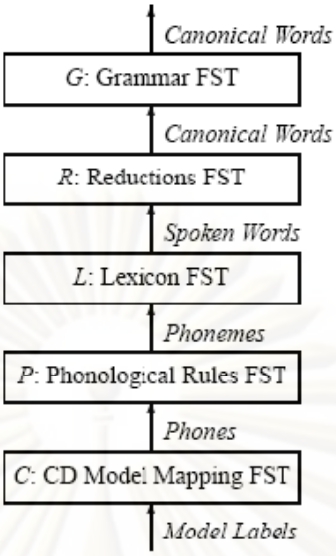
กล่าวคือถ้าจำนวนลำดับที่ดีที่สุดมีจำนวนมากขนาดของกราฟของเซกเมนต์ก็จะมีขนาดใหญ่ขึ้น ระบบรู้จำเสียงพูดในที่นี้จะสร้างโดยอาศัยแบบจำลองฮิดเดนมาร์คอฟ และจะค้นหาลำดับของหน่วยเสียงที่ดีที่สุดได้จากการค้นหาด้วยวิธีของไวเทอร์บิแบบไปข้างหน้า (Forward Viterbi) และวิธี *A** แบบย้อนกลับ (Backward *A**)



รูปที่ 2.9 การแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น [9]

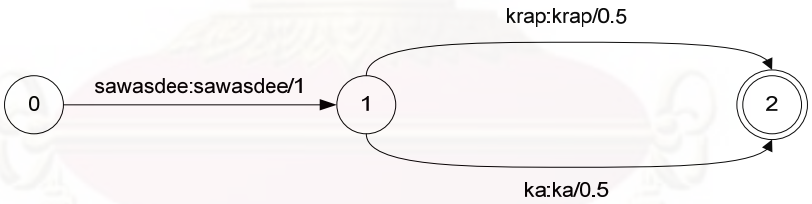
2.7 การสร้างกราฟการออกเสียงด้วยตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก

ในการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ได้มีการนำตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักมาประกอบกันเพื่อสร้างเป็นกราฟการออกเสียง เพื่อใช้เป็นตัวกำหนดเกี่ยวกับทางภาษาในขั้นตอนการค้นหาคำตอบ ซึ่งกราฟการออกเสียงจะประกอบไปด้วย $U = C \circ P \circ L \circ R \circ G$ โดยที่ U จะเป็นกราฟการออกเสียง C จะแทนแบบจำลองบริบทไม่อิสระ (Context-Dependent Model) P จะแทนกฎของระบบเสียง (Phonological Rule) L จะเป็นการแปลงจากการออกเสียงให้เป็นคำศัพท์ในพจนานุกรม (Lexicon) R จะแทนในส่วนของคำย่อต่างๆ และ G จะแทนแบบจำลองทางภาษาหรือไวยากรณ์ ซึ่งลักษณะของการนำตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักมาประกอบกันจะเป็นดังรูปที่ 2.10



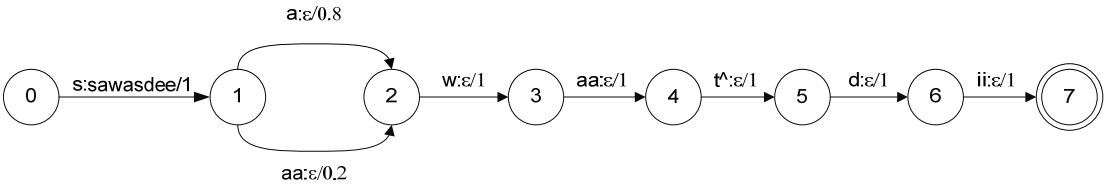
รูปที่ 2.10 การนำตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักมาประกอบกันเพื่อใช้สำหรับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ [6]

ตัวอย่างเช่น การรู้จำเสียงแบบหนึ่งกำหนดให้ลำดับของคำที่สามารถรู้จำได้เป็น “สวัสดีครับ” หรือ “สวัสดีค่ะ” รูปที่ 2.11 จะเป็นตัวอย่างของตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักที่ทำหน้าที่เป็นแบบจำลองทางภาษา

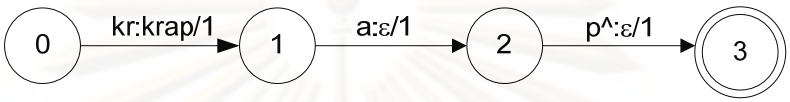


รูปที่ 2.11 ตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักที่ทำหน้าที่เป็นแบบจำลองทางภาษาของประโยค “สวัสดีครับ” และ “สวัสดีค่ะ”

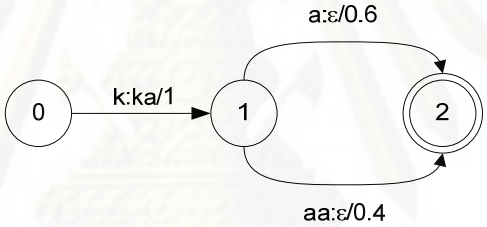
รูปที่ 2.12 รูปที่ 2.13 และรูปที่ 2.14 จะเป็นรูปของตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักที่ทำหน้าที่เป็นกฎของระบบเสียงจะสังเกตเห็นว่าคำว่า “สวัสดี” และคำว่า “ค่ะ” จะสามารถออกเสียงได้หลายแบบ



รูปที่ 2.12 ตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักที่ทำหน้าที่เป็นกฎของระบบเสียงของคำว่า “สวัสดี”

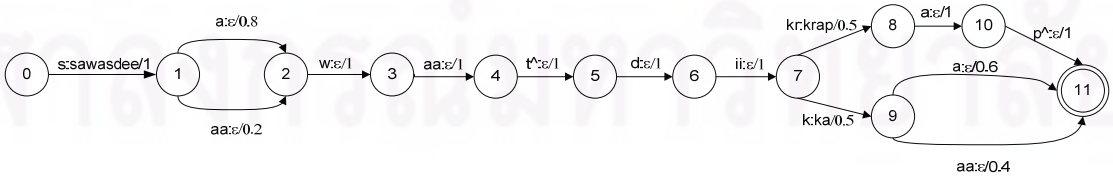


รูปที่ 2.13 ตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักที่ทำหน้าที่เป็นกฎของระบบเสียงของคำว่า “ครับ”



รูปที่ 2.14 ตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักที่ทำหน้าที่เป็นกฎของระบบเสียงของคำว่า “ค่ะ”

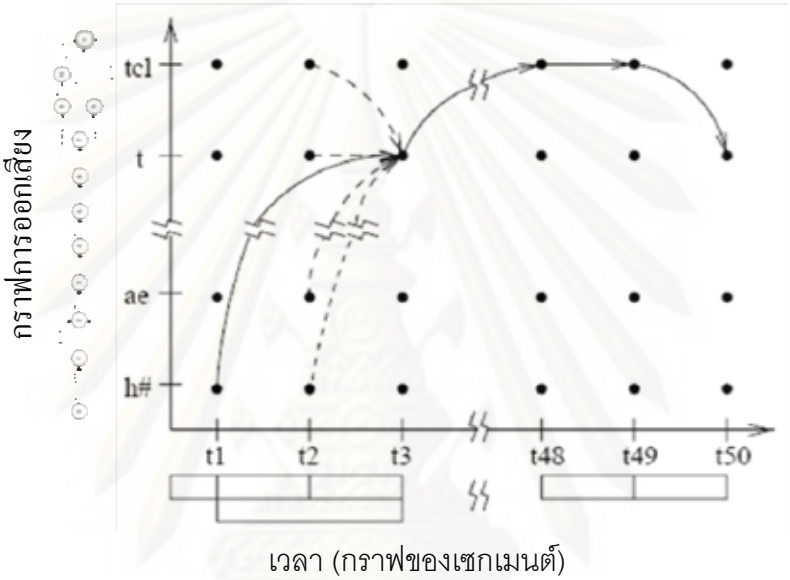
รูปที่ 2.15 จะเป็นการนำตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนักที่เป็นกฎของระบบเสียง (P) และแบบจำลองทางภาษา (G) มาประกอบกันเป็นกราฟการออกเสียง (U) โดยที่ $U = P \circ G$ ซึ่งในการค้นหาคำตอบของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์นั้นจะต้องมาค้นหาในกราฟการออกเสียงด้วย



รูปที่ 2.15 กราฟการออกเสียงของประโยค “สวัสดีครับ” และ “สวัสดีค่ะ”

2.8 การค้นหาแบบไวเทอร์บิสำหรับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

การรู้จำเสียงพูดแบบอาศัยเซกเมนต์จะใช้อัลกอริทึมของไวเทอร์บิในการค้นหาเส้นทางที่มีความน่าจะเป็นสูงสุดจากกราฟของเซกเมนต์ (แกนนอนในรูปที่ 2.16) และกราฟการออกเสียง (แกนตั้งในรูปที่ 2.16) เพื่อนำมาเป็นคำตอบ โดยที่กราฟของเซกเมนต์จะสร้างจากการแบ่งเสียงพูดเป็นเซกเมนต์ และกราฟการออกเสียงจะสร้างจากตัวเปลี่ยนแปรสถานะจำกัดแบบถ่วงน้ำหนัก หลักของการค้นหาโดยใช้อัลกอริทึมของไวเทอร์บิจะแสดงในรูปที่ 2.16



รูปที่ 2.16 การค้นหาแบบไวเทอร์บิสำหรับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ [9]

การค้นหาโดยใช้อัลกอริทึมของไวเทอร์บิสำหรับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์จะมีอัลกอริทึมดังนี้ [9]

for each boundary b_{to} in the utterance

let best_score(b_{to}) = -∞

for each segment s that terminates at boundary b_{to}

let b_{from} be the staring boundary that terminates at boundary b_{to}

let $\vec{x}$ be the measurement vector for segment s y_b

let $\overrightarrow{y_b}$ be the measurement vector from boundary b_{from}

let $\overrightarrow{y_i}$[] be the array of boundary measurement vectors for every frame from b_{from+1} to b_{to-1}

for each node n_{to} in the pronunciation network

for each pronunciation arc a arriving at node n_{to}

let n_{from} be the source node of arc a

let b be the pronunciation arc arriving at node n_{from}

if (n_{from}, n_{from}) has not been pruned from the Viterbi lattice

let α be the label on arc a

let $\overline{\alpha}$ be the anti-phone label

let β_b be the label for the transition boundary b → a

let β_i be the label for the internal boundary a → a

let acoustic_score = $p(\vec{x}|\alpha) - p(\vec{x}|\overline{\alpha}) + p(\overrightarrow{y_b}|\beta_b) + p(\overrightarrow{y_i}|\beta_i)$

let duration_score = stw if b ≠ a, or 0 if b = a

let language_score = $p(\beta_b)$

let score = acoustic_score + duration_score + language_score

if (score(n_{from}, b_{from}) + score > score(n_{to}, b_{to}))

score(n_{to}, b_{to}) = score(n_{from}, b_{from}) + score

make a back pointer from score(n_{to}, b_{to}) to score(n_{from}, b_{from})

if score(n_{to}, b_{to}) > best_score(b_{to})

let best_score(b_{to}) = score(n_{to}, b_{to})

for each node n_{to} in the pronunciation network

if best_score(b_{to}) - score(n_{to}, b_{to}) > thresh

prune node (n_{to}, b_{to}) from the Viterbi lattice

2.9 การสกัดพารามิเตอร์คุณสมบัติทางเสียง (Acoustic Parameters: APs)

2.9.1 ค่าพลังงาน (Energy)

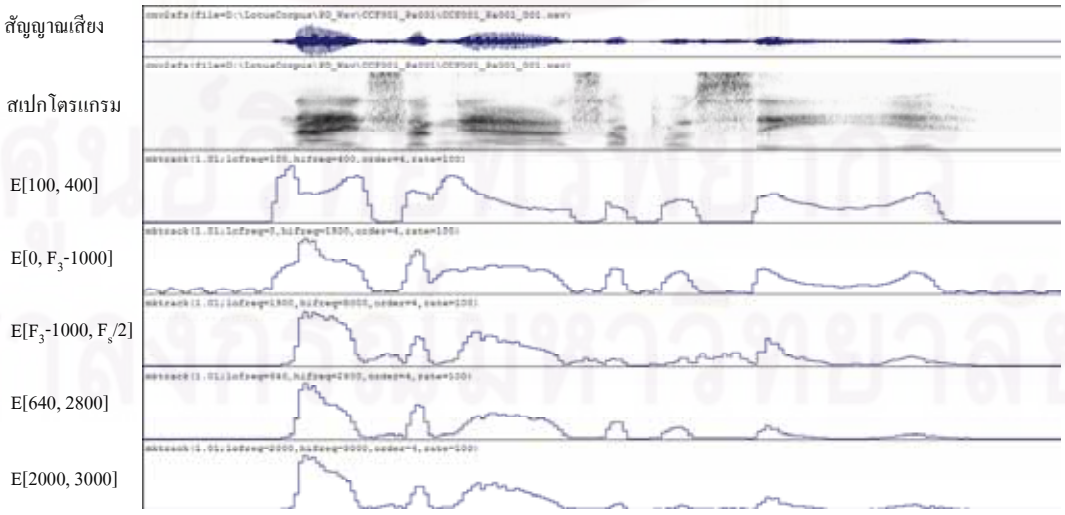
พลังงานของสัญญาณ [22] เป็นค่าลักษณะสำคัญที่นิยมนำมาใช้วิเคราะห์สัญญาณเสียง เนื่องจากเสียงพูดแต่ละชนิดจะมีค่าพลังงานของสัญญาณแตกต่างกันออกไป โดยค่าพลังงานของสัญญาณ $s(n)$ ใดๆที่แปรตามเวลาสามารถนิยาม ดังสมการต่อไปนี้

$$E = \sum_{n=-\infty}^{+\infty} s(n)^2 \tag{2.18}$$

โดยในการสกัดค่าพลังงานของสัญญาณเสียงจะต้องแบ่งสัญญาณออกมาพิจารณาเป็นกรอบเวลาด้วยฟังก์ชันหน้าต่าง $w(m)$ ที่มีขนาด N ดังนั้น ค่าพลังงานของสัญญาณเสียงที่กรอบเวลาที่ m เขียนแทนด้วยสัญลักษณ์ $E(m)$ จะสามารถคำนวณได้ดังสมการต่อไปนี้

$$E(m) = \sum_{n=0}^{N-1} [w(m)s(m-n)]^2 \tag{2.19}$$

อย่างไรก็ตามในการจำแนกเสียงต่างๆออกจากกัน มักจะใช้ค่าพลังงานในช่วงความถี่ต่างๆ ซึ่งในการหาค่าพลังงานในช่วงความถี่ต่างๆ จะดำเนินการโดยกรองความถี่ของสัญญาณเสียงโดยใช้ตัวกรองความถี่ชนิดแถบความถี่ผ่าน (Band-pass filter) ซึ่งจะยอมให้สัญญาณที่มีความถี่ในขอบเขตที่กำหนดผ่านไปได้และจะกรองสัญญาณที่ความถี่ส่วนอื่นๆทิ้ง แล้วจึงนำสัญญาณที่กรองได้นี้ไปคำนวณค่าพลังงานของสัญญาณเสียงในช่วงความถี่ที่ต้องการ ซึ่งผลการหาค่าพลังงานในช่วงความถี่ต่างๆ แสดงดังรูปที่ 2.17



รูปที่ 2.17 ค่าพลังงานของสัญญาณเสียงบนช่วงความถี่ต่างๆ

2.9.2 อัตราการตัดศูนย์ (Zero Crossing Rate: ZCR)

ค่าอัตราการตัดศูนย์ของสัญญาณเสียง [22] คือ จำนวนครั้งของการเปลี่ยนแปลงเครื่องหมายจากบวกเป็นลบหรือจากลบเป็นบวกของแอมพลิจูด (Amplitude) ของสัญญาณเสียงต่อหนึ่งหน่วยเวลา ค่าอัตราการตัดศูนย์นี้นิยมใช้กันมากในระบบรู้จำเสียงพูด โดยสามารถนำมาใช้วัดระดับเสียงรบกวน หรือนำมาไปประยุกต์วัดระดับความก้องของสัญญาณเสียงได้ อัตราการตัดศูนย์สามารถคำนวณได้จากสมการ ดังต่อไปนี้

$$zcr = \frac{1}{T}\sum_{t=0}^{T-1}\prod\{s_t s_{t-1} < 0\} \tag{2.20}$$

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

2.9.3 อัตสหสัมพันธ์ (Autocorrelation coefficients)

ค่าอัตสหสัมพันธ์ [22] เป็นค่าลักษณะสำคัญที่สามารถนำมาใช้วัดระดับความเป็นคาบของสัญญาณเสียง โดยพื้นฐานแล้วค่าอัตสหสัมพันธ์สามารถคำนวณได้จากการเปรียบเทียบสัญญาณเสียงที่เวลาหนึ่งกับสัญญาณเสียงเดียวกันที่ตำแหน่งซึ่งเยื้อง หรือหน่วงเวลาออกไปดังแสดงด้วยรูปที่ 2.18 จากกราฟด้านบนของรูปจะเป็นสัญญาณเสียงสระซึ่งมีลักษณะเป็นคาบ ซึ่งการหาค่าอัตสหสัมพันธ์จะพิจารณาสัญญาณเสียงที่ตำแหน่งที่ต้องการหาค่าอัตสหสัมพันธ์ว่ามีความคล้ายกันกับสัญญาณเสียงที่ตำแหน่งที่มีการหน่วงเวลาไปมากน้อยแค่ไหน

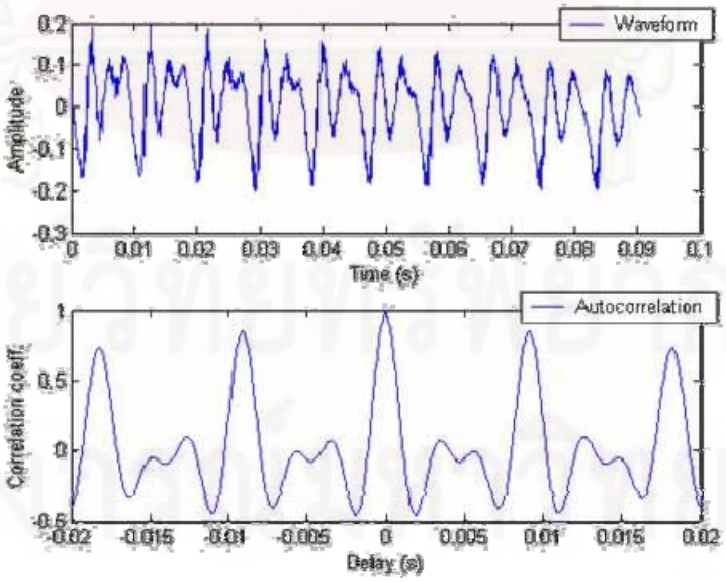
ค่าอัตสหสัมพันธ์ $autocorr(m,k)$ ของสัญญาณเสียงที่กรอบเวลาที่ m เมื่อหน่วงเวลาไปเป็นระยะเวลา k สามารถหาได้จากสมการต่อไปนี้

$$autocorr(m,k) = \frac{1}{N\sigma^2}\sum_{n=1}^{N}\left[s(n)w(m-n)-\mu\right]\left[s(n+k)w(m-n+k)-\mu\right] \tag{2.21}$$

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

รูปที่ 2.18 สัญญาณเสียงสระที่ระยะเวลาหน่วงต่างๆ

สัญญาณเสียงที่มีลักษณะเป็นคาบจะมีค่าอัตสหสัมพันธ์สูงเมื่อหน่วงเวลาไปเป็นจำนวนเท่าของคาบของสัญญาณเสียงนั้น ดังรูปที่ 2.19 จากกราฟในรูปจะสังเกตเห็นว่าค่าอัตสหสัมพันธ์จะมีค่าสูงสุดเมื่อไม่มีการหน่วงเวลา และค่าอัตสหสัมพันธ์สูงสุดอันดับต่อๆไปจะอยู่ที่ระยะหน่วงเวลาซึ่งเป็นจำนวนเท่าของคาบของสัญญาณเสียงนั้น



รูปที่ 2.19 สัญญาณเสียงที่มีลักษณะเป็นคาบ และค่าอัตสหสัมพันธ์ที่ระยะเวลาหน่วงต่างๆ

2.9.4 ค่าระดับความก้องของเสียง

ค่าระดับความก้องของเสียง [22] เป็นค่าหนึ่งที่นิยมใช้เพื่อจำแนกเสียงที่มีลักษณะเป็นเสียงก้อง โดยเฉพาะเสียงสระ เป็นต้น สัญญาณเสียงที่มีระดับความก้องสูงจะมีลักษณะสัญญาณเป็นคาบ มีค่าพลังงานสูง และมีอัตราการตัดศูนย์ต่ำ ในที่นี้การประมาณระดับความก้องของสัญญาณเสียงที่กรอบเวลา m เขียนแทนด้วยสัญลักษณ์ $vdegree(m)$ จะคำนวณโดยอาศัยค่าลักษณะสำคัญสามอย่างได้แก่ ค่าอัตสหสัมพันธ์ ค่าพลังงาน และอัตราการตัดศูนย์ของสัญญาณเสียง โดยคำนวณค่าระดับความก้องของสัญญาณเสียงจากอัตราส่วนระหว่าง ผลคูณของคะแนนที่ได้จากค่าอัตสหสัมพันธ์ คะแนนที่ได้จากค่าพลังงาน และคะแนนที่ได้จากอัตราการตัดศูนย์ เทียบกับคะแนนที่มีค่าต่ำสุด ดังสมการต่อไปนี้

$$vdegree(m) = \frac{ap(m) \times ep(m) \times zp(m)}{\min\{ap(m), zp(m), ep(m)\}} \tag{2.22}$$

โดยที่ $ap(m)$ คือ คะแนนที่ได้จากค่าอัตสหสัมพันธ์ $ep(m)$ คือ คะแนนที่ได้จากค่าพลังงาน และ $zp(m)$ คือ คะแนนที่ได้จากอัตราการตัดศูนย์คะแนนทั้งสามจะคำนวณจากสัญญาณเสียงที่กรอบเวลา m และจะมีค่าอยู่ในช่วง 0 ถึง 1 โดยคะแนนที่ได้จากค่าอัตสหสัมพันธ์จะคำนวณจากสมการต่อไปนี้

$$ap(m) = \frac{1}{1+e^{-\left(\frac{autocorr(m)-0.75}{0.1}\right)}} \tag{2.23}$$

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

ค่าคะแนนที่ได้จากค่าพลังงานของสัญญาณเสียงจะคำนวณจากสมการดังนี้

$$ep(m) = \frac{1}{1+e^{-\left(\frac{E(m)-maxenergy}{5}\right)}} \tag{2.24}$$

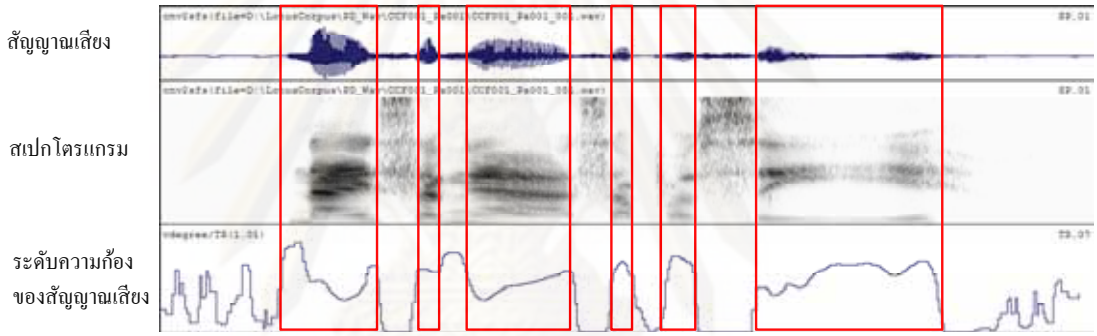
เมื่อกำหนดให้ $E(m)$ คือค่าพลังงานที่กรอบเวลา m และ $maxenergy$ คือค่าพลังงานมากสุดของสัญญาณเสียงนั้น โดยถ้าคะแนนของค่าพลังงานที่คำนวณออกมามีค่ามากก็จะหมายความว่าสัญญาณเสียงนั้นมีพลังงานมาก ซึ่งเป็นหนึ่งในลักษณะของสัญญาณเสียงที่มีความก้อง

ค่าคะแนนที่ได้จากอัตราการตัดศูนย์จะคำนวณจากสมการต่อไปนี้

$$zp(m) = \frac{1}{1+e^{\left(\frac{zcr(m)-1000}{200}\right)}} \tag{2.25}$$

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

รูปที่ 2.20 แสดงระดับความก้องของสัญญาณเสียงโดยเสียงสระจะมีความก้องของสัญญาณเสียงสูง (ส่วนที่มีการใส่กรอบ)



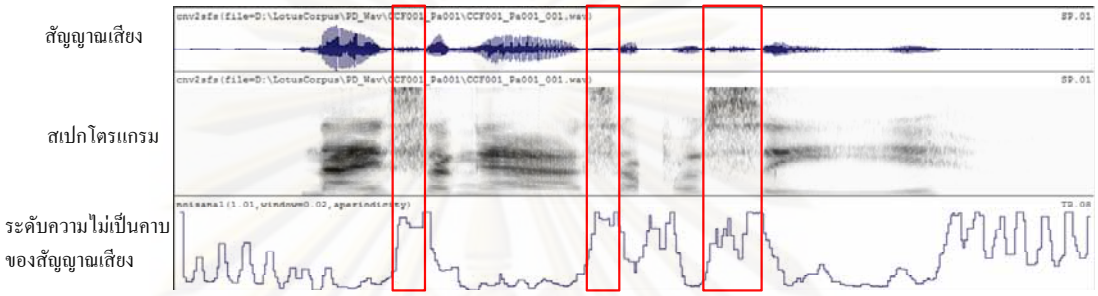
รูปที่ 2.20 ระดับความก้องของสัญญาณเสียง

2.9.5 ค่าระดับความไม่เป็นคาบของสัญญาณเสียง (Aperiodicity)

