การประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความ

นายณัฏฐ์ เลิศวงศ์คณากูล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2555

REAL-TIME AUTOMATIC SPEECH-TEXT SYNCHRONIZATION

Mr. Nat Lertwongkhanakool

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering Program in Computer Engineering

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2012

หัวข้อวิทยานิพนธ์	การประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความ
โดย	นายณัฏฐ์ เลิศวงศ์คณากูล
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

…….. คณบดีคณะวิศวกรรมศาสตร์

(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

……. ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.เศรษฐา ปานงาม)

…….. อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ)

…….. อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต)

…….. กรรมการภายนอกมหาวิทยาลัย

(ดร.ชัย วุฒิวิวัฒน์ชัย)

ณัฏฐ์ เลิศวงศ์คณากูล: การประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความ (Real-Time Automatic Speech-Text Synchronization) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.โปรดปราน บุณยพุกกณะ, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: รศ. ดร.อติวงศ์ สุชาโต, 71 หน้า

การประสานเวลาอัตโนมัติระหว่างเสียงและข้อความนั้น เป็นวิธีการที่แสดงเนื้อหาเดียวกันจากสื่อที่แตกต่างกัน ซึ่งในที่นี้คือเสียงและข้อความ ซึ่งโปรแกรมประยุกต์ส่วนใหญ่จะเป็นการประสานเวลาในระดับประโยค และใช้ข้อมูลของเสียงและข้อความทั้งหมดในการประสานเวลา แต่เนื่องด้วยความต้องการของโปรแกรมประยุกต์บางประเภท เช่น โปรแกรมการสร้างหนังสือเสียงซึ่งมีข้อความทั้งหมด และต้องการที่จะประสานเวลาในทันทีที่เสียงเข้ามาในระบบ อย่างไรก็ตาม ด้วยลักษณะของภาษาไทยซึ่งมีการแบ่งประโยคและคำไม่ชัดเจน ทำให้การประสานเวลานั้นมีความท้าทาย ดังนั้นวิทยานิพนธ์นี้จึงเสนอขั้นตอนวิธีในการประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความในระดับพยางค์ ขั้นตอนวิธีที่นำเสนอนั้นใช้หลักการในการตรวจหาพยางค์และตรวจหาความไม่ตรงกันของการถอดเสียง การทดลองได้ศึกษาการใช้ลักษณะเด่นต่าง ๆ และการปรับค่าพารามิเตอร์อย่างละเอียด ขั้นตอนวิธีที่นำเสนอถูกนำมาเปรียบเทียบกับระบบอ้างอิง 2 ระบบ ซึ่งได้ผลลัพธ์ดีกว่าระบบอ้างอิง 75% และ 41% ตามลำดับ และในแง่ของเวลาสามารถคำนวณได้ในทันที

ภาควิชา...วิศวกรรมคอมพิวเตอร์... ลายมือชื่อนิสิต.....

สาขาวิชา..วิศวกรรมคอมพิวเตอร์... ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก.....

ปีการศึกษา.....2555..... ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม.....

547 01886 21 : MAJOR COMPUTER ENGINEERING

KEYWORDS : AUTOMATIC SPEECH-TEXT SYNCHRONIZATION / SYLLABLE DETECTION / REAL-TIME ENDPOINT DETECTION / LIVE SPEECH AND TRANSCRIPTION SYNCHRONIZATION / TRANSCRIPTION ERRORS DETECTION

NAT LERTWONGKHANAKOOL : REAL-TIME AUTOMATIC SPEECH-TEXT SYNCHRONIZATION. ADVISOR: ASST.PROF. PROADPRAN PUNYABUKKANA, PH.D., CO-ADVISOR: ASSOC.PROF. ATIWONG SUCHATO, PH.D., 71 PP.

Most of the researches in synchronization of audio and text have been focusing on the synchronization at the level of utterance. However, to generate audio books in unstructed language like Thai from live speech, a finer lever of synchronization is necessary. We propose an algorithm to synchronize live speech with its corresponding transcription in real time at syllabic unit. The proposed algorithm employs the syllable detection concept and the transcription errors detection concept. The experiment was studied the features and the parameters empirically. The result were compared with 2 baselines and found that the proposed algorithm was better than 2 baselines 75% and 41% respectively. In term of processing time, the proposed algorithm was able to give the results in real-time.

Department : Computer Engineering Student's Signature……

Field of Study : Computer Engineering Advisor's Signature…….....

Academic Year : ……2012…… Co-advisor's Signature……...

กิตติกรรมประกาศ

ในการจัดทำวิทยานิพนธ์นี้ ข้าพเจ้าได้รับทุนอุดหนุนการศึกษาระดับบัณฑิตศึกษา จุฬาลงกรณ์มหาวิทยาลัยเพื่อเฉลิมฉลองวโรกาสที่พระบาทสมเด็จพระเจ้าอยู่หัวทรงเจริญพระชนมายุครบ 72 พรรษา ประจำปีการศึกษา 2554 ซึ่งข้าพเจ้าขอขอบพระคุณและรู้สึกยินดีเป็นอย่างยิ่งที่ได้รับเกียรติเป็นหนึ่งในผู้ที่ได้รับทุนนี้ และทำให้การดำเนินงานของข้าพเจ้าเป็นไปอย่างราบรื่น

ข้าพเจ้าขอขอบพระคุณอาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ และ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม รองศาสตราจารย์ ดร.อติวงศ์ สุชาโต ที่ให้คำแนะนำทั้งการดำเนินงานวิจัย ความรู้ต่าง ๆ ที่มีประโยชน์ และเรื่องอื่น ๆ จนสามารถทำให้ข้าพเจ้าสามารถดำเนินการทำวิทยานิพนธ์สำเร็จลุล่วงไปได้ ข้าพเจ้าขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.เศรษฐา ปานงาม และ อาจารย์ ดร.ชัย วุฒิวิวัฒน์ชัย ที่กรุณาสละเวลาอันมีค่ามาให้คำแนะนำอันเป็นประโยชน์ต่อการดำเนินการทำวิทยานิพนธ์

นอกจากนั้น ข้าพเจ้าขอขอบพระคุณ หน่วยปฏิบัติการวิจัยวิทยาการมนุษยภาษา ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ: NECTEC ที่เอื้อเฟื้อเครื่องมือต่าง ๆ ที่จำเป็นต่อการศึกษาและทดลองในการทำวิทยานิพนธ์นี้ และขอขอบคุณรุ่นพี่ รุ่นน้อง และเพื่อนร่วมงานในห้องปฏิบัติการระบบภาษาพูด ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่คอยให้กำลังใจ และคำแนะนำต่าง ๆ ทั้งในและนอกการประชุมเล็บเสมอมา

และสุดท้ายข้าพเจ้าขอกราบขอบพระคุณบิดา มารดา และครอบครัว ที่ให้กำลังใจแก่ข้าพเจ้าในยามที่ท้อแท้ และสนับสนุนให้ข้าพเจ้าสามารถศึกษาและจัดทำวิทยานิพนธ์นี้ให้สำเร็จลุล่วงได้

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

สารบัญรูป

หน้า

หน้า

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

การประสานเวลาอัตโนมัติของเสียงและข้อความ เป็นวิธีการในการที่จะเชื่อมโยงเนื้อหาจากสื่อที่ต่างกัน แต่ต้องการจะสื่อสารในเนื้อหาเดียวกัน ให้สามารถแสดงเนื้อหาออกมาได้ตรงกันแบบอัตโนมัติ โดยสื่อในที่นี้คือ ข้อความและเสียงพูด ซึ่งมีความสัมพันธ์กันเชิงเวลา การประสานเวลาของเสียงและข้อความนั้นมีความต้องการในการใช้เพื่อการสร้างฐานข้อมูลเสียง หรืออาจจะใช้ในการสร้างโปรแกรมประยุกต์บางอย่าง เช่น การสร้างโปรแกรมการประสานเวลาระหว่างคำร้องและเสียงร้องในเพลง [1], ระบบช่วยชี้คำในการเรียนรู้บทเรียนออนไลน์ [2] หรือ ระบบการแสดงคำบรรยายใต้ภาพในการแพร่สัญญาณข่าว [3], [4] เป็นต้น ซึ่งโดยปกติแล้ว การประสานเวลาของข้อความและเสียงโดยใช้คนเป็นผู้กระทำ จะทำให้เสียเวลาอย่างมาก จึงมีการคิดค้นการประสานเวลาเสียงและข้อความแบบอัตโนมัติขึ้นมา แต่อย่างไรก็ตาม การประสานเวลาที่กล่าวมาข้างต้น จะเป็นแนวคิดในการประสานเวลาระหว่างเสียงที่มีอยู่และข้อความที่สอดคล้องกัน และส่วนใหญ่จะประสานเวลาในระดับของประโยคหรือวลี แต่ในโปรแกรมประยุกต์บางประเภทต้องการแนวคิดในการประสานเวลาของข้อความและเสียงพูดที่กำลังถูกพูดออกมาในทันที ซึ่งจะมีลักษณะในการประมวลแบบทันที (Real-Time) โดยที่มีระดับการประสานเวลานั้นมีความละเอียดมากกว่าประโยคหรือวลี เช่น ระดับคำ หรือระดับพยางค์ ตัวอย่างเช่น การนำมาใช้การโปรแกรมในการสร้างหนังสือเสียง [5] เพราะการสร้างหนังสือเสียงนั้นจำเป็นจะต้องอ่านตามข้อความที่แสดงไว้อยู่แล้วให้ถูกต้อง ดังนั้น หากสามารถแสดงคำหรือพยางค์ที่ผู้อ่านกำลังอ่านอยู่ ก็สามารถทำให้เสียงที่ผู้อ่านอ่านตามข้อความประสานเวลากับข้อความได้อย่างอัตโนมัติ โดยที่ไม่ต้องทำเองด้วยมืออีกด้วย นอกจากนั้นก็ยังสามารถนำไปประยุกต์กับโปรแกรมประยุกต์ เช่น โปรแกรมสำหรับฝึกการเรียนรู้ภาษา หรือโปรแกรมคาราโอเกะก็ได้ ซึ่งโปรแกรมจะช่วยเน้นข้อความที่กำลังถูกอ่าน เพื่อให้ทราบตำแหน่งปัจจุบันที่ผู้อ่านกำลังอ่าน

ในวิทยานิพนธ์นี้จะนำเสนอขั้นตอนวิธี (Algorithm) ในการประสานเวลาอัตโนมัติของเสียงและข้อความแบบทันที เพื่อใช้ในโปรแกรมประยุกต์บางประเภทดังที่กล่าวข้างต้น ซึ่งจะมีความ

ละเอียดของการประสานเวลาในระดับพยางค์ เนื่องจากในภาษาไทยนั้น มีความกำกวมทั้งตัวของภาษาเอง เช่น ไม่มีจุดจบของประโยคที่แน่นอน และจุดจบของคำที่เห็นได้ชัดเจน [6] แต่ในขณะเดียวกัน พยางค์มีโครงสร้างในภาษาไทยที่ชัดเจนและสามารถตรวจหาได้ในทางเสียง และมีความละเอียดเพียงพอที่จะแสดงให้ผู้ใช้เห็นการประสานเวลาในทันทีที่ผู้พูดพูดออกมา โดยอาศัยหลักการของการตรวจหาจุดจบพยางค์และการหาแกนกลางของพยางค์ และเนื่องจากเราทราบข้อความที่จะต้องอ่านอยู่แล้ว ดังนั้น เราจะใช้หลักการตรวจหาความผิดพลาดของการถอดเสียงนำมาประยุกต์ด้วย เพื่อให้ผลลัพธ์มีความแม่นยำเพิ่มมากขึ้น

วัตถุประสงค์ของการวิจัย

1. นำเสนอขั้นตอนวิธีในการประสานเวลาเสียงและข้อความแบบประมวลผลทันที

ขอบเขตของการวิจัย

1. ผู้พูดจะต้องพูดด้วยความเร็วปกติ ไม่เร็วหรือช้าจนเกินไป
2. ผู้พูดจะต้องพูดตามข้อความที่กำหนดอย่างถูกต้อง
3. ระหว่างพูด ผู้พูดจะต้องอยู่ในสภาพแวดล้อมที่ไม่มีเสียงรบกวนมาก
4. ระบบรองรับเฉพาะภาษาไทยเท่านั้น

วิธีการดำเนินการวิจัย

1. ขั้นตอนการศึกษาเบื้องต้น
   1.1. ศึกษาข้อมูลและงานวิจัยที่เกี่ยวกับการประสานเวลาที่มีอยู่ในปัจจุบัน
   1.2. ศึกษาข้อมูลและงานวิจัยที่เกี่ยวกับการตรวจหาพยางค์
   1.3. ศึกษาข้อมูลและงานวิจัยที่เกี่ยวกับการตรวจสอบความผิดพลาดของการถอดเสียง
2. ขั้นตอนการออกแบบ พัฒนาและทดลองส่วนของการตรวจหาพยางค์
   2.1. ศึกษาข้อมูลเกี่ยวกับการตรวจหากึ่งกลางพยางค์และการตรวจหาจุดจบของพยางค์
   2.2. พัฒนาขั้นตอนวิธีในการประสานเวลาโดยใช้หลักการตรวจหาจุดจบของพยางค์และการตรวจหาแกนกลางพยางค์
   2.3. สรุปและวิเคราะห์ผลของขั้นตอนวิธีในส่วนนี้

3. ขั้นตอนการออกแบบ พัฒนา ทดลองส่วนของการตรวจหาความผิดพลาดของการถอดเสียง
   3.1. ศึกษาข้อมูลและออกแบบการตรวจหาความผิดพลาดของการถอดเสียง
   3.2. พัฒนาขั้นตอนวิธีในการตรวจหาความผิดพลาดของการถอดเสียง
   3.3. สรุปและวิเคราะห์ผลของขั้นตอนวิธี
4. ออกแบบวิธีการทดลอง ทดสอบวิธีการที่นำเสนอ และปรับค่าพารามิเตอร์ เพื่อให้ได้ผลลัพธ์ที่ดีที่สุดกับฐานข้อมูลเสียงที่ใช้ทดสอบ
5. ทดสอบการประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความ
6. วิเคราะห์และสรุปผล

ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ขั้นตอนวิธีในการประสานเวลาระหว่างข้อความและเสียงแบบทันทีสำหรับภาษาไทย

ผลงานตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้รับการตอบรับและตีพิมพ์ในบทความทางวิชาการ 1 หัวข้อ เรื่องได้แก่

1. “Real-time Synchronization of Live Speech with Its Transcription” โดย “Nat Lertwongkhanakool”, “Proadpran Punyabukkana” และ “Atiwong Suchato” ในงานประชุมวิชาการนานาชาติ “2013 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2013)” ณ โรงแรมมาริไทม์ ปาร์คแอนด์สปา จังหวัดกระบี่ ประเทศไทย ระหว่างวันที่ 15 พฤษภาคม 2556 ถึงวันที่ 17 พฤษภาคม 2556
2. “ChulaDAISY: An Automated DAISY Audio Book Generation” โดย “Proadpran Punyabukkana”, “Nat Lertwongkhanakool”, “Natthawut Kertkeidkachorn”, “Surapol Vorapatratorn”, “Pawanrat Hirankan” และ “Atiwong Suchato” ในงานประชุมวิชาการนานาชาติ “6th International Convention on Rehabilitation Engineering and

Assistive Technology (i-CREATe 2012)" ณ ITE College East ประเทศสิงคโปร์ ระหว่างวันที่ 24 กรกฎาคม 2555 ถึงวันที่ 26 กรกฎาคม 2555

ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์

ในวิทยานิพนธ์เล่มนี้ได้แบ่งการนำเสนอออกเป็น 5 ส่วนคือ

1. ในบทที่ 1 คือ บทนำซึ่งจะกล่าวถึงที่มาและความสำคัญของปัญหา วัตถุประสงค์ของการวิจัย ขอบเขตในการวิจัย ขั้นตอนในการวิจัย ประโยชน์ที่จะได้รับ รวมถึงผลงานที่ได้เผยแพร่ และลำดับในการจัดเรียงเนื้อหาของวิทยานิพนธ์
2. ในบทที่ 2 กล่าวถึง 2 ส่วน ได้แก่ แนวคิดและทฤษฎี และงานวิจัยที่เกี่ยวข้อง โดยแนวคิดและทฤษฎีจะประกอบไปด้วย ทฤษฎีทางด้านภาษาศาสตร์ ลักษณะของพยางค์ ทฤษฎีของเสียง ตลอดจนวรรณกรรมที่เกี่ยวข้องกับการประสานเวลา การตรวจหาพยางค์ และการตรวจหาความผิดพลาดของการถอดเสียง
3. ในบทที่ 3 นำเสนอขั้นตอนวิธีในการประสานเวลาอัตโนมัติของข้อความและเสียงแบบทันที
4. ในบทที่ 4 กล่าวถึงวิธีการทดสอบและวัดประสิทธิภาพของระบบในการประสานเวลาอัตโนมัติแบบทันที ในส่วนของการตรวจหาพยางค์เท่านั้น และทั้งส่วนของการตรวจหาพยางค์รวมกับการตรวจหาความผิดพลาดของการถอดเสียงรวมกัน โดยจะทดสอบในแง่ของความถูกต้องในการประสานเวลาและเวลาในการทำงานของขั้นตอนวิธีที่นำเสนอ
5. ในบทที่ 5 เป็นการสรุปผลการวิจัย วิเคราะห์ผลลัพธ์ของการประสานเวลา รวมถึงตัวอย่างของโปรแกรมประยุกต์ที่นำขั้นตอนวิธีที่นำเสนอไปใช้

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีที่เกี่ยวข้อง ซึ่งแบ่งเป็น 3 ส่วน คือ ส่วนแรกจะกล่าวถึง ทฤษฎีทางภาษาศาสตร์ที่เกี่ยวข้อกับวิทยานิพนธ์ ส่วนที่สองจะกล่าวถึง ทฤษฎีที่เกี่ยวกับการประมวลผลสัญญาณ ซึ่งส่วนใหญ่เป็นการดึงลักษณะเด่นของสัญญาณเสียง ส่วนที่สามจะกล่าวถึงแบบจำลองฮิดเดนมาร์คอฟและการประสานเวลาของเสียง ในส่วนสุดท้ายจะกล่าวถึงวรรณกรรมที่เกี่ยวข้องกับวิทยานิพนธ์นี้

ทฤษฎีที่เกี่ยวข้อง

1. ทฤษฎีทางภาษาศาสตร์

1.1. ระบบเสียงในภาษาไทย

ระบบเสียงในภาษาไทย [7] ประกอบด้วย 3 ส่วน ได้แก่ เสียงพยัญชนะ เสียงสระ และเสียงวรรณยุกต์ โดยเสียงแต่ละประเภทประกอบด้วยหน่วยเสียงดังนี้

1.1.1. เสียงพยัญชนะ (Consonant)

เสียงพยัญชนะ เป็นเสียงที่เกิดจากลมที่ผ่านเส้นเสียง แต่จะถูกอวัยวะต่าง ๆ ภายในปากดัดแปลงลมทำให้เกิดเป็นเสียงที่แตกต่างกันไป ซึ่งพยัญชนะในภาษาไทยนั้น ประกอบด้วย 21 หน่วยเสียง ในรูปพยัญชนะ 44 รูป ด้วยกัน หน่วยเสียงพยัญชนะในภาษาไทยถูกแสดงไว้ในตารางที่ 2.1 แยกตามประเภทของเสียงและตำแหน่งที่เกิดเสียงในปาก

หน่วยเสียงพยัญชนะในภาษาไทย ยังสามารถจำแนกตามตำแหน่งที่เกิดเสียงในพยางค์ได้ดังนี้

1.1.1.1. หน่วยเสียงพยัญชนะต้น (Initial Consonant) เป็นหน่วยเสียงพยัญชนะที่เกิดต้นคำหรือต้นพยางค์ สามารถเกิดได้ทั้ง 21 หน่วยเสียง ดังตารางที่ 2.2

1.1.1.2. หน่วยเสียงพยัญชนะท้าย (Final Consonant) เป็นหน่วยเสียงพยัญชนะที่เกิดท้ายคำหรือท้ายพยางค์ มีจำนวน 8 หน่วยเสียง คือ /p, t, k, m, n, ŋ, w, j/ ดังตารางที่ 2.3

1.1.1.3. หน่วยเสียงพยัญชนะควบกล้ำ (Consonant Cluster) เป็นหน่วยเสียงพยัญชนะต้น 2 ตัวอยู่ติดกันอ่านออกเสียงคู่กัน ในภาษาไทยมีกฎเกณฑ์ในการเกิดหน่วยเสียงควบกล้ำที่ตายตัว คือ พยัญชนะตำแหน่งแรกมี 6 หน่วยเสียง คือ /p, t, k, p^h, t^h, k^h/ และพยัญชนะตำแหน่งที่สองมี 3 หน่วยเสียง คือ /r, l, w/ เท่านั้น และเมื่อรวมกันแล้วเกิดได้เพียง 12 รูปแบบเท่านั้น คือ /pr, tr, kr, p^hr, t^hr, k^hr, pl, p^hl, kl, k^hl, kw, k^hw/ ดังตารางที่ 2.4 แต่อย่างไรก็ตาม อาจเพิ่มหน่วยเสียงยืมภาษาต่างประเทศซึ่งจะเกิดในหน่วยเสียงพยัญชนะควบกล้ำเช่นกัน คือ /br, bl, fr, fl, dr/

ตารางที่ 2.1 ตารางหน่วยเสียงพยัญชนะในภาษาไทยตามประเภทเสียงและตำแหน่งที่เกิดเสียง

ประเภทของเสียง / ที่เกิดเสียง	ริมฝีปาก	ริมฝีปากล่าง-ฟันบน	ปุ่มเหงือก	เพดานแข็ง-ปุ่มเหงือก	เพดานแข็ง	เพดานอ่อน	เส้นเสียง
เสียงระเบิด	/p/ /p^h/ /b/		/t/ /t^h/ /d/			/k/ /k^h/	/?/
เสียงนาสิก	/m/		/n/			/ŋ/	
เสียงกักเสียดแทรก				/c/ /c^h/			
เสียงเสียดแทรก		/f/	/s/				/h/
เสียงข้างลิ้น			/l/				
เสียงลิ้นรัว			/r/				
เสียงเปิด	/w/				/j/		

ตารางที่ 2.2 ตารางสัญลักษณ์หน่วยเสียงพยัญชนะต้นในภาษาไทย

พยัญชนะต้นในภาษาไทย	สัทอักษรสากล	สัญลักษณ์หน่วยเสียง
ก	/k/	k
ข ฃ ค ฅ ฆ	/k^h/	kh

พยัญชนะต้นในภาษาไทย	สัทอักษรสากล	สัญลักษณ์หน่วยเสียง
ง	/ŋ/	ng
จ	/c/	c
ฉ ช ฌ	/c^{h}/	ch
ซ ศ ษ ส	/s/	s
ญ ย	/j/	j
ฎ ด	/d/	d
ฏ ต	/t/	t
ฐ ฑ ฒ ถ ท ธ	/t^{h}/	th
ณ น	/n/	n
บ	/b/	b
ป	/p/	p
พ ภ ผ	/p^{h}/	ph
ฟ ฝ	/f/	f
ม	/m/	m
ร	/r/	r
ล ฬ	/l/	l
ว	/w/	w
ห ฮ	/h/	h
อ	/?/	z

ตารางที่ 2.3 ตารางสัญลักษณ์หน่วยเสียงพยัญชนะท้ายในภาษาไทย

พยัญชนะท้ายในภาษาไทย	สัทอักษรสากล	สัญลักษณ์หน่วยเสียง
แม่กก (ก ข ค ฆ)	/k*/	k^
แม่กด (จ ช ซ ฎ ฏ ฑ ฒ ฆ ด ต ถ ท ธ ศ ษ ส)	/t*/	t^
แม่กบ (บ ป ภ พ ฟ)	/p*/	p^
แม่กน (น ณ ญ ร ล ฬ)	/n/	n^

พยัญชนะท้ายในภาษาไทย	สัทอักษรสากล	สัญลักษณ์หน่วยเสียง
แม่กง (ง)	/ŋ/	ng^
แม่กม (ม)	/m/	m^
แม่เกอย (ย)	/j/	j^
แม่เกอว (ว)	/w/	w^

ตารางที่ 2.4 ตารางสัญลักษณ์หน่วยเสียงพยัญชนะควบกล้ำ

พยัญชนะควบกล้ำในภาษาไทย	สัทอักษรสากล	สัญลักษณ์หน่วยเสียง
ปร-	/pr/	pr
ปล-	/pl/	pl
ตร-	/tr/	tr
กร-	/kr/	kr
กล-	/kl/	kl
กว-	/kw/	kw
พร-	/p^hr/	phr
พล-, ผล-	/p^hl/	phl
ทร-	/t^hr/	thr
คร-, ขร-	/k^hr/	khr
คล-, ขล-	/k^hl/	khl
คว-	/k^hw/	khw

1.1.2. เสียงสระ (Vowel)

เป็นหน่วยเสียงที่สำคัญในทุกภาษา เสียงสระนั้นเกิดจากเสียงที่เปล่งออกมาผ่านเส้นเสียงและทำให้เส้นเสียงสั่นจนเกิดเสียงก้อง และลมจะถูกเปล่งออกมาทางช่องปากหรือช่องจมูก โดยไม่มีอวัยวะมาปิดกั้นทางลมเลย ทำให้เสียงสระสามารถออกเสียงได้ยาวนาน ซึ่งเสียงสระแต่ละเสียง

เกิดขึ้นจากรูปแบบท่าทางและตำแหน่งของอวัยวะในปากมีลักษณะต่างกันไป เช่น การห่อริมฝีปาก การยกลิ้นสูงต่ำ

เสียงสระในภาษาไทยนั้น มีทั้งหมด 24 หน่วยเสียง ซึ่งสามารถแบ่งได้เป็น สระเดี่ยว สระประสม ดังตารางที่ 2.5 และนอกจากนี้ยังมี สระเกิน ซึ่ง เป็นสระที่มีเสียงซ้ำกับสระแท้แต่มีเสียงพยัญชนะท้ายผสมอยู่ด้วย

1.1.2.1. สระเดี่ยว (Monophthong) ในภาษาไทยมีทั้งหมด 18 หน่วยเสียง ซึ่งสามารถแบ่งได้เป็น สระเสียงสั้น 9 หน่วยเสียง และสระเสียงยาว 9 หน่วยเสียง

1.1.2.2. สระประสม (Diphthong) ในภาษาไทยมีทั้งหมด 6 หน่วยเสียง ซึ่งสามารถแบ่งได้เป็น สระเสียงสั้น 3 หน่วยเสียง และสระเสียงยาว 3 หน่วยเสียง

1.1.2.3. สระเกิน คือ สระที่มีเสียงซ้ำกับสระแท้แต่มีพยัญชนะท้ายผสมอยู่ด้วย ซึ่งมีอยู่ 8 เสียงด้วยกัน ดังตารางที่ 2.6

ตารางที่ 2.5 ตารางแสดงสัญลักษณ์ของหน่วยเสียงสระในภาษาไทย

ประเภทสระ	เสียงสั้น		เสียงยาว	
	รูปสระ	สัญลักษณ์หน่วยเสียง	รูปสระ	สัญลักษณ์หน่วยเสียง
สระเสียงเดี่ยว	–ะ	a	–า	aa
	–ิ	i	–ี	ii
	–ึ	v	–ือ	vv
	–ุ	u	–ู	uu
	เ–ะ	e	เ–	ee
	แ–ะ	x	แ–	xx
	โ–ะ	o	โ–	oo
	เ–าะ	@	–อ	@@
	เ–อะ	q	เ–อ	qq
สระประสม	–ัวะ	ua	–ัว	uua
	เ–ียะ	ia	เ–ีย	iia