ค่าระดับความไม่เป็นคาบของสัญญาณเสียง [22] มักจะใช้ในการจำแนกเสียงที่มีความไม่เป็นคาบซึ่งจะเกิดจากเสียงที่เป็นเสียงรบกวน เช่น เสียงเสียดแทรก การประมาณระดับความไม่เป็นคาบของสัญญาณเสียงที่เวลาใดๆจะคำนวณโดยอาศัย ค่าอัตสหสัมพันธ์ ซึ่งในที่นี้ค่าระดับความไม่เป็นคาบของสัญญาณเสียงในกรอบเวลาที่ m เขียนแทนด้วยสัญลักษณ์ $aperiodicity(m)$ จะคำนวณได้จากอัตราส่วนระหว่างค่าอัตสหสัมพันธ์ต่ำสุดและค่าอัตสหสัมพันธ์เมื่อไม่มีการหน่วงเวลาซึ่งเป็นค่ามากสุด ได้ดังสมการต่อไปนี้

$$aperiodicity(m) = \frac{autocorr(m, k_{min})}{autocorr(m, 0)} \tag{2.26}$$

เมื่อกำหนดให้ $autocorr(m, k_{min})$ คือค่าอัตสหสัมพันธ์ที่มีค่าน้อยที่สุดที่มีระยะหน่วงเวลาเป็น k_{min} ซึ่งในที่นี้จะพิจารณาหาค่าอัตสหสัมพันธ์ที่มีค่าน้อยสุดในช่วงระยะหน่วงเวลาที่ความถี่ตั้งแต่ 50 ถึง 400 เฮิร์ตซึ่งเป็นช่วงความถี่ต่ำ ดังรูปที่ 2.21 จากรูปส่วนของสัญญาณเสียงเสียดแทรกซึ่งมีลักษณะของสัญญาณคล้ายเสียงรบกวนไม่เป็นคาบ ดังนั้น ค่าอัตสหสัมพันธ์ที่คำนวณออกมาได้จะมีค่ามาก



รูปที่ 2.21 ระดับความไม่เป็นคาบของสัญญาณเสียง

2.10 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM)

ซัพพอร์ตเวกเตอร์แมชชีน [23, 24] เป็นตัวจำแนกที่มีการแบ่งระยะขอบสูงสุด (Maximum Margin Classifier) สำหรับข้อมูลที่สามารถแบ่งแยกได้ด้วยเส้นตรงที่อยู่ใน $\mathbb{R}^n$ โดยจุดมุ่งหมายของซัพพอร์ตเวกเตอร์แมชชีน คือ การหาระนาบที่ทำหน้าที่แบ่งข้อมูลระหว่างข้อมูลทั้งสองประเภท ดังสมการ

$$w.x + b = 0, x \in \mathbb{R}^n \tag{2.27}$$

ซึ่งระยะห่างของการแบ่งแยกระหว่างข้อมูลทั้งสองประเภทเรียกว่า ระยะขอบ (Margin) คือ $2/\|w\|$ จะต้องมีค่ามากสุด รูปที่ 2.22 แสดงตัวจำแนกสำหรับข้อมูลที่สามารถแยกได้ด้วยเส้นตรง (a) ตัวจำแนกแบบเส้นตรงที่ไม่มีระยะขอบสูงสุด และ (b) ตัวจำแนกแบบเส้นตรงที่มีระยะขอบมากสุด จากรูปที่ 2.22 จะเห็นว่าตัวจำแนกใน (b) มีความทนทานต่อข้อมูลรบกวนมากกว่า

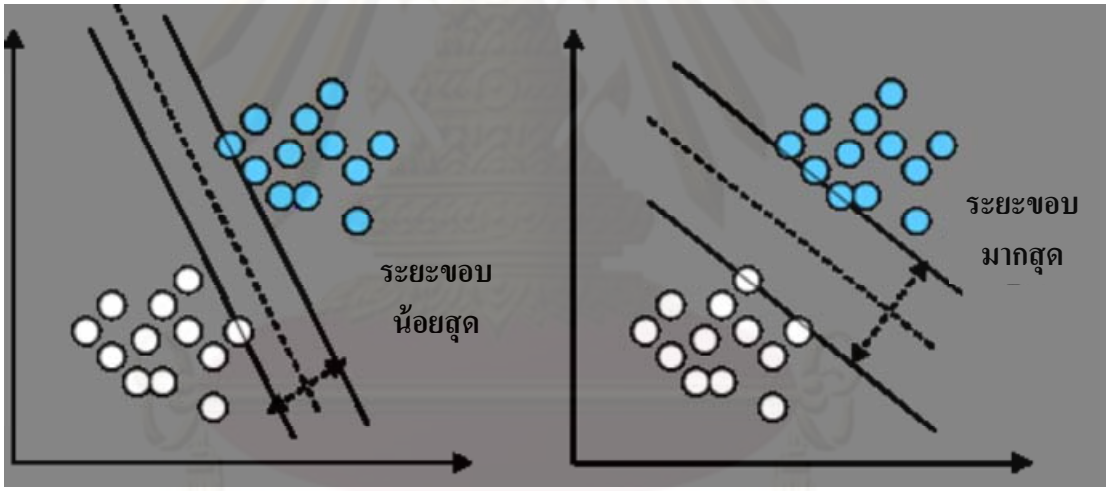
โดยปกติ ซัพพอร์ตเวกเตอร์แมชชีนจะเลือกเซตของ ซัพพอร์ตเวกเตอร์ (Support Vector) N_{SV} โดยที่ $\{x_i^{SV}\}_{i=1}^{N_{SV}}$ เป็นเซตย่อยของเวกเตอร์ l ในชุดข้อมูลฝึกฝน $\{x_i\}_{i=1}^{l}$ และคลาสกำกับ $\{y_i\}_{i=1}^{l}$ เพื่อหาระนาบแบ่งแยกที่ดีที่สุด $f(x)$ (เพื่อทำให้ระยะขอบมากสุด) ในปริภูมิระดับสูง F

$$f(x) = \sum_{i=1}^{N_{SV}} y_i \alpha_i K(x_i^{SV}, x) - b \qquad (2.28)$$

โดยปกติแล้วการแมป $\Phi :\rightarrow F$ จะต้องใช้การคำนวณที่สูงเกินไป ฟังก์ชันเคอร์เนล (Kernel Function) จะช่วยหลีกเลี่ยงปัญหาในการคำนวณหาฟังก์ชันในการแมปได้ จากสมการที่ (2.28) $K(x_i, x_j)$ เป็นฟังก์ชันเคอร์เนล ที่เป็นเส้นตรง หรือที่ไม่เป็นเส้นตรง, α_i เป็นตัวถ่วงน้ำหนัก (Weight), $\{x_i^{sv}\}_{i=1}^{N_{SV}}$ เป็นเซตของซัพพอร์ตเวกเตอร์ และ b เป็นความลำเอียง (Bias)

ฟังก์ชันเคอร์เนลมีด้วยกันหลายแบบด้วยกัน ซึ่งเคอร์เนลที่นิยมใช้กัน คือ

- เคอร์เนลแบบเส้นตรง (Linear)
- เคอร์เนลแบบพหุนาม (Polynomial)
- เคอร์เนลแบบอาร์บีเอฟ (Radial Basic Function: RBF)
- เคอร์เนลแบบซิกมอยด์ (Sigmoid)



รูปที่ 2.22 (a) ตัวจำแนกแบบระยะขอบน้อยสุด (b) ตัวจำแนกแบบระยะขอบมากสุด สำหรับข้อมูลที่ใช้เส้นตรงแบ่งแยกได้ [23]

2.10.1 การลดความเสี่ยงเชิงโครงสร้างให้ต่ำสุด (Structural Risk Minimization) [23, 24]

ในการรู้จำเราต้องการหาฟังก์ชันประมาณตัวจำแนกที่แท้จริง ซึ่งมีค่าความผิดพลาด คือ ผลต่างระหว่างค่าที่ได้จากฟังก์ชันประมาณกับค่าที่ได้จากฟังก์ชันที่แท้จริง ซึ่งโดยปกติเราจะต้องการลดความเสี่ยงจากการจำแนกประเภทผิดให้ต่ำที่สุด แต่ในทางปฏิบัติเราไม่รู้ฟังก์ชันที่

แท้จริง จึงไม่สามารถคำนวณหาค่าผิดพลาดที่แท้จริงได้ ซึ่งวิธีที่ทำได้ คือ ในการสอนเราจะพิจารณาค่าผิดพลาดจากกลุ่มตัวอย่างแทน และลดค่าของความเสี่ยงเชิงโครงสร้างซึ่งประกอบด้วยความเสี่ยงเชิงทดลอง (Empirical risk) กับช่วงความเชื่อมั่น (confidence interval) ให้ต่ำสุดแทน

สมมติว่ามีข้อมูลอยู่ l ตัว เซตของเวกเตอร์ที่ใช้ฝึกฝน $\{x_i\}_{i=1}^{l}$ และคลาสกำกับ $\{y_i\}_{i=1}^{l}$ โดยที่

$$y_i \in \{-1,+1\} \text{ และ } x_i \in \mathbb{R}^n, \tag{2.29}$$

สมมติว่าตัวอย่าง $\{x_i\}_{i=1}^{l}$ และคลาสกำกับ $\{y_i\}_{i=1}^{l}$ ถูกสร้างภายใต้การแจกแจงความน่าจะเป็นร่วม (Joint Probability Distribution) $P(x,y)$ (โดยให้ $dP(x,y) = p(x,y)dxdy$ ซึ่ง $p(x,y)$ เป็นความหนาแน่นของความน่าจะเป็น) สำหรับฟังก์ชันที่เป็นไปได้ $f(x,\alpha)$ ซึ่งพยายามหาคลาสคำตอบของเวกเตอร์ x ที่ได้มา ความเสี่ยงของฟังก์ชัน หรือความผิดพลาดที่แท้จริง เป็น

$$R(\alpha) = \int \frac{1}{2} | y - f(x,\alpha) | dP(x,y) \tag{2.30}$$

โดยที่ค่า $1/2|y-f(x,\alpha)|$ จะมีค่าเป็น 0 หรือ 1 ด้วยความน่าจะเป็น $\eta(0 \le \eta \le 1)$ ความผิดพลาดที่แท้จริงจะมีขอบเขต ดังนี้

$$R(\alpha) \le R_{emp}(\alpha) + \sqrt{\frac{h(\log(2l/h)+1-\log(\eta/4)}{l}} \tag{2.31}$$

ซึ่งเราเรียก h ว่า มิติวีซี (Vapnik Chervonenkis dimension – VC dimension) และเรียกพจน์ขวามือว่า ขอบเขตความเสี่ยง (Risk Bound) ซึ่งประกอบไปด้วย ช่วงความเชื่อมั่นวีซี (VC confidence) ในพจน์ที่สอง และ $R_{emp}(\alpha)$ เป็นความผิดพลาดโดยเฉลี่ยในข้อมูลสอน

$$R_{emp}(\alpha) = \frac{1}{2l} \sum_{i=1}^{l} | y_i - f(x_i,\alpha) | \tag{2.32}$$

มิติวีซี h จะขึ้นอยู่กับคลาสของฟังก์ชัน $f(x,\alpha)$ และความผิดพลาดโดยเฉลี่ยในข้อมูลสอน ตามหลักของทฤษฎีการลดความเสี่ยงเชิงโครงสร้างให้ต่ำสุดจะเป็นการหาคลาสของฟังก์ชันและฟังก์ชันที่เป็นสมาชิกอยู่ในคลาส ที่ทำให้ผลบวกของ ช่วงความเชื่อมั่นวีซี และความผิดพลาดโดยเฉลี่ยในข้อมูลสอนมีค่าน้อยที่สุด โดยการฝึกฝนซัพพอร์ตเวกเตอร์แมชชีนจะหาระบบแบ่งที่ทำ

ให้ระยะขอบระหว่างคลาสสองคลาสมีระยะห่างมากที่สุด และหาตัวจำแนกที่มีระยะขอบสูงสุดที่เป็นไปตามทฤษฎีการลดความเสี่ยงเชิงโครงสร้างให้ต่ำสุด

2.10.2 การหาค่าความน่าจะเป็นจากซัพพอร์ตเวกเตอร์แมชชีน

งานวิจัยนี้มีการใช้ค่าความน่าจะเป็นจากซัพพอร์ตเวกเตอร์แมชชีนที่ได้มาจาก LibSVM [25] ซึ่ง LibSVM ได้มีการดัดแปลงมาจากงานวิจัยของ Platt [26] ในงานวิจัยนี้ Platt ได้มีความพยายามที่จะทำให้ Class-Conditional Densities $P(f \mid y = \pm 1)$ ของซัพพอร์ตเวกเตอร์แมชชีนแบบเส้นตรง ที่ฝึกฝนจากชุดข้อมูล UCI Adult ให้อยู่ในรูปของซิกมอยด์ ดังที่แสดงในรูปที่ 2.23 ซึ่ง Platt แสดงให้เห็นว่าซิกมอยด์นั้นมีความเหมาะสม ซึ่งจะได้ว่าการหาค่าความน่าจะเป็นจากซิกมอยด์จะเป็นดังสมการต่อไปนี้

$$P(y = 1 \mid f) = \frac{1}{1 + \exp(Af + B)} \tag{2.33}$$

โดยที่ A และ B เป็นพารามิเตอร์ของซิกมอยด์ ซึ่งได้จากการฝึกฝน โดยวิธีการประมาณความควรจะเป็นสูงสุด (Maximum Likelihood Estimation) จากชุดข้อมูลฝึกฝน (f_i, y_i) โดยตอนแรกจะต้องประกาศชุดฝึกฝนใหม่ (f_i, t_i) ซึ่ง t_i เป็นความน่าจะเป็นเป้าหมาย (Target Probabilities) ดังสมการ

$$t_i = \frac{y_i + 1}{2} \tag{2.34}$$

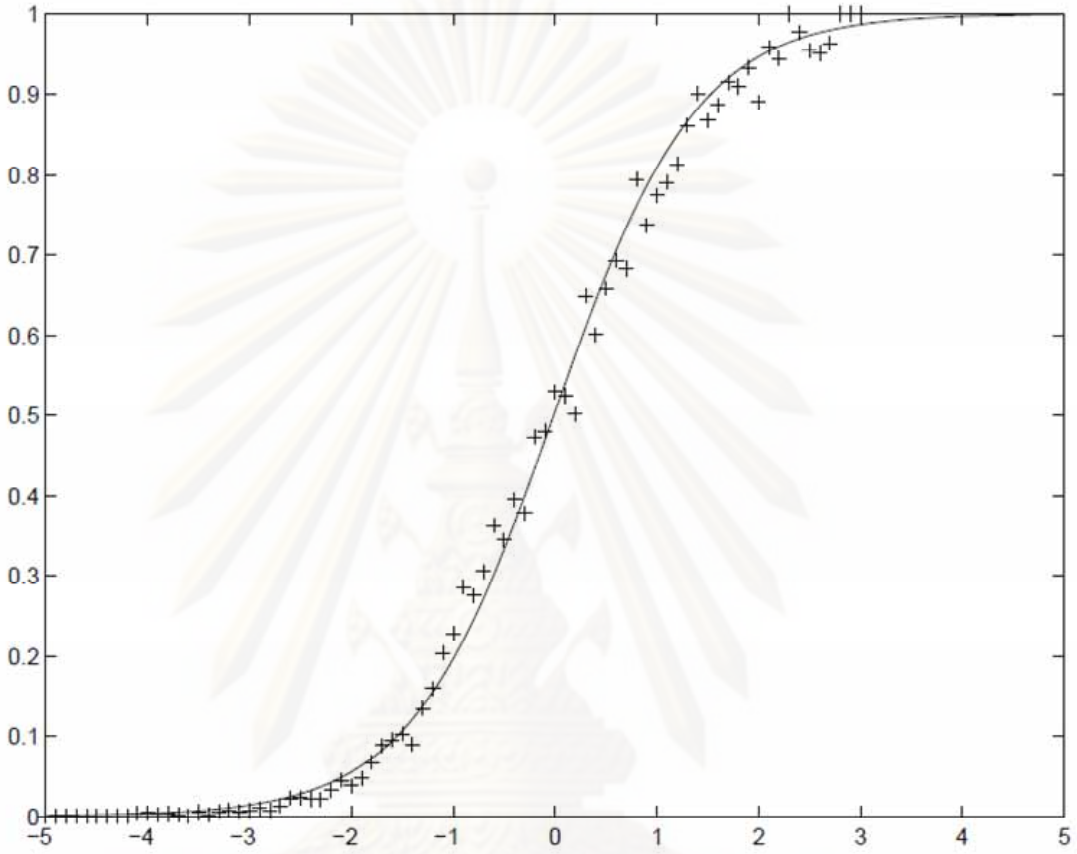
พารามิเตอร์ A และ B จะหาจากการทำให้ค่าลบของลอการิทึมของความควรจะเป็น (Negative Log Likelihood) ให้น้อยที่สุดจากข้อมูลฝึกฝน ซึ่งเป็น Cross-Entropy Function

$$\min -\sum_i t_i \log(p_i) + (1 - t_i)\log(1 - p_i), \tag{2.35}$$

ซึ่ง

$$p_i = \frac{1}{1 + \exp(Af_i + B)} \tag{2.36}$$

โดยที่ $f_i = f(x_i)$ ซึ่ง x_i เป็นตัวอย่างฝึกฝนตัวที่ i การทำให้มีค่าน้อยที่สุดตามสมการ (2.35) จะเป็นการทำให้ค่าพารามิเตอร์ A และ B น้อยสุด ซึ่งสามารถใช้อัลกอริทึมที่ใช้ในการหาค่าเหมาะที่สุดใดก็ได้ ที่ในงานวิจัยของ Platt ใช้อัลกอริทึม Model-Trust Minimization



รูปที่ 2.23 การทำให้อยู่ในรูปแบบของซิกมอยด์ของซัพพอร์ตเวกเตอร์แมชชีนที่มีเคอร์เนลแบบเส้นตรง [26]
(แต่ละเครื่องหมายบวกจะระบุที่ค่าความน่าจะเป็นภายหลังคำนวณจากตัวอย่างทั้งหมด)

2.11 งานวิจัยที่เกี่ยวข้อง

2.11.1 งานวิจัยที่เกี่ยวข้องกับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

Halberstadt และ Glass [3, 19] ได้พัฒนาการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ และเปรียบเทียบประสิทธิภาพในการรู้จำเสียงพูดในระบบหน่วยเสียงกับการรู้จำเสียงพูดแบบอาศัยกรอบเวลา ซึ่งผลลัพธ์ที่ได้ คือ การรู้จำเสียงพูดแบบอาศัยเซกเมนต์มีความผิดพลาดน้อยกว่าการรู้จำเสียงพูดแบบอาศัยกรอบเวลา ซึ่งผลลัพธ์ต่างๆนั้นแสดงในตารางที่ 2.7

ตารางที่ 2.7 ผลการรู้จำเสียงพูดในระดับหน่วยเสียงของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์เปรียบเทียบกับการรู้จำเสียงพูดแบบอาศัยกรอบเวลา [3, 19]

วิธีการรู้จำเสียงพูด	ความผิดพลาด (%)
การรู้จำเสียงพูดแบบอาศัยกรอบเวลาของ Lamel และ Gauvain [27]	27.1
การรู้จำเสียงพูดแบบอาศัยกรอบเวลาของ Ming และ Smith [28]	25.6
การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ใช้แบบจำลองเซกเมนต์เพียงอย่างเดียว	27.7
การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ใช้แบบจำลองเสียงของขอบเขตของหน่วยเสียงเพียงอย่างเดียว	24.9
การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ใช้แบบจำลองเซกเมนต์ และแบบจำลองเสียงของขอบเขตของหน่วยเสียง	24.4

2.11.2 งานวิจัยที่เกี่ยวข้องกับการแบ่งเสียงพูดเป็นเซกเมนต์ที่ใช้ในการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

งานวิจัยที่เกี่ยวกับการแบ่งเสียงพูดเป็นเซกเมนต์ที่ใช้สำหรับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์นั้นแบ่งเป็นสองกลุ่มด้วยกัน คือ การแบ่งเสียงพูดเป็นเซกเมนต์ด้วยการค้นหาตำแหน่งที่มีการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียง และการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น

สำหรับการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยการค้นหาตำแหน่งที่มีการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียง Glass และ Zue [20] ใช้การแบ่งเสียงพูดเป็นเซกเมนต์ด้วยการค้นหาตำแหน่งที่มีการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียงที่เรียกว่า การแบ่งเสียงพูดเป็นเซกเมนต์โดยอาศัยการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียงแบบหลายระดับ ดังที่กล่าวไว้ข้างต้น

นอกจากนี้ Wang และคณะ [29] ได้นำเสนอการแบ่งเสียงพูดเป็นประโยคโดยใช้การจำแนก เสียงสระ เสียงพยัญชนะ และเสียงหยุดระหว่างพูด (Pause) เพื่อสร้างขอบเขตของหน่วยเสียงที่มีโอกาสเป็นไปได้ หลังจากนั้นจะดึงเอา เสียงหยุดระหว่างพูด อัตราการออกเสียง (Rate of Speech: ROS) และสัทสัมพันธ์ (Prosody) เพื่อใช้ในการระบุความเป็นขอบเขตของหน่วยเสียงจากรายการขอบเขตของหน่วยเสียงที่มีโอกาสเป็นไปได้

Leelapatarakit และคณะ [22, 30] ได้มุ่งเน้นในการค้นหาตำแหน่งของขอบเขตของหน่วยเสียง ซึ่งอยู่บนสมมติฐานที่รู้จำนวนขอบเขตของหน่วยเสียงอยู่แล้ว โดยใช้คุณสมบัติสองแบบ คือ 1) การวัดการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียงจากค่าพลังงาน (Energy) โดยวิธีนี้จะนำค่าพลังงานจากช่วงความถี่ 5 ช่วงมาเปรียบเทียบกับค่าเฉลี่ยของพลังงานถ้าค่าพลังงานมีค่ามากกว่า

ค่าเฉลี่ยก็มีโอกาสที่จะเป็นขอบเขตของหน่วยเสียง 2) การวัดการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียงจากระยะห่างยูคลีเดียนของสัมประสิทธิ์เซปสตรัมบนสเกลเมล ระหว่างกรอบเวลาปัจจุบัน และกรอบเวลาถัดไป ซึ่งถ้าค่าระยะห่างยูคลีเดียนมีมากก็จะมีโอกาสที่จะเป็นขอบเขตของหน่วยเสียง โดยจากผลการทดลองของ Leelaphattarakij แสดงให้เห็นว่าการวัดการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียงจากค่าพลังงานมีโอกาสที่จะพบขอบเขตของเสียงที่ถูกต้องได้ดีกว่าการวัดการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียงจากระยะห่างยูคลีเดียนของสัมประสิทธิ์เซปสตรัมบนสเกลเมลเพียงเล็กน้อย แต่จะมีอัตราส่วนระหว่างจำนวนขอบเขตของหน่วยเสียงที่ถูกต้องกับจำนวนขอบเขตของหน่วยเสียงที่หาได้น้อยกว่าการวัดการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียงจากระยะห่างยูคลีเดียนของสัมประสิทธิ์เซปสตรัมบนสเกลเมลประมาณ 50% ดังนั้นวิธีการวัดการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียงจากระยะห่างยูคลีเดียนของสัมประสิทธิ์เซปสตรัมบนสเกลเมลจึงดีกว่า เนื่องจากมีการผลลัพธ์ที่ดี และมีความผิดพลาดแบบแทรกน้อยกว่า

Sainath และ Hazen [31] ได้พัฒนาการตรวจสอบหาตำแหน่งแลนด์มาร์ค และการแบ่งเสียงพูดเป็นเซกเมนต์โดยใช้ McAulay-Quatieri Sinusoidal Model ซึ่งมีความคงทนต่อเสียงรบกวน โดยมีการตรวจหาขอบเขตของเสียงก้องและเสียงไม่ก้องจากฮาร์มอนิกของคลื่นรูปไซน์ และตรวจหาแลนด์มาร์คจากจำนวนฮาร์มอนิกของคลื่นรูปไซน์ ซึ่งในการแบ่งเสียงพูดเป็นเซกเมนต์ จะมีการจำแนกแลนด์มาร์คออกเป็น แลนด์มาร์คหลักแบบถาวร (Hard major landmark) แลนด์มาร์คหลักแบบไม่ถาวร (Soft major landmark) และแลนด์มาร์ครอง (Minor landmark) จากนั้นก็จะสร้างกราฟของเซกเมนต์โดยใช้โดยนำแลนด์มาร์คมาต่อกันโดยใช้กฎสามข้อ ต่อมา Sainath และคณะ [32] ได้นำเสนอวิธีการแบ่งเสียงโดยเป็นเซกเมนต์โดยใช้ Extended Baum-Welch (EBW) transformation ซึ่งวิธีการนี้จะสามารถตรวจหาแลนด์มาร์คที่มีระยะเวลาที่สั้น และมีลดจำนวนแลนด์มาร์คที่เกิดมาโดยเพิ่มขั้นตอนการปรับปรุงโดยมีการวัดค่า Gradient Steepness T ซึ่งถ้าค่า T มีค่ามากกว่าค่าที่กำหนดก็จะมีการเพิ่มขอบเขตของหน่วยเสียง ณ บริเวณนั้น และเมื่อค่า T มีค่าต่ำกว่าค่าที่กำหนดก็จะมีการรวมเซกเมนต์เข้าด้วยกัน อย่างไรก็ตามงานของ Sainath ที่ได้กล่าวมาใช้วิธีการเปรียบเทียบกับค่าแบ่ง (Threshold) ซึ่งเป็นการยากที่จะหาค่าแบ่งที่เหมาะสม นอกจากนี้ Sainath และ Zue [33] ยังมีการเปรียบเทียบการใช้กลุ่มของหน่วยเสียงแบบกว้าง (Broad Phonetic Classes: BPCs) กับ กลุ่มของสวนศาสตร์แบบกว้าง (Broad Acoustic Classes: BACs) ในการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ โดยการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยการใช้กลุ่มของหน่วยเสียงแบบกว้าง และกลุ่มของสวนศาสตร์แบบกว้าง นั้นจะดำเนินการโดยมีการตรวจหากลุ่มของเสียงแบบกว้างในสัญญาณเสียงพูด และวางตำแหน่งของแลนด์มาร์คไปในที่มีการเปลี่ยนแปลงของกลุ่มของเสียงแบบกว้าง โดยผลการทดลองแสดงให้เห็นว่าการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยกลุ่มของเสียงแบบกว้างสามารถให้ผลดีกว่าการใช้การเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียง และใช้แบบจำลอง Sinusoidal