ประเภทสระ	เสียงสั้น		เสียงยาว	
สระประสม	รูปสระ	สัญลักษณ์หน่วยเสียง	รูปสระ	สัญลักษณ์หน่วยเสียง
	เ–ือะ	va	เ–ือ	vva

ตารางที่ 2.6 ตารางแสดงสระเกินในภาษาไทย

เสียงสั้น	เสียงยาว
อำ (อะ+ม)	-
ไอ (อะ+ย)	-
ใอ (อะ+ย)	-
เอา (อะ+ว)	-
ฤ (ร+อึ)	ฤา (ร+อื)
ฦ (ล+อึ)	ฦา (ล+อื)

1.1.3. เสียงวรรณยุกต์ (Tone)

เสียงวรรณยุกต์เป็นเสียงทีความสำคัญอีกอย่างหนึ่ง ซึ่งบอกลักษณะเสียงสูงต่ำ ซึ่งสามารถทำให้คำภาษาไทยมีความหมายแตกต่างกันออกไป ทำให้มีจำนวนคำเพิ่มมากขึ้น ภาษาบางภาษาเท่านั้นที่มีวรรณยุกต์ ซึ่งสามารถพบได้หลายภาษา แต่จำนวนหน่วยเสียงของวรรณยุกต์นั้นอาจแตกต่างกัน

สำหรับภาษาไทยนั้นสามารถแบ่งเสียงวรรณยุกต์ออกได้เป็น 5 เสียง คือ เสียงสามัญ เสียงเอก เสียงโท เสียงตรี และเสียงจัตวา

1.2. **โครงสร้างของพยางค์ในภาษาไทย**

พยางค์ (Syllable) คือเสียงที่เปล่งออกมาครั้งหนึ่ง อาจจะมีความหมายหรือไม่มีความหมายก็ได้ โดยเสียงที่เปล่งออกมานั้นจะมีเสียงที่เด่นดัง 1 เสียง และเสียงที่อยู่ข้างเคียงอย่างน้อย 2 เสียง โดยเสียงที่ดังเด่นกว่าเสียงอื่น ๆ นั้น จะเป็นแกนกลางของพยางค์ ส่วนเสียงอื่น ๆ ที่ไม่เด่นดังก็ทำหน้าที่เป็นส่วนประกอบ หรือส่วนเสริมของพยางค์นั้น ๆ โดยปกติ หน่วยเสียงสระจะเป็น

แกนกลางของพยางค์ สำหรับภาษาไทยนั้น โครงสร้างพยางค์ (Syllable Structure) สามารถเขียนเป็นสมการได้ดังนี้ [8]

$$S = C_i(C_i)V^T(V)(C_f) \quad (2.1)$$

โดยที่ S หมายถึง พยางค์

C_i หมายถึง พยัญชนะต้น

V หมายถึง สระ

C_f หมายถึง พยัญชนะท้าย

T หมายถึง วรรณยุกต์

จากสมการดังกล่าว หน่วยเสียงทุกหน่วยเสียงสามารถเป็นพยัญชนะต้น C_i ได้ทั้งหมด 21 หน่วยเสียง แต่พยัญชนะท้าย C_f มีได้เพียง 9 หน่วยเสียงดังที่กล่าวไว้ข้างต้น และการเกิดพยัญชนะควบกล้ำ C_iC_i สามารถเกิดได้ทั้งหมด 12 คู่ในภาษาไทย หรืออาจเกิดได้มากกว่าหากเป็นคำยืมจากภาษาอื่น

2. ทฤษฎีที่เกี่ยวข้องกับการวิเคราะห์สัญญาณเสียง

2.1. การคำนวณค่าพลังงานของสัญญาณ

ค่าพลังงานของสัญญาณเสียง [9] เป็นหนึ่งในค่าที่นิยมนำวิเคราะห์สัญญาณเสียงซึ่งในการคำนวณค่าพลังงานของสัญญาณเสียงโดยทั่วไปสามารถคำนวณได้ตามสมการที่ 2.2

$$E = \sum_{n=-\infty}^{\infty} s^2[n] \quad (2.2)$$

โดยที่ $s(n)$ คือสัญญาณเสียงตำแหน่งที่ n และ E คือค่าพลังงานของเสียง แต่เนื่องจากในการวิเคราะห์สัญญาณจริงเราวิเคราะห์สัญญาณในช่วงที่มีกรอบของสัญญาณเวลาสั้น ๆ ดังนั้นในการคิดค่าพลังงานจึงต้องพิจารณาผลของกรอบของสัญญาณเวลาด้วย ดังนั้นจะสามารถคำนวณค่าพลังงานในช่วงกรอบสัญญาณเวลาได้ตามสมการที่ 2.3

$$E(m) = \sum_{n=0}^{N-1}[w(m)s(m-n)]^2 \quad (2.3)$$

โดยที่ $E(m)$ คือค่าพลังงานที่คำนวณได้ของสัญญาณเสียงที่กรอบเวลาที่ m และ $s(n)$ คือสัญญาณเสียงตำแหน่งที่ n และ $w(m)$ คือฟังก์ชันหน้าต่างที่มีขนาดความกว้าง N

2.2. **การคำนวณค่าระดับความเข้มของเสียง** (Sound Pressure Level)

เป็นการปรับหน่วยการวัดของพลังงานของเสียงเพื่อให้เป็นหน่วยที่หูของมนุษย์สามารถได้ยินได้ ซึ่งปกติหูของมนุษย์จะต้องได้ยินเสียงที่มีพลังงานมากพอ ซึ่งมีความถี่อยู่ที่ 20-20,000 เฮิรตซ์ (Hz) อ้างอิงจากการคำนวณจากโปรแกรม “Praat” [10] เราสามารถคำนวณ ระดับความเข้มของเสียงได้จากสมการที่ 2.4

$$SPL = 10log_{10}\{1/(t_2 - t_1)\int_{t1}^{t2} x^2(t)dt/\ (2 \cdot 10^{-5}\ Pa)^2\} \qquad (2.4)$$

โดยที่ $x(t)$เป็นพลังงานของเสียงในช่วงเวลา t_1 ถึง t_2 ซึ่งก็คือ ค่าพลังงานของเสียงที่คำนวณได้ในกรอบเวลา m ดังในสมการที่ 2.3 และถูกทำให้เป็นบรรทัดฐาน (Normalize) ด้วยขนาดของกรอบสัญญาณเสียง m ซึ่งมีความกว้าง N นั้น นั่นเอง ระดับความเข้มของเสียงจึงคำนวณได้ดังสมการที่ 2.5

$$SPL = 10log_{10}\{\frac{E(m)\ (2\cdot10^{-5}\ Pa)^2}{N}\} \qquad (2.5)$$

2.3. **การคำนวณความเป็นรายคาบของสัญญาณ** (Periodicity)

ความเป็นรายคาบของสัญญาณ (Periodicity) คือ ค่าที่บ่งบอกลักษณะของสัญญาณว่าสัญญาณที่เข้ามานี้มีลักษณะที่ซ้ำกันทุก ๆ คาบเวลาใด ๆ หรือไม่ ซึ่งความเป็นรายคาบดังกล่าวสามารถนำมาใช้ในการแยกแยะเสียงก้องและเสียงไม่ก้องของสัญญาณเสียงพูดได้ โดยสามารถคำนวณได้จากค่าอัตสหสัมพันธ์ (Autocorrelation) [9]

การคำนวณค่าอัตสหสัมพันธ์ของสัญญาณ $X(t) = x_1, x_2, \dots, x_t$ เมื่อทำการหน่วงเวลาไปเป็นระยะ h จะสามารถคำนวณได้ตามสมการที่ 2.6

$$r_x(h;m) = \frac{1}{N}\sum_{n=-\infty}^{\infty} X(n)w(m-n)X(n+|h|)w\ (m-n+|h|) \qquad (2.6)$$

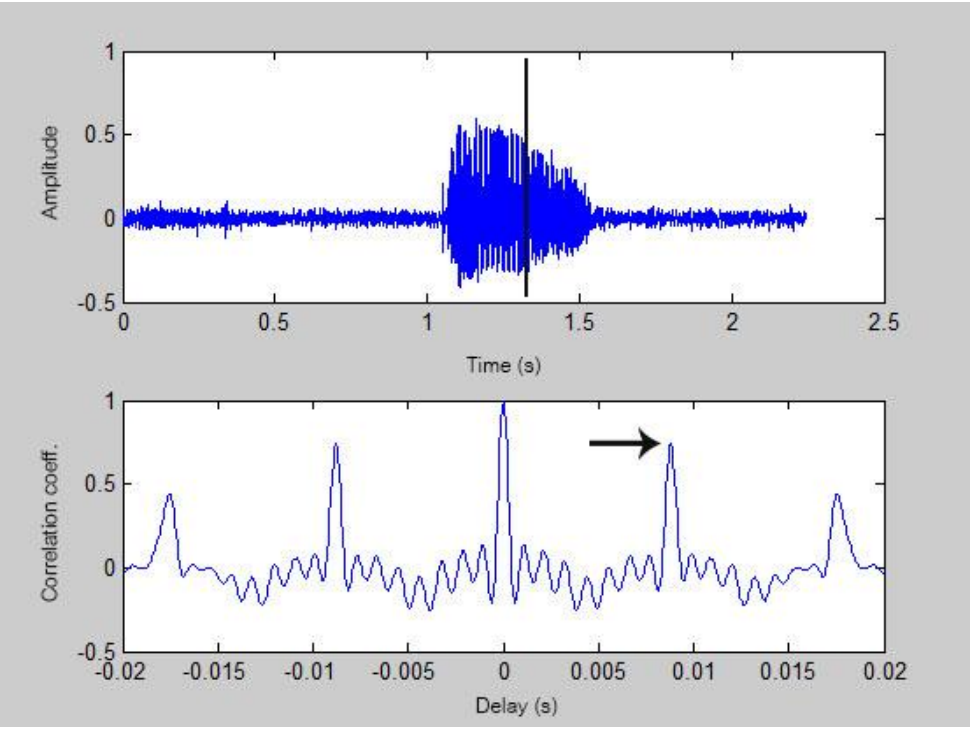
เมื่อ $r_x(h;m)$ คือ ค่าอัตสหสัมพันธ์ที่กรอบสัญญาณเสียง m เมื่อหน่วงเวลาไปเป็นระยะ h

$X(n)$ คือ สัญญาณเสียงที่ตำแหน่ง n

$w(n)$ คือ ฟังก์ชันหน้าต่าง

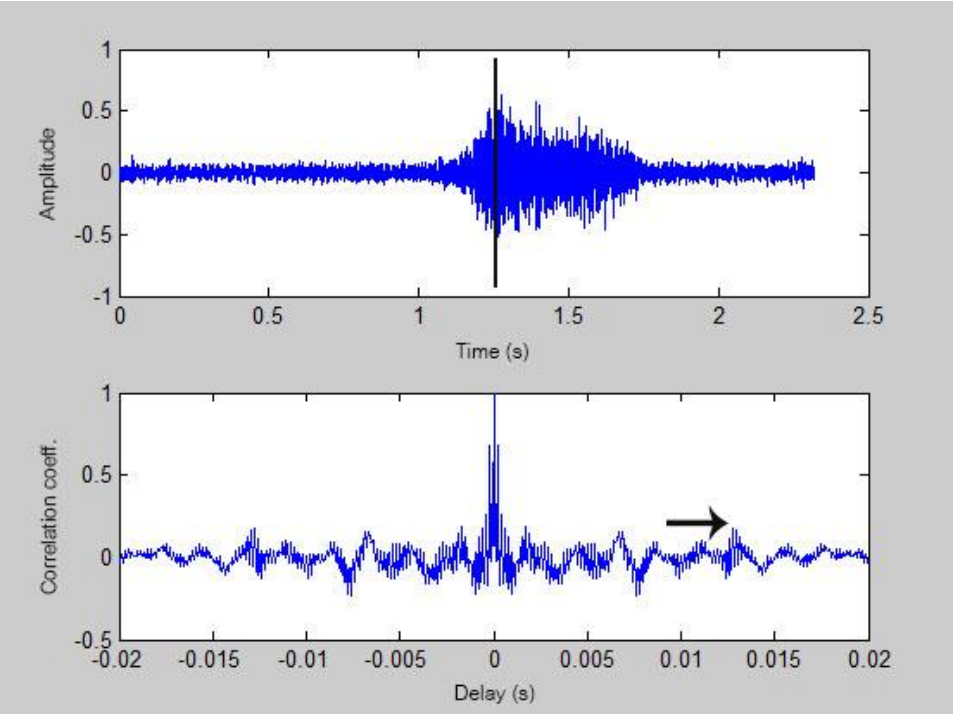
N คือ ความกว้างของกรอบสัญญาณเสียง

แต่โดยทั่วไปแล้วเมื่อค่าอัตสหสัมพันธ์ถูกนอร์มัลไลซ์ด้วยค่าสูงที่สุดซึ่งก็คือค่า ณ จุดกลาง และค่าอัตสหสัมพันธ์จะมีค่ามากที่สุดคือ 1 ต่อจากนั้น เพื่อให้ได้ค่าความเป็นรายคาบของสัญญาณ เราจะหาจุดที่มีค่าสูงที่สุดในค่าอัตสหสัมพันธ์ ในช่วง 60 ถึง 320 เฮิรตซ์ หรือที่ 3 ถึง 17 มิลลิวินาที เพื่อให้เป็นค่าความเป็นรายคาบของสัญญาณเสียงในกรอบสัญญาณหนึ่ง ๆ ดังที่ Dareeyoah [9] นำเสนอ ดังตัวอย่างในรูปที่ 2.1 ซึ่งหลังจากหาค่าอัตสหสัมพันธ์ของสัญญาณเสียงคำว่า “กา” /ga/ ซึ่งเป็นตัวอย่างของเสียงก้อง ณ เวลา 1.344 วินาทีได้แล้ว จะเห็นค่าระหว่างระยะเวลาที่หน่วงไปตั้งแต่ 3 มิลลิวินาที จนถึง 17 มิลลิวินาที นั้นค่าอัตสหสัมพันธ์ที่สูงที่สุด คือ จุดที่ระยะหน่วงประมาณ 9 มิลลิวินาที โดยมีค่าอยู่ 0.7389



รูปที่ 2.1 ค่าอัตสหสัมพันธ์ของสัญญาณเสียงคำว่า “กา” ณ เวลา 1.344 วินาที

ในขณะที่ค่าความเป็นรายคาบของเสียงไม่ก้องนั้น จะมีค่าอัตสหสัมพันธ์ที่ต่ำกว่าค่อนข้างสูง ดังรูปที่ 2.2 ซึ่งเป็นสัญญาณเสียงของหน่วยเสียง /s/ เป็นตัวแทนของเสียงไม่ก้อง ณ เวลา 1.3 วินาที เมื่อได้ค่าอัตสหสัมพันธ์แล้วพบว่า ค่าที่สูงที่สุดมีค่าเพียง 0.1717 ซึ่ง อยู่ที่ระยะหน่วงประมาณ 13 มิลลิวินาที



รูปที่ 2.2 ค่าอัตสหสัมพันธ์ของสัญญาณเสียงของหน่วยเสียง /z/ ณ เวลา 1.3 วินาที

อย่างไรก็ตามเพื่อลดความผิดพลาดที่อาจเกิดจากสัญญาณรบกวนที่มาก ดังนั้นในแต่ละกรอบสัญญาณจะทำการขลิบกลาง (Central Clipping) ก่อนที่จะนำมาหาค่าอัตสหสัมพันธ์ ซึ่งการขลิบการสามารถกระทำได้ดังสมการที่ 2.7

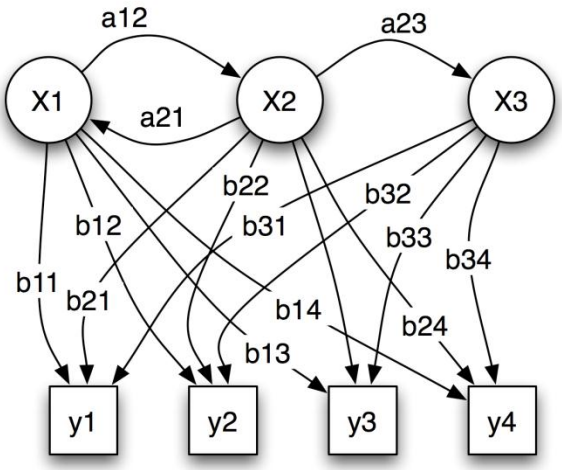
$$C\{X(n)\} = \begin{cases} X(n) - C^{+}, & X(n) > C^{+} \\ 0, & C^{-} \le X(n) \le C^{+} \\ X(n) - C^{-}, & X(n) < C^{-} \end{cases} \qquad (2.7)$$

ซึ่งโดยทั่วไปแล้วการขลิปสัญญาณ ค่า C^{-}, C^{+} จะมีค่าประมาณ 30% ของค่าสูงที่สุดของค่าสัมบูรณ์ในกรอบสัญญาณนั้น ๆ

3. **แบบจำลองฮิดเดนมาร์คอฟ** (Hidden Markov Model - HMM)

แบบจำลองฮิดเดนมาร์คอฟ เป็นวิธีในการจำแนกรูปแบบ โดยอาศัยวิธีทางสถิติ โดยการเก็บข้อมูลการกระจายของคุณลักษณะที่สำคัญของข้อมูลฝึกฝนในรูปแบบพารามิเตอร์ต่าง ๆ ของแบบจำลอง แบบจำลองนี้ถูกนำมาใช้อย่างแพร่หลาย ซึ่งส่วนใหญ่จะถูกนำมาใช้กับงานการรู้จำแบบอย่างที่ขึ้นกับทางเวลา เช่น งานทางด้านเสียงพูด, ลายมือ, การรู้จำท่าทาง เป็นต้น ยกตัวอย่างเช่น ในงานรู้จำเสียงพูด แบบจำลองฮิดเดนมาร์คอฟถูกนำมาใช้ในการสร้างแบบจำลองทางเสียง (Acoustic Model) เพื่อใช้เก็บค่าสถิติแทนแต่ละหน่วยเสียงในภาษา

แบบจำลองฮิดเดนมาร์คอฟ จะประกอบไปด้วยสถานะ (State) และการเปลี่ยนแปลงระหว่างสถานะ (Transition) ซึ่งในแต่ละการเปลี่ยนแปลงจะมีความน่าจะเป็นค่าหนึ่งกำกับไว้ทุก ๆ การเปลี่ยนแปลง และในแต่ละสถานะเองก็จะมีความน่าจะเป็นที่ค่าสังเกตแต่ละค่าอยู่ในแต่ละสถานะนั้น ๆ ดังรูปที่ 2.3

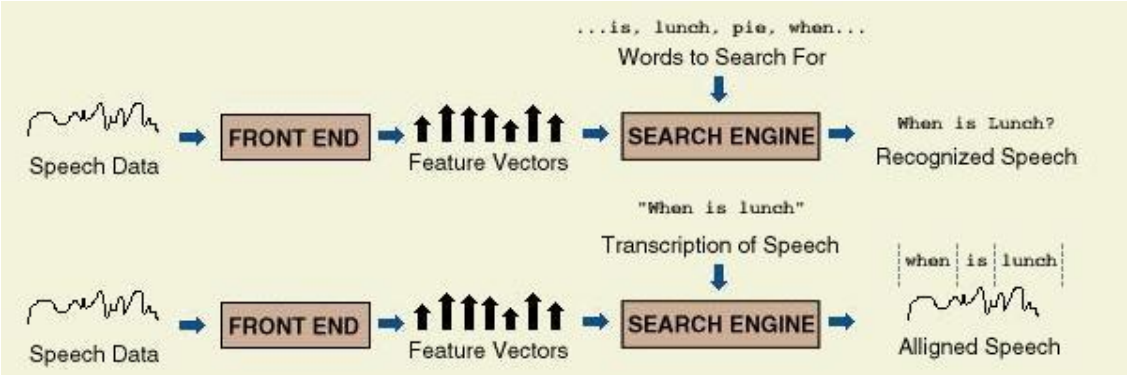


รูปที่ 2.3 แบบจำลองฮิดเดนมาร์คอฟ [11]

ค่าความน่าจะเป็นต่าง ๆ สามารถคำนวณได้จากการสอนแบบจำลองด้วยข้อมูลชุดฝึกฝน หลังจากนั้น เราสามารถจะใช้ค่าสังเกต (Observation Vector) ของข้อมูล เราจะสามารถผ่านค่าสังเกตนี้เข้าไปยังแบบจำลอง เพื่อหาค่าความน่าจะเป็นที่ค่าสังเกตดังกล่าวนั้น ว่ามีลักษณะเช่นเดียวกับข้อมูลชุดฝึกฝนหรือไม่

4. **การปรับแนวของเสียง** (Force Alignment)

การปรับแนวของเสียง เป็นวิธีการในการระบุขอบเขตของเวลาให้กับหน่วยของเสียงพูด โดยที่แต่ละหน่วยเสียงพูดนั้นจะถูกกำหนดขึ้นมา แล้วกระบวนการจะพิจารณาแต่ละหน่วยเสียงพูดที่กำหนดไว้ มาทำการค้นหาช่วงเวลาที่เหมาะสมที่สุดในเสียงพูดและความน่าจะเป็นของแต่ละหน่วยย่อยนั้น ๆ การปรับแนวของเสียงนั้น มีความคล้ายคลึงกับการรู้จำเสียงพูด ซึ่งคำนวณได้จากการวิเคราะห์ข้อมูลเสียงกับแบบจำลองเสียงของหน่วยเสียงย่อยนั้น ๆ แต่จะแตกต่างกันคือ การปรับแนวจะการหาความน่าจะเป็นที่หน่วยเสียงในคำบรรยายเสียงที่ถูกกำหนดขึ้นมาจะสอดคล้องกับข้อมูลเสียง ซึ่งจะแสดงการตัดแบ่งของเวลาในแต่ละหน่วยเสียงย่อยออกมาเป็นผลลัพธ์ตามคำบรรยายเสียงที่กำหนด ส่วนกระบวนการถอดรหัสการรู้จำเสียงพูดจะทำการหาหน่วยเสียงที่มีความน่าจะเป็นที่จะสอดคล้องกับข้อมูลเสียงมากที่สุดและกำหนดให้หน่วยเสียงนั้นเป็นผลลัพธ์ของกระบวนการถอดรหัสการรู้จำเสียงพูด



รูปที่ 2.4 ความแตกต่างของการรู้จำเสียงและการปรับแนวของเสียง [12]

งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้อง จะแบ่งเป็นงานวิจัยทางการการประสานเวลา ด้านการตรวจหาพยางค์และงานวิจัยทางด้านการตรวจหาความผิดพลาดของการถอดเสียง

1. **งานวิจัยทางด้านการประสานเวลา** (Speech-Text Alignment)

การประสานเวลาระหว่างเสียงและข้อความ (Speech-Text Alignment) นั้นส่วนใหญ่จะถูกนำไปประยุกต์กับโปรแกรมประยุกต์บางอย่างที่ต้องการแสดงเสียงและข้อความไปพร้อม ๆ กัน

ในทางเวลา เช่น งานวิจัยที่เกี่ยวกับระบบสร้างคำบรรยายบทพูด (Subtitling System) หรือ การประสานเวลาคำร้องกับเสียงร้อง (Lyrics Synchronization) เราจึงเริ่มศึกษาวิธีการในการประสานเวลาของระบบเหล่านี้เป็นหลัก

สำหรับระบบการประสานเวลาระหว่างบทพูดและคำบรรยายบทพูดนั้น ถูกพัฒนาขึ้นมาเพื่อให้ผู้พิการทางการได้ยิน (Hearing Impaired) สามารถเข้าถึงข่าวจากโทรทัศน์ได้ ในยุคที่โทรทัศน์เป็นสิ่งประดิษฐ์ที่สามารถกระจายข้อมูลข่าวสารได้รวดเร็ว การสร้างคำบรรยายบทพูดนั้นเริ่มมาจากในสมัยก่อน จะใช้คนซึ่งทำหน้าที่เป็น แค็พชันเนอร์ (Captioners) คือ เป็นผู้ถอดเสียงจากการรายงานข่าวออกมาจากการบันทึกเทปของข่าวนั้น ๆ ซึ่งการที่จะได้มาซึ่ง บทการรายงานข่าวที่ถูกต้อง แค็พชันเนอร์จะต้องเป็นผู้ที่มีความสามารถพิเศษในการพิมพ์ และรวมถึงการใช้แป้นพิมพ์พิเศษที่ถูกออกแบบขึ้นมาเพื่อใช้แค็พชันเนอร์สามารถถอดเสียงทันการรายงานข่าวนั้น ๆ ซึ่งวิธีการแบบนี้ต้องใช้กำลังทรัพย์เพื่อหาผู้ที่มีความสามารถในการเป็นแค็พชันเนอร์ได้ จำเป็นจะได้รับการฝึกฝนมาอย่างดี แต่ก็อาจเกิดความผิดพลาดขึ้นได้ง่าย เนื่องจากเป็นความผิดพลาดของมนุษย์ (Human Error) Ando, et al. [3] จึงเสนอการใช้เทคโนโลยีการรู้จำเสียงพูด (Speech Recognition) เข้ามาแทนแค็พชันเนอร์ เพื่อถอดเสียงจากการรายงานข่าวแบบทันที เพื่อลดค่าใช้จ่ายและความผิดพลาดที่กล่าวถึงออกไป เมื่อเราได้คำบรรยายบทพูดของข่าวนั้น ๆ แล้ว ขั้นตอนต่อไปคือการนำเสนอบทพูดแต่ละประโยคให้ตรงกับภาพและเสียงที่ถูกแพร่สัญญาณออกไป โดยในงานวิจัยส่วนใหญ่นั้น [5] บทพูดแต่ละประโยคจะถูกกำหนดเวลาในการแสดงว่า คำบรรยายบทพูดชิ้นนี้ จะเริ่มที่เวลาใด และจบที่เวลาใด ของการบันทึกการรายงานข่าว

มีงานวิจัยอยู่หลายงานที่เสนอวิธีการในการประสานเวลาระหว่างเสียงและข้อความซึ่งตรงกับเสียงแบบอัตโนมัติ โดยที่ไม่ต้องให้คนเป็นผู้กำหนดการประสานเวลา [13-15] เสนอการประสานเวลาระหว่างเสียงและข้อความโดยมีทั้งข้อความและเสียงครบทั้งหมด แต่บางครั้งข้อความที่ได้จากการถอดเสียง (Transcription) มักจะเป็นข้อความที่ไม่สมบูรณ์ ซึ่งเป็นผลลัพธ์ที่มักได้จากการระบบรู้จำเสียง ซึ่งสามารถใช้ได้ดีกับการประสานเวลาของระบบสร้างคำบรรยายบทพูดในข่าว เพราะบทบรรยายข่าวมักได้จากระบบรู้จำเสียงซึ่งเสียงนั้นมักมีสัญญาณรบกวนมาก ทำให้การถอดเสียงนั้นไม่สมบูรณ์ ในขณะเดียวกันงานประเภทประสานเวลาระหว่างคำร้องและเนื้อร้องก็เช่นกัน [16] แต่

จะใช้หลักการที่เกี่ยวกับดนตรีเข้าช่วย เช่นการตรวจหาท่อนทำนอง หรือการตรวจหาจังหวะเป็นต้น แต่อย่างไรก็ตาม การประสานเวลาประเภทนี้ ยังจำเป็นจะต้องรู้ข้อมูลของเสียงและข้อความทั้งหมด และอาศัยข้อมูลนี้มาประมวลผลในการประสานเวลา แต่ในงานวิจัยนี้ ข้อมูลของเสียง เป็นสิ่งที่ไม่ทราบมาก่อน จนกระทั่งพูดออกมา แล้วจึงประมวลผล