ถึงแม้วิธีการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยการค้นหาตำแหน่งที่มีการเปลี่ยนแปลงคุณสมบัติของสัญญาณเสียงจะทำงานได้อย่างรวดเร็ว แต่อย่างไรก็ตามการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีนี้เป็นเพียงการตรวจสอบไปที่สัญญาณเสียงโดยตรง ไม่ได้มีการใช้เงื่อนไขบังคับ (Constraint) อื่นๆเพื่อช่วยให้การแบ่งเสียงพูดเป็นเซกเมนต์ดีขึ้น เช่น การใช้แบบจำลองทางภาษาเป็น จึงทำให้ขนาดของกราฟของเซกเมนต์ก็มีขนาดใหญ่มากเกินความจำเป็น และมีจำนวนขอบเขตของหน่วยเสียงที่แทรกมาจำนวนมาก จึงอาจจะเหตุที่ทำให้ขั้นตอนการค้นหาช้าลง และความถูกต้องอาจจะลดลงด้วยเช่นกัน เนื่องจากการแบ่งเสียงพูดเป็นเซกเมนต์ที่ดีนั้นจะต้องทำให้เกิดจำนวนของเซกเมนต์ที่น้อยโดยที่ความผิดพลาดแบบแทรก และความผิดพลาดตัดออก ก็จะต้องน้อยด้วย ดังนั้น Lee และ Glass [8-10] จึงได้เสนอการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น วิธีการนี้จะทำได้โดยอาศัยระบบรู้จำเสียงพูดแบบอาศัยกรอบเวลาในระดับหน่วยเสียงมารู้จำเสียงพูดออกมาเป็นลำดับของหน่วยเสียง *N* ลำดับที่ดีที่สุด แล้วจึงนำแต่ละเซกเมนต์ของหน่วยเสียงที่รู้จำได้มาประกอบรวมกันเป็นกราฟของเซกเมนต์ ซึ่งวิธีการนี้จะมีข้อดีตรงที่ 1) สามารถใช้เงื่อนไขบังคับในระดับสูงเข้ามาช่วย ตัวอย่างเช่น แบบจำลองเสียงแบบบริบทไม่อิสระ และแบบจำลองทางภาษา 2) มีความยืดหยุ่นสูงเนื่องจากเราจะสามารถเลือกได้ว่าจะใช้วิธีการใดในขั้นตอนการรู้จำเสียงพูดในระดับหน่วยเสียงแบบอาศัยกรอบเวลาได้ ทั้งนี้จะอยู่กับประสิทธิภาพในการคำนวณของเครื่องที่มีอยู่ เช่น เราสามารถเลือกใช้แบบจำลองเสียงแบบบริบทไม่อิสระ หรือแบบจำลองทางภาษาได้ 3) มีความแม่นยำ และเลือกสมมุติฐานที่มีความคล้ายคลึงกับเซกเมนต์มากที่สุด จึงทำให้จำนวนเซกเมนต์ที่สร้างมีจำนวนน้อย และมีความผิดพลาดแบบแทรก และความผิดพลาดตัดออกน้อย ซึ่งในปัจจุบันระบบรู้จำเสียงพูด SUMMIT ใช้วิธีการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็นในการแบ่งเสียงพูดเป็นเซกเมนต์

อย่างไรก็ตามในกรณีที่ระบบรู้จำเสียงพูดแบบอาศัยกรอบเวลาที่ใช้โดยการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็นมีความแม่นยำต่ำ ก็จะทำให้การแบ่งเสียงพูดเป็นเซกเมนต์เกิดความผิดพลาดขึ้น ทั้งที่เป็นความผิดพลาดแบบแทรก และความผิดพลาดตัดออก ซึ่งจะเป็นส่วนสำคัญที่ต้องการการปรับปรุงแก้ไขกราฟของเซกเมนต์ เพื่อทำให้ความแม่นยำของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์สูง ดังนั้นเป้าหมายหนึ่งของวิทยานิพนธ์นี้ คือ พยายามเพิ่มจำนวนเซกเมนต์ที่ถูกต้องลงในกราฟของเซกเมนต์ โดยการลดความผิดพลาดของการแบ่งเสียงพูดเป็นเซกเมนต์

2.11.3 งานวิจัยที่เกี่ยวข้องกับการใช้หลักสวนสัทศาสตร์ (Acoustic Phonetics) และสมบัติลักษณะเฉพาะมาใช้ในการแบ่งเสียงพูดเป็นเซกเมนต์

ปัจจุบันได้มีความพยายามที่จะนำความรู้ทางด้านสวนสัทศาสตร์ และสมบัติลักษณะเฉพาะมาใช้ทั้งในการแบ่งเสียงพูดเป็นเซกเมนต์โดยการหาตรวจหาตำแหน่งของแลนด์มาร์ค ซึ่งมีงานวิจัยต่างๆ ดังนี้

Niyogi และคณะ [34] มีการใช้สมบัติลักษณะเฉพาะในการตรวจหาเสียงพยัญชนะกักโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน

Liu [35, 36] พัฒนาวิธีในการตรวจหาแลนด์มาร์คของเสียงสามอย่าง คือ แลนด์มาร์คของเสียง เสียงเกิดที่เส้นเสียงบริเวณคอหอย (Glottal) เสียงพูดที่เกิดขณะกำลังเปิดช่องทางเดินเสียง (Burst) และเสียงที่มีเสียงสั่น โดยเลือกวัดเวกเตอร์คุณสมบัติของเสียงจากสัมประสิทธิ์เมลฟรีเควนซีเคปสตรอล เพื่อนำคำนวณลักษณะเด่นของเสียงในแต่ละกรอบเวลาแล้วจึงจำแนกเอาว่าแต่ละกรอบเวลานั้นๆเป็นแลนด์มาร์คหรือไม่

Hosom [37] มีการใช้สารสนเทศสวนสัทศาสตร์ และสมบัติลักษณะเฉพาะในการจัดวางตำแหน่งของหน่วยเสียง ซึ่งในงานนี้ได้มีการเสนอสมบัติทางเสียงที่ใช้ในการจัดวางตำแหน่งของหน่วยเสียง และมีการเปลี่ยนหน่วยเสียงให้อยู่ในรูปแบบของ ลักษณะการออกเสียง และฐานในการออกเสียง นอกจากนี้ยังได้เสนอการนำข้อมูลที่มีการเปลี่ยนแปลงระหว่างหน่วยเสียงมาใช้ โดยงานวิจัยนี้จะมีการเปรียบเทียบกับการใช้แบบจำลองฮิดเดนมาร์คอฟในการจัดวางตำแหน่งของหน่วยเสียง

Schutte และ Glass [38] ยังมีการตรวจหาแลนด์มาร์คของเสียงสั่นที่อากาศผ่านช่องปากไปโดยไม่ได้ถูกกักเอาไว้เพียงพอต่อการสร้างเสียงรบกวนหรือกักการไหลของอากาศ โดยใช้ซัพพอร์ตเวกเตอร์แมชชีนเช่นกัน

Dareyoah และคณะ [39] มีการตรวจหาตำแหน่งของเสียงสระสำหรับภาษาไทย โดยใช้อัตราการตัดศูนย์

Leelaphattarakij [22] จะใช้สารสนเทศสวนสัทศาสตร์ในหาตำแหน่งขอบเขตของหน่วยเสียงเพื่อใช้ในการแบ่งเสียงพูดเป็นเซกเมนต์

นอกจากนี้ Boonsuk และคณะ [40, 41] ก็มีการใช้สารสนเทศสวนสัทศาสตร์ในการปรับหาตำแหน่งขอบเขตของเสียง โดยในงานนี้ได้มีการใช้สมบัติทางเสียงแต่ละแบบขึ้นอยู่กับชนิดของขอบเขตของหน่วยเสียงว่าเป็นขอบเขตของหน่วยเสียงชนิดใด เช่น ขอบเขตของหน่วยเสียงระหว่างเสียงสระกับเสียงนาสิก ก็จะใช้สมบัติทางเสียงอย่างหนึ่ง ขอบเขตของหน่วยเสียงระหว่างเสียงสระกับเสียงเสียดแทรกก็จะใช้สมบัติทางเสียงอีกอย่างหนึ่ง

จากงานวิจัยที่ได้กล่าวมา งานวิธีในกลุ่มนี้ส่วนใหญ่มีแนวคิดในการนำสมบัติทางเสียงที่เฉพาะเจาะจงกับเสียงชนิดต่างๆ มาใช้ในการตรวจหาตำแหน่งของแลนด์มาร์ค หรือใช้ในการหาตำแหน่งของขอบเขตของหน่วยเสียง ซึ่งในงานวิจัยนี้ก็ได้มีแนวความคิดในการใช้สมบัติทางเสียงที่เฉพาะเจาะจงกับชนิดของเสียงเช่นเดียวกัน

2.11.4 งานวิจัยที่เกี่ยวข้องกับการใช้หลักสวนสัทศาสตร์ และสมบัติลักษณะเฉพาะมาใช้ในการรู้จำเสียงพูด

นอกจากนี้ยังมีนักวิจัยอีกกลุ่มหนึ่งที่มีความเชื่อว่าในการรู้จำเสียงพูดจะต้องมีการใช้องค์ความรู้ทางด้านสวนสัทศาสตร์ ดังนั้น จึงมีการหลักสวนสัทศาสตร์ และสมบัติลักษณะเฉพาะมาใช้ในการรู้จำเสียงพูด ซึ่งตัวอย่างของงานวิจัยในกลุ่มนี้ได้แก่

Meng และคณะ [42] ได้มีการทดลองจำแนกเสียงสระภาษาอังกฤษจำนวน 16 เสียงโดยใช้สมบัติลักษณะเฉพาะ

Bitar และ Espy-Wilson [43] ได้นำเสนอพารามิเตอร์คุณสมบัติทางเสียง เพื่อใช้ในการรู้จำเสียงพูดให้อยู่ในกลุ่มของหน่วยเสียงแบบกว้าง โดยมีการเปรียบเทียบผลการรู้จำเสียงกับการใช้พารามิเตอร์แบบเคปสตรอล และใช้แบบจำลองฮิดเดนมาร์คอฟในการรู้จำเสียงพูด ซึ่งผลการทดลองแสดงให้เห็นว่าพารามิเตอร์คุณสมบัติทางเสียงให้ผลการรู้จำเสียงที่ดีกว่าการใช้พารามิเตอร์แบบเคปสตรอล

Salomon และ Espy-Wilson [44] ได้พัฒนาระบบจำแนกลักษณะการออกเสียงของเสียงพยัญชนะจากการวัดค่าพารามิเตอร์ตามเวลา (Temporal Parameters) ซึ่งงานนี้แสดงให้เห็นว่าการใช้พารามิเตอร์ 7 ตัวที่มีการคำนวณจากความแตกต่างของเวลาระหว่างการเริ่มต้น (Onset) และการสิ้นสุด (Offset) ของพลังงานทางสวนศาสตร์ที่สามารถกำหนดลักษณะการออกเสียง และความเป็นเสียงก้อง หลังจากนั้น Salomon [45] ก็ได้นำการใช้พารามิเตอร์จำนวน 4 ตัวมาใช้ในการรู้จำลักษณะการออกเสียงโดยจะเปรียบเทียบผลการรู้จำที่ใช้สัมประสิทธิ์เมลฟรีเคว็นซีเคปสตรอลซึ่งจำนวนพารามิเตอร์ 39 ตัว โดยพารามิเตอร์ทั้ง 4 ตัว คือ 1) ความเป็นคาบของสัญญาณเสียง 2) ความไม่คาบของสัญญาณเสียง 3) การเริ่มต้น และ 4) การสิ้นสุด ซึ่งผลการทดลองแสดงให้เห็นว่าการใช้พารามิเตอร์ 4 ตัวที่นำเสนอสามารถให้ผลการรู้จำเทียบเท่ากับการใช้สัมประสิทธิ์เมลฟรีเคว็นซีเคปสตรอลซึ่งจำนวนพารามิเตอร์ 39 ตัว

Kirchhoff และคณะ [46] ได้เสนอการแยกงานที่มีความยาก และซับซ้อนในการรู้เสียงพูดจากสัญญาณเสียง เป็นงานที่มีเล็ก และง่ายกว่าในการจำแนกแต่ละกรอบเวลาเป็นค่าบวก และลบของคุณสมบัติที่แสดงถึงกระบวนการออกเสียงเทียม (Pseudo-Articulatory Features) ซึ่งแสดงให้เห็นว่าการใช้คุณสมบัติที่แสดงถึงกระบวนการออกเสียงเทียมสามารถให้ผลการรู้จำเสียงที่ดีให้สภาพแวดล้อมที่มีเสียงรบกวน

Tang และคณะ [47, 48] มีการจำลองเสียงพูดให้อยู่ในรูปแบบของลักษณะทางภาษาศาสตร์ โดยมีการนำเสนอการรู้จำเสียงพูดแบบสองขั้นตอน ในขั้นตอนแรก สัญญาณเสียงจะถูกแบ่ง แล้วจำแนกเป็น ลักษณะการออกเสียงและตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียง จากนั้นในขั้นตอนที่สอง จะใช้หลักทางสัทศาสตร์ในการค้นหาคำตอบจากกลุ่มของข้อมูลที่ได้จากขั้นตอนแรก โดยจะมีการแก้ไขระบบรู้จำเสียงพูด SUMMIT ให้มีการเปลี่ยนแลนด์มาร์คระหว่างหน่วยเสียง

เป็นการใช้ตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียง และแลนด์มาร์คภายในหน่วยเสียงเป็นการใช้ลักษณะการออกเสียง ซึ่งจะมีการทดลองกับเสียงจากระบบ Jupiter ซึ่งเป็นระบบที่ใช้สอบถามข้อมูลสภาพอากาศ และระบบ Mercury ซึ่งเป็นระบบสอบถามแผนการบิน ในการทดลองได้ผลลัพธ์ว่าระบบที่นำเสนอมีความแม่นยำสูงขึ้นในทั้งการทดสอบกับระบบ Jupiter และ Mercury นอกจากนั้น

Pruthi และ Espy-Wilson [49] ได้เสนอพารามิเตอร์คุณสมบัติทางเสียงที่ใช้ในการจำแนกเสียงนาสิก และเสียงกึ่งสระ โดยพารามิเตอร์คุณสมบัติทางเสียงได้แก่ การเริ่มต้นและการสิ้นสุด อัตราส่วนของพลังงาน (Energy Ratio) ยอดของสเปกตรัม (Spectral Peak) และความหนาแน่นของความถี่ฟอร์แมนต์ (Formant Density) โดยมีการใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นตัวจำแนก ต่อมา Pruthi และ Espy-Wilson [50] ก็ได้นำเสนอพารามิเตอร์คุณสมบัติทางเสียงใหม่ ซึ่งมี การเริ่มต้นและสิ้นสุดของพลังงาน อัตราส่วนของพลังงาน ยอดของสเปกตรัม และพารามิเตอร์ความแปรปรวนชั้นนอก (Envelope Variance) ซึ่งมีการทดลองในการรู้จำเสียงตัวเลข จากการทดลองพบว่าการใช้พารามิเตอร์คุณสมบัติทางเสียงดังกล่าวช่วยลดความผิดพลาดลงไป 60%

Borys และ Hasegawa-Johnson [51] ได้ใช้สมบัติลักษณะเฉพาะที่ได้จากซัพพอร์ตเวกเตอร์แมชชีนในการรู้จำเสียงพูดในระดับหน่วยเสียง โดยใช้ทั้งหมด 33 ซัพพอร์ตเวกเตอร์แมชชีน ซึ่ง 10 ซัพพอร์ตเวกเตอร์แมชชีนใช้สำหรับการเปลี่ยนแปลงของลักษณะการออกเสียง และอีก 23 สำหรับแทนตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียง

Scanlon และคณะ [52] ได้นำกลุ่มของหน่วยเสียงแบบกว้างเข้าไปใช้เพื่อเพิ่มประสิทธิภาพในการรู้จำเสียงพูดในระดับหน่วยเสียงของการรู้จำเสียงเสียงพูดแบบผสมระหว่างโครงข่ายประสาทเทียม และแบบจำลองฮิดเดนมาร์คอฟ (Hybrid ANN/HMM) แต่อย่างไรก็ตาม Scanlon ได้ใช้ PLP จำนวน 351 มิติเป็นคุณสมบัติที่ใช้ในการจำแนกกลุ่มของหน่วยเสียงแบบกว้าง ซึ่งเป็นคุณสมบัติที่ไม่ใช่คุณสมบัติเฉพาะของกลุ่มของหน่วยเสียงแบบกว้าง และมีจำนวนมิติมากเกินไป

Chang และ Glass [53] ได้เสนอแบบจำลองเกาส์เซียนเชิงลำดับชั้นในการจำแนกหน่วยเสียง ซึ่ง Chang และ Glass ได้ใช้หลักการของการแบ่งปัญหาให้เป็นปัญหาย่อย โดย Change และ Glass เสนอการแบ่งลำดับชั้นเป็น 2 ลำดับชั้น โดยให้ลำดับชั้นแรกเป็นกลุ่มของลักษณะการออกเสียงจำนวน 9 กลุ่ม และลำดับชั้นที่สองเป็นหน่วยเสียง ซึ่งผลการจำแนกหน่วยเสียงได้ผลเป็นที่น่าพอใจ

Sainath และคณะ [54] ได้เสนอการรู้จำกลุ่มของหน่วยเสียงแบบกว้างโดยใช้ Extended Baum-Welch transformation โดย Sainath ได้รู้จำกลุ่มของหน่วยเสียงแบบกว้างจำนวน 7 กลุ่ม โดยใช้ PLP จำนวน 13 มิติ เป็นเวกเตอร์คุณสมบัติ และใช้แบบจำลองฮิดเดนมาร์คอฟแบบซ้ายไปขวาจำนวน 3 สถานะ เป็นแบบจำลองของกลุ่มของหน่วยเสียงแบบกว้าง Sainath ได้กล่าวถึง

ความสำคัญของกลุ่มของหน่วยเสียงแบบกว้างว่า กลุ่มของหน่วยเสียงแบบกว้างเป็นตัวสำคัญที่จะช่วยเพิ่มประสิทธิภาพของการรู้จำเสียงพูด การรู้จำภาษา และการตรวจหาตำแหน่งของแลนด์มาร์ค

Juneja และ Espy-Wilson [17, 55] ได้นำเสนอพารามิเตอร์คุณสมบัติทางเสียงที่มีการใช้ความรู้ทางสวนสัทศาสตร์ในการจำแนกเสียงพูดต่อเนื่องให้อยู่กลุ่มของหน่วยเสียงแบบกว้าง และมีการเปรียบเทียบผลลัพธ์กับแบบจำลองฮิดเดนมาร์คอฟที่มีการใช้สัมประสิทธิ์เมลฟรีเคว็นซีเคปสตรอลจำนวน 39 มิติ ซึ่งจากการทดลองแสดงให้เห็นว่าพารามิเตอร์คุณสมบัติทางเสียงให้ผลลัพธ์ที่ดีกว่าการใช้สัมประสิทธิ์เมลฟรีเคว็นซีเคปสตรอลในขณะที่มีจำนวนพารามิเตอร์น้อยกว่าถึงสามเท่า ต่อมา Juneja และ Espy-Wilson [18] ได้ขยายงานไปเป็นการรู้จำเสียงพูดตัวเลข ซึ่งในงานนี้จะมีการตรวจหาแลนด์มาร์คโดยใช้พารามิเตอร์คุณสมบัติทางเสียง จากนั้นก็จะใช้ซัพพอร์ตเวกเตอร์แมชชีนจำแนกเสียงให้อยู่ในรูปแบบของกลุ่มของหน่วยเสียงแบบกว้าง และค้นหาคำตอบจากการค้นหาผ่านผลลัพธ์ของซัพพอร์ตเวกเตอร์แมชชีนกับแบบจำลองการออกเสียง

จากงานวิจัยต่างๆที่ได้กล่าวถึงจะเห็นว่างานวิจัยทางด้านการรู้จำเสียงพูดค่อนข้างมีแนวโน้มที่จะมีการใช้หลักสวนสัทศาสตร์ และสมบัติลักษณะเฉพาะมาใช้มากขึ้น แต่อย่างไรก็ตาม ณ ปัจจุบันยังไม่มีงานวิจัยในด้านนี้ที่มีการใช้สวนสัทศาสตร์ และสมบัติลักษณะเฉพาะที่แท้จริงที่ประสบความสำเร็จในการรู้จำเสียงพูดต่อเนื่อง แม้แต่งานของ Juneja และ Espy-Wilson [18] ซึ่งได้รับการตีพิมพ์ในปี 2008 ก็ยังเป็นเพียงการรู้จำเสียงพูดแบบคำโดดเฉพาะเสียงตัวเลขเท่านั้น ดังนั้นในงานวิจัยนี้ จึงได้เลือกระบบรู้จำเสียงพูดซึ่งสามารถใช้งานได้จริงในการรู้จำเสียงพูดต่อเนื่องมาพัฒนา ซึ่งก็คือ วิธีการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ ซึ่งมีข้อดีในการเพิ่มความรู้ทางเสียงเข้าไปได้ง่าย ด้วยข้อดีข้อนี้งานวิจัยนี้จึงจะมีการเพิ่มความรู้ทางด้านหลักสวนสัทศาสตร์ และสมบัติลักษณะเฉพาะไปในการรู้จำเสียงพูดแบบอาศัยเซกเมนต์เพื่อเพิ่มความแม่นยำในการรู้จำเสียงพูดให้แก่การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ ซึ่งมีหลายงานวิจัยด้วยกันที่ใช้ลักษณะการออกเสียง และกลุ่มของหน่วยเสียงแบบกว้างเพื่อช่วยเพิ่มประสิทธิภาพ และความแม่นยำให้แก่การจำแนก และรู้จำเสียงพูด ดังนั้นในงานวิจัยนี้จึงมีแนวความคิดที่จะนำสมบัติเฉพาะที่เกี่ยวกับกลุ่มของหน่วยเสียงแบบกว้าง ซึ่งเป็นผลที่เกี่ยวโยงมาจากลักษณะการออกเสียง เข้าไปเป็นตัวถ่วงน้ำหนักในกระบวนการคิดคะแนนของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์เพื่อเพิ่มความแม่นยำให้แก่การรู้จำเสียงพูดแบบอาศัยเซกเมนต์

นอกจากนี้งานวิจัยต่างๆ แสดงให้เห็นว่าสวนสัทศาสตร์ และสมบัติลักษณะเฉพาะนั้นสามารถช่วยวัดความแตกต่างเพื่อระบุขอบเขตของหน่วยเสียง และสามารถแยกแยะหน่วยเสียงต่างๆได้เป็นอย่างดี ดังนั้นจึงเป็นเหตุผลหนึ่งที่ทำให้งานวิจัยนี้ได้นำสวนสัทศาสตร์ และสมบัติลักษณะเฉพาะมาใช้ในการแก้ไขความผิดพลาดของการแบ่งเสียงพูดเป็นเซกเมนต์

บทที่ 3

การทดลองเบื้องต้น

3.1 ฐานข้อมูลเสียงโลตัส

ฐานข้อมูลเสียงโลตัส [11-13] เป็นฐานข้อมูลเสียงพูดขนาดใหญ่สำหรับระบบรู้จำเสียงพูดต่อเนื่องภาษาไทยที่พัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (National Electronics and Computer Technology Center: NECTEC) โดยโลตัสจะประกอบด้วยชุดข้อมูลเสียงพูด 2 ส่วนด้วยกัน คือ

3.1.1 ชุดหน่วยเสียงสมดุล หรือ PD (Phonetically Distributed Set)

ชุดเสียงพูดชุดนี้จะออกแบบให้ครอบคลุมการเกิดของ “หน่วยเสียงคู่ (Biphone)” ในภาษาไทยทั้งภายในพยางค์ ระหว่างพยางค์ และระหว่างคำ ซึ่งชุดประโยค PD จะครอบคลุมหน่วยเสียงคู่จำนวน 1,628 คู่ ซึ่งคิดเป็น 90.90% ของหน่วยเสียงคู่ที่เกิดขึ้นได้ในภาษาไทย

3.1.2 ชุดประโยคที่ครอบคลุมคำศัพท์ที่มีสถิติการใช้สูงสุด 5,000 คำ

ชุดประโยคนี้จะมาจากการคัดเลือกประโยคที่ประกอบด้วยคำศัพท์ที่มีสถิติการใช้งานสูงสุด 5,000 ลำดับแรกจากคลังข้อความทั้งหมด โดยชุดประโยคนี้จะแบ่งข้อมูลออกเป็น 3 ชุดย่อย ดังต่อไปนี้ 1) ชุดฝึกฝน (TR: Training set) 2) ชุดทดสอบเพื่อพัฒนา (DT: Development test set) 3) ชุดทดสอบเพื่อประเมิน (ET: Evaluation test set)

ฐานข้อมูลเสียงโลตัสบันทึกเสียงพูดจากผู้พูด 48 คนแบ่งเป็นเพศชาย และหญิงในจำนวนที่เท่ากัน ในการบันทึกเสียงพูดบันทึกผ่านไมโครโฟน 2 ประเภทพร้อมกัน ประเภทแรกเป็นไมโครโฟนสำหรับพูดระยะใกล้ (Dynamic Close-talk) คุณภาพสูง และประเภทที่สองเป็นไมโครโฟนแบบทิศทางเดียว (Dynamic Unidirectional) ระดับคุณภาพปานกลาง และทำการบันทึกเสียงใน 2 สภาพแวดล้อม คือ สภาพแวดล้อมแบบห้องเงียบ และ สภาพแวดล้อมแบบสำนักงาน โดยเก็บข้อมูลเสียงผ่านแถบบันทึกเสียงดิจิตัล (Digital Audio Tape) ก่อนแปลงเป็นไฟล์อิเล็กทรอนิกส์ที่มีอัตราการสุ่มเป็น 16 กิโลเฮิร์ต (kHz)

ในงานวิจัยนี้จะใช้ข้อมูลเสียงที่ได้จากการอัดเสียงในสภาพแวดล้อมแบบห้องเงียบ และใช้ไมโครโฟนสำหรับพูดระยะใกล้ในการทดลองทั้งหมด ในส่วนของชุดเสียงพูด จะใช้ชุดข้อมูลเสียง TR และ PD เป็นข้อมูลสำหรับฝึกฝนแบบจำลองทางเสียง ชุดข้อมูลเสียง DT ไว้ในการทดสอบเพื่อพัฒนา และใช้ชุดข้อมูลเสียง ET เพื่อไว้ประเมินประสิทธิภาพ ในส่วนของแบบจำลองทางภาษา ใน

งานวิจัยนี้จะใช้แบบจำลองทางภาษาแบบสองหน่วย (Bigram Language Model) ซึ่งฝึกฝนจากไฟล์กำกับของ TR

3.2 เกณฑ์การเปรียบเทียบผล

ในการเปรียบเทียบระบบรู้จำเสียงพูดแบบต่างๆ นั้นจะมีหลักเกณฑ์ในการเปรียบเทียบดังนี้

3.2.1 ความผิดพลาดของเซกเมนต์

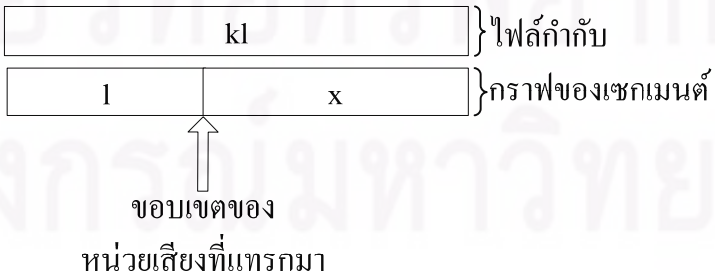
ความผิดพลาดของเซกเมนต์นี้จะเป็นค่าของจำนวนของเซกเมนต์ที่ไม่ปรากฏในกราฟของเซกเมนต์ ซึ่งจะเป็นอัตราส่วนของจำนวนเซกเมนต์ที่อยู่ในไฟล์กำกับแต่ไม่ได้อยู่ในกราฟของเซกเมนต์ ต่อจำนวนเซกเมนต์ทั้งหมดในไฟล์กำกับ โดยในการเปรียบเทียบนั้นจะเปรียบเทียบขอบเขตหน่วยเสียงทั้งสองข้างของเซกเมนต์ของไฟล์กำกับ และในกราฟของเซกเมนต์ ซึ่งจะอนุญาตให้ขอบเขตของหน่วยเสียงมีความแตกต่างทางเวลาได้ไม่เกิน 20 มิลลิวินาที (ms) ถ้าเซกเมนต์ของไฟล์กำกับไม่พบอยู่ในกราฟของเซกเมนต์จะถูกนับเป็นความผิดพลาด

$$\text{เปอร์เซ็นต์ความผิดพลาดของเซกเมนต์} = \frac{\text{จำนวนเซกเมนต์ที่อยู่ในไฟล์กำกับแต่ไม่อยู่ในกราฟของเซกเมนต์}}{\text{จำนวนเซกเมนต์ทั้งหมดในไฟล์กำกับ}} \times 100\% \quad (3.1)$$

ในงานวิจัยนี้เราจะแบ่งความผิดพลาดนั้นออกเป็น 2 ประเภท คือ

1) ความผิดพลาดของเซกเมนต์ที่เกิดจากการแทรกของขอบเขตของหน่วยเสียง

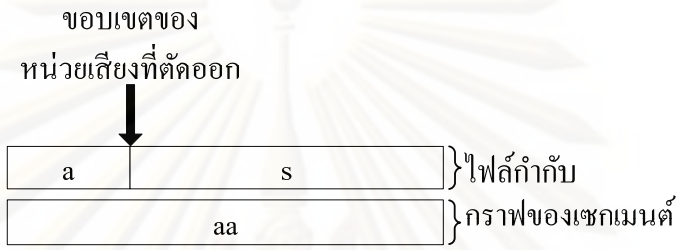
ความผิดพลาดประเภทนี้เกิดจากความผิดพลาดของการตัดแบ่งเสียงพูดเป็นเซกเมนต์ที่ทำให้มีขอบเขตของหน่วยเสียงแทรกมา จากรูปที่ 3.1 จะเห็นว่าเซกเมนต์ของหน่วยเสียง “kl” ในไฟล์กำกับไม่พบอยู่ในกราฟของเซกเมนต์ เนื่องจากมีขอบเขตของหน่วยเสียงแทรกขึ้นมา (ขอบเขตของหน่วยเสียงระหว่างหน่วยเสียง “l” และ “x”) จึงทำให้ในกราฟของเซกเมนต์มีเซกเมนต์ “l” และเซกเมนต์ “x” ซึ่งเป็นเซกเมนต์ที่ไม่ถูกต้อง



รูปที่ 3.1 ความผิดพลาดของเซกเมนต์ที่เกิดจากการแทรกของขอบเขตของหน่วยเสียง

2) ความผิดพลาดของเซกเมนต์ที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียง

ความผิดพลาดประเภทนี้เกิดจากความผิดพลาดของการตัดแบ่งเสียงพูดเป็นเซกเมนต์ ที่ทำให้มีขอบเขตของหน่วยเสียงหายไป จากรูปที่ 3.2 จะเห็นว่าเซกเมนต์ของหน่วยเสียง “a” และเซกเมนต์ของหน่วยเสียง “s” ในไฟล์กำกับไม่พบอยู่ในกราฟของเซกเมนต์ เนื่องจากขอบเขตของหน่วยเสียงระหว่างหน่วยเสียง “a” และ “s” หายไป จึงทำให้ในกราฟของเซกเมนต์มีเซกเมนต์ “aa” เพียงเซกเมนต์เดียว



รูปที่ 3.2 ความผิดพลาดของเซกเมนต์ที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียง

3.2.2 ความถูกต้อง

วิทยานิพนธ์นี้ใช้ความถูกต้องเป็นเปอร์เซ็นต์ในการเปรียบเทียบผลการจำแนกต่างๆ โดยค่าเปอร์เซ็นต์ความถูกต้องจะคำนวณได้ตามสมการ (3.2)

$$\text{เปอร์เซ็นต์ความถูกต้อง} = \frac{\text{จำนวนคำตอบที่ถูกต้อง}}{\text{จำนวนตัวอย่างทั้งหมดที่ใช้ในการทดสอบ}} \times 100\% \qquad (3.2)$$

3.2.3 ความแม่นยำ

วิทยานิพนธ์นี้ใช้ความแม่นยำเป็นเปอร์เซ็นต์ในการเปรียบเทียบผลการรู้จำเสียงพูดวิธีต่างๆ โดยค่าเปอร์เซ็นต์ความแม่นยำจะคำนวณได้ตามสมการ (3.3)

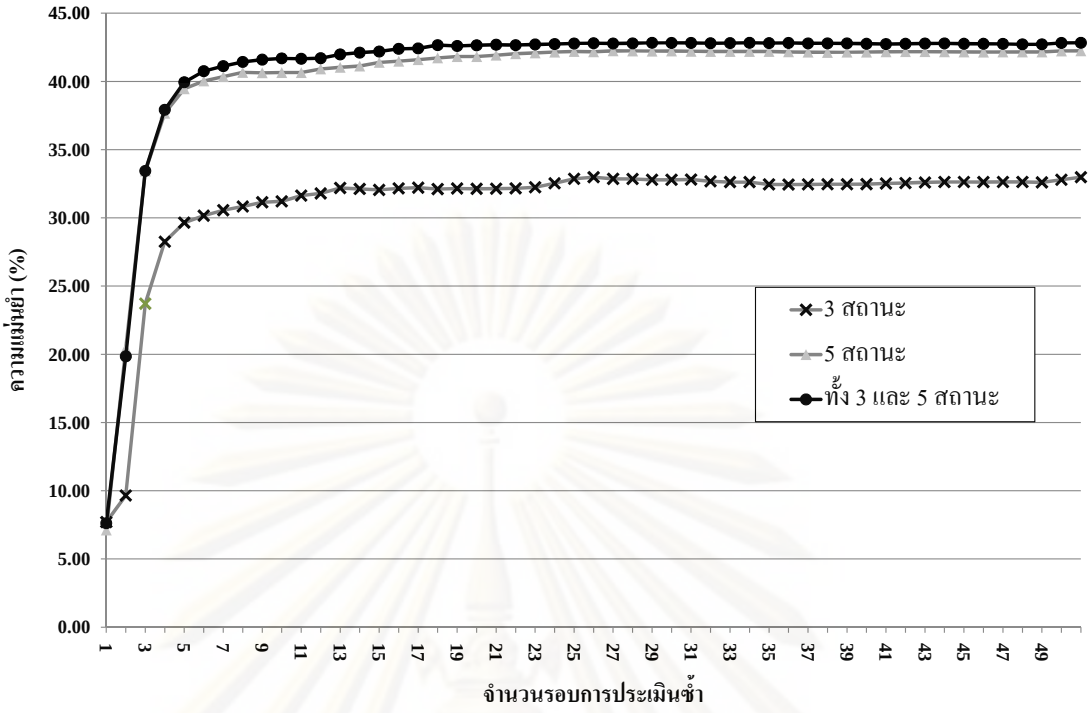
$$\text{เปอร์เซ็นต์ความแม่นยำ} = \frac{\text{จำนวนหน่วยเสียงที่ถูกต้อง} - \text{ความผิดพลาดแบบแทรก}}{\text{จำนวนหน่วยเสียงทั้งหมด}} \times 100\% \qquad (3.3)$$

โดยที่ จำนวนหน่วยเสียงที่ถูกต้อง = จำนวนหน่วยเสียงทั้งหมด – ความผิดพลาดแบบแทนที่ – ความผิดพลาดตัดออก

3.3 ระบบรู้จำเสียงพูดแบบอาศัยกรอบเวลา

ในงานวิจัยนี้เราจะใช้ระบบรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ ซึ่งเป็นการรู้จำเสียงพูดที่จัดอยู่ในการรู้จำเสียงพูดแบบอาศัยกรอบเวลา เพื่อเป็นระบบที่ใช้สำหรับเปรียบเทียบความแม่นยำในการรู้จำเสียงพูด ซึ่งในการวิจัยนี้ใช้ HTK (Hidden Markov Toolkit) [56] เป็นเครื่องมือสำหรับใช้ทดสอบความแม่นยำของระบบรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ ใช้สัมประสิทธิ์เมลฟรีเคว็นซีเคปสตรอล จำนวน 39 มิติเป็นเวกเตอร์คุณสมบัติของแต่ละกรอบเวลา และใช้แบบจำลองเกาส์เซียนที่มีเมทริกซ์ความแปรปรวนร่วมเกี่ยวตามแนวทแยง (Diagonal Covariance Gaussian Model) เป็นแบบจำลองที่ใช้ในการจำลองหน่วยเสียงภาษาไทยทั้งหมดที่มาจากฐานข้อมูลเสียงโลตัสจำนวน 75 หน่วยเสียง แบบจำลองทางเสียงจากฝึกฝนจากชุดข้อมูลเสียง TR และ PD ส่วนแบบจำลองทางภาษาจะใช้แบบจำลองทางภาษาแบบสองหน่วยซึ่งฝึกฝนจากไฟล์กำกับของชุดข้อมูลเสียง TR นอกจากนี้ในงานวิจัยนี้ได้มีความพยายามปรับแต่งความแม่นยำของระบบรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ โดยการปรับเปลี่ยนทอพอโลยีของแบบจำลองฮิตเดนมาร์คอฟ (HMM Topology) ออกเป็น 3 ชนิดด้วยกันคือ 1) ทอพอโลยีแบบ 3 สถานะจากซ้ายไปขวา 2) ทอพอโลยีแบบ 5 สถานะจากซ้ายไปขวา และ 3) ทอพอโลยีแบบที่ใช้ทั้ง 3 และ 5 สถานะจากซ้ายไปขวา โดยที่เราจะใช้ทอพอโลยีแบบสามสถานะกับสระเสียงสั้นและส่วนที่เหลือจะใช้ทอพอโลยีแบบห้าสถานะ นอกจากนี้ยังมีการปรับเปลี่ยนจำนวนรอบของการประมาณซ้ำ (Re-estimate) จาก 1 ถึง 50 รอบโดยจะเลือกแบบจำลองทางเสียงที่ดีที่สุดที่ได้ทดลองปรับเปลี่ยนทั้งทอพอโลยีและจำนวนรอบการประมาณซ้ำจากการทดสอบความแม่นยำของการรู้จำเสียงพูดระดับหน่วยเสียงบนชุดข้อมูลเสียง DT มาใช้ เป็นระบบที่ใช้สำหรับเปรียบเทียบความแม่นยำ

รูปที่ 3.3 จะแสดงถึงความแม่นยำของการรู้จำเสียงของระบบรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ ที่จำนวนรอบการประมาณซ้ำ และทอพอโลยีแบบต่างๆ จะพบว่าเมื่อจำนวนรอบการประมาณซ้ำมากกว่า 17 รอบแล้วความแม่นยำจะไม่แตกต่างกันมาก นอกจากนี้ทอพอโลยีแบบที่ใช้ทั้ง 3 และ 5 สถานะจะให้ผลการรู้จำเสียงพูดที่ดีที่สุดเมื่อเทียบกับทอพอโลยีต่างๆ



รูปที่ 3.3 ความแม่นยำของการรู้จำเสียงพูดระบบหน่วยเสียงของระบบรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ ที่จำนวนรอบการประมาณซ้ำ และทอพอโลยีแบบต่างๆ

3.4 ระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์

ในงานวิจัยนี้ เวกเตอร์คุณสมบัติที่ใช้สำหรับเซกเมนต์ และขอบเขตของหน่วยเสียงประยุกต์มาจากงานของ Halberstadt และ Glass [3, 4] ซึ่งได้มาจากการทดสอบความสามารถในการจำแนกกับชุดข้อมูลเสียง DT โดยมีการเลือกเวกเตอร์คุณสมบัติที่มีความสามารถในการจำแนกได้ดีที่สุด โดยที่เวกเตอร์คุณสมบัติของเซกเมนต์จะเกิดจากการนำสัมประสิทธิ์เมลฟรีเคว็นซีเคปสตรอล จำนวน 39 มิติ จาก 3 ส่วนของเซกเมนต์มาต่อกัน โดยที่ส่วนแรกจะเป็น 30% แรกของเซกเมนต์ ส่วนที่สองเป็น 40% ถัดมา และส่วนที่สามเป็น 30% ถัดมาเช่นกัน นอกจากนี้ยังมีการใช้แบบจำลองหน่วยที่ไม่ใช่หน่วยเสียงมาเพื่อจำลองเซกเมนต์ที่ไม่ได้อยู่ในเส้นทางที่ค้นหา สำหรับขอบเขตของหน่วยเสียงนั้นจะเกิดจากการนำสัมประสิทธิ์เมลฟรีเคว็นซีเคปสตรอล จำนวน 13 มิติที่ทุกๆ 20 มิลลิวินาทีของทั้งสองด้านของขอบเขตของหน่วยเสียงมาต่อกันจนเป็นเวกเตอร์คุณสมบัติขนาด 78 มิติ แบบจำลองเสียงทุกตัวใช้แบบจำลองเกาส์เซียนที่มีเมทริกซ์ความแปรปรวนร่วมเกี่ยวตามแนวทแยงในการจำลองแบบจำลองทางเสียงเช่นเดียวกับระบบรู้จำเสียงพูดแบบอาศัยกรอบเวลา ในการฝึกฝนแบบจำลองทางเสียงจะต้องมีการรู้ตำแหน่งของขอบเขตของหน่วยเสียงที่แน่นอน ในงานวิจัยนี้จะดำเนินการกำกับตำแหน่งของขอบเขตของหน่วยเสียงโดยการใช้ระบบรู้จำเสียงพูดแบบอาศัยกรอบเวลามารู้จำเสียงพูดโดยให้คำตอบไปที่เรียกว่า “Forced Alignment” เพื่อจะให้ได้ไฟล์กำกับที่

มีเวลาที่ระบุตำแหน่งของขอบเขตของหน่วยเสียง ในส่วนของแบบจำลองทางภาษานั้นจะใช้แบบจำลองทางภาษาแบบสองหน่วยเช่นเดียวกับระบบรู้จำเสียงพูดแบบอาศัยกรอกเวลาเช่นกัน กราฟของเซกเมนต์จะสร้างจากการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็นโดยจะใช้ผลจากการรู้จำเสียงพูดระดับหน่วยเสียงที่ดีสุดจำนวน 20 ลำดับจากระบบรู้จำเสียงพูดแบบอาศัยกรอบเวลาที่กล่าวไว้ในข้างต้น ซึ่งระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ใช้ในงานวิจัยนี้จะแสดงดังรูปที่ 3.4

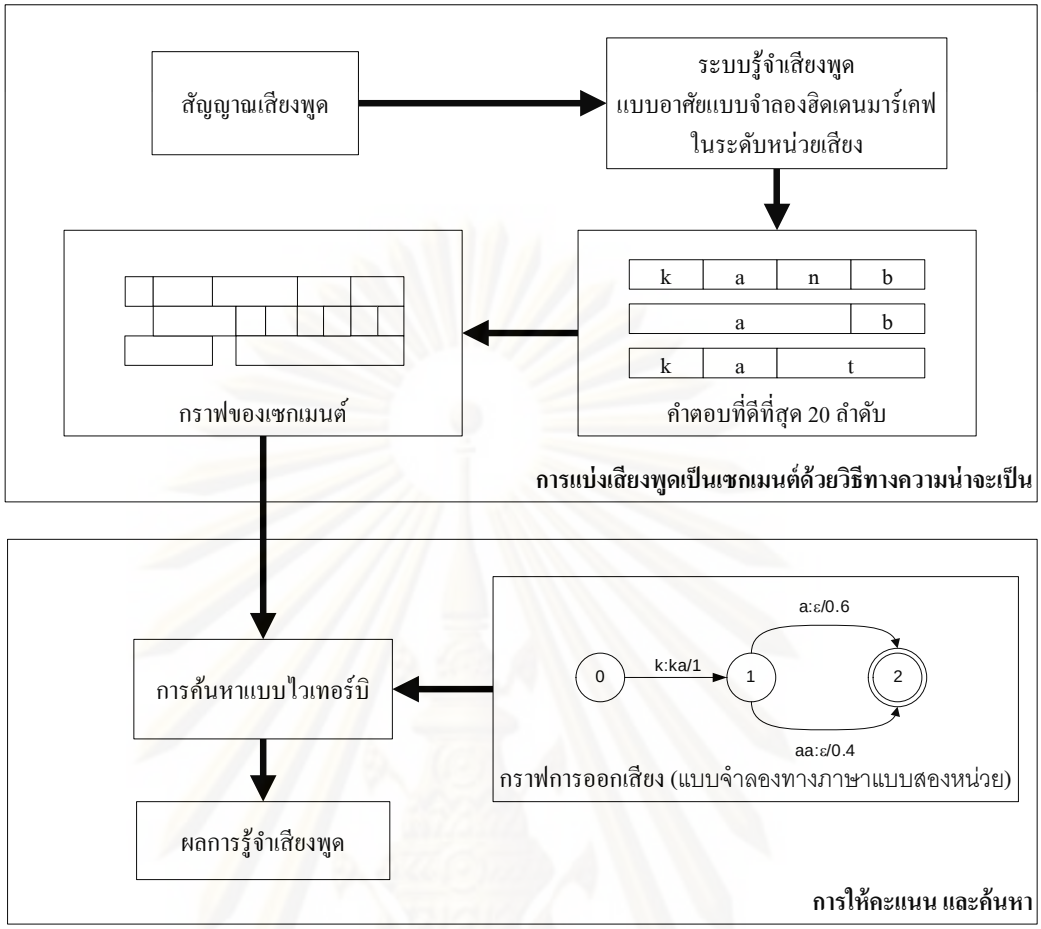
อย่างไรก็ตาม งานวิจัยนี้จะเป็นการทดสอบความแม่นยำในการรู้จำเสียงพูดในระดับหน่วยเสียง ดังนั้น จึงได้มีการปรับปรุงหลักความน่าจะเป็นที่ใช้สำหรับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่กล่าวในบทที่ 2 โดยเปลี่ยนจุดมุ่งหมายของการรู้จำเสียงพูดเป็นการค้นหาลำดับของหน่วยเสียง $U^* = u_1, \ldots, u_N$ ที่ทำให้ค่าความน่าจะเป็นภายหลังมากสุด ได้ดังสมการต่อไปนี้

$$U^* = \arg\max_{S,U} P(S,U \mid A) \tag{3.4}$$

เมื่อการใช้แบบจำลองหน่วยที่ไม่ใช่หน่วยเสียงตามวิธีของ Glass [2] ก็จะสามารถเขียนสมการใหม่ได้เป็นสมการที่ (3.5)

$$U^* = \arg\max_{S,U} \prod_{i=1}^{n} \frac{P(x_i \mid u_i)}{P(x_i \mid \overline{\alpha})} P(s_i \mid u_i) P(U) \tag{3.5}$$

โดยที่ $P(x_i|u_i)$ จะเป็นแบบจำลองทางเสียงของแต่ละหน่วยเสียง ขณะที่ $P(x_i|\overline{\alpha})$ เป็นแบบจำลองของเสียงที่ไม่ใช่หน่วยเสียง $P(s_i|u_i)$ เป็นแบบจำลองระยะเวลา และ $P(U)$ เป็นแบบจำลองทางภาษาแบบสองหน่วย



รูปที่ 3.4 การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ในงานวิจัยนี้

3.5 ความผิดพลาดของเซกเมนต์ในกราฟของเซกเมนต์

จากการทดสอบโดยการสร้างกราฟของเซกเมนต์ตามวิธีที่กล่าวไว้ในเรื่องระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ จากชุดข้อมูลเสียง ET ของฐานข้อมูลเสียงโลตัส จะพบว่ามีความผิดพลาดของเซกเมนต์ ดังนี้

ตารางที่ 3.1 รายการความผิดพลาดของเซกเมนต์

ความผิดพลาดของเซกเมนต์	เปอร์เซ็นต์ความผิดพลาด (%)
ความผิดพลาดของเซกเมนต์ทั้งหมด	27.70
ความผิดพลาดของเซกเมนต์ที่เกิดจากการแทรกของขอบเขตของหน่วยเสียง	15.80
ความผิดพลาดของเซกเมนต์ที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียง	11.90

จากตารางที่ 3.1 จะสังเกตได้ว่าจากความผิดพลาดของเซกเมนต์ทั้งหมดนั้น ความผิดพลาดของเซกเมนต์ที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงมีจำนวนมากกว่าความผิดพลาดของเซกเมนต์ที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียง

3.6 ผลการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ในกราฟของเซกเมนต์ต่างๆ

การทดลองนี้จะเป็นการนำระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์มาทดสอบการรู้จำเสียงพูดในระดับหน่วยเสียงเพื่อเปรียบเทียบความแม่นยำของการรู้จำเสียงพูดที่ค้นหาคำตอบจากกราฟของเซกเมนต์แบบต่างๆ คือ 1) กราฟของเซกเมนต์ที่ได้จากการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น 2) กราฟของเซกเมนต์ที่จากไฟล์กำกับ และ 3) กราฟของเซกเมนต์ที่ได้จาก 1) และ 2) รวมกัน นอกจากนี้ยังมีการเปรียบเทียบผลการรู้จำเสียงระหว่างระบบรู้จำเสียงพูดที่ใช้แบบจำลองทางเสียงของเซกเมนต์เพียงอย่างเดียว กับระบบรู้จำเสียงพูดที่ใช้ทั้งแบบจำลองทางเสียงของเซกเมนต์ และขอบเขตของหน่วยเสียง

ตารางที่ 3.2 ผลการรู้จำเสียงพูดระดับหน่วยเสียงของระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ในกราฟของเซกเมนต์ต่างๆ

กราฟของเซกเมนต์	**ความแม่นยำ (%)**	
	แบบจำลองเซกเมนต์	**แบบจำลองเซกเมนต์ และขอบเขตของหน่วยเสียง**
1) กราฟของเซกเมนต์ที่ได้จากการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น	47.70	51.47
2) กราฟของเซกเมนต์ที่จากไฟล์กำกับ	69.46	76.91
3) กราฟของเซกเมนต์ที่ได้จาก 1) และ 2) รวมกัน	59.25	65.46

จากผลการรู้จำเสียงพูดระดับหน่วยเสียงที่แสดงในตารางที่ 3.2 พบว่าระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์นั้นสามารถรู้จำเสียงพูดได้แม่นยำ 69.46% และ 76.91% เมื่อมีทุกเซกเมนต์ในกราฟของเซกเมนต์นั้นถูกต้อง ในขณะที่ผลการรู้จำเสียงพูดที่ดำเนินการบนกราฟของเซกเมนต์ที่ได้จากการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็นมีความแม่นยำเพียงแค่ 47.70% และ 51.47% เท่านั้น ดังนั้นเราจึงสามารถตั้งสมมุติฐานได้ว่าคุณภาพของกราฟของเซกเมนต์ ซึ่งจะเป็นตัวบอกจำนวนเซกเมนต์ที่ถูกต้องที่อยู่ในกราฟของเซกเมนต์ เป็นปัจจัยสำคัญที่มีผลกระทบต่อ