Gao, et al. [17] เสนอวิธีในการ Synchronization แบบ Real-time ในลักษณะที่เป็นการพูดสด โดยการตรวจสอบจะตรวจว่าบทบรรยายของประโยคที่กำลังถูกพูดออกมาอยู่นั้น ถึงจุดจบประโยคหรือไม่ (End Time Detection) หากตรวจสอบได้ว่าพูดจนถึงจุดจบประโยคแล้ว ก็นำคำบรรยายบทพูดประโยคใหม่ขึ้นมาแสดง แล้วตรวจสอบเช่นนี้ไปเรื่อย ๆ จนจบการรายงาน โดยการใช้ค่า Frame-Synchronous Likelihood Ratio Test เพื่อตรวจสอบสมมติฐาน 2 สมมติฐาน ได้แก่ H_0 คือสมมติฐานที่เวลา t เป็นจุดจบของของคำบรรยายบทพูดแล้ว และ H_1 คือสมมติฐานที่เวลา t ไม่ใช่จุดจบของคำบรรยายบทพูดนั้น ซึ่งได้ผลลัพธ์ประมาณ 85.6% เมื่อกำหนดให้ยอมรับความคลาดเคลื่อนของจุดจบของเวลาได้ 0.5 วินาที แต่อย่างไรก็ตาม ในงานวิจัยชิ้นนี้ จะทำการประสานเวลาในระดับที่เล็กลงไปกว่าระดับประโยคดังที่ Gao นำเสนอ ดังนั้นหลักการในการหาจุดจบและทดสอบสมมติฐานจึงถูกนำมาประยุกต์ใช้กับงานวิจัยชิ้นนี้ในระดับพยางค์ โดยการตรวจหาพยางค์ออกมาจากเสียงพูดให้ได้

2. **งานวิจัยทางด้านการตรวจหาพยางค์** (Syllable Detection)

ในการตรวจหาพยางค์จากสัญญาณเสียงนั้น ลักษณะเด่นทางเสียงที่เป็นที่นิยม คือ ค่าพลังงานของเสียง [18-20] โดยงานวิจัยที่เป็นต้นแบบและนิยมใช้จนถึงปัจจุบันในการระบุและตัดแบ่งพยางค์ก็คือ ขั้นตอนวิธีคอนเว็กซ์ฮัลล์ (Convex Hull Algorithm) ของ Mermelstein [18] โดย Mermelstein ระบุว่า พยางค์พยางค์ในทางภาษาศาสตร์คือส่วนของเสียงพูดที่มีพลังประจำเสียงสูงที่สุด ซึ่งอยู่ระหว่างส่วนที่มีประพลังประจำเสียงที่ต่ำที่สุด ซึ่งพลังประจำเสียงไม่สามารถคำนวณได้ Mermelstein จึงตีความออกมาว่าเป็น Power Spectrum Intensity ซึ่งก็คือค่าความเข้มของพลังงานของเสียงนั่นเอง โดยที่จุดสูงสุดของความเข้มนั้นก็คือแกนกลางของพยางค์ และจุดที่ค่าต่ำที่สุดระหว่างแกนกลางนี้ก็คือ จุดที่มีโอกาสเป็นขอบเขตของพยางค์ โดยที่ Mermelstein เลือก

ความถี่ในช่วง 500 ถึง 4,000 เฮิรตซ์ ของค่าความเข้มของพลังงานของเสียงมาคำนวณตามขั้นตอนวิธีคอนเว็กซ์ฮัลล์ จนได้จุดแบ่งของพยางค์ทั้งหมดในเสียงนั้น

อย่างไรก็ตามวิธีการของ Mermelstein นั้นเลือกความถี่ในช่วงที่ไม่เหมาะสม ซึ่งอาจทำให้เกิดความผิดพลาดค่อยข้างสูงจากเสียงอื่น ๆ ได้ที่ไม่ใช่แกนกลางของพยางค์ อย่างเช่น เสียงพยัญชนะอื่น ๆ ที่มีเสียงก้อง เสียงสัญญาณรบกวน เป็นต้น ต่อมา Pfitzinger, et al. [19] ได้เสนอการเลือกความถี่ที่แคบลง ก่อนที่จะดึงพลังงาน ซึ่งเป็นเป็นลักษณะเด่นในการตรวจหาพยางค์ เพื่อความโดยที่เสียงจะถูกนำมาผ่านตัวกรอกความถี่ก่อน 2 ช่วง คือ ช่วง 250 ถึง 2500 เฮิรตซ์ เพราะเป็นช่วงที่เป็นความถี่ที่มีความสัมพันธ์ร่วมกันของพยางค์อยู่ และความถี่ต่ำในช่วง 7 ถึง 13 เฮิรตซ์ ซึ่งเป็นช่วงของความถี่ F0 จากการทดลอง ได้ค่าความผิดพลาดประมาณ 12.87% สำหรับเสียงพูดที่เป็นลักษณะการอ่าน และ 21.03% สำหรับเสียงพูดปกติ Pfitzinger ศึกษาค่าพารามิเตอร์ต่าง ๆ เพื่อใช้ในปรับค่าขีดแบ่งในการตัดสินใจหาแกนกลางพยางค์ เช่นเดียวกันกับ Juneja, et al. [20] ก็เสนอการแยกคุณสมบัติของลักษณะเด่นของเสียงต่าง ๆ ซึ่งหนึ่งในลักษณะเด่นของพยางค์นั้น ก็คือพลังงานในช่วงของความถี่ 640 ถึง 2,800 เฮิรตซ์ และ พลังงานในช่วงความถี่ 2,000 ถึง 3,000 เฮิรตซ์

อีกสิ่งหนึ่งที่ Pfitzinger สรุปในงานวิจัย คือ ไม่มีทางเป็นไปได้ที่จะตรวจหาพยางค์ได้ 100% เนื่องจากคำตอบแกนกลางพยางค์นั้น ให้นักสัทศาสตร์เป็นผู้กำหนดขึ้น ซึ่งมีความกำกวมอยู่ เพราะขึ้นอยู่กับการนิยามพยางค์ในแง่ภาษาศาสตร์ว่า พยางค์คืออะไร เพราะตำแหน่งของแกนกลางพยางค์นั้น อาจจะปรากฏอยู่ในเสียงสระก็ได้ หรือเสียงพยัญชนะก็ได้ ยกตัวอย่างเช่น Please ซึ่งเป็นคำพยางค์เดียว แต่ออกเสียงเหมือนกับคำว่า Police ซึ่งเป็นคำที่มี 2 พยางค์ หรืออาจจะเกิดกับพยางค์ที่เป็นเสียงไม่เน้นพยางค์ (Unstressed Syllable) เช่น คำว่า Support เป็นคำที่มี 2 พยางค์ แต่ออกเสียงเหมือนกับ Sport ซึ่งเป็นคำพยางค์เดียว ดังนั้นคำตอบจึงขึ้นอยู่กับนักสัทศาสตร์ว่าจะให้มีกี่พยางค์ แต่ในภาษาไทยนั้นคำไทยส่วนใหญ่จะเป็นคำพยางค์เดียว และ โครงสร้างของพยางค์ต้องประกอบด้วยสระเป็นสำคัญ [21] จึงทำให้คำไม่มีเสียงเน้นพยางค์หรือไม่เน้นพยางค์ ดังนั้นสมมติฐานในงานวิจัยนี้คือ เสียงพยางค์ในภาษาไทยจะต้องออกเสียงสระชัดเจน ในงานวิจัยนี้ จึงตั้งสมมติฐานขึ้นมาว่า แกนกลางของพยางค์นั้น จะต้องเป็นเสียงสระภาษาไทย ดังนั้นก็จะสามารถลดความกำกวมโดยการกำหนดว่าแกนกลางของพยางค์นั้นคืออะไรได้

งานวิจัยในการตรวจหาพยางค์ในวิทยานิพนธ์นี้จึงสนใจงานทางด้านการตรวจหาสระต่อไป เป็นหลัก ซึ่งการตรวจหาตรวจสระนั้น เนื่องจากสระเสมือนเป็นแกนกลางของพยางค์ งานวิจัยก็ยังพยายามใช้พลังงานของเสียงเป็นลักษณะเด่นอยู่เช่นกัน เพียงแต่จะใช้หลักของการตรวจหาเสียงก้องและเสียงไม่ก้อง (Voicing Detection) มาช่วยในการพิจารณาเสียงสระ ป้องกันความผิดพลาดจากเสียงเสียดแทรก และเสียงรบกวน ดังเช่นในงานวิจัยของ Pfau,et al. [22] จึงเสนออัตราการตัดศูนย์ (Zero Crossing Rate) เป็นตัวตรวจสอบอีกอย่างหนึ่งเพื่อตัดเสียงเสียดแทรกออกไป นอกจากค่าความดังดัดแปรของเสียง (Modified Loudness) ซึ่งมีลักษณะคล้าย ๆ กับพลังงานของเสียง ซึ่งได้ผลลัพธ์ความถูกต้องประมาณ 77%

Xie, et al. [23] ยังเสนออีกลักษณะเด่นหนึ่งคือ ภาวะความเป็นรายคาบ (Periodicity) ของแต่ละกรอกสัญญาณ มาใช้ตรวจสอบความก้องไม่ก้องของเสียง ประกอบกับการขั้นตอนวิธีดังเดิม เช่น คอนเว็กซ์ฮัลล์ของ Mermelstein กับความเข้มของเสียง ซึ่งก็ผลลัพธ์ที่ได้ก็มีความถูกต้องมากขึ้น ซึ่งก็ตรงกับงานวิจัยของ Dareeyoah [9] ซึ่งเสนอการใช้ค่าอัตสหสัมพันธ์ (Autocorrelation) ของเสียงมาวัดความเป็นรายคาบของสัญญาณเช่นกัน ประกอบกับการใช้ขั้นตอนวิธีคอนเว็กซ์ฮัลล์กับพลังงานในช่วงความถี่มากกว่า 300 เฮิรตซ์ ซึ่งในงานวิจัยนี้ ได้ความถูกต้องประมาณ 84% ซึ่งทดสอบในฐานข้อมูลเสียงภาษาไทยโลตัส [24] ซึ่งเปรียบเทียบกับงานวิจัยของ Howitt [25] ซึ่งใช้ลักษณะเด่นคือ พลังงานในช่วงความถี่ 300 ถึง 900 เฮิรตซ์ เพียงอย่างเดียว แต่ได้ความถูกต้องประมาณ 75% เมื่อเปรียบเทียบในฐานข้อมูลเสียงภาษาไทยโลตัส

Rochkittichareon, W., et al. [26] เสนอลักษณะเด่นทางเสียงต่าง ๆ ซึ่งใช้แยกสมบัติทางสวนสัทศาสตร์ในภาษาไทย ซึ่งแบบเป็น [speech], [sonorant], [syllabic] และ [continuant] ซึ่ง [syllabic] คุณสมบัติทางเสียงที่ทำหน้าที่เป็นแกนกลางของพยางค์ ซึ่งเป็นลักษณะของเสียงสระในภาษาไทยนั้น ใช้ลักษณะเด่นหลายลักษณะด้วยกันในการแยก [syllabic] ออกมา ซึ่งได้แก่ค่าพลังงานของเสียงในช่วงความถี่ 640 ถึง 2,800 เฮิรตซ์ พลังงานของเสียงในช่วงความถี่ 2,000 ถึง 3,000 เฮิรตซ์ ค่าเข้มสูงสุดในช่วงความถี่ 0 ถึง 900 เฮิรตซ์ และ ค่าอัตราส่วนของพลังงานที่ช่วงความถี่ 0 ถึง 400 เฮิรตซ์ และช่วงความถี่ 400 ถึง 6,000 เฮิรตซ์ มาประกอบกัน ซึ่งได้ผลลัพธ์ในการแยก [syllabic] 84.47% โดยทดสอบบนฐานข้อมูลเสียงภาษาไทยโลตัสเช่นกัน

ดังนั้นในวิทยานิพนธ์นี้ จะนำหลักการตรวจหาพยางค์โดยการใช้ลักษณะเด่นทางเสียงซึ่งจะเป็นค่าระดับความเข้มของเสียงเป็นหลัก เพราะเป็นลักษณะเด่นที่นิยมใช้ในการตรวจหาแกนกลางพยางค์ และพิจารณาลักษณะเด่นอื่น ๆ ประกอบกันตามความต้องการของระบบที่มีลักษณะการคำนวณแบบทันที (Real-time) เพิ่มด้วย

3. **งานวิจัยทางด้านการตรวจหาความผิดพลาดของการถอดเสียง** (Transcription Errors Detection)

เนื่องจากงานวิจัยในการตรวจหาพยางค์ในทางเสียงแล้ว ส่วนใหญ่จะมีความถูกต้องเฉลี่ยประมาณ 70 ถึง 80% เนื่องจากเป็นการประสานเวลาแบบทันที การตัดสินใจตำแหน่งของการประสานเวลาหลังจากที่ผู้พูดพูดออกไปในระดับพยางค์ ซึ่งมีความละเอียดของการประสานเวลาค่อยข้างมาก ถ้าหากการประสานเวลาในแต่ละครั้งเกิดความผิดพลาดขึ้น ความผิดพลาดนั้นจะสะสมไปเรื่อย ๆ ถ้าหากไม่เกิดความผิดพลาดในทิศทางตรงกันข้ามขึ้นมา แต่เนื่องจาก เราทราบข้อมูลนำเข้าระบบในส่วนของข้อความที่ผู้พูดจะพูดตั้งแต่ต้นจนจบ จึงเกิดสมมติฐานในการวิจัยที่ว่า ถ้าหากสามารถตรวจสอบได้ว่า เสียงที่พูดออกมานั้น ตรงกับข้อความที่แสดงอยู่หรือไม่ และสามารถปรับให้ตำแหน่งที่การตรวจหาพยางค์นั้น ประสานเวลาผิดพลาดไป กลับมาอยู่ในตำแหน่งที่เสียงที่เพิ่งพูดไปขณะนั้น ก็จะสามารถลดความผิดพลาดซึ่งเกิดจากการตรวจหาพยางค์เพียงอย่างเดียวได้

งานวิจัยทางด้านการตรวจหาความผิดพลาดของการถอดเสียงนั้น ส่วนใหญ่มักจะถูกนำไปใช้ประโยชน์ในการตรวจสอบความถูกต้องของป้ายกำกับการถอดเสียง (Transcription Label) ในการสร้างฐานข้อมูลเสียง (Corpus) [27] เพื่อใช้พัฒนาให้ฐานข้อมูลเสียงมีความถูกต้องทำให้ระบบรู้จำเสียงมีความถูกต้องมากขึ้น หรือใช้ในการตรวจสอบผลลัพธ์ของการถอดรหัสการรู้จำเสียงพูดอัตโนมัติอีกครั้ง ว่าการถอดรหัสนั้นมีความผิดพลาดหรือไม่ [28,29,30] งานประเภทนี้มักใช้ค่าความน่าเชื่อถือของผลลัพธ์ (Confidence Measure) ในการตรวจสอบความน่าเชื่อถือของผลลัพธ์ จากงานวิจัยทบทวนวรรณกรรมของ Jiang [31] แบ่งเทคนิคการใช้งานเป็น 3 แนวทางด้วยกันคือ

1. ใช้เป็นค่าหนึ่งของลักษณะสำคัญในการทำนาย (CM as Combination of Predictor Features)

คือการใช้ค่า CM ที่ได้เป็นค่าสำคัญในการตัดสินใจผลลัพธ์เลย ซึ่งจะสามารถแยกแยะความแตกต่างระหว่างผลลัพธ์ที่ถูกต้องและผลลัพธ์ที่ไม่ถูกต้องได้

2. ใช้เป็นค่าความน่าจะเป็นของผลลัพธ์หลังจากที่ได้ค่าความน่าจะเป็นแล้ว (CM as Posterior Probability)

โดยปกติแล้ว จะใช้ในการตรวจสอบความน่าเชื่อถือของผลลัพธ์จากการรู้จำเสียงว่ามีความน่าเชื่อถือหรือไม่ ซึ่งโดยทั่วไปการถอดรหัสการรู้จำเสียงโดยใช้แบบจำลองฮิดเดนมาร์คอฟนั้น จะเลือกสมมติฐานที่ดีที่สุดที่ได้จากการคิดค่าน้ำหนักของคำตอบของโครงข่ายที่เป็นไปได้ของคำทั้งหมด แล้วนำมาเรียงลำดับค่าที่มีค่าที่ดีที่สุด N แบบ (N-Best) มาเป็นคำตอบของการรู้จำเสียง ซึ่งจริง ๆ แล้วการถอดรหัสนั้นอาจจะมีความผิดพลาดได้ ดังนั้น เราอาจคำนวณโดยใช้ลักษณะเด่นบางอย่างเข้ามาตรวจสอบสมมติฐานที่เหลือที่การถอดรหัสไม่ได้เลือกเพิ่มเติม แล้วคิดค่าความน่าจะเป็นของผลลัพธ์ใหม่ ค่าที่ได้หลังจากการคำนวณใหม่นี้จึงเรียกว่า Posterior Probability และนำมาเปรียบเทียบ หรือ เรียงลำดับใหม่ก็ตาม

3. ใช้ในการทวนสอบการพูด (CM as utterance verification)

ในการใช้งานแบบนี้จะนำค่าความเชื่อถือของผลลัพธ์ มาใช้เป็นตัววัดสมมติฐาน ว่าควรจะเป็นสมมติฐานใดโดยจะตั้งสมมติฐานขึ้นมา 2 สมมติฐานคือ สมมติฐาน H_0 หมายถึง สมมติฐานที่การถอดรหัสของการรู้จำเสียงนั้นถูกต้องและสมมติฐาน H_1 หมายถึง สมมติฐานที่การถอดรหัสของการรู้จำเสียงนั้นไม่ถูกต้อง จากนั้นจึงใช้ค่าที่ได้จาก 2 สมมติฐานนี้ มาให้คะแนนเปรียบเทียบกันว่าควรจะเป็นสมมติฐานใด

ในงานวิทยานิพนธ์นี้ เนื่องจากเราทราบการถอดเสียงของข้อความที่ผู้พูดจะต้องพูดอยู่แล้ว ดังนั้น เพื่อที่จะตรวจสอบว่าผู้พูดพูดถูกต้องหรือไม่นั้น จะประยุกต์การใช้ค่าความเชื่อถือของผลลัพธ์ในเทคนิคที่ 2 มาประยุกต์ใช้ในวิทยานิพนธ์นี้

บทที่ 3

ขั้นตอนวิธีการประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความ

ในบทนี้ จะกล่าวถึงขั้นตอนวิธีการในการประสานเวลาอัตโนมัติระหว่างเสียงและข้อความแบบทันที ซึ่งข้อมูลนำเข้านั้นก็จะต้องประกอบด้วยเสียงและข้อความ รูปแบบของเสียงและข้อความที่เข้ามาในวิทยานิพนธ์นี้ เราจะทราบเพียงแต่ข้อความทั้งหมดที่จะถูกประสานเวลา ในขณะที่ข้อมูลนำเข้าในส่วนของเสียงนั้น เราจะยังไม่ทราบเลย จนกระทั่งผู้พูดพูดข้อความนั้นออกมา ขั้นตอนวิธีนี้จึงจะเริ่มทำงานและทำการประสานเวลาในทันทีที่ผู้พูดพูดออกมา โดยหน่วยย่อยในทางภาษาศาสตร์ของการประสานเวลาที่นำเสนอก็คือ หน่วยพยางค์ โดยทุก ๆ พยางค์ที่ผู้พูดพูดออกมาจะถูกนำไปปรับแนวกับแต่ละพยางค์ในข้อความ จึงมีลักษณะการทำงานเป็นแบบทันที (Real-time)

จากรายละเอียดที่กล่าวมาข้างต้น เราจึงจำเป็นต้องนิยามพยางค์ให้ชัดเจนก่อน เพื่อที่จะนำไปใช้กับขั้นตอนวิธีที่นำเสนอได้

การนิยามของตำแหน่งพยางค์

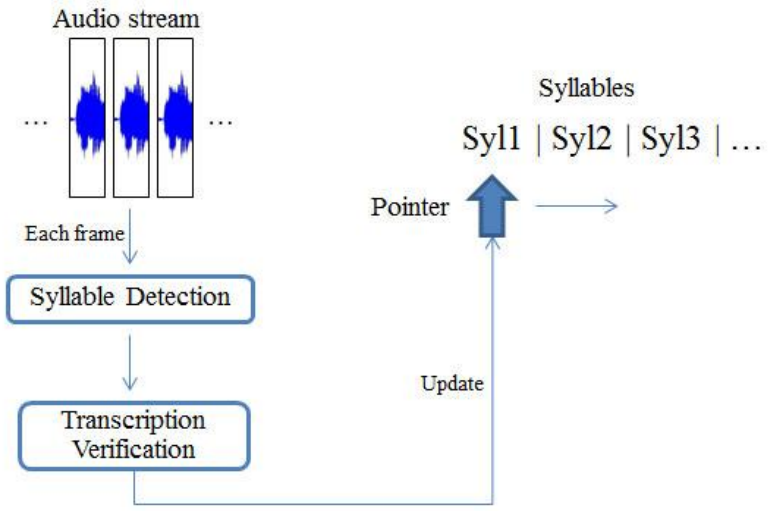
ดังที่กล่าวไว้ในบทที่ 2 ในทางภาษาศาสตร์ พยางค์ หมายถึง เสียงที่ถูกเปล่งออกมาหนึ่งครั้ง จะมีหรือไม่มีความหมายก็ได้ ซึ่งในภาษาไทยเองก็มีโครงสร้างของพยางค์อย่างชัดเจน และโครงสร้างหลักของพยางค์ในภาษาไทยก็คือสระ ดังนั้นการตัดสินใจตำแหน่งของพยางค์นั้น จะยึดช่วงของเสียงสระในพยางค์นั้น ๆ เป็นเกณฑ์

สำหรับข้อมูลนำเข้าที่เป็นเสียงพูดนั้น Mermelstein [18] ได้กล่าวถึงพยางค์เอาไว้ว่า เป็นชุดของเสียงพูดที่มีจุดสูงสุดของพลังประจำเสียง ซึ่งอยู่ระหว่างจุดต่ำที่สุดของพลังประจำเสียงนั้น จุดที่สูงที่สุดในชุดเสียงก็คือ แกนกลางของพยางค์ (Syllable Nucleus) และจุดที่ต่ำที่สุด 2 จุด ระหว่างจุดที่สูงที่สุด ก็คือ ขอบเขตของพยางค์ (Syllable Boundary) ดังนั้น ถ้าหากเราสามารถระบุได้ว่า จุดไหนเป็นที่จุดสูงและจุดที่ต่ำที่สุด ก็จะสามารถระบุพยางค์ออกมาจากเสียงได้เช่นกัน

ภาพรวมของขั้นตอนวิธี

ขั้นตอนวิธีในการประสานเวลาแบบทันทีตามลักษณะที่กล่าวเอาไว้ข้างต้นนั้น ออกมาให้ใช้กับหน่วยย่อยในระดับพยางค์ แนวคิดหลัก ก็คือ จะทำอย่างไรจึงจะระบุพยางค์ที่กำลังพูดออกให้ตรงกับพยางค์ในข้อความที่มีอยู่แล้ว เนื่องจากเราทราบข้อความทั้งหมดที่ผู้พูดจะต้องพูดอยู่แล้วจากนิยามตำแหน่งของพยางค์ การถอดเสียงจากข้อความทำให้เราได้พยางค์จากข้อความทั้งหมดมาเรียบร้อยก่อนที่ผู้พูดจะเริ่มพูดแล้ว

ความท้าทายจะอยู่ที่ในส่วนของการระบุพยางค์จากเสียงที่พูดออกมาในขณะนั้น หากเราทราบว่าผู้พูด พูดจบหนึ่งพยางค์แล้ว เราก็จะสามารถจับคู่พยางค์ที่ได้จากเสียงพูดกับพยางค์ในข้อความที่เกิดจากการถอดเสียงได้ไปเรื่อย ๆ ทีละพยางค์ตามเสียงที่กำลังพูดเข้ามา ขั้นตอนวิธีที่นำเสนอจะถูกแบ่งเป็น 2 ขั้นตอนใหญ่ ๆ ซึ่ง ได้แก่ ขั้นตอนการประสานเวลาโดยการตรวจหาพยางค์และขั้นตอนการตรวจหาและแก้ไขความไม่ตรงกันของการถอดเสียง ดังภาพที่ 3.1



รูปที่ 3.1 แนวคิดของขั้นตอนวิธีที่นำเสนอ

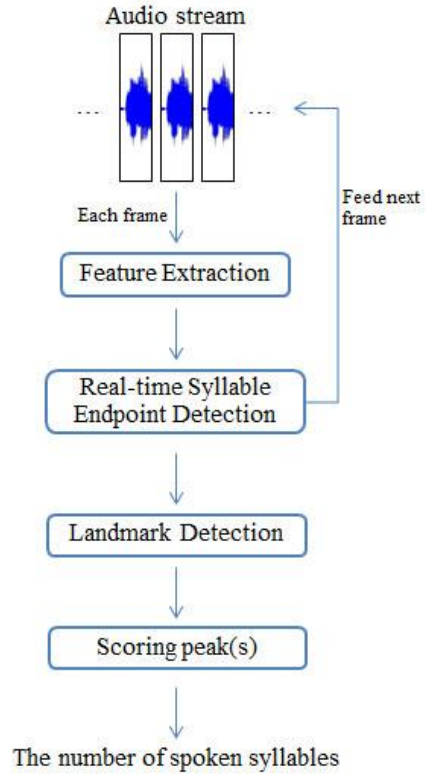
จากภาพ กระบวนการประสานเวลาจะเริ่มต้นเมื่อ ผู้พูดเริ่มพูด ทุก ๆ เฟรมของเสียงพูดที่รับเข้ามาผ่านทางไมโครโฟน ในส่วนของข้อความนั้น จะมีตัวชี้ตัวหนึ่งซึ่งเมื่อเริ่มต้น จะชี้อยู่ที่ตัวแรกคอยบอกว่า ในขณะนี้พูดถึงพยางค์ใดในข้อความ ในขั้นตอนแรกของการประสานเวลาผลลัพธ์จะสามารถบอกได้ว่า ผู้พูดพูดไปแล้วกี่พยางค์จากเสียงที่พูดออกมา ตัวชี้พยางค์ในข้อความก็จะถูก

เลื่อนไปตามจำนวนพยางค์ที่ตรวจหาได้จากเสียงที่พูดออกไปเรื่อย ๆ โดยกระบวนการนี้เสียงที่พูดเข้ามาจะถูกดึงลักษณะเด่นของเสียงออกมาทีละเฟรม จากนั้นจึงเริ่มกระบวนการตรวจหาพยางค์ ซึ่งจะแบ่งเป็นขั้นตอนย่อยอีก 2 ส่วนคือ ขั้นตอนการตรวจหาจุดจบของพยางค์แบบทันที และขั้นตอนการตรวจหาแกนกลางของพยางค์ ในส่วนนี้จะได้คำตอบออกมาเป็นจำนวนพยางค์ ว่าผู้พูดได้พูดไปแล้วกี่พยางค์ เพื่อทำการเลื่อนตัวชี้ดังที่กล่าวไป แต่อย่างไรก็ตาม กระบวนการแรกนั้น มีโอกาสผิดพลาดขึ้นได้ ซึ่งถ้าหากเกิดความผิดพลาดขึ้นมา จะเกิดความผิดพลาดสะสมไปเรื่อย ๆ ดังรูปที่ 3.1 ดังนั้น จึงต้องมีกระบวนการที่สอง ซึ่งเป็นกระบวรการตรวจหาความผิดพลาดที่อาจจะเกิดขึ้นได้จากการประสานเวลาในส่วนแรกและแก้ไขให้ถูกต้อง โดยการตั้งสมมติฐานความผิดพลาดขึ้นมา แล้วนำไปคำนวณเพื่อหาคำตอบสมมติฐานที่ดีที่สุดมาเป็นคำตอบสุดท้าย แล้วจึงค่อยเลื่อนตัวชี้ในข้อความไปยังพยางค์ที่สมมติฐานที่ถูกต้อง

วิธีการประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความ

1. **การตรวจหาของพยางค์**

เมื่อผู้พูดเริ่มพูดตามข้อความที่กำหนดไว้ สัญญาณเสียงก็จะเริ่มเข้ามาในระบบเช่นกัน การตรวจหาพยางค์จะเริ่มทันทีที่มีสัญญาณเสียงพูดเข้ามาในระบบ โดยจะแบ่งขั้นตอนใหญ่ ๆ ออกเป็น 3 ขั้นตอนด้วยกัน ซึ่งประกอบด้วย การตรวจหาจุดจบของพยางค์แบบทันที เพื่อหาจุดจบของพยางค์ การตรวจหาแกนกลางของพยางค์เพื่อหาตำแหน่งของแกนกลางพยางค์ที่ได้พูดออกไปแล้วจนถึงจุดจบนั้น และเมื่อได้ตำแหน่งของแกนกลางจึงเข้าส่วนของการให้คะแนนแกนกลางที่ตรวจหาได้ ว่าเป็นแกนกลางพยางค์จริงหรือไม่ ดังรูปที่ 3.2



รูปที่ 3.2 ขั้นตอนวิธีการในส่วนของการตรวจหาพยางค์

1.1. **การตรวจหาจุดจบของพยางค์แบบทันที** (Real-time Syllable endpoint detection)

ในวิทยานิพนธ์นี้ เสียงที่เป็นข้อมูลนำเข้านี้ เป็นสิ่งที่ไม่ทราบลักษณะมาก่อน เพื่อที่จะทำให้มีลักษณะเป็นการคำนวณแบบทันที จึงเริ่มจากการตรวจหาจุดจบของพยางค์ เพื่อให้เป็นเหตุการณ์ที่จะนำไปสู่การคำนวณขั้นต่อไป ขั้นตอนของการตรวจหาจุดจบของพยางค์แบบทันที มีดังนี้

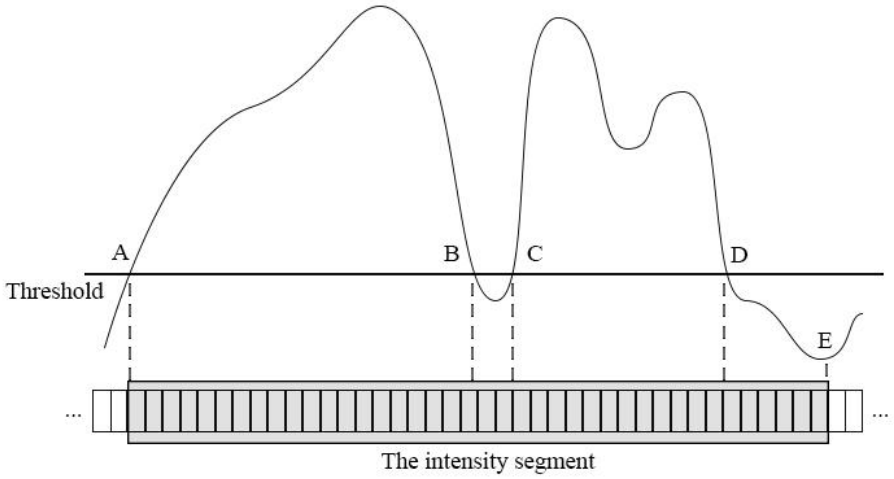
1.1.1. ติดตามค่าระดับความเข้มของเสียงเสียงทีละกรอบสัญญาณ (Window) ของเสียง ซึ่งคำนวณจากชอร์ตไทม์เอเนอร์ยี (Short-time energy) โดยมีพารามิเตอร์คือ ขนาดของกรอบสัญญาณ (Window Size) มีค่า 25 มิลลิวินาที และ ระยะห่างของแต่ละเฟรม (Interval) มีขนาด 10 มิลลิวินาที

1.1.2. กำหนดค่าขีดแบ่งไว้ค่าหนึ่ง เพื่อแยกระหว่างส่วนที่เป็นเสียงเงียบและส่วนที่เป็นพลังประจำเสียง เมื่อไรก็ตามที่ความดังของเสียง มีค่าเกินค่าขีดแบ่งที่กำหนดไว้

แล้ว จะถือว่าอยู่ในเป็นสถานะที่เป็นพลังประจำเสียง ซึ่งสามารถอนุมานได้ว่า นี่คือจุดเริ่มต้นของขอบเขตของพยางค์แล้ว

1.1.3. เมื่อใดก็ตามที่ระดับความเข้มเสียงมีค่าต่ำกว่าค่าขีดแบ่งอีกครั้งหนึ่ง จะถือว่าอยู่ในสถานะเสียงเงียบ ซึ่งก็สามารถอนุมานได้ว่า เฟรมที่ระดับความเข้มต่ำกว่าค่าขีดแบ่งนี้เป็นจุดสิ้นสุดของขอบเขตของพยางค์ที่เพิ่งอ่านไปนั่นเอง

1.1.4. แต่อย่างไรก็ตามในบางครั้ง เมื่อติดตามระดับความเข้มเสียงในแต่ละเฟรมไปเรื่อย ๆ แล้วระดับความเข้มเสียงนั้น มีค่าน้อยกว่าค่าขีดแบ่งอยู่เพียง 2-3 เฟรม (ซึ่งประมาณเป็นเวลา 35-45 มิลลิวินาที) แล้วระดับความเข้มเสียงก็กลับไปมีค่าเกินค่าขีดแบ่งอีกครั้ง ซึ่งเหตุการณ์เช่นนี้ จากการทดลองเบื้องต้น พบว่าอาจจะเกิดจากพลังงานภายในของพยางค์นั้น มีค่าลดลงจนต่ำกว่าค่าขีดแบ่งที่ตั้งไว้ ทำให้เกิดการเข้าใจผิดว่าจุดนั้นเป็นจุดจบของพยางค์แล้ว เพื่อให้มั่นใจว่าขั้นตอนวิธีที่นำเสนอนี้จะไม่ตัดที่จุดกึ่งกลางของพยางค์นั้น จึงทำการติดตามค่าพลังงานของเสียงหลังจากการค่าพลังงานลงลงต่ำกว่าค่าขีดแบ่งไปอีก 5 เฟรม (ซึ่งเป็นเวลาประมาณ 75 มิลลิวินาที) หลังจากนั้นจึงจะถือว่าจุดนี้เป็นจุดจบของพยางค์ที่แท้จริง ดังตัวอย่างในรูปที่ 3.3 จุด A เป็นจุดที่ค่าความเข้มของเสียงเกินค่าขีดแบ่งที่กำหนดเอาไว้ จากนั้นก็มีค่าเพิ่มขึ้นและลดลงจนกระทั่งมีค่าต่ำกว่าค่าขีดแบ่งที่จุด B แต่หลังจากนั้นเพียง 2-3 เฟรม กลับมีค่ามากกว่าค่าขีดแบ่งอีกที่จุด C จุด B จะไม่ถูกตัดสินใจว่าเป็นจุดจบของพยางค์จริง ๆ จนกระทั่งค่าระดับความเข้มเสียงกลับมามีค่าต่ำกว่าค่าขีดแบ่งอีกครั้งที่จุด D และเป็นเช่นนี้ไปอีก 5 เฟรมจนถึงจุด E จึงจัดสินใจให้จุด E เป็นจุดจบของพยางค์



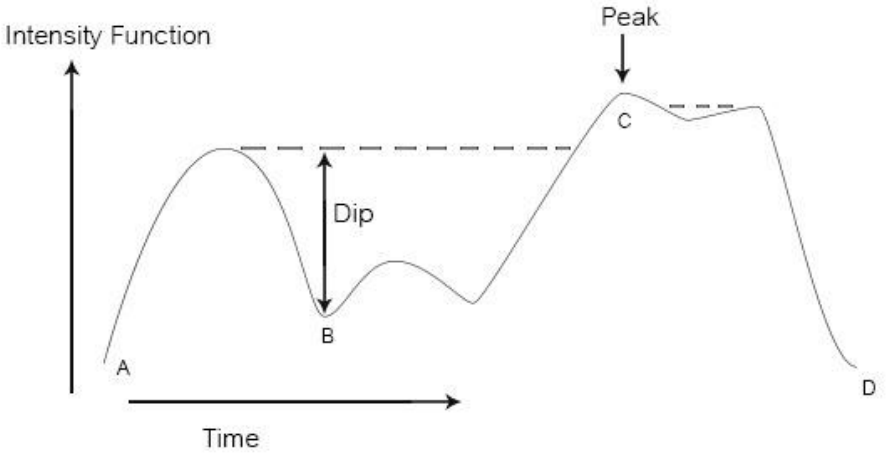
รูปที่ 3.3 ตัวอย่างการหาจุดจบของพยางค์แบบทันที

หลังจากที่ได้จุดจบของพยางค์มาแล้ว ตั้งแต่จุดที่เริ่มต้นขอบเขตพยางค์จนกระทั่งถึงจุดจบของขอบเขตพยางค์นั้น ไม่จำเป็นจะต้องมีพยางค์เดียวเสมอไปในสถานะที่อยู่ระหว่างช่วงที่เป็นพลังประจำเสียงของพยางค์ เนื่องจากการพูดปกติ เสียงของแต่ละพยางค์มักจะเชื่อมต่อกันไป จนกระทั่งเจอพยางค์ที่เป็นคำตายในภาษาไทย (คำที่มีพยัญชนะสะกดเป็นแม่ กก กด กบ) เพราะพยางค์ของคำตายจะมีช่วงระยะเวลาการออกเสียงที่สั้นและหยุดก่อนที่จะพูดพยางค์ตัวต่อไป ทำให้พลังงานหลังจากจบพยางค์ของคำตายลดลงอย่างทันที หรือเจอการเว้นวรรคสั้น ๆ เพื่อหยุดหายใจของการพูด ก็จะทำให้พลังงานในช่วงของการหยุดเว้นวรรคสั้น ๆ นี้มีค่าน้อยเช่นกัน ดังนั้น ในส่วนของกระบวนการตรวจหาจุดจบของพยางค์แบบทันที การติดตามค่าพลังงานของเสียงตั้งแต่เฟรมแรกที่มีค่าเกินค่าขีดแบ่งจนกระทั่งถึงจุดจบของพยางค์ได้ทำการเก็บค่าพลังงานเอาไว้เป็นกลุ่มของค่าพลังงานที่สะสมในส่วนที่เป็นพลังประจำเสียงของพยางค์ เพื่อนำไปใช้ในการคำนวณต่อในส่วนต่อไป เพื่อหาจำนวนพยางค์ที่แท้จริงที่พูดออกไป สรุปได้ว่า เมื่อกระบวนการหาจุดจบของพยางค์พบจุดจบแล้วจะเป็นเหตุการณ์ที่ทำให้เกิดการตอบให้ผู้ใช้เห็นว่า ได้พูดไปแล้วกี่พยางค์

1.2. **การตรวจหาแกนกลางของพยางค์** (Peak Detection)

จากผลลัพธ์ที่ได้จากส่วนของการตรวจหาจุดจบของพยางค์แบบทันที นอกจากจะได้จุดจบของพยางค์แล้ว ยังได้กลุ่มของค่าระดับความเข้มเสียงในส่วนที่เป็นพลังประจำเสียงของพยางค์มา

ด้วย ในส่วนนี้จะนำกลุ่มของค่าระดับความเข้มเสียงมาทำการคำนวณต่อ โดยใช้ขั้นตอนวิธีคอนเว็กซ์ฮัลล์ (Convex Hull Algorithm) ของ Mermelstein [18] ซึ่งมีหลักการทำงาน ดังรูปที่ 3.4 ดังนี้



รูปที่ 3.4 หลักการทำงานของขั้นตอนวิธีคอนเว็กซ์ฮัลล์

1.2.1. เมื่อเริ่มต้นการทำงานจะทำการสร้างคอนเว็กซ์ฮัลล์)Convex Hullขึ้นมาตั้งแต่ (จุดเริ่มต้นของกลุ่มค่าระดับความเข้มเสียงที่ได้มา ซึ่งถูกนิยามโดยฟังก์ชันที่มีค่าโดยมี (ซึ่งจะมีค่าน้อยก่อน) เพิ่มขึ้นตามค่าของระดับความเข้มเสียงตั้งแต่จุดเริ่มต้นค่าเพิ่มขึ้นในทางเดียว ไม่ลดลงระหว่างทางเลย จนกระทั่งถึงจุดสูงสุดในกลุ่มความเข้มเสียงนั้น จากนั้นก็จะมีค่าลดลงในทางเดียว ไม่มีค่าเพิ่มขึ้นระหว่างทางจนกระทั่งถึงจุดจบของกลุ่มค่าระดับความเข้มเสียง ดังเส้นประ ในรูปที่ 3.4

1.2.2. ในกลุ่มของระดับความเข้มเสียง คำนวณค่าความแตกต่างระหว่างคอนเว็กซ์ฮัลล์และค่าพลังงานทุก ๆ จุด จุดที่มีค่าความแตกต่างมากที่สุด (Dip) ดังจุด B ในรูป มีศักยภาพที่จะเป็นขอบเขตของพยางค์ได้ ซึ่งหากจุดดังกล่าวมีค่าความแตกต่างเกินค่าขีดแบ่งของความแตกต่างที่กำหนดไว้ (Dip Threshold) กลุ่มความเข้มเสียงในตอนแรก จะถูกตัดแบ่งออกเป็น 2 กลุ่มของความเข้มเสียง

1.2.3. กระทำขั้นตอนที่ 3.1 และ 3.2 ของกลุ่มค่าระดับความเข้มเสียงที่ถูกแบ่งออกมาจนกระทั่งไม่มีกลุ่มไหนเลยที่มีค่าความแตกต่างระหว่างคอนเว็กซ์ฮัลกับค่าระดับ

ความเข้มเสียงเกินค่าขีดแบ่งของความแตกต่างที่กำหนดไว้ ซึ่งหมายความว่าไม่สามารถตัดแบ่งได้อีกแล้ว เพียงเท่านี้ก็จะได้จำนวนพยางค์ที่พูดออกไปได้แล้ว

1.2.4. ส่วนแกนกลางของพยางค์สามารถทราบได้จากจุดที่มีค่าสูงที่สุด (Peak) ในแต่ละกลุ่มย่อยที่ถูกตัดแบ่งออกจากกลุ่มของค่าพลังงานของเสียง (ซึ่งมีจำนวนเท่ากับจำนวนกลุ่มย่อยที่ถูกแบ่ง)

หลังจากที่เราทราบจุดที่มีค่าสูงสุดแล้ว เพื่อความถูกต้องมากขึ้น จะสันนิษฐานว่าจุดเหล่านั้นมีศักยภาพที่จะเป็นแกนกลางของพยางค์ เราจะนำจุดเหล่าไปตรวจสอบให้คะแนนอีกครั้ง

1.3. การให้คะแนนของแกนกลางพยางค์ที่ตรวจหาได้

หลังจากได้ตำแหน่งของจุดที่มีค่าความเข้มสูงที่สุดแล้ว จะนำตำแหน่งแต่ละตำแหน่งมาผ่านซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine - SVM) ซึ่งเป็นตัวจำแนกประเภท (Classifier) ประเภทหนึ่ง โดยลักษณะเด่นทางเสียงที่ใช้เป็นข้อมูลชุดฝึกให้กับซัพพอร์ตเวกเตอร์แมชชีน ได้แก่ ค่าระดับความเข้มของเสียงที่ความถี่มากกว่า 300 เฮิรตซ์ ขึ้นไป และค่าอัตสหสัมพันธ์ที่มากที่สุดในช่วงความถี่ 60 ถึง 320 เฮิรตซ์ ของสัญญาณเสียงที่มีช่วงความถี่ต่ำกว่า 900 เฮิรตซ์ ตามที่ Dareeyoah เสนอ [9] เพื่อใช้จำแนกตำแหน่งที่มีค่าระดับความเข้มเสียงสูงที่สุดที่ได้มา ว่าตำแหน่งนั้นเป็นตำแหน่งที่ควรจะเป็นแกนกลางของพยางค์จริงหรือไม่

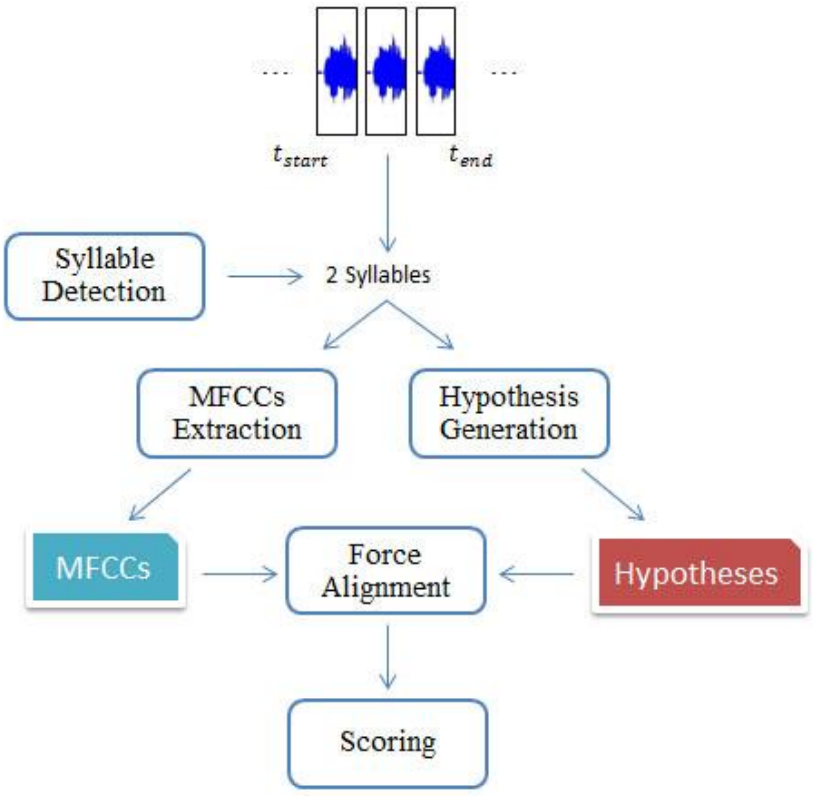
หลังจากที่เราได้แกนกลางที่แท้จริงของพยางค์แล้ว ก็ทำให้เราทราบอีกว่า ในกลุ่มของค่าระดับความเข้มเสียงที่ถูกสะสมไว้จนถึงจุดจบของพยางค์นั้น ได้พูดไปแล้วกี่พยางค์ เพียงเท่านี้ก็จะสามารถประสานเวลาเสียงให้ตรงกับข้อความได้ เพราะทราบจำนวนพยางค์ที่พูดออกไปแล้ว

2. การตรวจหาความผิดพลาดของการถอดเสียง

เนื่องจากกระบวนการที่นำเสนอจะได้จำนวนพยางค์ที่พูดออกไปแล้ว ก็ยังมีโอกาสเกิดความผิดพลาดขึ้นอีกเช่นกัน ดังนั้น เราจะใช้ประโยชน์จากข้อความที่ทราบอยู่แล้ว เพราะผู้ใช้จะต้องพูดตามข้อความทั้งหมด ทำให้เราทราบว่า ผู้พูดจะพูดอะไรบ้าง ดังนั้นเมื่อเราทราบจำนวนพยางค์ที่ผู้พูดพูดออกมาแล้ว เรายังสามารถตรวจสอบได้อีกว่า เสียงที่พูดออกมามีจำนวนพยางค์ ตรงกับข้อความที่ควรจะเป็นหรือไม่ แนวคิดที่สำคัญในส่วนนี้คือ ถ้าหากเราสามารถตรวจหาได้ว่า เสียงที่

พูดออกมานั้นมีเมื่อนับจำนวนพยางค์ที่พูดออกมาได้แล้ว ไม่ตรงกับพยางค์ที่ได้จากการถอดเสียงจากข้อความที่เลื่อนจำนวนพยางค์ไปเท่า ๆ กัน แล้วเป็นไปได้หรือไม่ที่จะสามารถปรับเลื่อนคำตอบจากที่ได้ออกมาให้ตรงกับพยางค์ในข้อความได้ ดังนั้น เราจะใช้หลักการตรวจหาความผิดพลาดของการถอดเสียง และพยายามแก้ไขโดยการเลื่อนตำแหน่งพยางค์จากการถอดเสียงของข้อความให้ตรงตามเสียงพูดจริง ๆ ที่ควรจะเป็น

หลักการนี้จะประยุกต์ใช้หลักของการปรับแนวของเสียง (Force Alignment) มาประยุกต์ใช้ แนวคิดก็คือ หากเสียงและหน่วยเสียงที่ได้จากการถอดเสียงของข้อความนั้นตรงกัน ความน่าจะเป็นที่ได้จากการประสานเวลาของเสียงก็จะมีค่าสูง ในขณะเดียวกัน หากเสียงของหน่วยเสียงไม่ตรงกัน ความน่าจะเป็นที่ได้จากการปรับแนวของเสียงก็จะมีค่าน้อยลงตามไปด้วย ดังแสดงในรูปที่ 3.5 ซึ่งมีขั้นตอนวิธี ดังนี้



รูปที่ 3.5 แผนภาพขั้นตอนวิธีในส่วนการตรวจหาความผิดพลาดของการถอดเสียง

1.3. ก่อนที่จะทำการปรับแนวของเสียงได้ จำเป็นจะต้องสร้างแบบจำลองเสียง (Acoustic Model) ขึ้นมาก่อน เช่นเดียวกันกับการสร้างแบบจำลองเสียงของระบบรู้จำเสียง รายละเอียดของแบบจำลองที่ใช้เป็นข้อมูลชุดฝึกสอนจะกล่าวละเอียดในบทที่ 4

1.4. เมื่อทำการเตรียมแบบจำลองเสียงเรียบร้อยแล้ว กลุ่มของพยางค์ที่ได้มาจากส่วนที่แล้ว ทำการตัดข้อความออกมาตามจำนวนพยางค์ที่เป็นผลลัพธ์จากส่วนที่แล้ว แล้วทำการถอดเสียงของข้อความให้อยู่ในรูปแบบหน่วยเสียงเดี่ยว (Mono-phone) เพื่อเตรียมใช้เป็นข้อมูลนำเข้าของระบบการปรับแนวของเสียง ส่วนข้อมูลนำเข้าที่เป็นเสียง จะนำเสียงเฉพาะส่วนที่ตัดกลุ่มของค่าระดับความเข้มเสียงที่เก็บไว้จากส่วนแรกนั้นมาวิเคราะห์ โดยสกัดลักษณะเด่นของเสียงเป็น 39 MFCCs ตามศาสตร์แห่งศิลป์ของการรู้จำเสียงพูดทั่วไป

1.5. สร้างสมมติฐานของคำตอบ (H_n) ของกลุ่มพยางค์ที่พูดออกมาว่ามีโอกาสเป็นสมมติฐานใดได้บ้าง ซึ่งจะสร้าง $2n+1$ สมมติฐานขึ้นมา โดยที่ n คือขอบเขตจำนวนพยางค์ที่จะสร้างสมมติฐานเพิ่ม

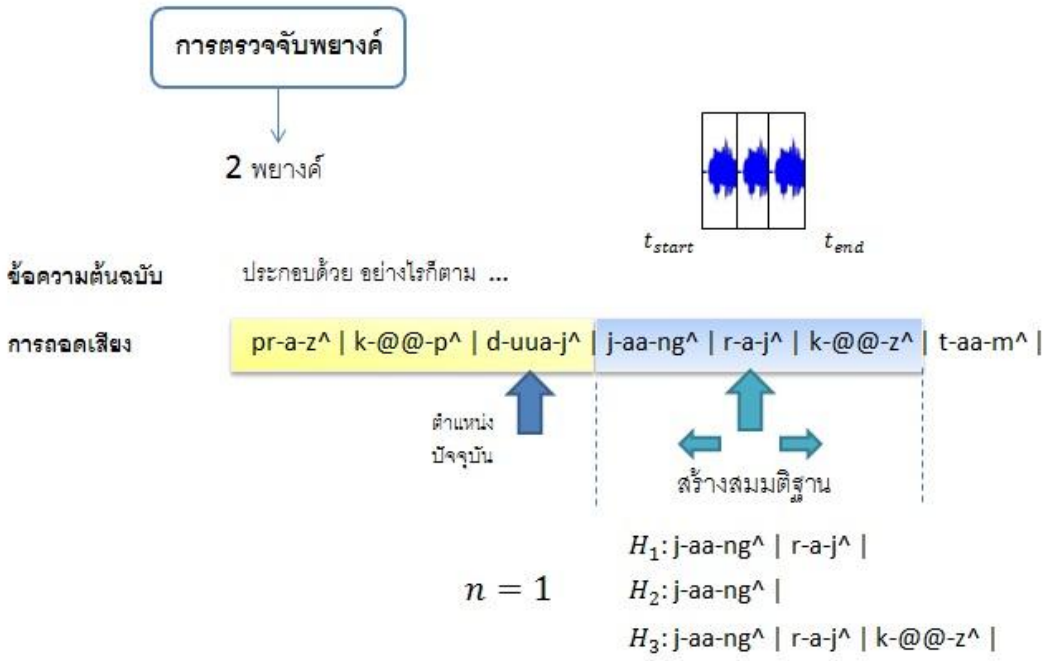
ยกตัวอย่างในรูปที่ 3.6 สมมติให้ $n=1$ จะได้สมมติฐานจำนวน 3 สมมติฐานด้วยกัน คือ

H_1 คือ สมมติฐานจำนวนพยางค์เท่าผลลัพธ์ที่ได้จากการขั้นตอนตรวจหาพยางค์

H_2 คือ สมมติฐานจำนวนพยางค์น้อยกว่าผลลัพธ์ที่ได้จากการขั้นตอนตรวจหาพยางค์อยู่ 1 พยางค์

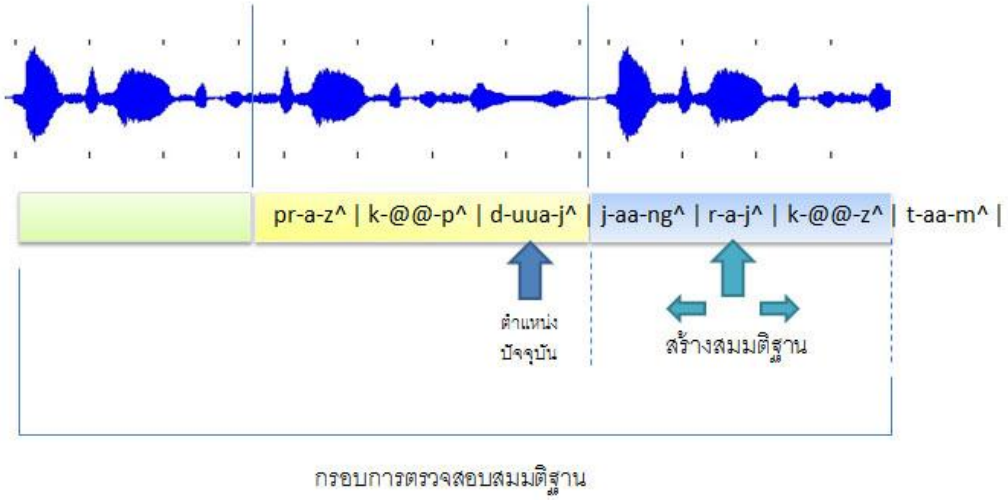
H_3 คือ สมมติฐานจำนวนพยางค์มากกว่าผลลัพธ์ที่ได้จากการขั้นตอนตรวจหาพยางค์อยู่ 1 พยางค์