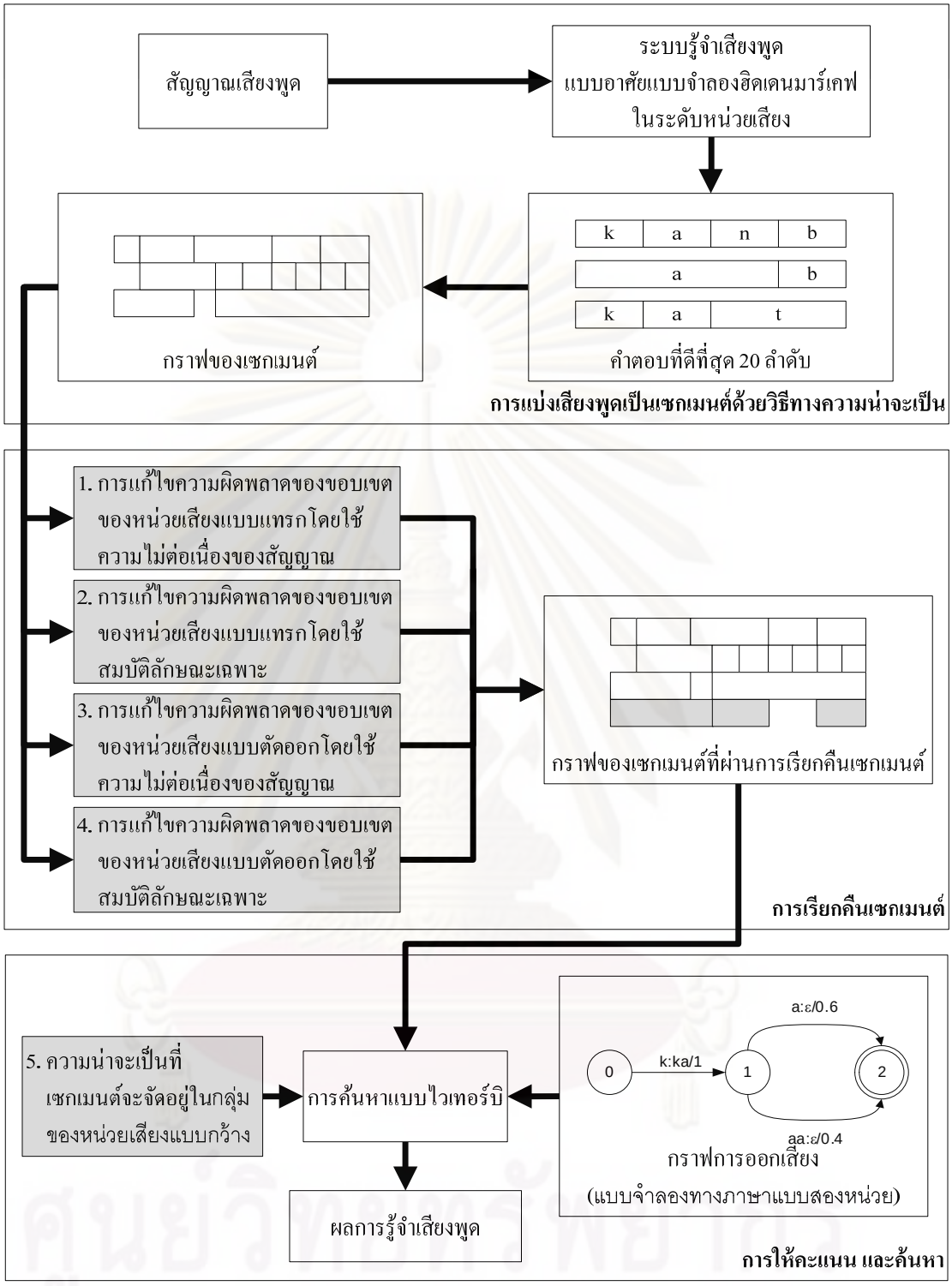
ความแม่นยำในการรู้จำเสียงพูดของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ จากผลในตารางที่ 3.2 จะเห็นว่าระหว่างผลการรู้จำของ 1) กราฟของเซกเมนต์ที่ได้จากการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น และ 2) กราฟของเซกเมนต์ที่จากไฟล์กำกับ ยังมีช่องว่างในการปรับปรุงความแม่นยำได้อีก (แต่ในความเป็นจริงแล้วความแม่นยำที่จะมาจากการเรียกคืนเซกเมนต์นั้นจะมีโอกาสได้ถึงกรณีที่ 3) ซึ่งเป็นกรณีที่มีเซกเมนต์ที่ถูกต้องทั้งหมดอยู่ปนกับเซกเมนต์ที่ไม่ถูกต้องด้วย ดังนั้นในทางอุดมคติแล้วถ้าเราสามารถเรียกคืนเซกเมนต์ที่ถูกต้องทั้งหมดได้ ความแม่นยำก็ไม่อาจจะสูงกว่าผลการรู้จำของกรณีที่ 3 ได้) ดังนั้น ในงานวิจัยนี้ก็จะมีการมุ่งเน้นไปที่การปรับปรุงคุณภาพของกราฟของเซกเมนต์ โดยการพยายามเรียกคืนเซกเมนต์ที่ถูกต้องลงในกราฟของเซกเมนต์ด้วยการแก้ไขความผิดพลาดของเซกเมนต์ที่เกิดจากการแทรกของขอบเขตของหน่วยเสียง และความผิดพลาดของเซกเมนต์ที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียง

บทที่ 4

การปรับปรุงความแม่นยำของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

วิทยานิพนธ์ฉบับนี้ได้นำเสนอวิธีการปรับปรุงความแม่นยำของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์โดยมุ่งเน้นไปที่ 2 ส่วนหลักๆด้วยกัน คือ 1) การปรับปรุงคุณภาพของกราฟของเซกเมนต์ และ 2) การปรับปรุงการให้คะแนนโดยเพิ่มคะแนนความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้าง ในส่วนของการปรับปรุงคุณภาพของกราฟของเซกเมนต์ เราได้เพิ่มขั้นตอนการเพิ่มจำนวนเซกเมนต์ที่ถูกต้องลงในกราฟของเซกเมนต์ที่สร้างมาจากการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น โดยการแก้ไขความผิดพลาดที่เกิดจากขอบเขตของหน่วยเสียงที่แทรกมา และความผิดพลาดที่เกิดจากขอบเขตของหน่วยเสียงหายไป วิธีการที่จะนำมาแก้ไขความผิดพลาดนั้นจะใช้การวัดความไม่ต่อเนื่องของสัญญาณ และใช้หลักสมบัติลักษณะเฉพาะ ถึงแม้ว่าการเพิ่มจำนวนเซกเมนต์ที่ถูกต้องเข้าไปจะทำให้ขนาดของกราฟของเซกเมนต์ใหญ่ขึ้น แต่ก็จะทำให้การรู้จำเสียงพูดมีความแม่นยำมากขึ้น

นอกจากการปรับปรุงคุณภาพของกราฟของเซกเมนต์แล้ว ในวิทยานิพนธ์นี้จะมีการปรับปรุงการให้คะแนนของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์โดยการนำความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างมาใช้ในขั้นตอนการให้คะแนน และค้นหาคำตอบ วิธีการที่วิทยานิพนธ์นี้นำเสนอจะแสดงในรูปที่ 4.1 ซึ่งได้กำกับเป็นตัวเลข 1 ถึง 5 และจะได้มีการอธิบายรายละเอียดของแต่ละวิธีดังต่อไปนี้



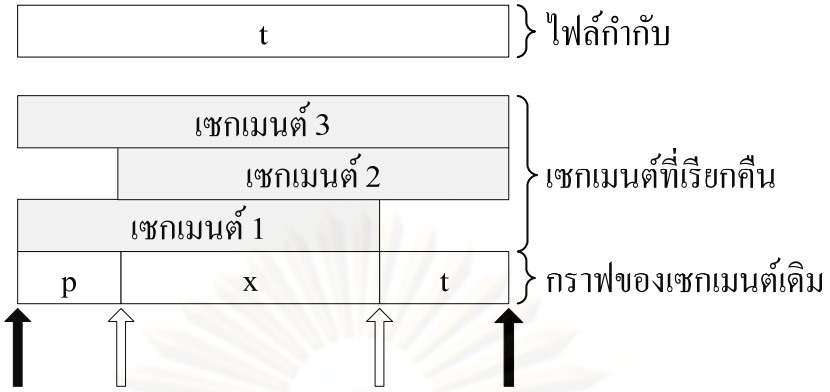
รูปที่ 4.1 การปรับปรุงความแม่นยำของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่นำเสนอในวิทยานิพนธ์นี้

4.1 การแก้ไขความผิดพลาดของขอบเขตของหน่วยเสียงแบบแทรกโดยใช้ความไม่เนื่องของสัญญาณ

การปรับปรุงคุณภาพของกราฟของเซกเมนต์ของวิธีทำโดยตรวจสอบที่ขอบเขตของหน่วยเสียงทุกขอบเขตในกราฟของเซกเมนต์ว่ามีโอกาสที่จะเป็นขอบเขตของหน่วยเสียงจริง หรือเป็นขอบเขตของหน่วยเสียงที่เกิดจากความผิดพลาดแบบแทรกของการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น ในที่นี้จะใช้ค่าความไม่ต่อเนื่องของสัญญาณเป็นตัวระบุความเป็นขอบเขตของหน่วยเสียง ซึ่งโดยปกติแล้วบริเวณขอบเขตของหน่วยเสียงจะมีการเปลี่ยนแปลงของสัญญาณอย่างฉับพลัน และจะทำให้มีค่าความไม่ต่อเนื่องของสัญญาณสูง โดยที่ในงานวิจัยนี้จะวัดค่าความไม่ต่อเนื่องของสัญญาณจากการหาระยะห่างยูคลีเดียนของสัมประสิทธิ์เซปสตรัมบนสเกลเมลระหว่างกรอบเวลาที่อยู่ด้านซ้าย และกรอบเวลาที่อยู่ด้านขวาของขอบเขตของหน่วยเสียงที่เราต้องการตรวจสอบจำนวน 3 คู่กรอบเวลา ต่อจากนั้นจะนำค่าความไม่ต่อเนื่องของสัญญาณมาสร้างเป็นเวกเตอร์ขนาด 3 มิติ และใช้ตัวจำแนกว่าขอบเขตของหน่วยเสียงในกราฟของเซกเมนต์นี้เป็นขอบเขตของหน่วยเสียงจริงหรือไม่ ซึ่งตัวจำแนกขอบเขตของหน่วยเสียงดังกล่าวจะใช้แบบจำลองเกาส์เซียนเป็นแบบจำลองของขอบเขตของหน่วยเสียงจริง และส่วนที่ไม่ใช่ขอบเขตของหน่วยเสียง (ตรงกลางของเซกเมนต์)

ถ้าตัวจำแนกขอบเขตของหน่วยเสียงจำแนกได้ว่าขอบเขตของหน่วยเสียงในกราฟของเซกเมนต์เป็นขอบเขตของหน่วยเสียงที่แทรกมา วิธีการนี้จะดำเนินการเพิ่มเซกเมนต์ที่มีขนาดครอบคลุมของเซกเมนต์ที่ขั้นด้วยขอบเขตของหน่วยเสียงดังกล่าวลงไปในกราฟของเซกเมนต์ และจะมีการตรวจสอบเช่นเดียวกันนี้กับเซกเมนต์ที่เพิ่มเข้าไปใหม่ด้วย จนกว่าจะไม่มีการเพิ่มเซกเมนต์ใหม่ลงไปในกราฟของเซกเมนต์

ตัวอย่างของอัลกอริทึมที่กล่าวมาได้แสดงอยู่ในรูปที่ 4.2 สมมุติว่าขอบเขตของหน่วยเสียงระหว่างเซกเมนต์ “p” และเซกเมนต์ “x” (ขอบเขตของหน่วยเสียง “p-x”) นั้นถูกจำแนกเป็นขอบเขตของหน่วยเสียงที่แทรกมา (ลูกศรหัวโปร่ง) “เซกเมนต์ 1” ซึ่งเป็นเซกเมนต์เปรียบเสมือนการนำเซกเมนต์ “p” และเซกเมนต์ “x” มารวมกัน จะถูกเพิ่มลงไปในกราฟของเซกเมนต์ ขณะที่ขอบเขตของหน่วยเสียง “x-t” นั้นถูกจำแนกเป็นขอบเขตของหน่วยเสียงที่แทรกมา “เซกเมนต์ 2” ก็จะถูกเพิ่มลงในกราฟของเซกเมนต์เช่นกัน นอกจากนี้ “เซกเมนต์ 3” ก็จะถูกเพิ่มลงในกราฟของเซกเมนต์ เนื่องจากขอบเขตของหน่วยเสียงระหว่างเซกเมนต์ “p” และเซกเมนต์ “เซกเมนต์ 2” ถูกจำแนกเป็นขอบเขตของหน่วยเสียงที่แทรกมา



รูปที่ 4.2 การแก้ไขความผิดพลาดของขอบเขตของหน่วยเสียงแบบแทรกโดยใช้ความไม่เนื่องของสัญญาณ

(ลูกศรหัวทึบจะเป็นขอบเขตของหน่วยเสียงที่จำแนกได้ว่าเป็นขอบเขตของหน่วยเสียงจริง ส่วนลูกศรหัวโปร่งจะเป็นขอบเขตของหน่วยเสียงที่จำแนกได้ว่าเป็นขอบเขตของหน่วยเสียงที่แทรกมา)

4.2 การแก้ไขความผิดพลาดของขอบเขตของหน่วยเสียงแบบแทรกโดยใช้สมบัติลักษณะเฉพาะ

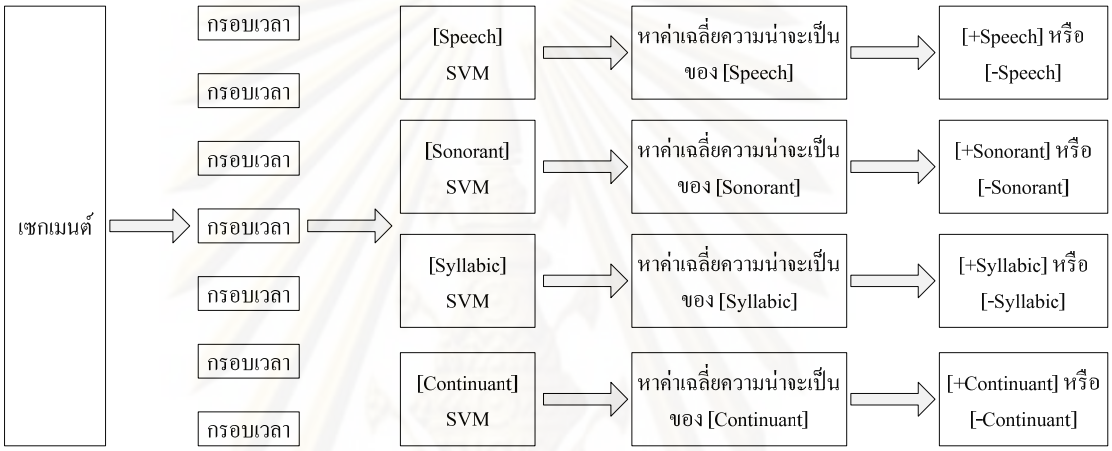
ในงานวิจัยนี้จะใช้สมบัติลักษณะเฉพาะที่เกี่ยวกับลักษณะการออกเสียงของกลไกการออกเสียงของมนุษย์ โดยลักษณะการออกเสียงที่ใช้ในงานวิจัยนี้ ได้แก่ [Speech] [Sonorant] [Syllabic] และ [Continuant] โดยในงานวิจัยนี้จะมีการประยุกต์ใช้พารามิเตอร์คุณสมบัติทางเสียงจากงานของ Juneja [18] โดยเราจะนำพารามิเตอร์คุณสมบัติทางเสียงดังกล่าวมาคัดเลือกโดยใช้ ANOVA เพื่อดูว่าพารามิเตอร์ต่างๆมีความเหมาะสมที่จะมาใช้แบ่งแยกความแตกต่างของของลักษณะการออกเสียงแต่ละประเภทเท่าไร โดยเราจะเลือกพารามิเตอร์ที่สามารถแบ่งแยกความแตกต่างได้ดี ซึ่งพารามิเตอร์คุณสมบัติทางเสียงที่เลือกมานั้นจะแสดงในตารางที่ 4.1

ตารางที่ 4.1 พารามิเตอร์คุณสมบัติทางเสียงที่ใช้จำแนกลักษณะการออกเสียง
(F_3 คือความถี่สั่นพ้องลำดับที่ 3 และ f_s ความถี่การชักตัวอย่าง (Sampling Frequency))

ลักษณะการออกเสียง	พารามิเตอร์คุณสมบัติทางเสียง
[Speech]	พลังงานในช่วงความถี่ 100 ถึง 400 เฮิร์ต พลังงานในช่วงความถี่ 0 ถึง F_3-1000 เฮิร์ต พลังงานในช่วงความถี่ F_3 ถึง f_s/2 เฮิร์ต พลังงานในช่วงความถี่ 640 ถึง 2800 เฮิร์ต พลังงานในช่วงความถี่ 2000 ถึง 3000 เฮิร์ต ระดับความก้องของเสียง ระดับความไม่เป็นคาบของสัญญาณเสียง
[Sonorant]	พลังงานในช่วงความถี่ 100 ถึง 400 เฮิร์ต พลังงานในช่วงความถี่ 0 ถึง F_3-1000 เฮิร์ต พลังงานในช่วงความถี่ 640 ถึง 2800 เฮิร์ต ระดับความก้องของเสียง ระดับความไม่เป็นคาบของสัญญาณเสียง
[Syllabic]	พลังงานในช่วงความถี่ 0 ถึง F_3-1000 เฮิร์ต พลังงานในช่วงความถี่ F_3-1000 ถึง f_s/2 เฮิร์ต พลังงานในช่วงความถี่ 640 ถึง 2800 เฮิร์ต พลังงานในช่วงความถี่ 2000 ถึง 3000 เฮิร์ต
[Continuant]	พลังงานในช่วงความถี่ 100 ถึง 400 เฮิร์ต พลังงานในช่วงความถี่ 0 ถึง F_3-1000 เฮิร์ต พลังงานในช่วงความถี่ F_3 ถึง f_s/2 เฮิร์ต ระดับความก้องของเสียง ระดับความไม่เป็นคาบของสัญญาณเสียง

งานวิจัยนี้จะใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นตัวจำแนก ค่าบวก ลบ ของแต่ละลักษณะการออกเสียง ซึ่งซัพพอร์ตเวกเตอร์แมชชีนเป็นตัวจำแนกที่สามารถจำแนกสิ่งสองสิ่งได้อย่างดี ในการฝึกฝนซัพพอร์ตเวกเตอร์แมชชีนจะฝึกฝนจากตัวอย่างเสียงจากชุดข้อมูลเสียง TR และ PD โดยจะมีจำนวนตัวอย่าง 500,000 ตัวอย่างที่มีการดึงเอาพารามิเตอร์คุณสมบัติทางเสียงออกจากแต่ละกรอบเวลา แต่อย่างไรก็ตามในงานวิจัยนี้จะใช้ซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกว่าเซกเมนต์ถูกจำแนกอยู่ในลักษณะการออกเสียงแบบใด ดังนั้นงานวิจัยนี้จึงใช้การหาค่าเฉลี่ยของความน่าจะเป็น

ที่ได้มาจากซัพพอร์ตเวกเตอร์แมชชีนของแต่ละกรอบเวลาเพื่อนำมาตัดสินใจในการจำแนกเซกเมนต์ ดังที่แสดงในรูปที่ 4.3 ในการเลือกเคอร์เนลของซัพพอร์ตเวกเตอร์แมชชีนจะมีการเลือกจากการทดสอบใช้เคอร์เนลแบบเส้นตรง แบบพหุนาม แบบอาร์บีเอฟ แบบซิกมอยด์ บนชุดข้อมูลเสียง DT ซึ่งจะได้ผลว่าเคอร์เนลแบบอาร์บีเอฟ สามารถจำแนกลักษณะการออกเสียงได้ดีที่สุด ดังนั้น งานวิจัยนี้เลือกใช้เคอร์เนลแบบฟังก์ชันอาร์บีเอฟ และในงานวิจัยนี้จะใช้ LIBSVM [25] เป็นชุดเครื่องมือในการดำเนินการฝึกฝนซัพพอร์ตเวกเตอร์แมชชีน และจำแนกโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน



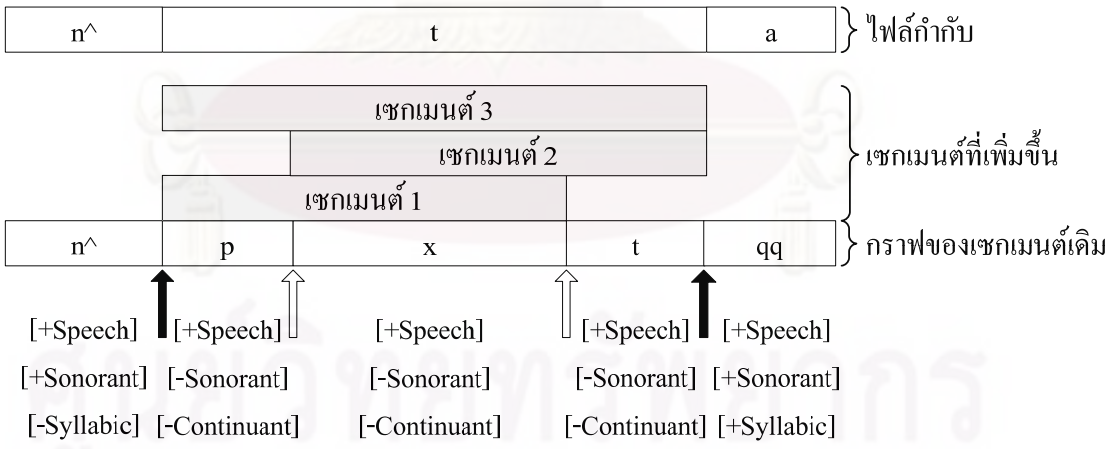
รูปที่ 4.3 การจำแนกลักษณะการออกเสียงโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน

เนื่องจากในไฟล์กำกับของฐานข้อมูลเสียงโลตัส ช่วงกัก และช่องระเบิดของเสียงพยัญชนะระเบิดจะมีการกำกับให้เป็นหน่วยเสียงเพียงหน่วยเดียว ดังนั้นในการตรวจสอบลักษณะการออกเสียงแบบ [-Continuant] เราจึงต้องตรวจสอบหาช่วงกัก ซึ่งในที่นี้เรามีการตรวจสอบเมื่อกับที่เราตรวจสอบเสียงเงียบ [-Speech] ในส่วนแรกของเซกเมนต์ ซึ่งถ้าตรวจสอบเจอช่วงกักในเซกเมนต์เซกเมนต์นั้นก็จะถูกจำแนกเป็นเสียงพยัญชนะระเบิด [-Continuant]

ตามหลักการการออกเสียงเมื่อมีการเปลี่ยนแปลงลักษณะของการออกเสียงก็จะถือว่าเป็นหน่วยเสียงที่ต่างกัน ดังนั้นจึงมีการใช้หลักการนี้ในการเรียกคืนเซกเมนต์ที่ถูกต้องที่เกิดจากความผิดพลาดของขอบเขตของหน่วยเสียงแบบแทรก ซึ่งในงานวิจัยนี้ตั้งสมมุติฐานว่าถ้าไม่มีการเปลี่ยนแปลงของลักษณะของการออกเสียงของเซกเมนต์ที่อยู่ด้านซ้าย และด้านขวาของขอบเขตของหน่วยเสียง ก็จะถือว่าขอบเขตของหน่วยเสียงนั้นเป็นขอบเขตของหน่วยเสียงที่เกิดจากความผิดพลาดแบบแทรก และจะเพิ่มเซกเมนต์ที่เปรียบเสมือนเซกเมนต์ด้านซ้าย และด้านขวาของขอบเขตของหน่วยเสียงรวมกัน ซึ่งเป็นกลไกเดียวกับการแก้ไขความผิดพลาดของขอบเขตของหน่วยเสียงแบบแทรกโดยใช้ความไม่เนื่องของสัญญาณ แต่ก่อนที่จะมีการตรวจสอบการ

เปลี่ยนแปลงของลักษณะการออกเสียงได้นั้น ขั้นแรก พารามิเตอร์คุณสมบัติทางเสียงของทุกๆ กรอบเวลาของสัญญาณเสียงที่เราต้องการตรวจสอบจะถูกดึงออกมา และนำมาจำแนกว่ามีลักษณะการออกเสียงอย่างไร โดยใช้ซัพพอร์ตเวกเตอร์แมชชีน จากนั้นนำค่าเฉลี่ยของความน่าจะเป็นที่ได้จากผลลัพธ์ของซัพพอร์ตเวกเตอร์แมชชีนมาเป็นตัวตัดสินใจว่าแต่ละเซกเมนต์ในกราฟของเซกเมนต์มีลักษณะการออกเสียงอย่างไร จึงสามารถตรวจหาการเปลี่ยนแปลงของลักษณะการออกเสียงได้