ดังนั้น n จึงเป็นอีกหนึ่งพารามิเตอร์หนึ่งที่สำคัญที่จะต้องหาคำตอบว่าควรจะตั้งสมมติฐานให้ครอบคลุมจำนวนพยางค์ที่เป็นไปได้เท่าไร จึงจะเหมาะสม



รูปที่ 3.6 การสร้างสมมติฐานขอบเขต $n = 1$ เมื่อส่วนการตรวจจับพยางค์ให้คำตอบเป็น 2 พยางค์

1.6. สร้างกรอบของการตรวจสอบสมมติฐานขึ้นมา เพราะจากการศึกษาเบื้องต้น หากความยาวของหน่วยเสียงที่นำมาต่อกันมีความยาวมากขึ้น น่าจะมีผลทำให้ผลลัพธ์ของการปรับแนวของเสียงนั้นแม่นยำมากขึ้น ดังนั้นเราจะนำหน่วยเสียงจากผลลัพธ์ของช่วงที่ตัดสินใจไปแล้ว มาพิจารณาร่วมกับช่วงเสียงปัจจุบันที่กำลังพิจารณาอยู่ด้วย ดังนั้น กรอบของการตรวจสอบสมมติฐาน คือ เราจะนำช่วงเสียงกี่ช่วงก่อนหน้าที่ตัดสินแล้วกี่ช่วง จนถึงช่วงเสียงปัจจุบันที่กำลังพิจารณา ดังรูปที่ 3.7



รูปที่ 3.7 กรอบการตรวจสอบสมมติฐานจำนวน 2 กรอบย้อนกลับ

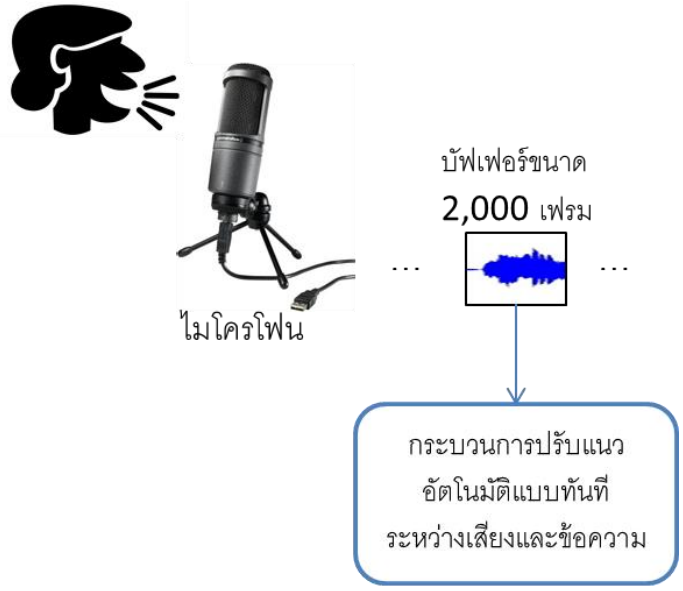
1.7. จากนั้นนำแต่ละสมมติฐานเข้ากระบวนการปรับแนวของเสียงทุก ๆ สมมติฐาน ผลลัพธ์ของการปรับแนวนั้น จะได้ค่าความน่าจะเป็นของแต่ละหน่วยเสียงในแต่ละสมมติฐาน จากนั้นทำการรวมความน่าจะเป็นของทุก ๆ หน่วยเสียงเข้าด้วยกัน รวมเป็น ค่าความน่าจะเป็นของสมมติฐานนั้น ๆ

1.8. เปรียบเทียบสมมติฐานกันทุก ๆ สมมติฐาน ถ้าสมมติฐานใดมีค่าความน่าจะเป็นมากที่สุดก็จะเลือกสมมติฐานนั้นเป็นคำตอบจริง ๆ ว่าได้พูดถึงพยางค์ไหนแล้วในข้อความ เพื่อที่จะแสดงออกมาเป็นผลลัพธ์ที่แท้จริงต่อไป

บทที่ 4

การเตรียมการทดลองและวิธีการวัดผล

ในบทนี้ จะกล่าวถึงการเตรียมข้อมูลสำหรับการทดลองขั้นตอนวิธีที่นำเสนอ และวิธีการวัดผลของการประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความ โดยจะนำไปทดสอบกับไฟล์เสียงที่มีอยู่แล้ว เนื่องจากการทดลอบกับคนจริงนั้นทำได้ยาก แต่จะทำการจำลองสถานการณ์เสมือนกับมีผู้พูดใช้ไมโครโฟนจริง ๆขณะนั้น โดยการตัดแบ่งไฟล์เสียงที่เป็นข้อมูลทดลองออกเป็นไฟล์เสียงย่อย ๆ ก่อน เป็นเสียงขนาด 2,000 เฟรมของบัฟเฟอร์ของไมโครโฟน แล้วจึงค่อย ๆ นำไฟล์เสียงย่อย ๆ นั้นป้อนเข้าไปในระบบอย่างต่อเนื่องเพื่อให้ระบบทำงานแบบทันที เหมือนกับสถานการณ์จริง ดังรูปที่ 4.1



รูปที่ 4.1 การจำลองสถานการณ์การทดลอง

ฐานข้อมูลเสียง

ฐานข้อมูลเสียงนั้น จะนำฐานข้อมูลเสียงขนาดใหญ่สำหรับระบบรู้จำเสียงพูดต่อเนื่องภาษาไทย “โลตัส” (Large Vocabulary Thai Continuous Speech Recognition Corpus: LOTUS) [24] ซึ่งเป็นฐานข้อมูลเสียงพูดภาษาไทยขนาดใหญ่ที่ได้รออกแบบและพัฒนาตาม

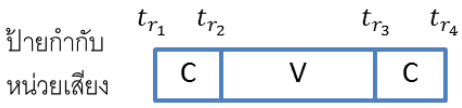
มาตรฐานสากล อีกทั้งฐานข้อมูลเสียงโลตัสยังถูกนำไปประยุกต์ใช้งานเพื่อที่สร้างโปรแกรมประยุกต์มากมาย

ฐานข้อมูลเสียงโลตัส ประกอบด้วยชุดเสียง 4 ชุดย่อย ได้แก่ ชุดหน่วยเสียงสมดุล (Phonetically Distribution set: PD), ชุดฝึกฝน (Training set: TR), ชุดทดสอบเพื่อพัฒนา (Development Test set: DT) และชุดทดสอบเพื่อประเมิน (Evaluation Test set: ET) เป็นเสียงสัญญาณเสียงมีความถี่สุ่ม 16,000 เฮิรตซ์ เก็บข้อมูลแบบ 16 บิต พีซีเอ็ม มีอัตราสัญญาณเสียงต่อคลื่นรบกวน (Signal to Noise) ประมาณ 30 เดซิเบล

การเตรียมข้อมูล

สำหรับในวิทยานิพนธ์นี้ การเตรียมข้อมูลที่ใช้จะแบ่งเป็นข้อมูลชุดฝึกฝน ข้อมูลชุดทดสอบและพัฒนา และข้อมูลชุดประเมินผล การเตรียมข้อมูลจะแบ่งเป็น 3 ขั้นตอนด้วยกัน คือ

1. ในขั้นตอนการตรวจหาพยางค์ จะใช้ข้อมูลชุดหน่วยเสียงสมดุล (PD) เป็นทั้งข้อมูลชุดฝึกฝน จำนวน 400 ไฟล์เสียง ซึ่งเป็นเสียงผู้หญิง 200 ไฟล์เสียง และเสียงผู้ชาย 200 ไฟล์เสียงแรก คิดเป็น 23.8% ข้อมูลชุดทดสอบเพื่อการพัฒนา จำนวน 440 ไฟล์เสียง แบ่งเป็นเสียงผู้หญิงและเสียงผู้ชายจำนวนเท่า ๆ กัน คิดเป็น 26.2% และข้อมูลชุดสำหรับทดสอบเพื่อประเมินผล 840 ไฟล์เสียง หรืออีก 50% ที่เหลือ เนื่องจากชุดหน่วยเสียงสมดุล ฐานข้อมูลเสียงเพียงชุดเดียวที่มีการระบุขอบเขตเวลาของแต่ละหน่วยเสียงไว้อย่างถูกต้องโดยนักภาษาศาสตร์ และงานวิจัยนี้จำเป็นจะต้องวัดผลความแม่นยำของการระบุตำแหน่งแกนกลางของพยางค์ด้วย ดังนั้นป้ายกำกับหน่วยเสียง (Phoneme Label) ของแต่ละหน่วยเสียงนั้น จะถูกเปลี่ยนมาเป็นรูปแบบ พยัญชนะ-สระ (Consonant-Vowel Label) โดยการเปลี่ยนหน่วยเสียงที่เป็นพยัญชนะทุกตัวให้อยู่ในรูปแบบ C (Consonant) และเปลี่ยนหน่วยเสียงที่เป็นสระทุกตัวให้อยู่ในรูปแบบ V (Vowel) ดังรูปที่ 4.2 เนื่องจากเราสนใจเฉพาะแกนกลางของพยางค์ ซึ่งก็คือช่วงเวลาที่เป็นสระนั่นเอง



รูปที่ 4.2 ตัวอย่างของป้ายกำกับหน่วยเสียงรูปแบบพยัญชนะ-สระ

2. ในขั้นตอนการใช้หลักการตรวจหาความผิดพลาดของการถอดเสียงรวมกับขั้นตอนแรกจะใช้ข้อมูลชุดฝึกฝน (TR) เป็นข้อมูลชุดฝึกฝน จำนวน 3,007 ประโยคในการสร้างแบบจำลองทางเสียง (Acoustic Model) ซึ่งเป็นแบบจำลองฮิดเดนมาร์คอฟ โดยการใช้เครื่องมือเอชทีเค (Hidden Markov Model Toolkit: HTK) [32] ในการเรียนรู้แบบจำลอง แบบจำลองฮิดเดนมาร์คอฟเป็นแบบจำลองฮิดเดนมาร์คอฟ จากซ้ายไปขวา 5 สถานะ เรียนรู้แบบจำลองเสียงพูดแบบที่ไม่ขึ้นกับบริบทรอบข้าง (Context-independent Phone Model) และประมาณค่าความน่าจะโดยใช้เกาส์เซียนมิกเจอร์โมเดล (Gaussian Mixture Model) จำนวน 1 มิกเจอร์ และใช้แบบจำลองทางภาษาหน่วยเสียงเดี่ยว (Mono-phone) จากนั้น ใช้ข้อมูลชุดหน่วยเสียงสมดุล (PD) เป็นข้อมูลชุดสำหรับทดสอบเพื่อพัฒนาจำนวน 504 ประโยคทดสอบ หรือ 30% ของชุดหน่วยเสียงสมดุล ในการปรับค่าพารามิเตอร์ในส่วนของขั้นตอนการตรวจหาความผิดพลาดของการถอดเสียง และข้อมูลชุดทดสอบเพื่อการประเมินผล จำนวน 1176 ประโยคทดสอบหรือ 70% เพราะมีการระบุขอบเขตเวลาและป้ายกำกับหน่วยเสียงของแต่ละหน่วยเสียงไว้อย่างถูกต้องอยู่แล้วเช่นกัน สำหรับการให้คะแนนประสานเวลาเสียง (Force Alignment) นั้น ใช้เครื่องมือเอชทีเคในการประสานเวลาเช่นกัน
3. ในขั้นสุดท้ายจะนำขั้นตอนวิธีที่นำเสนอไปทดสอบกับข้อมูลทดสอบเพื่อประเมินผลอีกหนึ่งชุด ซึ่งสร้างขึ้นมาโดยคัดลอกบทความหนึ่งจากอินเตอร์เน็ต [33] เพื่อทดสอบในลักษณะการใช้งานจริง เสมือนการอ่านหนังสือประมาณ 1 หน้า ซึ่งมีจำนวนคำ 328 คำ มีจำนวนพยางค์ 411 พยางค์ ผู้พูดทำการบันทึกเสียงก่อน โดยบันทึกเสียงผ่านไมโครโฟนคอนเดนเซอร์ (Audio-Technica AT2020 Cardioid Condenser Microphone) สัญญาณเสียงมีความถี่สุ่ม 16,000 กิโลเฮิรตซ์ รูปแบบเวฟ 16 บิต พีซีเอ็ม (WAV 16 bits PCM) ในสภาพแวดล้อมห้องปกติที่ไม่มีเสียงรบกวน จากนั้นก็

ทำการถอดเสียงด้วยมือ และระบุเวลาของแต่ละป้ายกำกับหน่วยเสียงลงไป โดยใช้โปรแกรมเวฟเซิฟเฟอร์ (Wave Suffer)

ระบบอ้างอิง (Baseline)

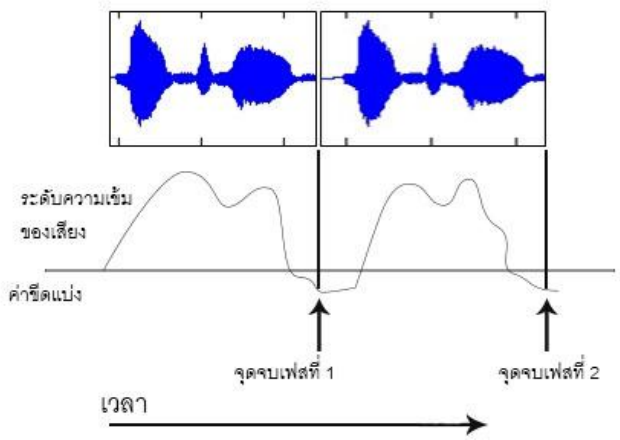
ในวิทยานิพนธ์นี้ ขั้นตอนวิธีที่นำเสนอนั้น จะถูกนำมาเปรียบเทียบกับบรรทัดฐานในการประสานเวลาข้อความและเสียงแบบทันทีกับการประสานเวลาโดยใช้หลักการตรวจหาความผิดพลาดของการถอดเสียงเพียงอย่างเดียว และเปรียบเทียบกับการใช้หลักของการตรวจหาพยางค์เพียงอย่างเดียวในการประสานเวลา

ระบบอ้างอิงแรก ใช้หลักการตรวจหาพยางค์มาประสานเวลา โดยขั้นตอนวิธีจะเหมือนกับขั้นตอนวิธีที่นำเสนอในบทที่ 3 เพราะสามารถใช้ในการประสานเวลาอัตโนมัติแบบทันทีได้ โดยไม่ต้องใช้ข้อมูลจากข้อความ กล่าวคือ การติดตามค่าของพลังงานไปเรื่อย ๆ จนพบจุดจบของพยางค์แล้วจึงใช้วิธีหาตำแหน่งของแกนกลางพยางค์เพื่อนับพยางค์ที่พูดออกไปแล้วว่ามีจำนวนกี่พยางค์ โดยจะใช้ลักษณะเด่นทางเสียงเช่นเดียวกันกับขั้นตอนวิธีที่นำเสนอ นั่นคือ พลังงานที่มีความถี่มากกว่า 300 เฮิรตซ์ และค่าอัตสหสัมพันธ์

ระบบอ้างอิงที่สอง คือการใช้หลักการตรวจหาความผิดพลาดของการถอดเสียง มาประมวลผลไปเรื่อย ๆ เมื่อข้อมูลเสียงเพิ่มขึ้นเรื่อย ๆ ซึ่งมีลักษณะคล้ายกับงานวิจัยที่เกี่ยวข้องของ Gao [17] แต่ประยุกต์ใช้ในระดับพยางค์ โดยวิธีการคือ ตั้งสมมติฐานว่า ทุก ๆ 1 วินาที คนทั่วไป ๆ จะพูดได้ประมาณ 3 พยางค์ ดังนั้น เมื่อข้อมูลเสียงเริ่มถูกบันทึกผ่านไมโครโฟนแล้ว เมื่อครบทุก ๆ 1 วินาที ส่วนการตรวจสอบความไม่ตรงกันของการถอดเสียงจะเริ่มทำงาน โดยมีลักษณะการทำงานดังที่กล่าวไว้ในบทที่ 3 อีกเช่นกัน คือการสร้างสมมติฐานของข้อความว่าพูดไปกี่พยางค์แล้วขึ้นมา โดยต้องกำหนดพารามิเตอร์คือ ขอบเขตจำนวนพยางค์ที่จะตรวจสอบ และ กรอบการตรวจสอบของสมมติฐาน แล้วประสานเวลาตามสมมติฐานที่มีค่าผลลัพธ์ที่มีมากที่สุด

วิธีการวัดผลของการประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความ (Evaluation Method)

ดังที่กล่าวถึงในตอนต้น เราได้เตรียมข้อมูลเสียงและข้อความที่จะใช้ในการทดลองแล้ว เสียงที่เตรียมมาจะเป็นไฟล์เสียงที่จะพูดตรงตามข้อความทั้งหมด จะถูกตัดแบ่งเป็นเสียงย่อย ๆ และค่อย ๆ ป้อนเข้าไปในระบบให้เหมือนกับวิธีการรับเสียงจากไมโครโฟน กระบวนการทำงานของขั้นตอนวิธีที่นำเสนอจะเริ่มทำงานทันทีที่เสียงเริ่มมีการถูกป้อนเข้ามาในระบบ ในส่วนของการตรวจหาจุดจบของพยางค์นั้น นอกจากจะได้กลุ่มค่าระดับความเข้มของเสียงที่ถูกตัดออกมาแล้ว เพื่อนำไปเข้าในส่วนของการตรวจหาแกนกลางพยางค์เพื่อแบ่งละเอียดอีกครั้งนั้น เราจะสามารถทราบเวลาที่เป็นจุดจบของพยางค์นั้นในไฟล์เสียงทดสอบนั้นอีกด้วย เวลาของไฟล์เสียงทดสอบตั้งแต่จุดเริ่มต้น จนถึงจุดจบของพยางค์แต่ละจุดที่สามารถตรวจหาได้ จะถือว่าเป็นหนึ่งเฟส (p_i) และจุดจบของพยางค์ในเฟสนั้นที่ตรวจหาได้ ก็จะถือให้เป็นจุดเริ่มต้นของเฟสต่อไป ดังนั้นทุก ๆ ครั้ง เฟส p_i จะต้องให้คำตอบออกมาเสมอว่าในเฟสนั้น ได้พูดออกมาแล้วกี่พยางค์ ดังรูปที่ 4.3



รูปที่ 4.3 ตัวอย่างการแยกเฟสแต่ละเฟสในแกนเวลา

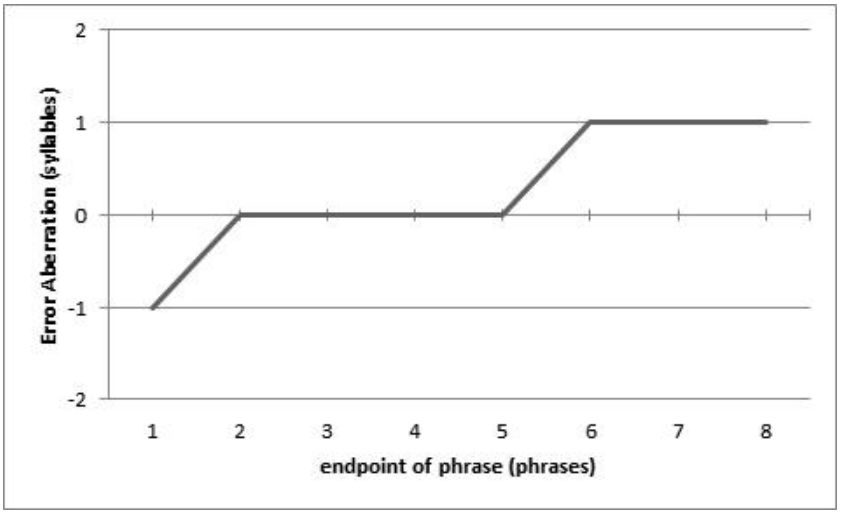
การตัดสินใจว่าการประสานเวลาถูกหรือไม่นั้น จะอาศัยการนับพยางค์จากการถอดเสียงอ้างอิง ซึ่งมีอยู่ข้อมูลเสียงในฐานข้อมูลเสียงโลตัสในชุดหน่วยเสียงสมดุล โดยเกณฑ์การจัดสินใจจะถือว่าแกนกลางของพยางค์ในภาษาไทยคือสระ และอาศัยเวลาที่เป็นจุดจบของพยางค์ในแต่ละเฟส p_i เปรียบเทียบในเวลาของการถอดเสียงอ้างอิง แต่ละเฟส p_i จะมีการระบุคำตอบออกมาว่าได้

พูดออกมาแล้วกี่พยางค์และสะสมคำตอบไปเรื่อย ๆ เราเรียกว่า S_{p_i} ส่วนทางด้านการถอดเสียงอ้างอิงก็จะนับพยางค์สะสมทั้งหมดคล้ายคลึงกัน โดยนับจำนวนพยางค์ (จำนวนสระ) ตั้งแต่เวลาเริ่มต้นของเสียงทดสอบนั้น จนไปถึงเวลาที่เป็นจุดจบของเฟส p_i ว่ามีจำนวนพยางค์ทั้งหมดกี่พยางค์ เราเรียกว่า S_{r_i}ดังนั้นเมื่อเราทราบจำนวนพยางค์ทั้งหมดที่ทั้งที่ขั้นตอนวิธีนับได้ และการถอดเสียงอ้างอิงนับได้ จนถึงเวลาที่เป็นจุดจบของพยางค์ในเฟส p_i ผลต่างของจำนวนพยางค์ของทั้งคู่ ก็คือ จำนวนพยางค์ที่ขั้นตอนวิธีที่นำเสนอนั้นผิดพลาดไป เรียกว่า Error Aberration (E_{p_i}) ดังสมการที่ 4.1

$$E_{p_i} = S_{p_i} - S_{r_i} \tag{4.1}$$

E_{p_i} นั้นแสดงถึงความผิดพลาดของการประสานเวลาโดยบอกว่าผิดพลาดไปกี่พยางค์ด้วยกัน ซึ่งหาก E_{p_i} มีค่าเท่ากับ 0 หมายความว่า ณ จุดจบของเฟส p_i นั้นมีการประสานเวลาที่ตรงตามพยางค์ที่ควรจะเป็น ถ้าหาก E_{p_i}มีค่าน้อยกว่า 0 หมายความว่า ณ จุดจบของเฟส p_i มีการประสานเวลาผิดพลาดล่าช้ากว่าพยางค์ที่ควรจะเป็น เรียกว่า Deletion Error และหาก E_{p_i} มีค่ามากกว่า 0 หมายความว่า ณ จุดจบของเฟส p_i มีการประสานเวลาผิดพลาดเร็วกว่าพยางค์ที่ควรจะเป็น เรียกว่า Insertion Error ส่วนค่าที่เป็นตัวเลขของ E_{p_i} หมายถึงว่ามีค่าผิดพลาดในการประสานเวลาออกจากการถอดเสียงอ้างอิงที่ควรจะเป็นกี่พยางค์

ดังตัวอย่างในรูปที่ 4.4 จะเห็นว่า หลังจากที่เจอจุดจบของพยางค์จุดแรกสมมติให้เป็นเวลา t_1 ในประโยคทดสอบนั้น เกิด Deletion Error ขึ้น 1 พยางค์ ซึ่งหมายถึง การประสานเวลานั้นล่าช้ากว่าที่ควรจะเป็นไป 1 พยางค์ แต่เมื่อตรวจพบจุดจบของพยางค์จุดที่ 2 สมมติให้เป็นเวลา t_2 นั้นการประสานเวลากลับมาถูกต้องตามผลลัพธ์อ้างอิง แต่พอถึงจุดที่ 5 นั้นเกิด Insertion Error ขึ้น ซึ่งหมายถึงการประสานเวลานั้นเร็วกว่าผลลัพธ์อ้างอิงอยู่ 1 พยางค์นั่นเอง



รูปที่ 4.4 กราฟแสดงความผิดพลาดของการประสานเวลาของประโยคทดสอบ

แต่สำหรับการวัดผลการประสานเวลาสำหรับประโยคทดสอบทั้งประโยคนั้น คำนวณโดยการใช้ Total Error Aberration ($Err(S)$) ของประโยคทดสอบ S เป็นตัวประเมินความถูกต้องของขั้นตอนวิธีการประสานเวลาอัตโนมัติแบบทันที ซึ่งคำนวณได้จากค่าสัมบูรณ์ของ E_{p_i} ในทุก ๆ เฟสรวมกัน ดังสมการที่ 4.2

$$Err(S) = \sum_{i=1}^{n} |E_{p_i}| \tag{4.2}$$

เมื่อ n คือ จำนวนเฟสทั้งหมดในประโยค S

ส่วนในแง่ของเวลาในการคำนวณนั้น จะวัดเวลาหลังจากที่ขั้นตอนวิธีที่เสนอนั้นตัดสินใจได้ว่าคืนจุดจบของพยางค์แล้ว จนกระทั่งกระบวนการตอบคำตอบออกมาว่าพูดไปแล้วกี่พยางค์ ว่าใช้เวลาคำนวณในจุดนี้เท่าไร ซึ่งไม่ควรจะมีค่าเกินที่ผู้พูดพูดพยางค์ต่อไปจบ เพราะจะเกิดการสะสมความล่าช้าของการคำนวณไป เรื่อย ๆ โดยที่ประสิทธิภาพในเรื่องเวลาของประโยคทดสอบ S นั้นคำนวณได้จากสมการที่ 4.3

$$t(S) = \frac{\sum_{i=1}^{n} t_{p_i}}{n} \tag{4.3}$$

เมื่อ $t(S)$ คือ เวลาในการคำนวณเฉลี่ยของแต่ละเฟสในประโยคทดสอบ S

t_{p_i} คือ เวลาที่ใช้คำนวณของเฟส p_i ตั้งแต่เริ่มตรวจพบจุดจบพยางค์จนกระทั่งให้คำตอบในการประสานเวลาออกมา

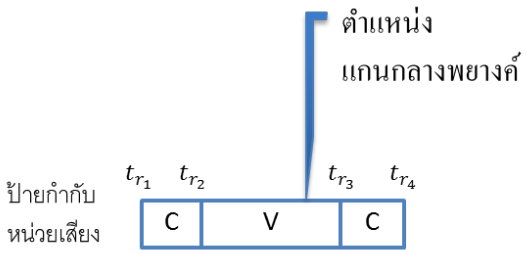
n คือ จำนวนเฟสทั้งหมดในประโยค S

นอกจากนั้นแล้ว ยังเพิ่มการวัดค่าเบี่ยงเบนจากตำแหน่งอ้างอิง (Reference Deviation) เพื่อดูการกระจายตัวของ E_{p_i} ในประโยคทดสอบจากเส้นอ้างอิง (เส้นที่ Error Aberration เป็น 0 จากภาพ 4.4) ประกอบกันไป ดังสมการที่ 4.4

$$Ref\ Deviation = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_{p_i})^2} \qquad (4.4)$$

วิธีการวัดผลการระบุตำแหน่งแกนกลางพยางค์

ดังที่กล่าวไว้ข้างต้น การตรวจหาพยางค์นั้นเป็นหนึ่งในขั้นตอนหนึ่งในขั้นตอนวิธีที่นำเสนอในการประสานเวลา ดังนั้น จึงจำเป็นต้องวัดความแม่นยำในการระบุตำแหน่งแกนกลางด้วย ความมีความแม่นยำเพียงใด โดยตำแหน่งแกนกลางของพยางค์ที่ถูกนั้น ในวิทยานิพนธ์นี้จะถือว่า เพียงแค่ตำแหน่งที่ขั้นตอนวิธีในการตรวจหาแกนกลางของพยางค์สามารถระบุตำแหน่งเวลาของแกนกลางพยางค์ให้ตกอยู่ในช่วงของสระ (หน่วยเสียง V)ในหน่วยเสียงอ้างอิง ก็ถือว่า ตำแหน่งที่ขั้นตอนวิธีตรวจหาแกนกลางของพยางค์นั้นถูกต้อง แต่ถ้าหากว่าตำแหน่งที่ตัดสินใจนั้น ตกอยู่ในช่วงที่เป็นพยัญชนะ (หน่วยเสียง C) ในหน่วยเสียงอ้างอิง ก็จะถือว่า ตำแหน่งนั้นผิด ดังภาพที่ 4.5