รูปที่ 4.4 แสดงตัวอย่างของการแก้ไขความผิดพลาดของขอบเขตของหน่วยเสียงแบบแทรก โดยใช้การตรวจสอบการเปลี่ยนแปลงของลักษณะการออกเสียง จากตัวอย่างจะเห็นว่าเซกเมนต์ “n^” จะถูกจำแนกเป็น [+Speech][+Sonorant][-Syllabic] (พยัญชนะนาสิก) ในขณะที่เซกเมนต์ “p” ถูกจำแนกเป็น [+Speech][-Sonorant][-Continuant] (พยัญชนะหยุด) ขอบเขตของหน่วยเสียง “n^-p” ที่อยู่ระหว่างเซกเมนต์ทั้งสองนี้มีการเปลี่ยนแปลงของลักษณะการออกเสียง จึงถูกจำแนกได้เป็นขอบเขตของหน่วยเสียงจริง (ลูกศรหัวทึบ) แต่ในขณะที่เซกเมนต์ที่อยู่ด้านซ้ายและขวาของขอบเขตของหน่วยเสียง “p-x” และ “x-t” (ลูกศรหัวโปร่ง) ถูกจำแนกได้เป็น [+Speech][-Sonorant][-Continuant] ซึ่งขอบเขตของหน่วยเสียงดังกล่าวจึงไม่มีการเปลี่ยนแปลงของลักษณะการออกเสียง ดังนั้นขอบเขตของหน่วยเสียงทั้งสอง จึงถูกจำแนกเป็นขอบเขตของหน่วยเสียงที่เกิดจากความผิดพลาดแบบแทรก ดังนั้นจึงมีการเพิ่มเซกเมนต์ “เซกเมนต์ 1” “เซกเมนต์ 2” และ “เซกเมนต์ 3” ลงไปในกราฟของเซกเมนต์



รูปที่ 4.4 การแก้ไขความผิดพลาดของขอบเขตของหน่วยเสียงแบบแทรกโดยใช้สมบัติลักษณะเฉพาะ

4.3 การแก้ไขความผิดพลาดของขอบเขตของหน่วยเสียงแบบตัดออกโดยใช้ความไม่เนื่องของสัญญาณ

ในการจัดการตรวจหาขอบเขตของหน่วยเสียงที่หายไปที่เกิดจากความผิดพลาดแบบตัดออกของการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็นจะตรวจหาจากความไม่ต่อเนื่องของสัญญาณเสียง ตั้งอยู่บนสมมติฐานที่ว่าบริเวณขอบเขตของหน่วยเสียงจะมีความไม่ต่อเนื่องของสัญญาณมาก ดังนั้นเราจะตรวจหาขอบเขตของหน่วยเสียงที่หายไปโดยหาค่าความไม่ต่อเนื่องของสัญญาณที่สูงที่สุดภายในเซกเมนต์ เมื่อตรวจเจออัลกอริทึมจะดำเนินการตรวจสอบที่ตำแหน่งที่ตรวจเจออีกทีว่าเป็นขอบเขตของหน่วยเสียงหรือไม่ โดยใช้ตัวจำแนกเดียวกับที่ใช้ในการแก้ไขความผิดพลาดของขอบเขตของหน่วยเสียงแบบแทรกโดยใช้ความไม่เนื่องของสัญญาณ ถ้าตัวจำแนกจำแนกว่าตำแหน่งที่ตรวจเจอนั้นเป็นขอบเขตของหน่วยเสียงจริง จะมีการเพิ่มเซกเมนต์ที่เหมือนเซกเมนต์ที่แบ่งโดยตำแหน่งที่ตรวจเจอขอบเขตของหน่วยเสียงลงในกราฟของเซกเมนต์จำนวน 2 เซกเมนต์

ค่าความไม่ต่อเนื่องของสัญญาณก็จะวัดจากการหาจากระยะห่างยูคลีเดียนของสัมประสิทธิ์เซปสตรัมบนสเกลเมลระหว่างกรอบเวลาที่อยู่ด้านซ้าย และกรอบเวลาที่อยู่ด้านขวาของขอบเขตของหน่วยเสียงเช่นเดียวกับวิธีการหาค่าความไม่ต่อเนื่องของสัญญาณในการแก้ไขความผิดพลาดของขอบเขตของหน่วยเสียงแบบแทรกโดยใช้ความไม่เนื่องของสัญญาณ

4.4 การแก้ไขความผิดพลาดของขอบเขตของหน่วยเสียงแบบตัดออกโดยใช้สมบัติลักษณะเฉพาะ

ด้วยสมมติฐานเดียวกันกับการแก้ไขความผิดพลาดของขอบเขตของหน่วยเสียงแบบแทรกโดยใช้สมบัติลักษณะเฉพาะ ในการแก้ไขความผิดพลาดของขอบเขตของหน่วยเสียงแบบตัดออกโดยใช้สมบัติลักษณะเฉพาะก็จะมีการตรวจหาขอบเขตของหน่วยเสียงจากการเปลี่ยนแปลงของลักษณะการออกเสียง โดยจะมีขั้นตอนดังนี้ 1) ดึงพารามิเตอร์คุณสมบัติทางเสียงออกจากทุกๆ กรอบเวลา 2) จำแนกแต่ละกรอบเวลาว่ามีลักษณะการออกเสียงเป็นแบบใด 3) ค้นหาการเปลี่ยนแปลงของลักษณะการออกเสียงภายในแต่ละเซกเมนต์ เนื่องจากอาจจะมีความผิดพลาดจากการจำแนกลักษณะการออกเสียง เพื่อป้องกันความผิดพลาดดังกล่าว เราจึงมีการตรวจสอบกรอบเวลาข้างเคียงไปจำนวนอีก 2 กรอบเวลาของแต่ละข้างของกรอบเวลาที่มีการตรวจพบการเปลี่ยนแปลงของลักษณะการออกเสียง นอกจากในขั้นตอนที่ 4) ยังมีการตรวจสอบความเป็นขอบเขตของหน่วยเสียงอีกทีด้วยการวัดค่าความไม่ต่อเนื่องของสัญญาณ และใช้ตัวจำแนกที่ใช้จำแนกความเป็นขอบเขตของหน่วยเสียง เช่นเดียวกับการแก้ไขความผิดพลาดของขอบเขตของ

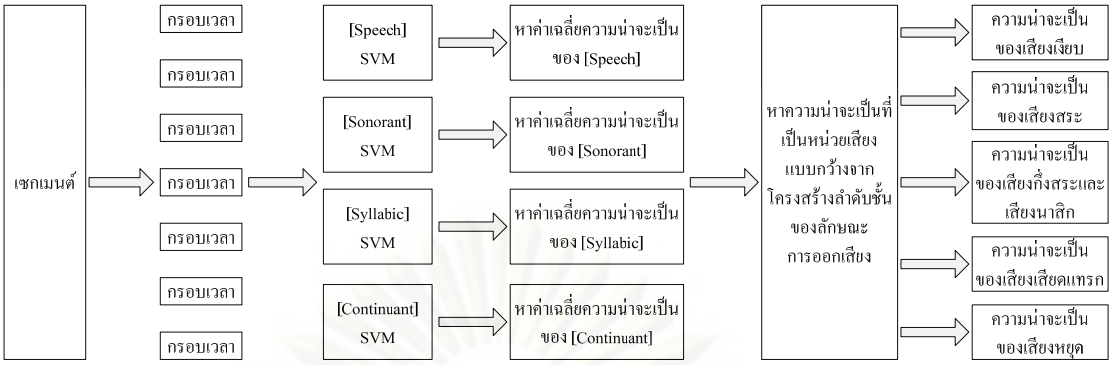
หน่วยเสียงแบบแทรกโดยใช้ความไม่ต่อเนื่องของสัญญาณ มาตรวจสอบอีกครั้ง ถ้าตัวจำแนกดังกล่าวจำแนกได้ว่าเป็นขอบเขตของหน่วยเสียงจริง อัลกอริทึมจะแยกเซกเมนต์ออกเป็นสองส่วน ณ ตำแหน่งที่มีการตรวจพบการเปลี่ยนแปลงของลักษณะการออกเสียง และเพิ่มลงไปในกราฟของเซกเมนต์

4.5 การปรับปรุงการให้คะแนนโดยใช้ความน่าจะเป็นที่เซกเมนต์จะจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้าง

ในการปรับปรุงการให้คะแนนโดยใช้ความน่าจะเป็นที่เซกเมนต์จะจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างจะเป็นการนำค่าความน่าจะเป็นที่เซกเมนต์จะจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างมาเป็นค่าน้ำหนักที่จะถ่วงน้ำหนักในการคิดคะแนนจากแบบจำลองทางเสียง โดยจะมีการปรับเปลี่ยนการคิดคะแนนในการค้นหาคำตอบที่ดีที่สุดของลำดับของหน่วยเสียง U ดังสมการที่ (3.4) เมื่อ S เป็นการแบ่งเสียงพูดที่เป็นที่เป็นไปได้ (ตัวอย่างเช่น เส้นทางที่เป็นไปในการค้นหาในกราฟของเซกเมนต์ และให้ x_i เป็นเวกเตอร์คุณสมบัติของ s_n ซึ่งจะเป็นเซกเมนต์ลำดับที่ i^{th} ใน S ซึ่งในกราฟของเซกเมนต์จะมีเซกเมนต์อยู่จำนวน n เซกเมนต์ เราได้กำหนดคะแนนของความเป็นกลุ่มของหน่วยเสียงแบบกว้าง $BC(s_i, u_i)$ ซึ่งจะเป็นตัวบอกว่าเซกเมนต์ s_i มีความน่าจะเป็นที่จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างของ u_i เท่าใด ซึ่ง u_i ก็เป็นหน่วยเสียงลำดับที่ i^{th} ใน U เมื่อมีนำคะแนน $BC(s_i, u_i)$ มาเป็นตัวถ่วงน้ำหนักของคะแนนของหน่วยเสียงมารวมในสมการ (3.5) จะได้สมการในการค้นหาคำตอบที่ดีที่สุดดังนี้

$$U^* = \arg\max_{S,U} \prod_{i=1}^{n} \frac{BC(s_i, u_i)P(x_i \mid u_i)}{P(x_i \mid \overline{\alpha})} P(s_i \mid u_i)P(U) \qquad (4.1)$$

ในวิทยานิพนธ์ฉบับนี้ $BC(s_i, u_i)$ สำหรับแต่ละเซกเมนต์ s_i จะใช้เป็น $P(b_i|s_i,A)$ ซึ่งเป็นความน่าจะเป็นที่เซกเมนต์ s_i จะเป็นสมาชิกในกลุ่มของหน่วยเสียงแบบกว้าง b_i โดยคะแนนดังกล่าวจะได้มาจากผลคูณของความน่าจะเป็นของลักษณะการออกเสียงตามโครงสร้างลำดับชั้นของลักษณะการออกเสียง ซึ่งเป็นผลลัพธ์ของซัพพอร์ตเวกเตอร์แมชชีนดังที่กล่าวไว้แล้วในบทที่ 2 โดยจะมีขั้นตอนการหาความความน่าจะเป็นที่เซกเมนต์จะอยู่จัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้าง ดังรูปที่ 4.5



รูปที่ 4.5 การหาความน่าจะเป็นที่เซกเมนต์จะอยู่จัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างจากซัพพอร์ตเวกเตอร์แมชชีน

บทที่ 5

รายละเอียดการทดลอง

ในการทดลองจะมีการใช้ฐานข้อมูลเสียงแบบเดียวกับการทดลองเบื้องต้น นอกจากนี้การทำงานของระบบรู้เสียงพูดแบบอาศัยเซกเมนต์ต่างๆยังทำงานในสภาวะแวดล้อมเดียวกัน เช่น เวกเตอร์คุณสมบัติ พารามิเตอร์ และการตั้งค่าของการแบ่งเสียงพูดเป็นเซกเมนต์ แบบจำลองทางเสียง แบบจำลองทางภาษา

วิทยานิพนธ์ฉบับนี้จะแบ่งการทดลองออกเป็น 3 การทดลองด้วยกันคือ

5.1 การจำแนกลักษณะการออกเสียง

การทดลองนี้จะเป็นการทดลองความสามารถในการจำแนกค่าบวก ลบของลักษณะการออกเสียงที่ใช้พารามิเตอร์คุณสมบัติทางเสียงที่ได้ระบุไว้ในตารางที่ 4.1 โดยเราจะฝึกฝนซัพพอร์ตเวกเตอร์แมชชีนของแต่ละลักษณะการออกเสียงจากตัวอย่างจำนวน 500,000 ตัวอย่างซึ่งมีการดึงพารามิเตอร์คุณสมบัติออกจากกรอบเวลาจากชุดข้อมูล TR และ PD ของฐานข้อมูลเสียงโลตัส ในส่วนของการทดสอบจะเป็นการจำแนกเซกเมนต์จากชุดข้อมูล DT ว่ามีลักษณะการออกเสียงเป็นแบบใด ซึ่งในการหาคำตอบเป็นเซกเมนต์นั้นจะทำได้โดยจำแนกแต่ละกรอบเวลาที่อยู่ในเซกเมนต์ที่ต้องการจำแนก และนำค่าความน่าจะเป็นซึ่งเป็นผลลัพธ์จากซัพพอร์ตเวกเตอร์แมชชีนของทุกกรอบเวลาในเซกเมนต์มาหาค่าเฉลี่ย แล้วนำค่าเฉลี่ยนี้มาตัดสินใจว่าเซกเมนต์ดังกล่าวจะมีลักษณะการออกเสียงเป็นอย่างไร ซึ่งการวัดผลจะเป็นการวัดเปอร์เซ็นต์ความถูกต้อง

5.2 การแก้ไขความผิดพลาดในกราฟของเซกเมนต์

การทดลองนี้จะเป็นการนำวิธีที่นำเสนอในการแก้ไขความผิดพลาดของกราฟของเซกเมนต์เพื่อศึกษาว่าวิธีที่นำเสนอสามารถเรียกคืนเซกเมนต์ที่ถูกต้องมายังกราฟของเซกเมนต์ได้มากน้อยแค่ไหน โดยจะแบ่งเป็นความผิดพลาดของกราฟของเซกเมนต์ที่เกิดจากความผิดพลาดของขอบเขตของหน่วยเสียงแบบแทรก และความผิดพลาดของขอบเขตของหน่วยเสียงแบบแทรกตัดออก โดยการวัดผลจะเป็นการเปรียบเทียบค่าเปอร์เซ็นต์ความผิดพลาดของเซกเมนต์ นอกจากนี้ในการทดลองนี้จะแสดงถึงขนาดของกราฟของเซกเมนต์ที่เพิ่มขึ้นจากอีกด้วย แต่อย่างไรก็ตามกราฟของเซกเมนต์ที่ได้มีการแก้ไขความผิดพลาดจะให้ผลการรู้จำเสียงดีขึ้นหรือไม่ อย่างไรก็แสดงให้เห็นในการทดลองถัดไป

5.3 ผลการรู้จำเสียงพูดในระดับหน่วยเสียง

การทดลองนี้จะเป็นการนำกราฟของเซกเมนต์ที่ผ่านการแก้ไขความผิดพลาดแบบต่างๆ มาผ่านกระบวนการรู้จำเสียงพูด เพื่อแสดงให้เห็นว่ากราฟของเซกเมนต์ที่มีจำนวนเซกเมนต์ที่ถูกต้องเพิ่มขึ้นจะให้ความแม่นยำที่ดีขึ้นอย่างไร นอกจากนี้ยังจะมีการแสดงถึงผลการรู้จำเสียงพูดที่ได้มีการนำค่าความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงกว้างไปเป็นตัวถ่วงน้ำหนักในการคิดคะแนนของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ เพื่อศึกษาว่าการใช้ค่าความน่าจะเป็นดังกล่าวมีผลต่อความแม่นยำในการรู้จำเสียงพูดอย่างไร ซึ่งในการทดลองนี้จะเป็นการเปรียบเทียบค่าเปอร์เซ็นต์ความแม่นยำ นอกจากนี้เรายังจะนำค่าความแม่นยำของการรู้จำเสียงพูดแบบอาศัยกรอบเวลามาเปรียบเทียบด้วย เพื่อแสดงให้เห็นว่าการรู้จำเสียงพูดแบบอาศัยเซกเมนต์นั้นมีความแม่นยำสูงกว่าการรู้จำเสียงพูดแบบอาศัยกรอบเวลาเท่าใด

บทที่ 6

ผลการทดลอง และการอภิปรายผลการทดลอง

ในบทนี้จะแสดงผลการทดลองของการทดลองต่างๆ และมีการอภิปรายถึงผลลัพธ์ที่เกิดขึ้น โดยจะเป็นจะมีการแบ่งการทดลองออกเป็น 3 การทดลองตามจำนวนการทดลองที่วิทยานิพนธ์นี้ได้ดำเนินการดังนี้

6.1 การจำแนกลักษณะการออกเสียง

ผลการจำแนกลักษณะออกเสียงของเซกเมนต์จะแสดงดังตารางที่ 6.1 ซึ่งจากผลการจำแนกที่ได้จะแสดงให้เห็นว่าพารามิเตอร์คุณสมบัติทางเสียงที่นำเสนอมีความสามารถที่จะจำแนกลักษณะการออกเสียงได้เป็นอย่างดี อย่างไรก็ตามจะสังเกตได้ว่าผลการจำแนกของลักษณะการออกเสียง [Continuant] ได้ผลออกมาไม่ค่อยดีนัก เนื่องจากในของเสียงพยัญชนะกักที่เป็นพยัญชนะที่เป็นตัวสะกด มักจะมีพลังงานน้อย และสามารถสังเกตได้ยาก นอกจากนั้นเสียงกักกับเสียงเสียดแทรกแบบเบาซึ่งมีพลังงานน้อยจะมีความใกล้เคียงกันมาก และทำเห็นเกิดความสับสนในการจำแนกได้ ซึ่งสิ่งเหล่านี้ทำให้เกิดความผิดพลาดขึ้นได้ และความผิดพลาดดังกล่าวนั้นยากที่จัดการได้ แต่อย่างไรก็ตามถ้านับจากจำนวนเซกเมนต์ทั้งหมดที่นำมาทดลองก็จะได้ผลการจำแนกโดยรวมเป็น 88.55% ซึ่งเป็นระดับความถูกต้องที่จะสามารถนำตัวจำแนกดังกล่าวมาใช้ในการแก้ไขความผิดพลาดในวิธีที่นำเสนอได้

ตารางที่ 6.1 ผลการจำแนกลักษณะการออกเสียงของเซกเมนต์

ลักษณะการออกเสียง	ความถูกต้อง (%)
[Speech]	93.47
[Sonorant]	93.54
[Syllabic]	80.66
[Continuant]	73.92

6.2 การแก้ไขความผิดพลาดในกราฟของเซกเมนต์

ตารางที่ 6.2 ผลการแก้ไขความผิดพลาดในกราฟของเซกเมนต์

กราฟของเซกเมนต์	ความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียง		ความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียง		อัตราส่วนขนาดกราฟของเซกเมนต์
	ความผิดพลาดของเซกเมนต์ (%)	ดีขึ้น (%)	ความผิดพลาดของเซกเมนต์ (%)	ดีขึ้น (%)	
ไม่มีการปรับปรุงคุณภาพกราฟของเซกเมนต์	15.80		11.90		1
1) แก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยใช้ความไม่ต่อเนื่องของสัญญาณ	11.71	25.89	11.90		1.73
2) แก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยใช้สมบัติลักษณะเฉพาะ	12.10	23.42	11.90		1.60
แก้ไขความผิดพลาดด้วยวิธี 1) และ 2)	10.34	34.56	11.90		1.93
3) แก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยใช้ความไม่ต่อเนื่องของสัญญาณ	15.80		10.96	7.90	1.53
4) แก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยใช้สมบัติลักษณะเฉพาะ	15.80		11.06	7.06	1.51
แก้ไขความผิดพลาดด้วยวิธี 3) และ 4)	15.80		10.18	14.45	1.78

ตารางที่ 6.2 จะแสดงเปอร์เซ็นต์ความผิดพลาดของกราฟของเซกเมนต์ที่มีการปรับปรุงคุณภาพของกราฟด้วยวิธีต่างๆ และกราฟของเซกเมนต์ที่ไม่มีการปรับปรุงคุณภาพ จากผลการทดลองจะเห็นว่าทุกวิธีที่นำเสนอจะสามารถลดความผิดพลาดในกราฟของเซกเมนต์ได้ สำหรับกรณีของการลดความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียง การใช้ความไม่ต่อเนื่องของสัญญาณจะสามารถลดความผิดพลาดได้มากกว่าการใช้สมบัติลักษณะเฉพาะ อย่างไรก็ตามวิธีการใช้สมบัติลักษณะเฉพาะก็จะให้ขนาดของกราฟที่เล็กกว่าการใช้ความไม่ต่อเนื่องของสัญญาณ เมื่อรวมเซกเมนต์ของวิธีการทั้งสองเข้าด้วยกันจะทำให้แก้ไขความผิดพลาดได้มากขึ้น แต่อย่างไรก็ตามกราฟของเซกเมนต์ก็จะมีขนาดเพิ่มขึ้นเกือบสองเท่า แต่ในส่วนของการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงจะพบว่ามีจะสามารถแก้ไขความผิดพลาดได้น้อย ด้วยกราฟของเซกเมนต์ที่ใช้วิธีทั้งสองรวมกันก็สามารถทำให้ความผิดพลาดลดลงไป 15% ซึ่งผลของการแก้ไขความผิดพลาดในกราฟของเซกเมนต์ในการรู้จำเสียงพูดในระดับหน่วยเสียงจะแสดงให้เห็นในการทดลองถัดไป

6.3 ผลการรู้จำเสียงพูดในระดับหน่วยเสียง

การทดลองนี้จะแสดงถึงผลของการแก้ไขความผิดพลาดในกราฟของเซกเมนต์ต่อการรู้จำเสียงพูดในระดับหน่วยเสียง และผลของการนำค่าความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างมาใช้เป็นตัวถ่วงน้ำหนักในกระบวนการให้คะแนนและค้นหาคำตอบของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ นอกจากนี้ยังได้การนำผลการรู้จำเสียงพูดของการรู้จำเสียงพูดแบบอาศัยกรอบเวลามาเปรียบเทียบอีกด้วย ซึ่งผลต่างๆที่กล่าวมาจะแสดงในตารางที่ 6.3 จากผลการทดลองจะพบว่าเมื่อมีการแก้ไขความผิดพลาดของเซกเมนต์ที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงจะทำให้ความแม่นยำของการรู้จำเสียงพูดในระดับหน่วยเสียงเพิ่มขึ้นเปรียบเทียบกับการรู้จำเสียงพูดที่ไม่ได้มีการปรับปรุงคุณภาพกราฟของเซกเมนต์ เมื่อมีการใช้การแก้ไขความผิดพลาดทั้งสองวิธีจะทำให้การรู้จำเสียงพูดดีขึ้น 12.03% และ 13.04% เมื่อมีการใช้แบบจำลองทางเสียงที่ใช้แบบจำลองของเซกเมนต์อย่างเดียว และ ใช้ทั้งแบบจำลองของเซกเมนต์และแบบจำลองของขอบเขตของหน่วยเสียง ตามลำดับ จากผลการรู้จำเสียงพูดจะเห็นว่าการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ทั้งที่ไม่มีการปรับปรุงคุณภาพกราฟของเซกเมนต์ และมีการปรับปรุงคุณภาพกราฟของเซกเมนต์โดยแก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียง จะมีความแม่นยำสูงกว่าการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

ตารางที่ 6.3 ผลการรู้จำเสียงพูดในระดับหน่วยเสียงของแต่ละวิธีการ

ระบบรู้จำเสียงพูด	วิธีการ	% ความแม่นยำ		% ความแม่นยำที่มีการใช้คะแนนที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงกว้าง	
		แก้ไขความผิดพลาดที่เกิดจากการแทรก	แก้ไขความผิดพลาดที่เกิดจากการตัดออก	แก้ไขความผิดพลาดที่เกิดจากการแทรก	แก้ไขความผิดพลาดที่เกิดจากการตัดออก
การรู้จำเสียงพูดแบบอาศัยกรอบเวลา	-	47.21	47.21	-	-
แบบจำลองเซกเมนต์	ไม่มีการปรับปรุงคุณภาพกราฟของเซกเมนต์	47.70	47.70	47.72	47.72
แบบจำลองเซกเมนต์	ความไม่ต่อเนื่องของสัญญาณ	52.92	41.88	52.92	42.16
แบบจำลองเซกเมนต์	สมบัติลักษณะเฉพาะ	52.21	44.79	52.32	45.29
แบบจำลองเซกเมนต์	ทั้งความไม่ต่อเนื่องของสัญญาณ และสมบัติลักษณะเฉพาะ	53.44	38.50	53.54	39.25

ระบบรู้จำเสียงพูด	วิธีการ	% ความแม่นยำ		% ความแม่นยำที่มีการใช้คะแนนที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงกว้าง	
		แก้ไขความผิดพลาดที่เกิดจากการแทรก	แก้ไขความผิดพลาดที่เกิดจากการตัดออก	แก้ไขความผิดพลาดที่เกิดจากการแทรก	แก้ไขความผิดพลาดที่เกิดจากการตัดออก
แบบจำลองเซกเมนต์และขอบเขตของหน่วยเสียง	ไม่มีการปรับปรุงคุณภาพกราฟของเซกเมนต์	51.47	51.47	51.58	51.58
แบบจำลองเซกเมนต์และขอบเขตของหน่วยเสียง	ความไม่ต่อเนื่องของสัญญาณ	57.50	48.74	57.59	48.91
แบบจำลองเซกเมนต์และขอบเขตของหน่วยเสียง	สมบัติลักษณะเฉพาะ	56.91	49.69	56.96	49.86
แบบจำลองเซกเมนต์และขอบเขตของหน่วยเสียง	ทั้งความไม่ต่อเนื่องของสัญญาณ และสมบัติลักษณะเฉพาะ	58.18	46.84	58.26	47.20

ถึงแม้ว่าการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงจะให้ความผิดพลาดของเซกเมนต์ลดลง แต่ว่าผลการรู้จำเสียงพูดนั้นจะให้ค่าความแม่นยำที่ลดลง จากการวิเคราะห์จากเมทริกซ์ของความผิดพลาดในระดับกลุ่มของหน่วยเสียงแบบกว้างของผลการรู้จำที่ไม่มีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียง ดังตารางที่ 6.4 เปรียบเทียบกับเมทริกซ์ของความผิดพลาดในระดับกลุ่มของหน่วยเสียงแบบกว้างของผลการรู้จำที่ไม่มีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยใช้ความไม่