รูปที่ 4.5 ตัวอย่างตำแหน่งแกนกลางพยางค์ที่ถือว่าถูกต้อง

ดังนั้น การจำแนกตำแหน่งของแกนกลางพยางค์นั้น แบ่งออกได้เป็น 3 ประเภท คือ

1. ตำแหน่งที่ถูกต้อง (Correction) คือ ตำแหน่งของแกนกลางพยางค์นั้นตกอยู่ในช่วงของหน่วยเสียงสระ
2. ตำแหน่งเกิน (Insertion) คือ ตำแหน่งของแกนกลางพยางค์นั้น อยู่ในขอบเขตของหน่วยเสียงที่เป็นพยัญชนะ หรือตำแหน่งที่ตกอยู่ในขอบเขตของหน่วยเสียงสระ แต่หน่วยเสียงสระอ้างอิงนั้น เคยมีตำแหน่งที่ตกในหน่วยเสียงเสียงอ้างอิงนั้นแล้ว ก็จะถือว่าตำแหน่งที่ตรวจหาได้นั้น เป็นตำแหน่งเกิน
3. ตำแหน่งขาด (Deletion) คือ ขอบเขตทางเวลาของหน่วยเสียงสระอ้างอิงนั้น ๆ ไม่เคยมีตำแหน่งที่ตรวจหาแกนกลางตกไปอยู่ในช่วงนั้นเลย ก็จะถือว่าแกนกลางของพยางค์นี้ ไม่เคยถูกตรวจพบ ในให้เกิดความผิดพลาดขึ้น เป็นตำแหน่งขาด

ส่วนวิธีการวัดผลของนั้นจะใช้ค่า จำนวนความผิดพลาดที่เกิดขึ้นทั้งในแบบตำแหน่งเกินและตำแหน่งขาดมาวิเคราะห์ประสิทธิภาพของตำแหน่งแกนกลางพยางค์ ซึ่งในวิทยานิพนธ์นี้รายงานผลเป็นค่าต่าง ๆ ดังนี้

1. เปอร์เซ็นต์ตำแหน่งเกิน สามารถคำนวณได้จาก สมการที่ 4.5 เมื่อ Total คือตำแหน่งของแกนกลางพยางค์ทั้งหมด

$$\%Insertion = \frac{Insertion}{Total} x100 \quad (4.5)$$

2. เปอร์เซ็นต์ตำแหน่งขาด สามารถคำนวณได้จาก สมการที่ 4.6

$$\%Deletion = \frac{Deletion}{Total} x100 \quad (4.6)$$

3. เปอร์เซ็นต์ความผิดพลาด (Error Rate) สามารถคำนวณได้จาก เปอร์เซ็นต์ความผิดพลาดตำแหน่งเกิน รวมกับเปอร์เซ็นต์ความผิดพลาดตำแหน่งขาด ดังสมการที่ 4.7

$$Error\ Rate = \%Deletion + \%Insertion \quad (4.7)$$

4. เปอร์เซ็นต์ความแม่นยำ (Accuracy) สามารถคำนวณได้จากสมการที่ 4.8

$$Accuracy = 100\% - Error\ Rate \quad (4.8)$$

บทที่ 5

การทดลองและสรุปผลการทดลอง

ในบทนี้จะกล่าวถึงขั้นตอนการสร้างขั้นตอนวิธีที่นำเสนอ การทดลอง และผลการทดลองทั้งหมด ตามขั้นตอนการทดลอง ดังต่อไป

1. ทดสอบประสิทธิภาพของเฉพาะส่วนการตรวจหาพยางค์
2. ประเมินผลการประสานเวลาอัตโนมัติแบบทันที

ทดสอบประสิทธิภาพของส่วนการตรวจหาพยางค์

ในการทดลองนั้น เราจะเริ่มจากการสร้างส่วนการตรวจหาพยางค์ขึ้นมาก่อน ซึ่งประกอบไปด้วยขั้นตอนการตรวจหาจุดจบของพยางค์ การตรวจหาแกนกลางของพยางค์และการให้คะแนนตำแหน่งแกนกลางของพยางค์

1. การทดลองการให้คะแนนตำแหน่งแกนกลางของพยางค์

อันดับแรก เราจะทำการสร้างตัวจำแนกประเภทซึ่งเป็นซัพพอร์ตเวกเตอร์แมชชีนจาก SVMLIB [34] มาใช้ในงานวิจัยชิ้นนี้เพื่อให้คะแนนตำแหน่งแกนกลางของพยางค์ของแต่ละตำแหน่ง โดยจะใช้ข้อมูลชุดทดสอบ 400 ประโยคจากชุดเสียงสมดุล (PD) โดยคัดเลือกเสียงผู้หญิง 200 ไฟล์แรก และคัดเลือกเสียงผู้ชาย 200 ไฟล์แรกจากฐานข้อมูลเสียงทดสอบ มาเป็นข้อมูลชุดฝึกสอน ซึ่งทั้งหมดประกอบด้วยกรอบเวลาของสัญญาณเสียงที่เป็นและไม่เป็นแกนกลางของพยางค์จำนวน 265,184 กรอบสัญญาณ

ส่วนลักษณะเด่นทางเสียงที่ใช้ทดสอบในการให้คะแนนนั้น ประกอบด้วย 2 แบบด้วยกันคือ

1.1 ค่าระดับความเข้มของเสียงที่ความถี่มากกว่า 300 เฮิรตซ์ และค่าอัตสหสัมพันธ์ที่มากที่สุดในช่วงความถี่ 60 ถึง 320 เฮิรตซ์ หรือภาวะความเป็นรายคาบของสัญญาณนั่นเอง [9]

1.2 ค่าพารามิเตอร์ทางเสียงสำแหรับภาษาไทยของสมบัติทางสวนสัทศาสตร์ที่เสนอโดย Rochkittichareon, et al. [26] ที่มีคุณสมบัติเป็น [Syllabic] ซึ่งประกอบด้วย

- ค่าพลังงานในช่วงความถี่ 640 ถึง 2800 เฮิรตซ์
- ค่าพลังงานในช่วงความถี่ 2000 ถึง 3000 เฮิรตซ์
- ค่าเข้มสูงสุดในช่วงความถี่ 0 ถึง 900 เฮิรตซ์
- ค่าอัตราส่วนของพลังงานที่ช่วงความถี่ 0 ถึง 400 เฮิรตซ์ และช่วงความถี่ 400 ถึง 6000 เฮิรตซ์

หลังจากนั้นจึงนำมาทดสอบกับชุดทดสอบเพื่อการพัฒนา ซึ่งแบ่งเป็นกรอบสัญญาณเสียงทั้งหมด 8,543 กรอบสัญญาณ เป็นกรอบสัญญาณที่เป็นแกนกลางของพยางค์ 6,782 และเป็นกรอบสัญญาณที่ไม่เป็นแกนกลางของพยางค์ 1,761 กรอบสัญญาณ ซึ่งได้ผลลัพธ์ ดังตารางที่ 5.1

ตารางที่ 5.1 ผลลัพธ์ของการทดสอบตัวจำแนกประเภทของการให้คะแนนตำแหน่งของแกนกลางพยางค์

	ชุดพารามิเตอร์ทางเสียงแบบแรก E มากกว่า 300 Hz, Periodicity	ชุดพารามิเตอร์ทางเสียงแบบที่สอง E[640,2800], E[2000,3000], Spectral Peak Ratio of E[0-400 Hz to E[400-6000]
ความแม่นยำ (%)	90.22%	86.23%
จำนวนกรอบเวลาที่ตอบว่าเป็นแกนกลางและอยู่ในช่วงของแกนกลาง	6574	6421
จำนวนกรอบเวลาที่ตอบว่าเป็นแกนกลางแต่ไม่อยู่ในช่วงของแกนกลาง	627	815

	ชุดพารามิเตอร์ทางเสียงแบบแรก E มากกว่า 300 Hz, Periodicity	ชุดพารามิเตอร์ทางเสียงแบบที่สอง E[640,2800], E[2000,3000], Spectral Peak Ratio of E[0-400 Hz to E[400-6000]
จำนวนกรอบเวลาที่ตอบว่าไม่เป็นแกนกลางแต่อยู่ในช่วงของแกนกลาง	208	361
จำนวนกรอบเวลาที่ตอบว่าไม่เป็นแกนกลางและไม่อยู่ในช่วงของแกนกลาง	1134	946

จากผลลัพธ์จะเห็นได้ค่อนข้างชัดเจนว่าตัวจำแนกประเภทที่เป็นชุดพารามิเตอร์ในแบบแรกนั้นให้ผลลัพธ์ที่ดีกว่า

2. การทดสองการระบุตำแหน่งของแกนกลางพยางค์

การทดลองนี้ จะทำการทดสอบประสิทธิภาพของระบุตำแหน่งแกนกลางพยางค์ โดยก่อนอื่นจะทำการปรับค่าพารามิเตอร์สำหรับขั้นตอนการหาจุดจบของพยางค์และวิธีคอนเว็กซ์ฮัลล์นั่น ก็คือ ค่าขีดแบ่งของพลังงานเพื่อใช้แยกระหว่างพลังประจำเสียงกับเสียงเงียบ และค่าขีดแบ่งของผลต่างระหว่างคอนเว็กซ์ฮัลล์ โดยใช้ข้อมูลชุดสำหรับทดสอบเพื่อการพัฒนาจำนวน 440 ไฟล์เสียง แบ่งเป็นเสียงผู้หญิงและเสียงผู้ชายอย่างละครึ่ง และหลังจากนั้นจึงประเมินผลด้วยข้อมูลสำหรับทดสอบเพื่อการประเมินผล 840 ไฟล์เสียงที่เหลืออยู่

การทดสอบนี้จะแบ่งการทดสอบตามลักษณะเด่นที่ใช้ออกเป็น 3 รูปแบบด้วยกันคือ

2.1 แบบที่ 1 ใช้ระดับความเข้มของเสียงที่ช่วงความถี่ 300 เฮิรตซ์ขึ้นไป เพื่อค้นหาตำแหน่งที่มีความเข้มที่มากที่สุด (Peak) และตัดสินใจให้ตำแหน่งนั้นเป็นแกนกลางของพยางค์ทันที

2.2 แบบที่ 2 ใช้ระดับความเข้มของเสียงที่ช่วงความถี่ 300 เฮิรตซ์ขึ้นไป เพื่อค้นหาความเข้มที่มากที่สุด (Peak) แล้วให้คะแนนตำแหน่งนั้นด้วยการสกัดลักษณะเด่นตามชุดพารามิเตอร์ทางเสียงแบบแรก

2.3 แบบที่ 3 ใช้ระดับความเข้มของเสียงที่ช่วงความถี่ 300 เฮิรตซ์ขึ้นไป เพื่อค้นหาความเข้มที่มากที่สุด (Peak) แล้วให้คะแนนตำแหน่งนั้นด้วยการสกัดลักษณะเด่นตามชุดพารามิเตอร์ทางเสียงแบบที่สอง

ก่อนเริ่มการทดสอบประสิทธิภาพนั้น จะต้องทราบพารามิเตอร์ 4 ค่า ซึ่งก็คือค่า Energy Threshold, Dip Threshold, Window Length และ Interval ดังตารางที่ 5.2 ค่าพารามิเตอร์ 2 ค่าแรก คือ ค่าขีดแบ่งซึ่งเป็นพารามิเตอร์ที่ใช้ในขั้นตอนวิธีคอนเว็กซ์ฮัลล์ เพื่อค้นหาตำแหน่งที่ค่าระดับความเข้มของเสียงมีค่าสูงที่สุดและตัดแบ่งค่าระดับความเข้มออกเป็นส่วนย่อยดังที่อธิบายไว้ในบทที่ 3 ในส่วนของ Window Length และ Interval นั้น ใช้สำหรับการคำนวณลักษณะเด่นต่าง ๆ ของเสียงตามแกนเวลา ซึ่งค่าที่ใช้กันทั่วไป คือ 25 มิลลิวินาที และมีช่วงซ้อนทับกัน 15 มิลลิวินาที ตามการตั้งค่าของ Xie [23]

ตารางที่ 5.2 คำอธิบายของพารามิเตอร์ในการระบุแกนกลางพยางค์

ค่าพารามิเตอร์	คำอธิบาย
Energy Threshold	ค่าขีดแบ่งของระดับความเข้มของเสียงระหว่างพลังประจำเสียง (Sonority) และเสียงเงียบ (Silence)
Dip Threshold	ค่าขีดแบ่งของผลต่างระหว่างคอนเว็กซ์ฮัลล์ (Convex hull) และค่าระดับความเข้มของเสียง
Window Length	ขนาดของกรอบเวลาสั้น ๆ ของสัญญาณเสียง
Interval	ช่วงทับซ้อนของกรอบสัญญาณ

หลังจากนั้น เมื่อได้ค่าพารามิเตอร์ที่ดีที่สุดแล้ว จึงนำมาทดสอบการระบุตำแหน่งแกนกลางของพยางค์ ซึ่งได้ผลลัพธ์ดังตารางที่ 5.3

ตารางที่ 5.3 ค่าพารามิเตอร์ที่ดีที่สุดสำหรับลักษณะเด่นแต่ละแบบ

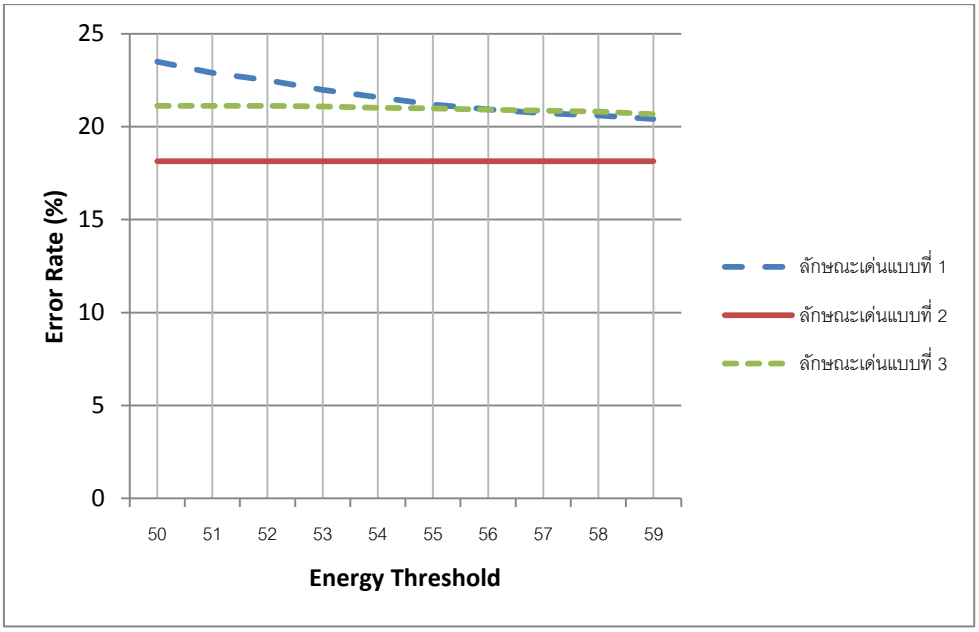
ลักษณะเด่น	ค่าพารามิเตอร์	ค่า
แบบที่ 1	Energy Threshold	60 เดซิเบล
	Dip Threshold	10 เดซิเบล
แบบที่ 2	Energy Threshold	56 เดซิเบล
	Dip Threshold	6 เดซิเบล
แบบที่ 3	Energy Threshold	60 เดซิเบล
	Dip Threshold	8 เดซิเบล

จากนั้นใช้ข้อมูลที่เหลือในชุด PD จำนวน 840 ไฟล์เสียง ในการทดสอบประสิทธิภาพในส่วนการตรวจหาพยางค์ ซึ่งได้ผลลัพธ์ดังตารางที่ 5.4

ตารางที่ 5.4 ผลลัพธ์ของการระบุตำแหน่งแกนกลางพยางค์

ลักษณะเด่น	เปอร์เซ็นต์ความแม่นยำ (Accuracy)	เปอร์เซ็นต์ตำแหน่งขาด (Deletion)	เปอร์เซ็นต์ตำแหน่งเกิน (Insertion)	เปอร์เซ็นต์ความผิดพลาด (Error Rate)	เวลาในการคำนวณเฉลี่ยต่อหนึ่งไฟล์เสียง (มิลลิวินาที)
แบบที่ 1	82.02%	11.66%	6.30%	17.98%	64.39
แบบที่ 2	84.70%	7.86%	7.44%	15.30%	166.62
แบบที่ 3	82.65%	10.97%	6.38%	17.35%	152.79

จากตารางที่ 5.4 จะเห็นว่า พารามิเตอร์แบบที่ 2 นั้นให้เปอร์เซ็นต์ความถูกต้องมากที่สุด ถึงแม้ว่าค่าความถูกต้องจะไม่แตกต่างกันมากนัก แต่เมื่อนำค่าที่ใช้เป็นข้อมูลทดสอบเพื่อการพัฒนามาพล็อตเป็นกราฟได้ข้อมูลหนึ่งที่น่าสนใจ ดังรูปที่ 5.1



รูปที่ 5.1 กราฟแสดงเปอร์เซ็นต์ความผิดพลาดจากการระบุพยางค์และ Energy Threshold

จากภาพจะเห็นว่าลักษณะเด่นในแบบที่ 2 นั้น มีค่าคงที่ ไม่ว่าจะปรับค่าพารามิเตอร์ Energy Threshold เป็นค่าใดก็ตาม เมื่อเทียบกับลักษณะเด่นในแบบที่ 1 เมื่อไม่มีการให้คะแนนทำให้ผลลัพธ์นั้นเร็วขึ้นมาก แต่ต้องอาศัยการปรับพารามิเตอร์ที่เหมาะสมเข้าช่วย ซึ่งมีข้อเสียในเรื่องของความดังเบาของผู้พูด เพราะต้องคอยปรับ ส่วนลักษณะเด่นในแบบที่ 3 นั้น ถึงแบบจะมีจำนวนลักษณะเด่นเยอะ แต่ได้ผลลัพธ์ออกมาไม่ดีเท่าที่ควร เพราะความแม่นยำในส่วนของการให้คะแนนน้อยกว่า จึงสรุปได้ว่าลักษณะเด่นแบบที่ 2 นั้นได้ความถูกต้องมากที่สุด และลักษณะเด่นนี้ไม่ขึ้นกับความดังเบาของเสียงมากเท่าไร เพราะมีการใช้ความเป็นรายคาบเข้าช่วย จึงน่าจะมีความเป็นอิสระต่อผู้พูด และทนทานต่อสัญญาณรบกวนได้ดีกว่า ดังนั้น เราจะใช้ลักษณะเด่นแบบที่ 2 ในขั้นตอนวิธีของการประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความต่อไป

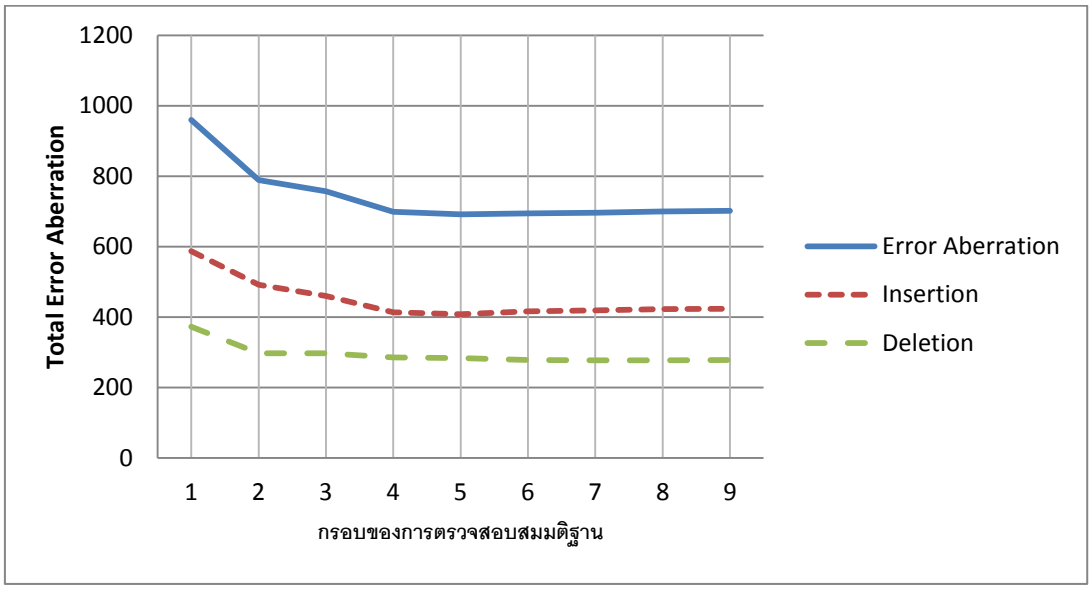
การประเมินผลการประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความ

1. **การศึกษาค่าพารามิเตอร์**

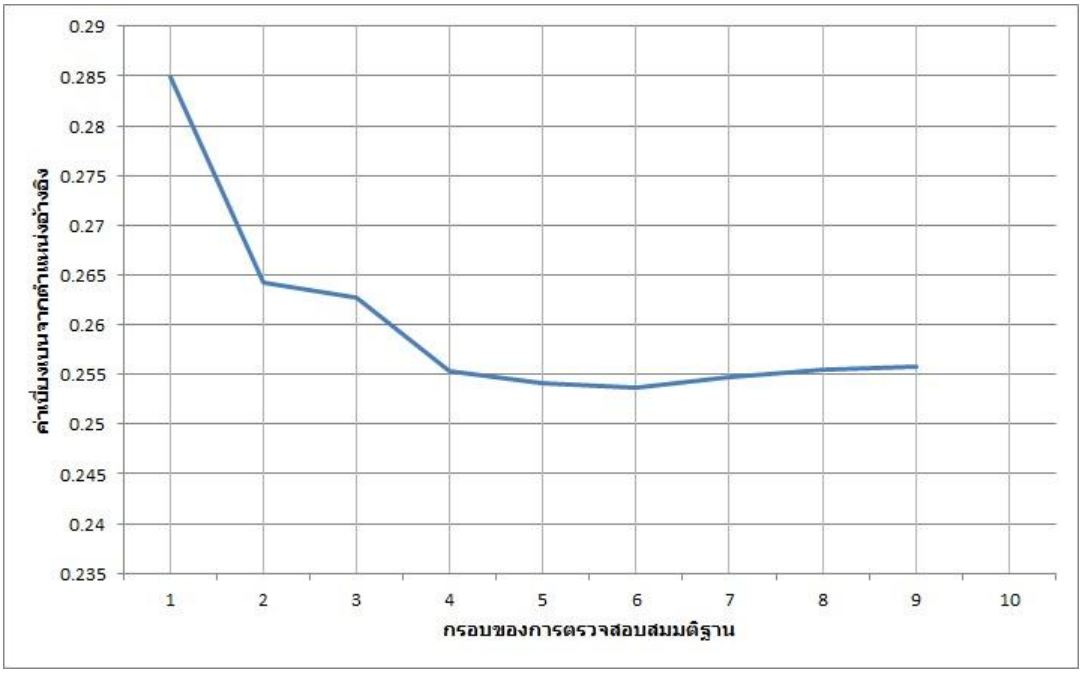
ดังที่กล่าวไว้ในบทที่ 4 ขั้นตอนนี้ การประเมินผลในขั้นนี้ ในอันดับแรก จำเป็นจะต้องรู้ค่าพารามิเตอร์ 2 ค่าเพิ่มเติม คือ กรอบของการตรวจสอบสมมติฐาน และขอบเขตจำนวนพยางค์

ของสมมติฐาน (H_n) โดยจะนำฐานข้อมูลเสียงชุดหน่วยเสียงสมดุล 30% ซึ่งประกอบด้วยเสียงผู้ชายและผู้หญิงจำนวน 504 ไฟล์เสียงแรก เพื่อใช้เป็นข้อมูลชุดสำหรับทดสอบเพื่อการพัฒนา

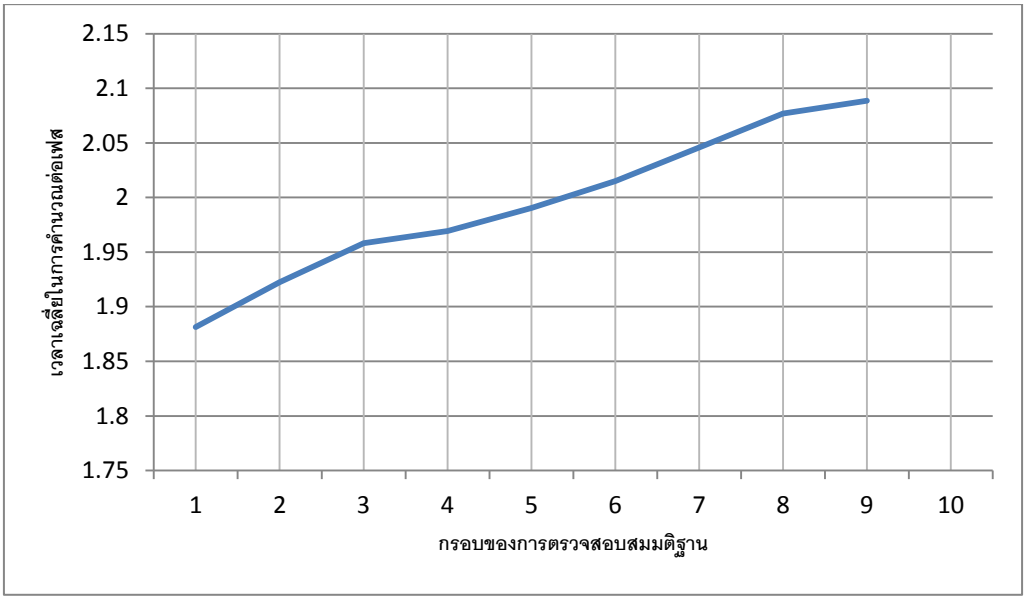
ในขั้นแรก กรอบของการตรวจสอบสมมติฐานได้ถูกนำมาทดสอบ ตั้งแต่ 1 ถึง 9 กรอบย้อนหลัง ซึ่งได้ผลลัพธ์ดังรูปที่ 5.2, 5.3, 5.4



รูปที่ 5.2 กราฟแสดงค่าความผิดพลาดของการประสานเวลาและกรอบของการตรวจสอบสมมติฐาน



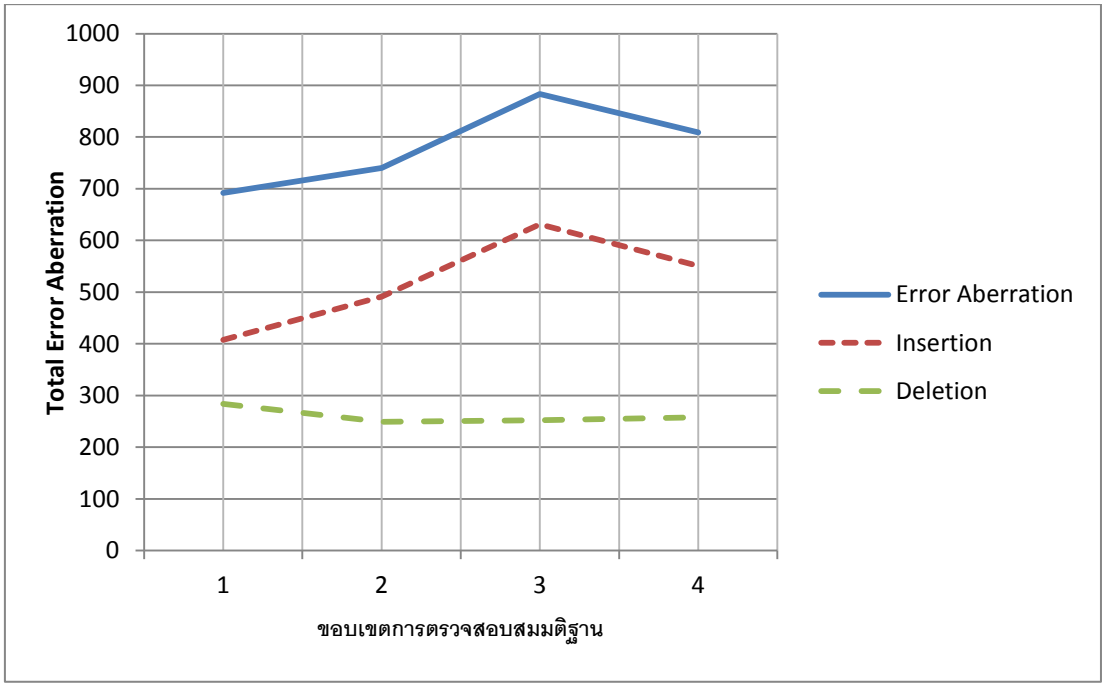
รูปที่ 5.3 กราฟแสดงค่าเบี่ยงเบนจากตำแหน่งอ้างอิงและกรอบของการตรวจสอบสมมติฐาน



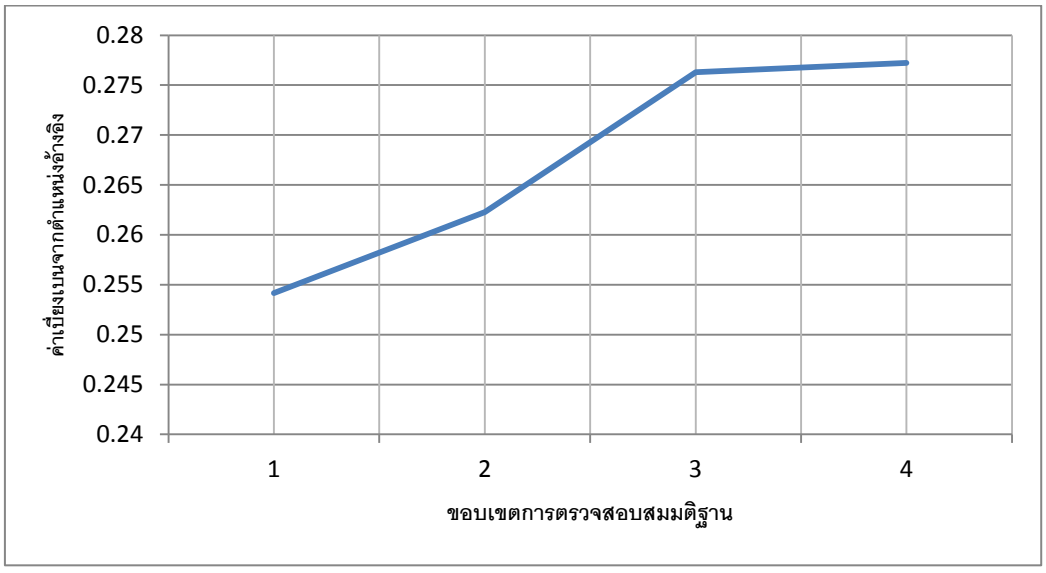
รูปที่ 5.4 กราฟแสดงเวลาเฉลี่ยในการคำนวณต่อเฟสและกรอบของการตรวจสอบสมมติฐาน

จากรูปที่ 5.2 จะเห็นว่า เมื่อกรอบของการตรวจสอบสมมติฐานยิ่งใช้กรอบย้อนกลับไป ซึ่งจะทำให้ความยาวของหน่วยเสียงของข้อความที่นำมาต่อกันนั้นยาวมากขึ้น มีผลทำให้ผลของการประสานเวลานั้นมีความถูกต้องมากขึ้น โดยค่าความผิดพลาดนั้นมีค่าลดลงทั้งในส่วนของความผิดพลาดแบบ Deletion และ Insertion จนกระทั่งลู่เข้าหาค่าหนึ่งที่กรอบการตรวจสอบประมาณ 5 กรอบย้อนหลัง ซึ่งสอดคล้องกับผลจากรูปที่ 5.3 ซึ่งค่าความเบี่ยงเบนจากตำแหน่งอ้างอิงนั้นก็ลดลงจนลู่เข้าสู่ค่าหนึ่งที่กรอบการตรวจสอบสมมติฐาน 5 กรอบเช่นกัน และจากรูปที่ 5.4 จะเห็นได้ว่า ถ้าหากยิ่งเพิ่มกรอบการตรวจสอบของสมมติฐานให้มากขึ้น เวลาที่ใช้ในการคำนวณก็จะยิ่งมากขึ้นตามไปด้วย

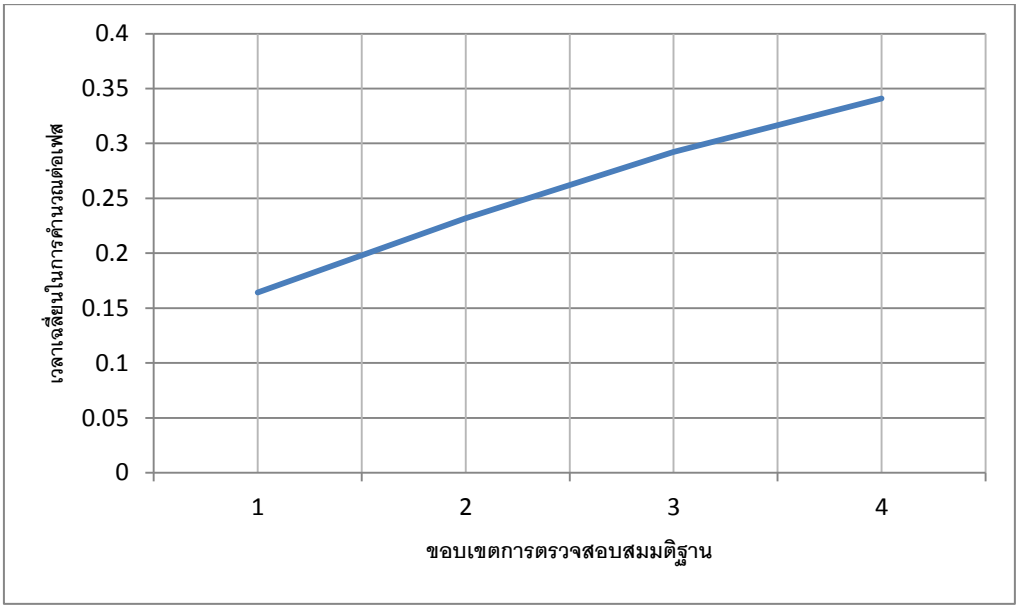
ในส่วนของขอบเขตการตรวจสอบสมมติฐานนั้น เมื่อทดลองตั้งแต่ $n = 1$ จนถึง $n = 4$ พบว่าได้ผลลัพธ์ดังรูปที่ 5.5, 5.6, 5.7



รูปที่ 5.5 กราฟแสดงค่าความผิดพลาดของการประสานเวลาและขอบเขตการตรวจสอบสมมติฐาน



รูปที่ 5.6 กราฟแสดงค่าเบี่ยงเบนจากตำแหน่งอ้างอิงและขอบเขตการตรวจสอบสมมติฐาน



รูปที่ 5.7 กราฟแสดงเวลาเฉลี่ยในการคำนวณและขอบเขตการตรวจสอบสมมติฐาน

จากรูปที่ 5.5, 5.6, 5.7 จะเห็นได้ว่า ขอบเขตการตรวจสอบสมมติฐานมีค่าเป็น $n = 1$ จะให้ผลลัพธ์ดีที่สุดทั้งในแง่ของความถูกต้องและเวลา

ดังนั้นค่าพารามิเตอร์ทั้งสองที่จะใช้ทดสอบนั้น จะมีค่าพารามิเตอร์ดังตาราง มีค่ากรอบการตรวจสอบสมมติฐาน 5 กรอบย้อนกลับ และขอบเขตการตรวจสอบสมมติฐานเป็น $n = 1$

ส่วนระบบอ้างอิงแบบที่ 2 นั้น ใช้ค่ากรอบของการตรวจสอบสมมติฐานเป็น 5 กรอบเช่นเดียวกัน แต่ขอบเขตการตรวจสอบสมมติฐานเป็น $n = 3$ เพราะได้ผลลัพธ์ที่ดีกว่า

2. การทดสอบการประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความ

เมื่อนำพารามิเตอร์ที่ได้ ไปทดลองกับชุดสำหรับทดสอบเพื่อการประเมินผล ซึ่งข้อมูลชุดนี้คือ ชุดหน่วยเสียงสมดุลที่เหลืออีก 70% หรืออีกจำนวน 1176 ไฟล์เสียง ซึ่งเป็นเสียงผู้ชายที่เหลือจำนวน 588 ไฟล์เสียงและเสียงผู้หญิงที่เหลือจำนวน 588 ไฟล์เสียงเปรียบเทียบกับระบบอ้างอิงทั้ง 2 แบบ ได้แก่

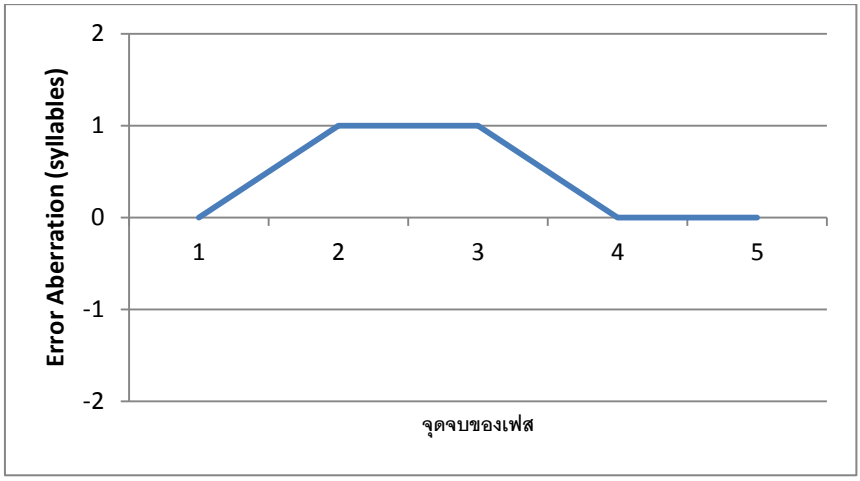
- แบบที่ 1 ที่เป็นแบบใช้หลักการตรวจจับพยางค์เพียงอย่างเดียว
- แบบที่ 2 ใช้หลักการตรวจจับความไม่ตรงกันของการถอดเสียงเพียงอย่างเดียว

ซึ่งได้ผลลัพธ์ดังตารางที่ 5.5

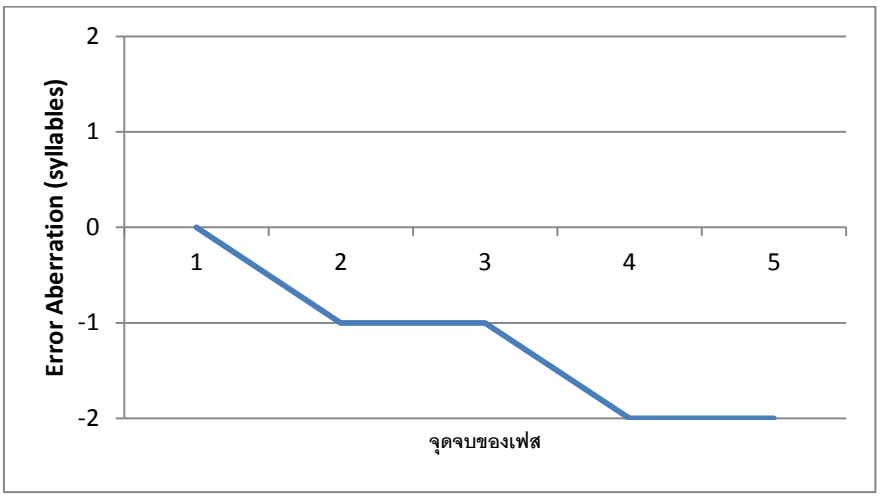
ตารางที่ 5.5 ผลลัพธ์ของการประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความ

	Total Error Aberration	Total Deletion	Total Insertion	Average Ref. Deviation	Max / Min Error Aberration	Average Runtime per Phrase (seconds)
ขั้นตอนวิธีที่นำเสนอ	2,431	1,363	1,068	0.2904	19 / -6	0.168
ระบบอ้างอิงแบบที่ 1	9,926	4,031	5,895	0.7294	9 / -11	0.020
ระบบอ้างอิงแบบที่ 2	2,445	1,216	1,229	0.4386	20 / -3	0.265

จากตารางที่ 5.5 จะเห็นได้ว่า ขั้นตอนวิธีที่นำเสนอมีค่า Total Error Aberration น้อยที่สุดเมื่อเทียบกับระบบอ้างอิงแบบที่ 1 และแบบที่ 2 สำหรับระบบอ้างอิงแบบที่ 1 นั้น ที่ค่า Total Error Aberration เยอะเพราะว่าเมื่อเกิดความผิดพลาดขึ้นแล้ว ความผิดพลาดนั้นจะถูกสะสมไปเรื่อย ๆ ถ้าหากไม่เกิดความผิดพลาดในทางกลับกัน ดังรูปที่ 5.8 คือตัวอย่างเมื่อเกิดความผิดพลาดขึ้นในทางใดทางหนึ่ง แล้วจากนั้นจึงเกิดความผิดพลาดในทางตรงกันข้ามขึ้น ทำให้ผลลัพธ์ของการประสานเวลานั้น กลับมาอยู่ที่ตำแหน่งที่ถูกต้องอีกครั้ง แต่ถ้าหากเมื่อเกิดความผิดพลาดขึ้น แล้วเกิดความผิดพลาดขึ้นอีกในทิศทางเดียวกัน ดังรูปที่ 5.9 ค่า Total Error Aberration ก็จะยิ่งถูกสะสมให้เยอะมากขึ้น ดังนั้น จึงสรุปได้ว่า การใช้ข้อมูลทางเสียงในการประสานเวลานั้นไม่เพียงพอ แต่เมื่อมีส่วนในการตรวจจับความไม่ตรงกันของการถอดเสียง ซึ่งอาศัยข้อมูลจากข้อความเข้ามาช่วยในการให้คะแนนสมมติฐานที่สร้างขึ้นมาใหม่ ทำให้การประสานเวลานั้นมีความถูกต้องมากขึ้นถึง 75.51% ด้วยกัน



รูปที่ 5.8 ตัวอย่างเมื่อเกิดความผิดพลาดแล้วเกิดความผิดพลาดกลับไปในทิศทางตรงกันข้าม



รูปที่ 5.9 ตัวอย่างเมื่อเกิดความผิดพลาดแล้วเกิดความผิดพลาดไปในทางทิศเดียวกัน

ในขณะเดียวกันเมื่อเปรียบเทียบกับระบบอ้างอิงแบบที่ 2 ที่ใช้การตรวจจับความไม่ตรงกันของการถอดเสียงเพียงอย่างเดียว จะเห็นว่าผลลัพธ์ที่ได้นั้นมีค่าใกล้เคียงกันมาก แต่เนื่องจากระบบอ้างอิงแบบที่ 2 นั้นจะทำการตรวจจับทุก ๆ 1 วินาที ซึ่งจะทำให้ได้จำนวนของการคำนวณ นั้นมีจำนวนน้อยกว่าขั้นตอนวิธีที่นำเสนอ ยกตัวอย่างเช่น ประโยคทดสอบประโยคหนึ่งมีความยาว 9 วินาทีระบบอ้างอิงแบบที่ 2 จะมีการคำนวณการประสานเวลาจำนวน 9 ครั้ง ในขณะที่ประโยคเดียวกัน ขั้นตอนวิธีที่นำเสนออาจจะมีจำนวนของการคำนวณมากกว่าหรือน้อยกว่า ขึ้นอยู่กับการตรวจหาจุดจบของพยางค์ว่าทำได้ละเอียดแค่ไหน ซึ่งถ้าหากนำ Total Error Aberration โดยการ

นำมาหารด้วยจำนวนครั้งที่เกิดการคำนวณ ก็จะทำให้เป็นบรรทัดฐานเดียวกันได้ แล้วรวมค่านี้เป็นค่าความผิดพลาดใหม่ ซึ่งจะได้ดังตารางที่ 5.6

ตารางที่ 5.6 ผลลัพธ์ที่ปรับบรรทัดฐานของการประสานเวลาอัตโนมัติระหว่างเสียงและข้อความแบบทันที

	Total Error Aberration	Total Error Aberration per Number of Phrases
ขั้นตอนวิธีที่นำเสนอ	2,431	165.10
ระบบอ้างอิงแบบที่ 1	9,926	686.20
ระบบอ้างอิงแบบที่ 2	2,445	284.23

จากที่รางที่ 5.6 ก็จะเห็นได้ว่า ขั้นตอนวิธีที่นำเสนอนั้นมีความผิดพลาดน้อยกว่าระบบอ้างอิงแบบที่ 2 อยู่ 41.56% เพราะระบบอ้างอิงแบบที่ 2 นั้นเป็นการคำนวณโดยคาดเดาจากอัตราเร็วในการพูด แล้วพยายามสร้างจำนวนรูปแบบขอสมมติฐานให้ครอบคลุม ทำให้ผลลัพธ์ที่ได้นั้นใช้เวลาค่อนข้างมาก เพราะใช้การสร้างสมมติฐานเป็นค่า $n = 3$ เพื่อให้มั่นใจว่าจะครอบคลุมคำตอบซึ่งที่เป็นได้ในสมมติฐานที่ตั้งขึ้น ในขณะที่ขั้นตอนวิธีที่นำเสนอนั้น ใช้หลักการตรวจหาพยางค์เพื่อจำนวนที่ได้ก่อน ซึ่งใช้เวลาน้อยมาก เมื่อเทียบกับขั้นตอนการตรวจจับความไม่ตรงกันของการถอดเสียง ซึ่งวิธีที่นำเสนอนั้นสร้างค่าสมมติฐานที่ $n = 1$ เท่านั้นก็เพียงพอ ทำให้เวลาการคำนวณนั้นเร็วกว่าระบบอ้างอิงแบบที่ 2 ค่อนข้างมาก

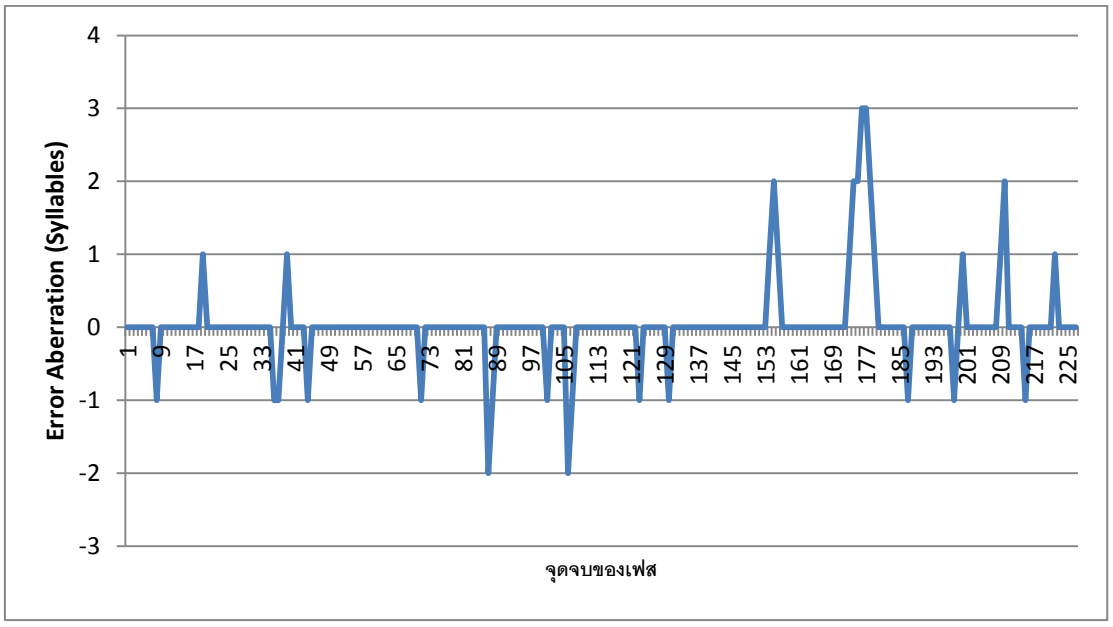
ในส่วนของค่าเบี่ยงเบนจากตำแหน่งอ้างอิง (Reference Deviation) นั้นสามารถบอกได้ว่าความผิดพลาดที่เกิดขึ้นนั้น มีค่าเฉลี่ยที่ผิดไปจากตำแหน่งอ้างอิงอยู่กี่พยางค์ ซึ่งก็ได้ผลลัพธ์ที่เป็นไปในทางเดียวกันกับ Total Error Aberration คือ ขั้นตอนวิธีที่นำเสนอนั้น ระบบอ้างอิงแบบที่ 1 และระบบอ้างอิงแบบที่ 2 มีค่าเฉลี่ยอยู่ที่ 0.29, 0.73 และ 0.44 ตามลำดับ แต่ในขณะเดียวกัน ค่า Error Aberration สูงที่สุดนั้น มีค่าที่ค่อนข้างใกล้เคียงกัน คือ ขั้นตอนวิธีที่นำเสนอนั้น ระบบอ้างอิงแบบที่ 1 และระบบอ้างอิงแบบที่ 2 มีค่าที่มากที่สุดอยู่ที่ 19, 11 และ 20 พยางค์ตามลำดับ (ไม่ว่าจะเกิดความผิดพลาดไปในทิศทางไหนก็ตาม) ซึ่งเป็นความผิดพลาดของการประสานเวลาที่มาก

ผิดปกติ ดังนั้นจากผลการทดลองของขั้นตอนวิธีที่นำเสนอ จึงทำการค้นหาประโยคทดสอบที่มีค่าเบี่ยงเบนจากตำแหน่งอ้างอิงเกิน 1 ขึ้นไป เพื่อค้นหาประโยคที่เกิดการประสานเวลาที่ผิดปกติมาก ซึ่งทำให้ทราบว่า มีประโยคทดสอบจำนวน 41 ประโยค จาก 1,176 ประโยค ซึ่งคิดเป็น 3.49% ของประโยคทดสอบทั้งหมดซึ่งมีค่าเบี่ยงเบนจากตำแหน่งอ้างอิงเกิน 1 ขึ้นไป ส่วนประโยคทดสอบที่เหลือนั้น มีค่า Error Aberration ที่มากที่สุดไม่เกิน 3 พยางค์และสามารถถูกปรับกลับมาสู่ตำแหน่งอ้างอิงได้อย่างถูกต้อง ภายในไม่กี่ครั้งของการคำนวณครั้งถัดไป

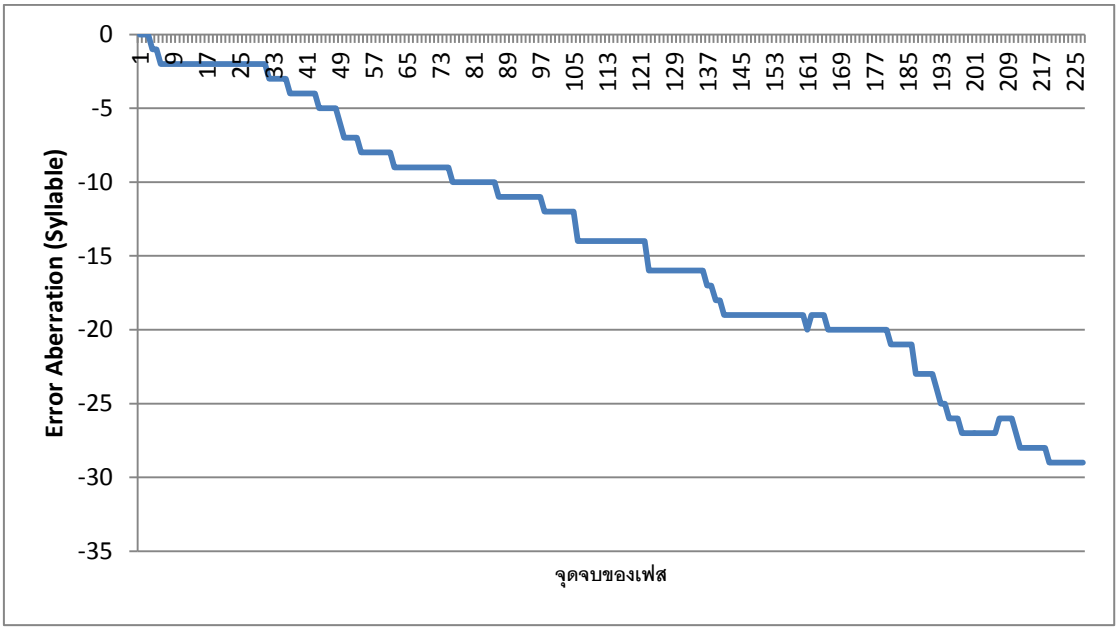
ในเรื่องของเวลาการคำนวณนั้น ทั้ง 3 แบบนั้นสามารถคำนวณได้ในทันที (Real-time) เพราะข้อมูลชุดทดสอบเพื่อการประเมินผลชุดนี้ มีอัตราเร็วในการพูดต่อหนึ่งพยางค์อยู่ที่ 404.20 มิลลิวินาที ซึ่งแม้ว่าขั้นตอนวิธีที่นำเสนอนั้น จะใช้เวลามากที่สุด คือ ประมาณ 265 มิลลิวินาที ก็ยังสามารถคำนวณได้ทันก่อนที่ผู้พูดจะพูดพยางค์ต่อไปจบ

3. การทดสอบกับข้อมูลชุดทดสอบสภาพแวดล้อมจริง

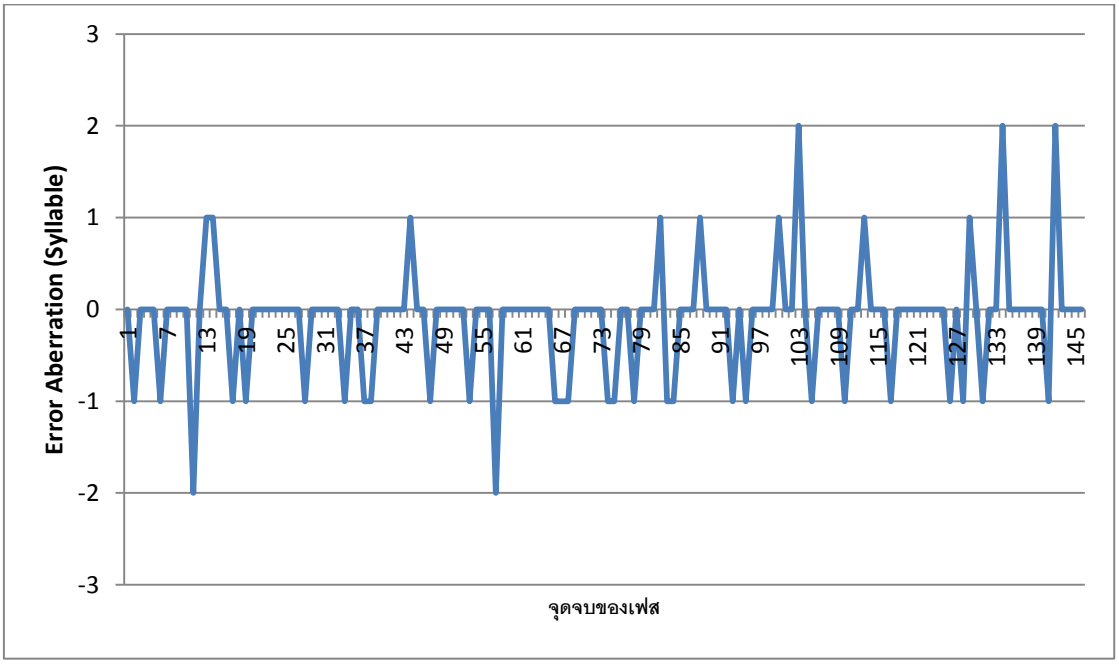
ดังที่กล่าวไว้ในบทที่ 4 ข้อมูลชุดทดสอบชุดสุดท้ายถูกเตรียมไว้เพื่อทดสอบลักษณะข้อมูลที่จะถูกพบจริง นั่นคือ ข้อมูลที่มีความยาวมาก เพื่อทดสอบดูว่าขั้นตอนวิธีที่นำเสนอจะสามารถนำไปใช้ในข้อมูลจริง ๆ ได้หรือไม่ ซึ่งเป็นข้อความที่คัดลอกบทความหนึ่งจากอินเตอร์เน็ต มีความยาวประมาณหนึ่งหน้าหนังสือ ซึ่งได้ผลลัพธ์ดังนี้