ต่อเนื่องของสัญญาณ ดังตารางที่ 6.5 และเมทริกซ์ของความผิดพลาดในระดับกลุ่มของหน่วยเสียงแบบกว้างของผลการรู้จำที่ไม่มีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยใช้สมบัติลักษณะเฉพาะ ดังตารางที่ 6.6 พบว่าเมื่อมีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงจะทำให้จำนวนความผิดพลาดแบบแทรกมีจำนวนมากขึ้น โดยเฉพาะในกลุ่มของหน่วยเสียงแบบกว้างที่เป็นเสียงหยุด

ตารางที่ 6.4 เมทริกซ์ของความผิดพลาดในระดับกลุ่มของหน่วยเสียงแบบกว้าง เมื่อไม่มีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียง

	SIL	V	SC	FRI	ST	Del	[%c/%e]
SIL	1008	0	0	0	0	0	
V	39	11918	6	0	23	245	[99.4/0.2]
SC	125	83	8009	75	672	1021	[89.3/2.9]
FRI	2	0	49	1798	236	79	[86.2/0.9]
ST	39	21	180	152	7177	527	[94.8/1.2]
Ins	1079	900	502	168	**1342**		

(SIL หมายถึง เสียงเงียบ, V หมายถึง เสียงสระ, SC หมายถึง เสียงพยัญชนะที่ช่องทางเสียงเปิด เช่น เสียงกึ่งสระ และเสียงนาสิก, FRI หมายถึง เสียงเสียดแทรก, ST หมายถึง เสียงหยุด, Ins หมายถึง ความผิดพลาดแบบแทรก, Del หมายถึง ความผิดพลาดแบบตัดออก, %c หมายถึง เปอร์เซ็นต์ความถูกต้อง, %e หมายถึง เปอร์เซ็นต์ความผิดพลาด)

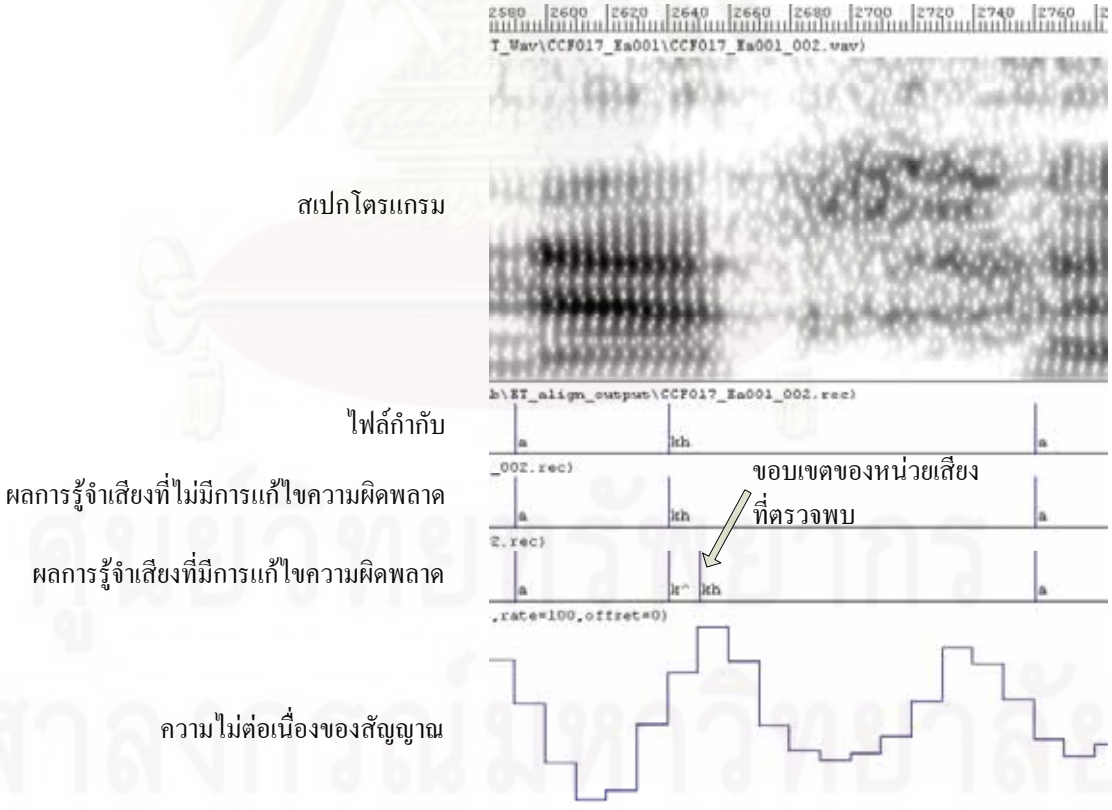
ตารางที่ 6.5 เมทริกซ์ของความผิดพลาดในระดับกลุ่มของหน่วยเสียงแบบกว้าง เมื่อมีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยใช้ความไม่ต่อเนื่องของสัญญาณ

	SIL	V	SC	FRI	ST	Del	[%c/%e]
SIL	1008	0	0	0	0	0	
V	30	11991	1	0	18	191	[99.6/0.1]
SC	115	81	8304	61	750	674	[89.2/3.0]
FRI	2	0	46	1814	243	59	[86.2/0.9]
ST	28	21	193	130	7384	340	[95.2/1.1]
Ins	1166	1185	691	207	**2003**		

ตารางที่ 6.6 เมทริกซ์ของความผิดพลาดในระดับกลุ่มของหน่วยเสียงแบบกว้าง เมื่อมีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยใช้สมบัติลักษณะเฉพาะ

	SIL	V	SC	FRI	ST	Del	[%c/%e]
SIL	1008	0	0	0	0	0	
V	34	12019	0	0	12	166	[99.6/0.1]
SC	108	64	8435	51	750	577	[89.7/2.9]
FRI	2	0	42	1836	240	44	[86.6/0.8]
ST	26	16	202	135	7430	287	[95.1/1.1]
Ins	1181	1453	874	234	**2305**		

รูปที่ 6.1 แสดงถึงสาเหตุของความผิดพลาดแบบแทรกที่เกิดจากการแทรกกลุ่มของหน่วยเสียงแบบกว้างที่เป็นเสียงหยุด โดยจะสังเกตเห็นว่าวิธีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยใช้ความไม่ต่อเนื่องของสัญญาณ จะมีการเพิ่มขอบเขตของหน่วยเสียงที่บริเวณที่มีค่าความไม่ต่อเนื่องของสัญญาณสูงสุดภายในเซกเมนต์ (ที่ลูกศรชี้) ซึ่งถ้าดูจากสเปกโตรแกรมแล้วน่าจะเป็นบริเวณที่เป็นขอบเขตของหน่วยเสียงจริง แต่อย่างไรก็ตามวิธีการที่นำเสนอจึงได้มีการแบ่งเซกเมนต์ออกเป็นสองเซกเมนต์ ซึ่งมีผลทำให้การรู้จำเสียงพูดมีหน่วยเสียง “k^” แทรกมาซึ่งเป็นความผิดพลาดแบบแทรก



รูปที่ 6.1 การเปรียบเทียบผลการรู้จำเสียงพูดที่ไม่มี และมีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยใช้ความไม่ต่อเนื่องของสัญญาณ

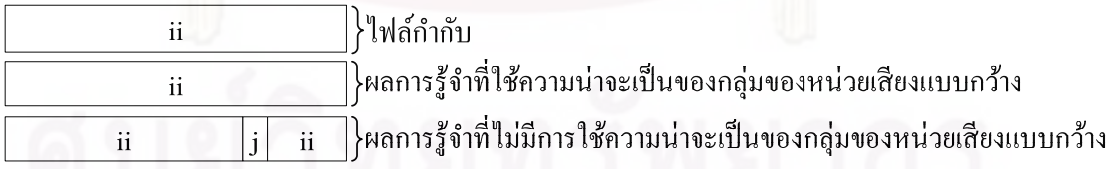
จากรูปที่ 6.2 แสดงถึงสาเหตุของความผิดพลาดแบบแทรกที่เกิดจากการแทรกกลุ่มของหน่วยเสียงแบบกว้างที่เป็นเสียงหยุด จากสเปกโตรแกรมจะเห็นว่ามีช่วงที่สัญญาณเสียงขาดหายไป ซึ่งอาจจะเกิดจากการพักหายใจจึงทำให้ช่วงนั้นถูกจำแนกออกเป็นเสียงเงียบ [-Speech] จึงมีผลทำให้วิธีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยใช้สมบัติลักษณะเฉพาะมีการตรวจสอบพบการเปลี่ยนแปลงของลักษณะการออกเสียง [Speech] จึงได้แยกเซกเมนต์ออกเป็นสองเซกเมนต์ที่ตำแหน่งที่ลูกศรชี้ และส่งผลทำให้การรู้จำเสียงพูดแบบอาศัยเซกเมนต์นั้นรู้จำหน่วยเสียง “t^” เกินมา ซึ่งเป็นความผิดพลาดแบบแทรก

รูปที่ 6.2 การเปรียบเทียบผลการรู้จำเสียงพูดที่ไม่มี และมีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยใช้สมบัติลักษณะเฉพาะ

สองคอลัมน์หลังสุดของตารางที่ 6.3 (% ความแม่นยำที่มีการใช้คะแนนที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงกว้าง) แสดงถึงผลการรู้จำเสียงพูดของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่มีการนำค่าความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างมาใช้เป็นตัวถ่วงน้ำหนักของคะแนนทางเสียง เพื่อเปรียบเทียบความแม่นยำกับการไม่ได้นำค่าความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างมาใช้ จากผลการทดลองแสดงให้เห็นว่าเมื่อมีการนำค่าความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างมาใช้ทำให้

การรู้จำเสียงพูดมีความแม่นยำเพิ่มขึ้นในทุกกรณี เหตุผลที่ความแม่นยำเพิ่มขึ้นนั้นเกิดจากการนำค่าความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างมาใช้นั้นจะช่วยลดความผิดพลาดที่มีลักษณะการออกเสียงต่างกันแทรกเข้ามา ดังตัวอย่างในรูปที่ 6.3 เห็นว่าเมื่อมีไม่การนำค่าความน่าจะเป็นของกลุ่มของหน่วยเสียงแบบกว้างมาใช้ก็จะเกิดหน่วยเสียง “j” ออกมาอยู่ในคำตอบ ซึ่งจะเป็นความผิดพลาดแบบแทรก แต่ถ้ามีการนำค่าความน่าจะเป็นของกลุ่มของหน่วยเสียงแบบกว้างมาใช้ คะแนนที่ “j” จะเกิดขึ้นมาก็จะน้อยลง เนื่องจากถูกถ่วงด้วยคะแนนของความน่าจะเป็นที่เซกเมนต์ดังกล่าวจะอยู่ในกลุ่มของหน่วยเสียงแบบกว้างที่เป็นเสียงกึ่งสระ ซึ่งจะมีค่าน้อย เนื่องจากเซกเมนต์ดังกล่าวควรจะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างที่เป็นเสียงสระ ตารางที่ 6.7 แสดงเมทริกซ์ของความผิดพลาดในระดับกลุ่มของหน่วยเสียงแบบกว้าง เมื่อไม่มีการใช้ค่าความน่าจะเป็นที่เสียงจะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้าง และตารางที่ 6.8 แสดงเมทริกซ์ของความผิดพลาดในระดับกลุ่มของหน่วยเสียงแบบกว้าง เมื่อมีการใช้ค่าความน่าจะเป็นที่เสียงจะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้าง เมื่อเปรียบเทียบผลกัน จะพบว่าเมื่อความผิดพลาดแบบแทรกจะลดลงเมื่อมีการนำค่าความน่าจะเป็นที่เสียงจะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างมาใช้ในขั้นตอนการให้คะแนน และค้นหาคำตอบของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ ซึ่งเป็นผลทำให้ความแม่นยำเพิ่มขึ้น

อย่างไรก็ตาม การนำค่าความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในหน่วยเสียงแบบกว้างมาใช้ก็ยังเพิ่มความแม่นยำได้ไม่มากนัก ทั้งนี้เนื่องจากความผิดพลาดส่วนใหญ่มักจะมีความผิดพลาดอยู่ภายกลุ่มของหน่วยเสียงแบบกว้างเดียวกัน ดังที่แสดงในตารางที่ 6.7 ซึ่งจะเห็นว่าถ้าพิจารณาผลการรู้จำเสียงพูดในระดับกลุ่มของหน่วยเสียงแบบกว้างแล้วจะพบว่ามีความถูกต้องสูง ซึ่งสอดคล้องกับ Scanlon [52] ที่ได้กล่าวว่าความผิดพลาดประมาณ 75% จะอยู่ภายในกลุ่มของหน่วยเสียงแบบกว้าง



รูปที่ 6.3 การเปรียบเทียบผลการรู้จำเสียงพูดที่มีการใช้ และไม่ใช้ความน่าจะเป็นของกลุ่มของหน่วยเสียงแบบกว้าง

ตารางที่ 6.7 เมทริกซ์ของความผิดพลาดในระดับกลุ่มของหน่วยเสียงแบบกว้าง เมื่อไม่มีการใช้ค่าความน่าจะเป็นที่เสียงจะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้าง

	SIL	V	SC	FRI	ST	Del	[%c/%e]
SIL	1008	0	0	0	0	0	
V	34	12019	0	0	12	166	[99.6/0.1]
SC	108	64	8435	51	750	577	[89.7/2.9]
FRI	2	0	42	1836	240	44	[86.6/0.8]
ST	26	16	202	135	7430	287	[95.1/1.1]
Ins	1181	**1453**	**874**	234	**2305**		

ตารางที่ 6.8 เมทริกซ์ของความผิดพลาดในระดับกลุ่มของหน่วยเสียงแบบกว้าง เมื่อมีการใช้ค่าความน่าจะเป็นที่เสียงจะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้าง

	SIL	V	SC	FRI	ST	Del	[%c/%e]
SIL	1008	0	0	0	0	0	
V	35	12017	0	0	12	167	[99.6/0.1]
SC	108	62	8430	49	753	583	[89.7/2.9]
FRI	2	0	42	1836	240	44	[86.6/0.8]
ST	26	17	197	139	7425	292	[95.1/1.1]
Ins	1181	**1416**	**857**	234	**2260**		

6.4 การเปรียบเทียบเวลาที่ใช้ในการรู้จำเสียงพูด

ในการดำเนินการปรับปรุงคุณภาพกราฟของเซกเมนต์ด้วยวิธีที่ได้นำเสนอจะต้องใช้เวลาในการประมวลผลเพิ่มขึ้น ในวิทยานิพนธ์นี้จึงได้มีการแสดงถึงเวลาที่ใช้ในการรู้จำเสียงพูดเสียงของการรู้จำเสียงพูดในแต่ละวิธีเพื่อเปรียบเทียบเวลาที่ใช้ในการดำเนินการรู้จำเสียงพูด โดยเวลาที่วัดนั้นจะเป็นเวลาในการรู้จำเสียงพูดจากชุดข้อมูล ET ซึ่งมีจำนวน 504 ไฟล์เสียง โดยมีความยาวทั้งหมดประมาณ 70 นาที เวลาที่ใช้ในการรู้จำเสียงจะแสดงดังตารางที่ 6.9

ตารางที่ 6.9 การเปรียบเทียบเวลาในการรู้จำเสียงพูดของแต่ละการรู้จำเสียงพูดที่มีวิธีการปรับปรุงกราฟของเซกเมนต์ด้วยวิธีต่างๆของวิทยานิพนธ์

วิธีการ	เวลา
1. การรู้จำเสียงพูดแบบอาศัยกรอบเวลา	1 นาที 12 วินาที

วิธีการ	เวลา
2. การรู้จำเสียงพูดแบบอาศัยเซกเมนต์	
2.1. การแบ่งเสียงพูดเป็นเซกเมนต์จากผลการรู้จำเสียงพูดแบบอาศัยกรอบเวลาจำนวน 20 คำตอบ	147 นาที 44 วินาที
2.2. การให้คะแนนและค้นหาของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์	340 นาที 13 วินาที
รวมทั้งหมด	487 นาที 57 วินาที
3. การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่มีการแก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยความไม่ต่อเนื่องของสัญญาณ	
3.1. การแบ่งเสียงพูดเป็นเซกเมนต์จากผลการรู้จำเสียงพูดแบบอาศัยกรอบเวลาจำนวน 20 คำตอบ	147 นาที 44 วินาที
3.2. การแก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยความไม่ต่อเนื่องของสัญญาณ	11 วินาที
3.3. การให้คะแนนและค้นหาของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์	469 นาที 43 วินาที
รวมทั้งหมด	617 นาที 38 วินาที
4. การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่มีการแก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยสมบัติลักษณะเฉพาะ	
4.1. การแบ่งเสียงพูดเป็นเซกเมนต์จากผลการรู้จำเสียงพูดแบบอาศัยกรอบเวลาจำนวน 20 คำตอบ	147 นาที 44 วินาที
4.2. การจำแนกแต่ละกรอบเวลาให้อยู่ในรูปแบบลักษณะการออกเสียง	
4.2.1. [Speech]	149 นาที 37 วินาที
4.2.2. [Sonorant]	86 นาที 46 วินาที
4.2.3. [Syllabic]	150 นาที 49 วินาที
4.2.4. [Continuant]	202 นาที 58 วินาที
4.3. การแก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยสมบัติลักษณะเฉพาะ	1 วินาที
4.4. การให้คะแนนและค้นหาของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์	429 นาที 34 วินาที
รวมทั้งหมด (*เวลาจาก 4.2 ใช้ค่าสูงสุดจาก 4.2.4)	780 นาที 17 วินาที

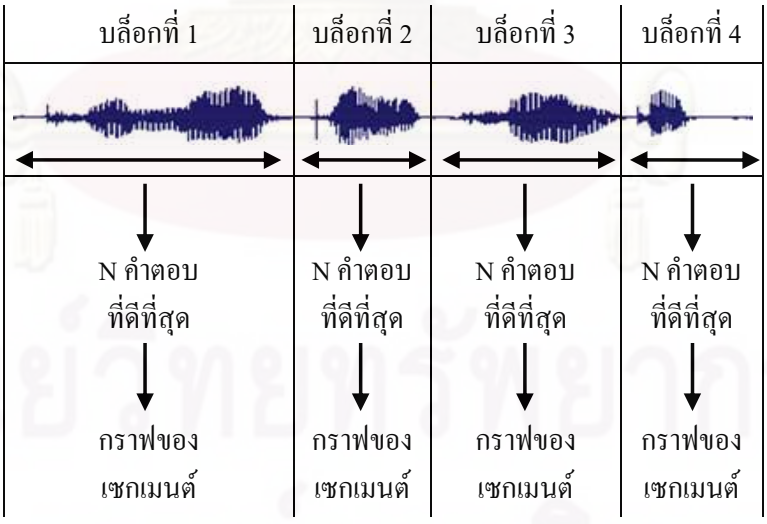
วิธีการ	เวลา
5. การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ทีมีการแก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยความไม่ต่อเนื่องของสัญญาณ และสมบัติลักษณะเฉพาะ	
5.1. การแบ่งเสียงพูดเป็นเซกเมนต์จากผลการรู้จำเสียงพูดแบบอาศัยกรอบเวลาจำนวน 20 คำตอบ	147 นาที 44 วินาที
5.2. การจำแนกแต่ละกรอบเวลาให้อยู่ในรูปแบบลักษณะการออกเสียง	
5.2.1. [Speech]	149 นาที 37 วินาที
5.2.2. [Sonorant]	86 นาที 46 วินาที
5.2.3. [Syllabic]	150 นาที 49 วินาที
5.2.4. [Continuant]	202 นาที 58 วินาที
5.3. การแก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยสมบัติลักษณะเฉพาะ	1 วินาที
5.4. การแก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยความไม่ต่อเนื่องของสัญญาณ	11 วินาที
5.5. การให้คะแนนและค้นหาของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์	517 นาที 46 วินาที
<u>รวมทั้งหมด</u> (*เวลาจาก 5.2 ใช้ค่าสูงสุดจาก 5.2.4 และสมมุติว่า 5.3 และ 5.4 ทำงานพร้อมกัน)	<u>866 นาที 28 วินาที</u>
6. การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ทีมีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยความไม่ต่อเนื่องของสัญญาณ	
6.1. การแบ่งเสียงพูดเป็นเซกเมนต์จากผลการรู้จำเสียงพูดแบบอาศัยกรอบเวลาจำนวน 20 คำตอบ	147 นาที 44 วินาที
6.2. การแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยความไม่ต่อเนื่องของสัญญาณ	18 วินาที
6.3. การให้คะแนนและค้นหาของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์	434 นาที 57 วินาที
<u>รวมทั้งหมด</u>	<u>582 นาที 59 วินาที</u>

วิธีการ	เวลา
7. การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ทีมีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยสมบัติลักษณะเฉพาะ	
7.1. การแบ่งเสียงพูดเป็นเซกเมนต์จากผลการรู้จำเสียงพูดแบบอาศัยกรอบเวลาจำนวน 20 คำตอบ	147 นาที 44 วินาที
7.2. การจำแนกแต่ละกรอบเวลาให้อยู่ในรูปแบบลักษณะการออกเสียง	
7.2.1. [Speech]	149 นาที 37 วินาที
7.2.2. [Sonorant]	86 นาที 46 วินาที
7.2.3. [Syllabic]	150 นาที 49 วินาที
7.2.4. [Continuant]	202 นาที 58 วินาที
7.3. การแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยสมบัติลักษณะเฉพาะ	12 วินาที
7.4. การให้คะแนนและค้นหาของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์	408 นาที 21 วินาที
รวมทั้งหมด (*เวลาจาก 7.2 ใช้ค่าสูงสุดจาก 7.2.4)	759 นาที 3 วินาที
8. การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ทีมีการแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยความไม่ต่อเนื่องของสัญญาณ และสมบัติลักษณะเฉพาะ	
8.1. การแบ่งเสียงพูดเป็นเซกเมนต์จากผลการรู้จำเสียงพูดแบบอาศัยกรอบเวลาจำนวน 20 คำตอบ	147 นาที 44 วินาที
8.2. การจำแนกแต่ละกรอบเวลาให้อยู่ในรูปแบบลักษณะการออกเสียง	
8.2.1. [Speech]	149 นาที 37 วินาที
8.2.2. [Sonorant]	86 นาที 46 วินาที
8.2.3. [Syllabic]	150 นาที 49 วินาที
8.2.4. [Continuant]	202 นาที 58 วินาที
8.3. การแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยสมบัติลักษณะเฉพาะ	12 วินาที
8.4. การแก้ไขความผิดพลาดที่เกิดจากการตัดออกของขอบเขตของหน่วยเสียงโดยความไม่ต่อเนื่องของสัญญาณ	18 วินาที
8.5. การให้คะแนนและค้นหาของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์	477 นาที 21 วินาที
รวมทั้งหมด (*เวลาจาก 8.2 ใช้ค่าสูงสุดจาก 8.2.4 และสมมุติว่า 8.3 และ 8.4 ทำงานพร้อมกัน)	828 นาที 15 วินาที

จากตารางที่ 6.9 จะเห็นว่าการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ผ่านกระบวนการปรับปรุงคุณภาพกราฟของเซกเมนต์จะใช้เวลาในการรู้จำเสียงมากกว่าการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ไม่มีการปรับปรุงคุณภาพกราฟของเซกเมนต์จำนวน 26.69% เมื่อมีการแก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยความไม่ต่อเนื่องของสัญญาณ 60.16% เมื่อมีการแก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยสมบัติลักษณะเฉพาะ และ 77.82% เมื่อมีการแก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยใช้ทั้งความไม่ต่อเนื่องของสัญญาณ และสมบัติลักษณะเฉพาะ แต่อย่างไรก็ตามการรู้จำเสียงพูดแบบอาศัยเซกเมนต์จะมีใช้เวลาในการรู้จำเสียงพูดมากกว่าการรู้จำเสียงพูดแบบอาศัยกรอบเวลาค่อนข้างมาก ทั้งนี้ เนื่องจากการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ในงานวิจัยนี้ไม่ได้ปรับจูนเรื่องเวลาที่ใช้ในการประมวลผล เมื่อมีการวิเคราะห์ถึงเวลาที่ใช้การประมวลผลมากนั้น มีสาเหตุมาจาก

1) วิธีการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็นจะต้องมีการนำเสียงพูดไปผ่านการรู้จำเสียงพูดแบบอาศัยกรอบเวลาและนำคำตอบจำนวน 20 คำตอบที่ดีที่สุดมาใช้สร้างกราฟของเซกเมนต์ ซึ่งสามารถปรับปรุงเวลาในการแบ่งเสียงพูดเป็นเซกเมนต์ได้ 2 แนวทาง คือ

1.1) การปรับปรุงเวลาในการแบ่งเสียงพูดเป็นเซกเมนต์โดยใช้วิธีการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็นแบบขนาน [10] ซึ่งวิธีการนี้จะมีการแบ่งเสียงพูดออกเป็นบล็อกก่อนและนำเสียงพูดในแต่ละบล็อกไปหาคำตอบที่ดีที่สุด N จำนวนพร้อมๆกัน โดยวิธีการนี้เป็นวิธีที่ SUMMIT ใช้อยู่



รูปที่ 6.4 การแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็นที่มีการแบ่งเป็นบล็อก [10]

1.2) การปรับปรุงเวลาในการแบ่งเสียงพูดเป็นเซกเมนต์โดยการสร้างกราฟของเซกเมนต์ที่ได้วิธีการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็นจากการรู้จำเสียงพูดแบบอาศัยเซกเมนต์เพิ่ง 1 คำตอบ จากตารางที่ 6.9 พบว่าเวลาในการแบ่งเสียงพูดเป็นเซกเมนต์ที่สร้างกราฟของเซกเมนต์จากผลการรู้จำเสียงพูดแบบอาศัยกรอบเวลาจำนวน 20 คำตอบนั้นใช้เวลาในการประมวลผลค่อนข้างนาน อย่างไรก็ถ้าเราสามารถสร้างกราฟของเซกเมนต์ที่ได้จากคำตอบ 1 คำตอบที่ดีที่สุดจากการรู้จำเสียงพูดแบบอาศัยกรอบเวลา และนำมาผ่านวิธีการแก้ไขความผิดพลาดของเซกเมนต์ที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยความไม่ต่อเนื่องของสัญญาณ (ซึ่งใช้เวลาในการดำเนินการเพียง 11 วินาที) โดยการเปรียบเทียบความแม่นยำ และเวลาที่ใช้ในการรู้จำเสียงจะแสดง ดังตารางต่อไปนี้