รูปที่ 5.10 กราฟแสดงผลลัพธ์การประสานเวลาของขั้นตอนวิธีที่นำเสนอ



รูปที่ 5.11 กราฟแสดงผลลัพธ์การประสานเวลาของระบบอ้างอิงที่ 1



รูปที่ 5.12 กราฟแสดงผลลัพธ์การประสานเวลาของระบบอ้างอิงที่ 2

จากผลลัพธ์ของกราฟทั้ง 3 ระบบนั้น สรุปได้ว่า ขั้นตอนวิธีที่นำเสนอนั้นและระบบอ้างอิงที่ 2 นั้น สามารถนำมาใช้ในการประสานเวลาอัตโนมัติแบบทันทีระหว่างเสียงและข้อความได้จริงกับ

ข้อมูลจริง แต่ระบบอ้างอิงที่ 1 นั้นไม่สามารถใช้ได้จริง เพราะผลลัพธ์ของการประสานเวลานั้นหลุดไปจากเส้นแนวอ้างอิงไปมาก ทำให้เกิดความผิดพลาดมหาศาล ในขณะที่ขั้นตอนวิธีที่นำเสนอและระบบอ้างอิงที่ 2 นั้นสามารถประสานเวลาตามเส้นอ้างอิงไปได้เรื่อย ๆ ถึงแม้จะเกิดความผิดพลาดขึ้นบ้างเล็กน้อย

ถ้าเทียบประสิทธิภาพระหว่างขั้นตอนวิธีที่นำเสนอและระบบอ้างอิงที่ 2 ในบรรทัดฐานเดียวกันแล้ว ได้ผลลัพธ์ดังตารางที่ 5.7

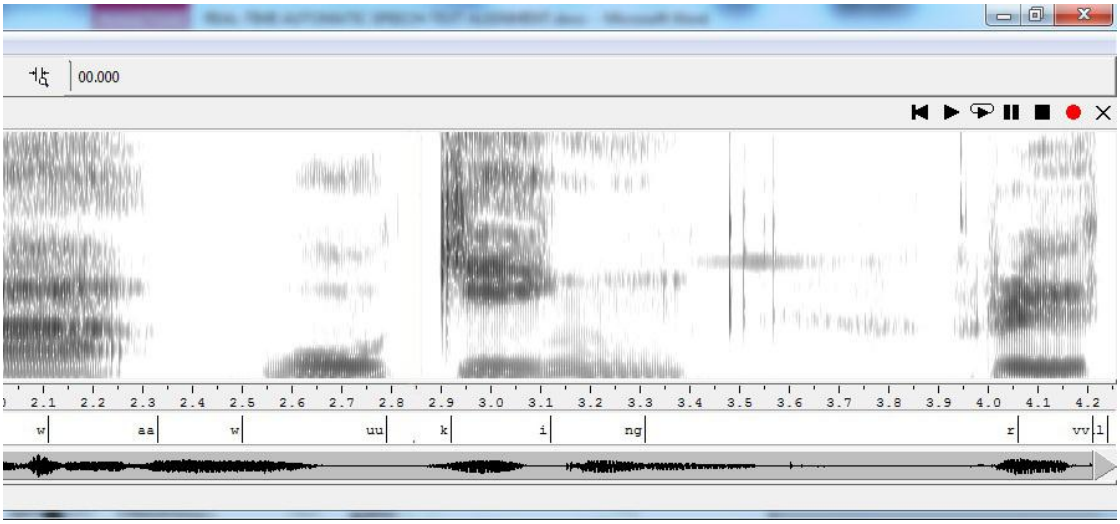
ตารางที่ 5.7 ผลลัพธ์ของการประสานเวลาของข้อมูลชุดทดสอบสภาพแวดล้อมจริง

	Total Error Aberration	Total Error Aberration per Number of Phrases
ขั้นตอนวิธีที่นำเสนอ	42	0.185
ระบบอ้างอิงแบบที่ 2	45	0.308

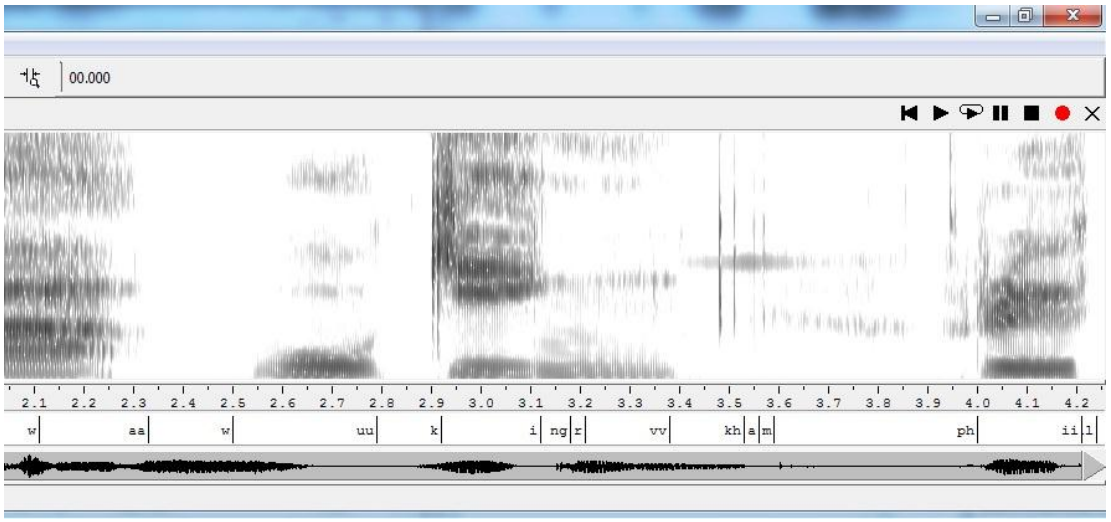
จากตารางที่ 5.7 จะพบว่าขั้นตอนวิธีที่นำเสนอนั้นมีประสิทธิภาพดีกว่าระบบอ้างอิงแบบที่ 2 อยู่ 39.93% ซึ่งใกล้เคียงกับความแตกต่างของผลลัพธ์ที่ได้ในการทดสอบจากข้อมูลชุดทดสอบเพื่อประเมินผลจากฐานข้อมูลเสียงโลตัส และการให้คำตอบของการประสานเวลานั้น ขั้นตอนวิธีที่นำเสนอสามารถให้ผลลัพธ์ที่ละเอียดกว่า เพราะขั้นตอนวิธีที่นำเสนอนั้นมีจำนวนเฟส 227 เฟส ส่วนระบบอ้างอิงแบบที่ 2 นั้นมีจำนวนเฟสเพียง 146 เฟส ซึ่งเกิดจากการประมวลผลลัพธ์ทุก ๆ 1 วินาที หมายความว่าผู้ใช้จะเห็นผลลัพธ์ของการประสานเวลาทุก ๆ 1 วินาทีเช่นกัน ในขณะที่ขั้นตอนวิธีที่นำเสนอจะเห็นผลลัพธ์ได้ละเอียดกว่า

4. วิเคราะห์ความผิดพลาดของผลลัพธ์

เมื่อเกิดความผิดพลาดขึ้นนั้น ส่วนใหญ่ มักจะเกิดจากส่วนของการตรวจหาความไม่ตรงกันของการถอดเสียงนั้นทำการประสานเวลาเสียงและให้คะแนนที่ผิดพลาด ซึ่งจุดที่มักเกิดความผิดพลาดในลักษณะนี้ จะเป็นจุดที่ผู้อ่านนั้นหยุดหายใจ ซึ่งสามารถอธิบายได้จากรูปที่ 5.13 และ 5.14



รูปที่ 5.13 ผลลัพธ์ของการประสานเวลาเสียงของสมมติฐานที่ถูกต้องใน WaveSurfer



รูปที่ 5.14 ผลลัพธ์ของการประสานเวลาเสียงของสมมติฐานที่ผิดใน WaveSurfer

รูปที่ 5.13 และรูปที่ 5.14 เป็นตัวอย่างเฟสหนึ่งในประโยคทดสอบในข้อมูลชุดทดสอบเพื่อการประเมินผล ข้อมูลเสียงนั้นพูดว่า “ว่า วูกิง หรือ” เมื่อกำหนดขอบเขตของสมมติฐานให้ $n = 1$ ซึ่งจะได้ 3 สมมติฐาน ผลลัพธ์ปรากฏว่า ข้อมูลเสียงระหว่าง “วูกิง” และ “หรือ” เกิดการหยุดหายใจของผู้อ่านขึ้น โดยที่เราไม่ทราบว่าผู้พูดจะหยุดหายใจที่ไหน การถอดคำจึงไม่มีจังหวะการหยุดหายใจ ช่วงที่หยุดหายใจนั้น ป้ายกำกับหน่วยเสียงของพยัญชนะตัวแรกหลังจากการหยุดหายใจ ซึ่งก็คือ ‘r’ ของคำว่า “หรือ” นั้น เมื่อพิจารณารูปที่ 5.13 ซึ่งเป็นสมมติฐานที่ควรจะถูกต้อง จะเห็นว่าป้ายกำกับหน่วยเสียงของ ‘r’ มีช่วงเวลาที่ยาวมาก ซึ่งผิดวิสัยของลักษณะหน่วยเสียงพยัญชนะที่

ควรจะมีช่วงเวลาในการออกเสียงที่สั้น จึงทำให้คะแนนที่ได้ออกมาจากการประสานเวลาเสียงในจุดนี้มีค่าน้อยมาก ในขณะที่รูปที่ 5.14 ซึ่งเป็นสมมติฐานที่ควรจะผิด แต่หน่วยเสียงแต่ละหน่วยเสียงถูกบีบอัดเข้าไปเนื่องจากกระบวนการประสานเวลาเสียง หน่วยเสียงพยัญชนะจึงมีค่าที่สั้นดังที่ควรจะเป็น เมื่อรวมผลรวมของค่าคะแนนของทุกหน่วยเสียงแล้ว จึงมีผลให้สมมติฐานที่ควรจะถูกต้องมีค่าน้อยกว่าสมมติฐานที่ผิด เป็นเหตุให้เกิดความผิดพลาดขึ้น

บทที่ 6

ข้อสรุปและข้อเสนอแนะ

ข้อสรุป

งานวิจัยชิ้นนี้ได้นำเสนอวิธีการในการประสานเวลาอัตโนมัติระหว่างเสียงและข้อความแบบทันทีที่ได้รับข้อมูลเสียงพูดเข้ามาในระบบ โดยวิธีการที่นำเสนอนั้นจะทำการประสานเวลาในความละเอียดระดับพยางค์ในภาษาไทย โดยการศึกษาเบื้องต้นเกี่ยวกับการตรวจหาพยางค์จากเสียงพูด เพื่อที่จะสามารถระบุจำนวนพยางค์ที่พูดออกไป แต่เนื่องจากเราทราบข้อมูลที่เป็นข้อความที่จะพูดทั้งหมด จึงใช้หลักการตรวจจับความไม่ตรงกันของการถอดเสียง เข้ามาประยุกต์ใช้เพื่อตรวจสอบและให้คะแนนคำตอบของส่วนการตรวจหาพยางค์อีกครั้ง เพื่อให้ได้ผลลัพธ์ที่แม่นยำมากขึ้น ขั้นตอนวิธีที่นำเสนอสามารถนำไปประยุกต์ใช้ได้กลับโปรแกรมประยุกต์ที่มีลักษณะดังกล่าว เช่น โปรแกรมการสร้างหนังสือเสียง เพราะมีข้อความอยู่ทั้งหมด และต้องการประสานเวลาบันทึกเสียงพูดให้ตรงกับข้อความที่แสดงอยู่ เนื่องจากภาษาไทยไม่ทราบจุดจบของประโยคและคำที่ชัดเจน จึงเลือกใช้การประสานเวลาในระดับพยางค์ เพราะพยางค์เป็นสิ่งที่สามารถระบุได้ทั้งในเสียงและข้อความ เพื่อให้ได้ความละเอียดการประสานเวลาและยังสามารถแสดงให้ผู้ใช้เห็นความเคลื่อนไหวของการประสานเวลาแบบทันทีได้อีกด้วย

ขั้นตอนวิธีในการประสานเวลาในวิทยานิพนธ์นี้ ประกอบไปด้วย 2 ส่วนใหญ่ ๆ คือ

1. การตรวจหาพยางค์

ในขั้นตอนนี้เริ่มจากการศึกษาหลักการตรวจหาพยางค์จากสัญญาณเสียง เนื่องจากภาษาไทยเป็นภาษาที่หน่วยที่เล็กที่สุดของการเปล่งเสียงออกมาคือพยางค์ และทุกพยางค์ต้องประกอบไปด้วยสระ ดังนั้นการตรวจหาพยางค์จึงใช้หลักการเดียวกันกับการตรวจหาสระได้ด้วย ซึ่งถ้าหากเราทราบว่าผู้พูดพูดออกมากี่พยางค์ ก็จะสามารถบอกได้ว่า ณ ขณะนั้นผู้พูดพูดถึงตำแหน่งใดของข้อความแล้วไปเรื่อย ๆ โดยขั้นตอนวิธีที่นำเสนอจะแบ่งเป็น 3 ขั้นตอนย่อย ๆ ด้วยกัน คือ

1.1. การตรวจหาจุดจบของพยางค์แบบทันที

เนื่องจากเสียงที่เป็นข้อมูลนำเข้ามานั้นเป็นเสียงแบบต่อเนื่องซึ่งต้องประมวลผลทันทีที่เสียงเข้ามา จึงต้องอ่านค่าของเสียงไปเรื่อย ๆ จนกระทั่งพบจุดจบของพยางค์ โดยดึงลักษณะเด่นทางเสียงคือ ค่าระดับความเข้มของเสียง โดยมีค่าขีดแบ่งเพื่อแยกระหว่างพลังประจำเสียง และ เสียงเงียบ

1.2. การตรวจหาแกนกลางของพยางค์

หลังจากที่พบจุดจบของพยางค์แล้ว หมายความว่าผู้พูดพูดพยางค์ออกไปแล้ว แต่ไม่ใช่ทุก ๆ พยางค์ที่จะพบจุดจบของพยางค์ ดังนั้นจึงต้องใช้หลักการตรวจหาแกนกลางพยางค์นับจำนวนพยางค์ที่พูดออกไปแล้วอีกเพื่อความแม่นยำเพิ่มมากขึ้นโดยใช้ขั้นตอนวิธีคอนเว็กซ์ฮัลล์ในการระบุตำแหน่งแกนกลางของพยางค์ก่อน โดยใช้ระดับความเข้มของเสียงที่ได้จากการหาจุดจบของพยางค์ตั้งแต่จุดจบก่อนหน้าจนถึงจุดจบปัจจุบัน

1.3. การให้คะแนนตำแหน่งแกนกลาง

เพื่อความแม่นยำที่มากขึ้น ขั้นตอนวิธีที่เสนอจะใช้ระดับความเข้มของเสียงที่ความถี่มากกว่า 300 เฮิรตซ์ ร่วมกับค่าความเป็นรายคาบของสัญญาณที่ความถี่ต่ำกว่า 900 เฮิรตซ์ มาตัดสินตำแหน่งของกรอบสัญญาณที่ได้จากการตรวจหาแกนกลางของพยางค์ โดยการใช้ตัวจำแนกประเภทซัพพอร์ตเวกเตอร์แมชชีน ในการให้คะแนนตำแหน่งนั้น ซึ่งได้ผลลัพธ์ที่ดีที่สุด

2. การตรวจจับความไม่ตรงกันของการถอดเสียง

เป็นการใช้ข้อมูลของข้อความร่วมกับข้อมูลเสียงที่ได้รับมา ทำการสร้างสมมติฐานและให้คะแนนแต่ละสมมติฐานว่า ผู้พูดพูดถึงตำแหน่งไหนแล้ว โดยการถอดเสียงออกมาจากข้อความแล้วนำมาใช้การปรับแนวเสียงกับเสียงพูดในช่วงนั้น เพื่อให้ได้ค่าความน่าจะเป็นออกมาในแต่ละสมมติฐานแล้วตอบคำตอบสมมติฐานที่มีค่าความน่าจะเป็นมากที่สุด ซึ่งน่าจะเป็นข้อความที่ตรงกับเสียงพูดมากที่สุด นอกจากนั้นยังศึกษาค่าพารามิเตอร์ที่มีส่วนช่วยในการประสานเวลาเสียง ซึ่งทำให้มีผลต่อความแม่นยำแต่ในขณะเดียวกันก็มีผลต่อเวลาในการคำนวณด้วยเช่นกัน

ประสิทธิภาพของวิธีที่นำเสนอนั้น ถูกนำมาเปรียบเทียบกับประสิทธิภาพของระบบอ้างอิง ซึ่งก็คือการประสานเวลาโดยใช้หลักความไม่ตรงกันของการถอดเสียงเพียงอย่างเดียว และการประสานเวลาโดยใช้หลักการตรวจหาพยางค์เพียงอย่างเดียว โดยทำการทดสอบกับเสียงในฐานข้อมูลโลตัส ในชุดหน่วยเสียงสมดุล เพราะมีการระบุขอบเขตของเวลาและป้ายกำกับหน่วยเสียงไว้ และทดสอบกับหนังสือความยาว 1 หน้า เพื่อดูความเป็นไปได้ในการใช้งานจริง ซึ่งผลลัพธ์ที่ได้นั้นสามารถประมวลผลแบบทันทีได้ทั้งหมด แต่มีเพียงระบบอ้างอิงแบบที่ 2 และขั้นตอนวิธีที่นำเสนอสามารถนำมาใช้ในการประสานเวลาจริงได้เท่านั้น แสดงให้เห็นว่าการใช้หลักการตรวจหาพยางค์เพียงอย่างเดียวนั้น ไม่สามารถนำมาใช้ได้ แต่หลักการตรวจหาความผิดพลาดของการถอดเสียงสามารถใช้ในการประสานเวลาได้ เพราะอาศัยทั้งข้อมูลของเสียงและข้อความที่มีครบให้เป็นประโยชน์ ส่วนการประยุกต์ใช้หลักของการตรวจหาพยางค์นั้น สามารถช่วยให้ผลลัพธ์นั้นมีความละเอียดมากขึ้นและมีความถูกต้องมากขึ้นได้ 75.51% และ 41.56% ตามลำดับ

ข้อเสนอแนะ

เนื่องจากการวิเคราะห์ความผิดพลาดของระบบที่เกิดขึ้นนั้นพบว่า เกิดจากจังหวะเว้นหายใจของผู้พูด ซึ่งทำให้ผลลัพธ์การถอดเสียงนั้นผิดพลาดไป เพราะในการถอดเสียงจากข้อความนั้นไม่สามารถรับรู้ได้ว่า ผู้พูดจะหยุดหายใจเมื่อไร ดังนั้นถ้าหากสามารถทำการแทรกจังหวะหายใจในส่วนของการถอดเสียงเพื่อให้ตรงตามลักษณะของเสียงพูดให้มากขึ้น ก็มีความเป็นไปได้ที่จะทำให้ผลลัพธ์ในส่วนนั้นมีความแม่นยำมากขึ้นด้วย

รายการอ้างอิง

[1] Loscos, A., Cano, P., and Bonada, J. Low-Delay Singing Voice Alignment to Text. In Proceedings of the International Computer Music Conference. (1999).

[2] Li, S., Lin, H., and Chen, H. How speech/text alignment benefits web-based learning. In Proceedings of the 13th annual ACM international conference on Multimedia. (2005) : 259-260.

[3] Ando, A., Imai, T., Kobayashi, A., Isono, H., and Nakabayashi, K. Real-time Transcription System For Simultaneous Subtitling of Japanese Broadcast News Programs. IEEE Transactions on Broadcasting. (2000) : 189-196.

[4] Garcia, J.E., Ortega, A., Lleida, E., Lozano, T., Bernues, E., and Sanchez, D. Audio and text synchronization for TV news subtitling based on Automatic Speech Recognition. In IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB '09. (2009) : 1-6.

[5] Punyabukkana, P., Lertwongkhanakool, N., Kertkeidkachorn, N., Vorapatratorn, S. Hirankan, P., and Suchato, A. ChulaDaisy: The Implementation and A Case Study. In Proceedings of the Sixth International Convention for Rehabilitation Engineering and Assistive Technology (i-CREATe 2012). (2012).

[6] Sornlertlamvanich, V., Potipiti, T., Wutiwiwatchai, C., and Mittrapiyanuruk, P. The state of the art in Thai language processing. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. (2000) : 1-2.

[7] กาญจนา นาคสกุล. ระบบเสียงภาษาไทย. พิมพ์ครั้งที่ 4. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์ มหาวิทยาลัย, 2541.

[8] Luksaneeyanawin, S. Three-DimentionalPhonology : A Historical Implication. In The Third International Symposium on Language and Linguistics. (1992) : 75-90.

[9] เพียงจิต ดารีเย๊าะ. การตรวจหาสระในเสียงพูดต่อเนื่องภาษาไทย. วิทยานิพนธ์ปริญญามหาบัณฑิต, สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2549.

[10] Paul Boersma. Sound Pressure Level [Online]. 2004. Available from: http://www.fon.hum.uva.nl/praat/manual/sound_pressure_level.html [2013, April 18]

[11] Wikipedia. Hidden Markov Model [Online]. *2012*. Available from: http://en.wikipedia.org/wiki/Hidden_Markov_model [2013, April 18]

[12] ISIP. Force Alignment [Online].2012.Available from: http://www.isip.piconepress.com/projects/speech/software/tutorials/production/fundamentals/v1.0/section_04/s04_04_p01.html [2013, April 18]

[13] Damm, D., Grohganz, H., Kurth, F., Ewert, S., and Clausen M. SyncTS: Automatic Synchronization of Speech and Text Documents. In Proceedings *of the AES 42nd International Conference. (2011).*

[14] Katsamanis, A., Black, M.P., Georgiou, P.G., Goldstein, L., and Narayanan, S. SailAlign: Robust Long Speech-Text Alignment. In Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research. (2011).

[15] Haubold, A., and Kender, J.R. Alignment of Speech to Highly Imperfect Text Transcriptions. In IEEE International Conference on Multimedia and Expo. (2007) : 224-227.

[16] Kan, M.Y., Wang, Y., Iskandar, D., Nwe, T.L.,and Shenoy, A. LyricAlly: Automatic Synchronization of Textual Lyrics to Acoustic Music

Signals. IEEE Transactions on Audio, Speech, and Language Processing. (2008) : 338-349.

[17] Gao J,. Zhao, Q., and Yan, Y. Automatic Synchronization of live speech and its Transcripts based on a frame-synchronous likelihood ratio test. In IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP'2010). (2010) : 1622-1625.

[18] Mermelstein, P. Automatic segmentation of speech into syllabic units. Journal of the Acoustical Society of America. (1975) : 880-883.

[19] Pfitzinger, H.R., Burger, S., and Heid, S. Syllable Detection in Read and Spontaneous Speech. In Proceeding of Fourth International Conference on Spoken Language (ICSLP'96). (1996) : 1261-1264.

[20] Juneja, A., and Espy-Wilson, C. Speech Segmentation Using Probabilistic Phonetic Feature Hierarchy and Support Vector Machines. In Proceedings of the International Joint Conference on Neural Networks. (2003) : 675-679.

[21] Koanantakool, H.T., Karoonboonyanan, T., and Wutiwiwatchai, C. IEEE Annals of the History of Computing. (2009) : 46-61.

[22] Pfau, T., and Ruske, G. Estimating the Speaking Rate using Vowel Detection. In Proceedings of ICASSP. (1998) : 945-948.

[23] Xie, Z., and Niyogi, P. Robust Acoustic-Based Syllable Detection. In INTERSPEECH-2006. (2006).

[24] Kasuriya, S. Sornlertlamvanich, V. Cotsomrong, P. Kanokphara, S. and Thatphithakkul, N. Thai Speech Corpus for Thai Speech Recognition. The Oriental COCOSDA. (2003) : 54-61.

[25] Howitt, A.W. Automatic Syllable Detection for Vowel Landmarks. Doctoral dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2000.

[26] Rochkitticharеon, W., Suchato, A., and Punyabukkana, P. Broad Phonetic Class Segmentation Study for Thai Automatic Speech Recognition. In Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). (2012) : 1-4.

[27] Pitz, M., Molau, S., Schlüter, R., and Ney, H. Automatic Transcription Verification of Broadcast News and Similar Speech Corpora. In Proc. DARPA Broadcast News Workshop. (1999).

[28] Jeong, M. Using Higher-Level Linguistic Knowledge For Speech Recognition Error Correction in a Spoken Q/A Dialog. In Proceedings of the HLT-NAACL special workshop on Higher-Level Linguistic Information for Speech Processing. (2004) : 48-55.

[29] Wang, L., Zhao, Y., Chu, M., Soong, F.K., and Cao, Z. Phonetic Transcription Verification with Generalized Posterior Probability. In Proceedings of INTERSPEECH. (2005) : 1949-1953.

[30] Gubian, M., Schuppler, B., Doremalen, J.V., Sanders, E., and Boves, L. Novelty Detection as a Tool for Automatic Detection of Orthographic Transcription Errors. In Proceedings in the 13th International Conference on Speech and Computer SPECOM. (2009) : 509-514.

[31] Jiang, H. Confidence measures for speech recognition: A survey. Journal of Speech Communication. (2005) : 455-470.

[32] Young, S., Evermann, G., Galse, M., Kershaw, D., Moore, G. Hidden Markov model toolkit – speech recognition toolkit. [Online]. 2012. Available from: http://htk.eng.cam.ac.uk [2013, April 18]

[33] คุณน้อยนราธิวาส. เรื่องของกบตัวเล็ก ๆ ตัวหนึ่ง [*ออนไลน์*]. 2555. Available from: http://www.bbtfamily.com/index.php?lay=boardshow&ac=webboard_show&WBntype=1&No=1531388 [18 เมษายน 2556]

[34] Chang and C.C. and Lin, C.J. LIBSVM: a library for support vector machines 2001.

ประวัติผู้เขียนวิทยานิพนธ์

นายณัฏฐ์ เลิศวงศ์คณากูล เกิดเมื่อวันอาทิตย์ที่ 19 กุมภาพันธ์ พ.ศ.2532 ที่จังหวัดกรุงเทพมหานคร สำเร็จการศึกษาระดับมัธยมศึกษาจาก โรงเรียนมารีย์วิทยา นครราชสีมา สำเร็จการศึกษาระดับปริญญาบัณฑิต จากคณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพมหานคร ในปีการศึกษา 2553