ตารางที่ 6.10 การเปรียบเทียบความแม่นยำ และเวลาที่ใช้ในการรู้จำเสียงของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

จำนวนคำตอบที่ดีที่สุดจากการรู้จำเสียงพูดแบบอาศัยกรอบเวลา	วิธีการปรับปรุงกราฟของเซกเมนต์	ความแม่นยำ (%)	เวลาที่ใช้ในการรู้จำเสียง
20	ไม่มีการปรับปรุง	51.47	487 นาที 57 วินาที
1	ไม่มีการปรับปรุง	50.86	320 นาที 27 วินาที
1	แก้ไขความผิดพลาดที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยใช้ความไม่ต่อเนื่องของสัญญาณ	57.26	431 นาที 7 วินาที

ตารางที่ 6.10 แสดงให้เห็นว่าผลการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่กราฟของเซกเมนต์สร้างมาจากคำตอบ 1 คำตอบที่ดีที่สุด และผ่านวิธีการแก้ไขความผิดพลาดของเซกเมนต์ที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงโดยความไม่ต่อเนื่องของสัญญาณมีความแม่นยำสูงกว่าการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่กราฟของเซกเมนต์สร้างมาจากคำตอบ 20 คำตอบที่ดีที่สุดที่ไม่มีการปรับปรุงคุณภาพกราฟของเซกเมนต์ถึง 5.79% และใช้เวลาน้อยกว่าประมาณ 56 นาที ซึ่งจากผลการทดลองจะเห็นว่าวิธีการนี้ก็เป็นวิธีการหนึ่งที่ช่วยลดเวลาในการรู้จำเสียงพูดของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

2) การให้คะแนนและค้นหาของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

เวลาในการขั้นตอนการให้คะแนนและค้นหาคำตอบของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์จะประกอบด้วยสองส่วน คือ เวลาในการให้คะแนน และวลาในการค้นหา สำหรับการให้คะแนนวิทยานิพนธ์นี้ยังไม่ได้มีการลดการคำนวณในการหาค่าความควรจะเป็น (Likelihood) จากแบบจำลองเสียง ซึ่งเป็นการดำเนินการโอเปอเรชันทางเมทริกซ์ต่างๆ โดยงานวิจัยนี้ได้มีการใช้แบบจำลองเกาส์เซียนที่มีเมทริกซ์ความแปรปรวนร่วมเกี่ยวตามแนวทแยงแต่ในการคำนวณนั้นเป็นการคำนวณเมทริกซ์ที่มีมิติเป็น $N \times N$ ถ้ามีการปรับลดการคำนวณให้เป็นแบบเมทริกซ์ที่มีความแปรปรวนร่วมเกี่ยวตามแนวทแยงจะสามารถลดความซับซ้อนในการคำนวณจาก $O(n^n)$ เป็น $O(n)$ ซึ่งจะทำให้เวลาในการรู้จำเสียงพูดของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ลดลง

ในส่วนของเวลาในการค้นหาคำตอบ การค้นหาแบบไวเทอร์บิที่ใช้ในงานวิจัย ยังไม่ได้มีการตัดเส้นทาง (Prune) บางเส้นทางที่มีค่าคะแนนต่ำกว่าที่กำหนด ซึ่งถ้ามีการตัดเส้นทางออก โดยมีการหาค่าขีดแบ่งที่เหมาะสม ก็จะช่วยลดการคำนวณลงไปได้อีก

6.5 ผลการกับฐานข้อมูลเสียงพูดภาษาอังกฤษ TIMIT

ในงานวิจัยนี้ได้มีการทดลองนำวิธีการรู้จำเสียงพูดแบบอาศัยเซกเมนต์มาทดสอบกับเสียงข้อมูลเสียงพูดภาษาอังกฤษ และได้มีการทดลองใช้วิธีที่นำเสนอในการปรับปรุงคุณภาพกราฟของเซกเมนต์กับเสียงพูดภาษาอังกฤษด้วยเช่นกัน เพื่อแสดงให้เห็นว่าวิธีที่งานวิจัยนี้ได้นำเสนอสามารถนำไปใช้กับการรู้จำเสียงพูดภาษาอังกฤษได้เช่นกัน

เสียงพูดภาษาอังกฤษที่ใช้ในการทดลองใช้เสียงพูดจากฐานข้อมูลเสียงพูด TIMIT ซึ่งเป็นฐานข้อมูลเสียงพูดต่อเนื่องภาษาอังกฤษสำเนียงอเมริกัน โดยมีลักษณะเป็นการอ่านบทความต่างๆ โดยฐานข้อมูล TIMIT ได้พัฒนาโดยได้รับการสนับสนุนจาก Defense Advanced Research Projects Agency (DARPA) โดยฐานข้อมูลเสียง TIMIT จะมีทั้งหมด 6,300 ประโยค ซึ่งใช้ผู้พูดจำนวน 630 คน อ่านคนละ 10 ประโยค โดยผู้พูดจะแบ่งเป็นชายจำนวน 438 คน หญิงจำนวน 192 คน โดยผู้พูดจะมาจาก 8 ภูมิภาคในสหรัฐอเมริกา ฐานข้อมูลเสียง TIMIT เป็นฐานข้อมูลเสียงที่นิยมใช้ในการประเมินผลการรู้จำเสียงพูดสำหรับงานวิจัยต่างๆ

ผลการทดลองในการรู้จำเสียงพูดในระดับหน่วยเสียงของแต่ละวิธีกับฐานข้อมูลเสียงพูด TIMIT จะแสดงดังตารางที่ 6.11 จากตารางจะสังเกตได้ว่าการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่มีการใช้ทั้งแบบจำลองเซกเมนต์และขอบเขตของหน่วยเสียงมีความแม่นยำสูงกว่าการรู้จำเสียงพูดแบบอาศัยกรอบเวลา และเมื่อมีการนำวิธีที่วิทยานิพนธ์นี้ได้เสนอมาใช้แก้ไขความผิดของเซกเมนต์ที่เกิดจากการแทรกของขอบเขตของหน่วยเสียง จะพบว่าความผิดพลาดของเซกเมนต์ที่เกิดจากการแทรกของขอบเขตของหน่วยเสียงลดลงจาก 11.31% เป็น 9.72% และเมื่อนำมาผ่านการรู้จำเสียงพูดพบว่าเมื่อมีการลดความผิดพลาดของเซกเมนต์แล้วจะได้ความแม่นยำในการรู้จำเสียงสูงขึ้นจาก

56.18% เป็น 57.33% ซึ่งจากผลการทดลองนี้ สามารถสรุปได้ว่าวิธีการปรับปรุงความผิดพลาดที่ได้นำเสนอในวิทยานิพนธ์นี้สามารถนำไปใช้ได้ในการรู้จำเสียงพูดภาษาอังกฤษได้

ตารางที่ 6.11 ผลการรู้จำเสียงพูดในระดับหน่วยเสียงของแต่ละวิธีการเมื่อทดลองกับฐานข้อมูลเสียงพูด TIMIT

ระบบรู้จำเสียงพูด	วิธีการ	ความผิดพลาดของเซกเมนต์ที่เกิดจากการแทรกของขอบเขตของหน่วยเสียง (%)	ความแม่นยำ (%)
การรู้จำเสียงพูดแบบอาศัยกรอบเวลา	-	-	52.85
การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่มีการใช้ทั้งแบบจำลองเซกเมนต์และขอบเขตของหน่วยเสียง	ไม่มีการปรับปรุงคุณภาพกราฟของเซกเมนต์	11.31	56.18
การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่มีการใช้ทั้งแบบจำลองเซกเมนต์และขอบเขตของหน่วยเสียง	ปรับปรุงคุณภาพกราฟของเซกเมนต์โดยใช้ความไม่ต่อเนื่องของสัญญาณ	9.72	57.33

บทที่ 7

สรุปผลการวิจัย

วิทยานิพนธ์ฉบับนี้ได้นำเสนอวิธีการที่จะเพิ่มความแม่นยำให้แก่การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ ซึ่งสำหรับภาษาไทยมีความแม่นยำในการรู้จำเสียงพูดในระดับหน่วยเสียงได้เพียงประมาณ 50% ซึ่งเกิดจากมีทรัพยากรไม่เพียงพอที่จะทำให้ระบบรู้จำเสียงพูดแบบอาศัยกรอบเวลาซึ่งใช้ในการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น มีความสมบูรณ์ถูกต้อง จึงทำเกิดความผิดพลาดอยู่ในกราฟของเซกเมนต์ และส่งผลกระทบต่อกระบวนการรู้จำเสียงพูดของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ ดังนั้น ในงานวิจัยนี้จะมุ่งเน้นที่จะปรับปรุงคุณภาพกราฟของเซกเมนต์ที่ได้จากการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีทางความน่าจะเป็น นอกจากนี้ยังมีการปรับปรุงการให้คะแนนในขั้นตอนการให้คะแนนและค้นหาคำตอบโดยนำค่าความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างมาใช้เป็นตัวถ่วงน้ำหนักในการคิดคะแนนทางเสียง จากผลการทดลองในการรู้จำเสียงพูดระดับหน่วยเสียงแสดงให้เห็นว่าการรู้จำเสียงพูดที่มีการปรับปรุงคุณภาพกราฟของเซกเมนต์ และนำความน่าจะเป็นที่เซกเมนต์จะถูกจัดอยู่ในกลุ่มของหน่วยเสียงแบบกว้างมาใช้จะให้ผลการรู้จำเสียงพูดที่ดีขึ้นประมาณ 15% เมื่อเทียบกับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ไม่มีการใช้วิธีที่นำเสนอ และได้ผลดีการรู้จำเสียงพูดดีกว่าการรู้จำเสียงพูดแบบอาศัยกรอบเวลาประมาณ 25%

อย่างไรก็ตามการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ยังสามารถที่จะปรับปรุงความแม่นยำได้อีกโดยนำความรู้ด้านสมบัติลักษณะเฉพาะที่เหลือมาเป็นคะแนนที่ใช้ในขั้นตอนการให้คะแนนและค้นหาคำตอบที่ดีที่สุด เช่น ตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียง เพื่อแก้ไขความผิดพลาดของคำตอบที่อยู่ภายในกลุ่มของหน่วยเสียงแบบกว้างเดียวกัน ซึ่งเป็นการใช้ข้อดีของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่สามารถเพิ่มความรู้ต่างๆเข้าไปได้อย่างง่ายดายอีกด้วย

รายการอ้างอิง

[1] Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77 (1989): 257-286.

[2] Glass, J. R. A Probabilistic Framework for Segment-Based Speech Recognition. Computer Speech & Language 17 (2003): 137-152.

[3] Halberstadt, A. K., and Glass, J. R. Heterogeneous acoustic measurements for phonetic classification. In Eurospeech, pp. 401-404. Rhodes, Greece, 1997.

[4] Halberstadt, A. K., and Glass, J. R. Heterogeneous Measurements and Multiple Classifiers for Speech Recognition. In ICSLP, pp. 995-998. Sydney, Australia, 1998.

[5] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallet, D., and Dahlgren, N. The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM NTIS order number PB91-505065. [CDROM]. 1990.

[6] Hazen, T. J., Hetherington, L., Shu, H., and Livescu, K. Pronunciation Modeling Using a Finite-State Transducer Representation. Speech Communication 46 (2005): 189-203.

[7] Likitsupin, K., Tangruamsub, S., Punyabukkana, P., and Suchato, A. Phoneme Recognition from Thai Continuous Speech using a Segment-based Approach. In The 11th National Computer Science and Engineering Conference (NCSEC2007), pp. 218-222. 2007.

[8] Chang, J. W. and Glass, J. R. Segmentation and Modeling in Segment-Based Recognition. In Eurospeech, pp. 1199-1202. Rhodes, Greece, 1997.

[9] Lee, S. C. Probabilistic Segmentation for Segment-Based Speech Recognition. Doctoral dissertation, Faculty of Electrical Engineering and Computer Science Massachusetts Institute of Technology, 1998.

[10] Lee, S. C. and Glass, J. R. Real-time Probabilistic Segmentation for Segment-Based Speech Recognition. In The Fifth International Conference on Spoken Language Processing (ICSLP 1998), pp. 1803-1806. Sydney, Australia, 1998.

[11] Wutiwiwatchai, C., Cotsomrong, P., Suebvisai, S., and Kanokphara, S. Phonetically Distributed Continuous Speech Corpus for Thai Language. In The Third International Conference on Language Resources and Evaluation (LREC 2002), pp. 869-872. Las Palmas, Spain, 2002.

[12] Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Kanokphara, S., and Thatphithakkul, N. Thai Speech Corpus for Thai Speech Recognition. In The Oriental COCOSDA 2003, pp. 54-61. Singapore, 2003.

[13] Cotsomrong, P., Sunpetchniyom, T., Kasuriya, S., Thatphithakku, N., and Wutiwiwatchai, C., "LOTUS: Large vOcabulary Thai continUous Speech Recognition Corpus," in NSTDA Annual Conference S&T in Thailand (NAC 2005). Thailand, 2005.

[14] กาญจนา นาคสกุล. ระบบเสียงภาษาไทย. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2541.

[15] หน่วยปฏิบัติการวิจัยวิทยาการมนุษยภาษา. เอกสารประกอบฐานข้อมูล LOTUS Corpus. ปทุมธานี: ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ, 2548.

[16] บุญเสริม กิจศิริกุล และณัฐกร ทับทอง. การพัฒนาระบบรู้จำเสียงพูดภาษาไทย. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2548.

[17] Juneja, A. and Espy-Wilson, C. Y. Speech Segmentation using Probabilistic Phonetic Feature Hierarchy and Support Vector Machines. In International Joint Conference on Neural Networks (IJCNN 2003), pp. 675-679. Oregon, USA, 2003.

[18] Juneja, A. and Espy-Wilson, C. Y. A Probabilistic Framework for Landmark Detection Based on Phonetic Features for Automatic Speech Recognition. The Journal of the Acoustical Society of America 123 (2008): 1154-1168.

[19] Glass, J. R. A Probabilistic Framework for Segment-Based Speech Recognition. Computer Speech and Language 17 (2003): 137–152.

[20] Glass, J. R. and Zue, V. Multi-level Acoustic Segmentation of Continuous Speech. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1988), pp. 429–432. New York, 1988.

[21] Zue, V., Glass, J. R., Pilips, M., and Seneff, S. Acoustic Segmentation and Phonetic Classification in the SUMMIT System. In Inernational Conference on Acoustics, Speech, and Signal Processing (ICASSP 1989), pp. 389-392. Glasgow, UK, 1989.

[22] Leelaphattarakij, P. Speech Segmentation for Thai Segment-Based Speech Recognition Using Acoustic-Phonetic Information. Master's Thesis, Faculty of Department of Computer Engineering Faculty of Engineering Chulalongkorn University, 2006.

[23] Juneja, A. Speech Recognition Based on Phonetic Features and Acoustic Landmarks. Doctoral dissertation, Faculty of Electrical and Computer Engineering University of Maryland, 2004.

[24] บุญเสริม กิจศิริกุล. การเรียนรู้ของเครื่อง. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2549.

[25] Chang, C.-C. and Lin, C.-J. A Library for Support Vector Machines. [Online]. 2001. Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm/ [2010, March]

[26] Platt, J. C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. USA: MIT Press, 1999.

[27] Lamel, L. F. and Gauvain, J. L. High Performance Speaker-Independent Phone Recognition Using CDHMM. In The Third European Conference on Speech Communication and Technology (Eurospeech), pp. 121-124. Berlin, Germany, 1993.

[28] Ming, J. and Smith, F. J. Improved Phone Recognition Using Bayesian Triphone Models. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1998), pp. 409-412. USA, 1998.

[29] Wang, D., Lu, L., and Zhang, H.-J. Speech Segmentation without Speech Recognition. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), pp. 468-471. Hong Kong, 2003.

[30] Leelaphattarakij, P., Punyabukkana, P., and Suchato, A. Locating Phone Boundaries from Acoustic Discontinuities using a Two-staged Approach. In The Ninth International Conference on Spoken Language Processing (Interspeech 2006), pp. 673-676. Pennsylvania, USA, 2006.

[31] Sainath, T. N. and Hazen, T. J. A Sinusoidal Model Approach to Acoustic Landmark Detection and Segmentation for Robust Segment-Based Speech Recognition. In IEEE International Conference on Acoustics, Speech ,and Signal Processing (ICASSP 2006), pp. 525-528. Toulouse, France, 2006.

[32] Sainath, T. N., Kanevsky, D., and Iyengar, G. Unsupervised Audio Segmentation Using Extended Baum-Welch Transformations. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), pp. 209-212. Hawaii, USA, 2007.

[33] Sainath, T. N. and Zue, V. A Comparison of Broad Phonetic and Acoustic Units for Noise Robust Segment-Based Phonetic Recognition In The Ninth Annual Conference of the International Speech Communication Association (Interspeech 2008), pp. 2378-2381. Brisbane, Australia, 2008.

[34] Niyogi, P., Burges, C., and Ramesh, P. Distinctive Feature Detection Using Support Vector Machines. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1999), pp. 425-428. Arizona, USA, 1999.

[35] Liu, S. A. Landmark Detection for Distinctive Feature-Based Speech Recognition. Doctoral dissertation, Faculty of Electrical Engineering and Computer Science Massachusetts Institute of Technology, 1995.

[36] Liu, S. A. Landmark Detection for Distinctive Feature-Based Speech Recognition. The Journal of the Acoustical Society of America 100 (1996): 3417-3430.

[37] Hosom, J. P. Automatic Phoneme Alignment Based on Acoustic-Phonetic Modeling. In The Seventh International Conference on Spoken Language Processing (ICSLP2002), pp. 357-360. Colorado, USA, 2002.

[38] Schutte, K. and Glass, J. R. Robust Detection of Sonorant Landmarks. In The Ninth European Conference on Speech Communication and Technology (Interspeech 2005), pp. 1005-1008. Lisbon, Portugal, 2005.

[39] Dareyoah, P., Suchato, A., and Punyabukkana, P. A Study of Acoustic Measurements for Voicing Detection in Speech with Room-level SNR. In The Sixth Symposium of Natural Language Processing (SNLP 2005), pp. 109-114. Chiang Rai, Thailand, 2005.

[40] Boonsuk, S., Punyabukkana, P., and Suchato, A. Phone Boundary Detecting using Selective Refinements and Context-dependent Acoustic Features. In The Tenth International Conference on Spoken Language Processing (Interspeech 2007). Antwerp, Belgium, 2007.

[41] Boonsuk, S., Punyabukkana, P., and Suchato, A. Precise Phone Boundary Detection using Selective Context-dependent Acoustic Refinement. In The Fourth International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2007). Chiang Rai, Thailand, 2007.

[42] Meng, H. M., Zue, V., and Leung, H. C. Signal Representation, Attribute Extraction and, the Use of Distinctive Features for Phonetic Classification. In The Speech and Natural Language Workshop, pp. 176-181. Pacific Grove, California, 1991.

[43] Bitar, N. N. and Espy-Wilson, C. Y. Knowledge-Based Parameters for HMM Speech Recognition. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1996), pp. 29-32. Georgia, USA, 1996.

[44] Salomon, A. and Espy-Wilson, C. Y. Automatic Detection of Manner Events Based on Temporal Parameters. In The Sixth European Conference on Speech Communication and Technology (EUROSPEECH 1999), pp. 2797-2800. Budapest, Hungary, 1999.

[45] Salomon, A., Espy-Wilson, C. Y., and Deshmukh, O. Detection of Speech Landmarks: Use of Temporal Information. The Journal of the Acoustical Society of America 115 (2004): 1296-1305.

[46] Kirchhoff, K., Fink, G. A., and Sagerer, G. Combining Acoustic and Articulatory Feature Information for Robust Speech Recognition. Speech Communication 37 (2002): 303-319.

[47] Tang, M., Seneff, S., and Zue, V. Two-Stage Continuous Speech Recogntion Using Feature-Based Models: A Preliminary Study. In IEEE Workshop on ASRU, pp. 49-54. 2003.

[48] Tang, M., Seneff, S., and Zue, V. Modeling Linguistic Features in Speech Recognition. In The Eighth European Conference on Speech Communication and Technology (EUROSPEECH 2003), pp. 2585-2588. Geneva, Switzerland, 2003.

[49] Pruthi, T. and Espy-Wilson, C. Y. Automatic Classification of Nasals and Semivowels. In The Fifteenth International Congress of Phonetic Sciences (ICPhS 2003), pp. 3061-3064. Barcelona, Spain, 2003.

[50] Pruthi, T. and Espy-Wilson, C. Y. Acoustic Parameters for Automatic Detection of Nasal Manner. Speech Communication 43 (2004): 225-239.

[51] Borys, S. and Hasegawa-Johnson, M. Distinctive Feature Based SVM Discriminant Features for Improvements to Phone Recognition on Telephone Band Speech. In The Ninth European Conference on Speech Communication and Technology (Interspeech 2005), pp. 697-700. Lisbon, Portugal, 2005.

[52] Scanlon, P., Ellis, D. P. W., and Reilly, R. B. Using Broad Phonetic Group Experts for Improved Speech Recognition. IEEE Transactions on Audio Speech and Language Processing 15 (2007): 803-812.

[53] Chang, H.-A. and Glass, J. R. Hierarchical Large-Margin Gaussian Mixture Models for Phonetic Classification. In IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2007), pp. 272-277. Kyoto, Japan, 2007.

[54] Sainath, T. N., Kanevsky, D., and Ramabhadran, B. Broad Phonetic Class Recognition in a Hidden Markov Model Framework Using Extended Baum-Welch Transformations. In The tenth biannual IEEE workshop on Automatic Speech Recognition and Understanding (ASRU 2007), pp. 306-311. Kyoto, Japan, 2007.

[55] Juneja, A. and Espy-Wilson, C. Y. Segmentation of Continuous Speech using Acoustic-Phonetic Parameters and Statistical Learning. In The Ninth International Conference on Neural Information Processing (ICONIP2002), pp. 726-730. Singapore, 2002.

[56] Young, S., et al. The HTK Book (for HTK Version 3.4). Cambridge: Cambridge University, 2006.

ภาคผนวก

ภาคผนวก ก

ในภาคผนวกนี้จะแสดงความแม่นยำของการรู้จำเสียงพูดระบบหน่วยเสียงของระบบรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ ที่จำนวนรอบการประมาณซ้ำ และทอพอโลยีแบบต่างๆ โดยผลลัพธ์ที่มีความแม่นยำสูงสุดที่ทอพอโลยีแบบนั้นๆจะเน้นด้วยตัวอักษรที่ขีดเส้นใต้

ตารางที่ ก.1 ผลการรู้จำเสียงพูดในระดับหน่วยเสียงของระบบรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟที่จำนวนรอบการประมาณซ้ำ และทอพอโลยีต่างๆ

จำนวนรอบการประมาณซ้ำ	ทอพอโลยี		
	สามสถานะ	ห้าสถานะ	ทั้งสามและห้าสถานะ
1	7.70	7.12	7.62
2	9.65	20.44	19.85
3	23.71	33.43	33.44
4	28.25	37.66	37.92
5	29.65	39.46	39.94
6	30.16	40.03	40.75
7	30.56	40.35	41.12
8	30.83	40.66	41.43
9	31.14	40.63	41.60
10	31.21	40.65	41.69
11	31.63	40.65	41.66
12	31.79	40.91	41.70
13	32.19	41.03	41.98
14	32.12	41.13	42.11
15	32.05	41.39	42.19
16	32.16	41.48	42.39
17	32.21	41.59	42.42
18	32.11	41.72	42.66
19	32.15	41.82	42.60
20	32.13	41.82	42.65

จำนวนรอบการประมาณซ้ำ	ทอพอโลยี		
	สามสถานะ	ห้าสถานะ	ทั้งสามและห้าสถานะ
21	32.14	41.92	42.69
22	32.16	42.03	42.66
23	32.24	42.09	42.71
24	32.53	42.15	42.74
25	32.86	42.19	42.78
26	**32.98**	42.17	42.80
27	32.86	**42.24**	42.78
28	32.85	42.23	42.79
29	32.79	42.22	**42.83**
30	32.79	42.23	42.82
31	32.81	42.21	42.82
32	32.68	42.19	42.80
33	32.62	42.19	42.81
34	32.62	42.19	42.83
35	32.45	42.19	42.81
36	32.44	42.17	42.82
37	32.45	42.14	42.79
38	32.46	42.12	42.79
39	32.46	42.15	42.79
40	32.47	42.15	42.76
41	32.52	42.17	42.74
42	32.55	42.18	42.76
43	32.59	42.18	42.78
44	32.63	42.17	42.79
45	32.63	42.16	42.76
46	32.62	42.15	42.76
47	32.63	42.16	42.75
48	32.63	42.16	42.72
49	32.60	42.16	42.72

จำนวนรอบการประมาณซ้ำ	ทอพอโลยี		
	สามสถานะ	ห้าสถานะ	ทั้งสามและห้าสถานะ
50	32.79	42.23	42.82

ประวัติผู้เขียนวิทยานิพนธ์

นายเกริกศักดิ์ ลิขิตสุภิณ เกิดเมื่อวันที่ 16 กรกฎาคม พ.ศ. 2522 ที่กรุงเทพมหานคร สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จากมหาวิทยาลัยบูรพา ในปีการศึกษา 2543 และสำเร็จการศึกษาในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการสารสนเทศ คณะเทคโนโลยีสารสนเทศ จากสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ในปีการศึกษา 2545