บทที่ 4
การทดลองแบ่งเสียงพูดเป็นเซกเมนต์

บทนี้จะนำเสนอเกี่ยวกับการทดลองการแบ่งเสียงพูดเป็นเซกเมนต์ โดยจะเริ่มจากนำเสนอฐานข้อมูลเสียงที่ใช้ในการทดลอง องค์ประกอบและประสิทธิภาพของเครื่องรู้จำเสียงพูดที่นำมาใช้ในการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด การทดลองและผลการทดลองของการแบ่งเสียงพูดเป็นเซกเมนต์ และสรุปผลการทดลอง ตามลำดับ

ฐานข้อมูลเสียงเพื่อการแบ่งเสียงพูดเป็นเซกเมนต์

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

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

ตารางที่ 4.1 สถิติเกี่ยวกับจำนวนขอบเขตหน่วยเสียงต่อการเปล่งเสียงหนึ่งครั้ง สำหรับข้อมูลเสียงในชุดหน่วยเสียงสมดุล (PD Set)

	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน	ต่ำสุด	สูงสุด
จำนวนขอบเขตหน่วยเสียงต่อการเปล่งเสียงหนึ่งครั้ง	52.7	30.4	12	190

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

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

องค์ประกอบและประสิทธิภาพของเครื่องรู้จำเสียงพูด

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

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

ตารางที่ 4.2 ตารางสรุปองค์ประกอบและประสิทธิภาพของเครื่องรู้จำเสียงพูดที่นำมาใช้เปรียบเทียบการตรวจหาขอบเขตของหน่วยเสียง

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

การทดลองและผลการทดลอง

การทดลองการแบ่งเสียงพูดเป็นเซกเมนต์ในที่นี้จะทดลองเพื่อวัดประสิทธิภาพของวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ที่เสนอไว้ในบทที่ 3 เปรียบเทียบกับวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด โดยจะประกอบไปด้วยการทดลองทั้งหมด 3 การทดลอง คือ การทดลองเพื่อวัดประสิทธิภาพการจำแนกลักษณะการออกเสียง การทดลองเพื่อเปรียบเทียบประสิทธิภาพการตรวจหาขอบเขตของหน่วยเสียง และการทดลองเพื่อเปรียบเทียบประสิทธิภาพการสร้างกราฟของเซกเมนต์ ซึ่งการทดลองแต่ละส่วนมีรายละเอียดดังต่อไปนี้

1. การทดลองเพื่อวัดประสิทธิภาพการจำแนกลักษณะการออกเสียง

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

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

1.1 การวัดประสิทธิภาพ

การวัดประสิทธิภาพการจำแนกลักษณะการออกเสียงแบบอาศัยเอสวีเอ็ม จะพิจารณาแยกตามเครื่องแบ่งแยกสัทลักษณ์เอสวีเอ็มแต่ละตัว โดยดูจากเปอร์เซ็นต์ความแม่นยำของการแบ่งแยกสัทลักษณ์ เขียนแทนด้วยสัญลักษณ์ $\%Accuracy$

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

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

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

1.2 ผลการทดลองการจำแนกลักษณะการออกเสียง

ผลการจำแนกลักษณะการออกเสียงของเสียงพูดจากตัวแบ่งแยกสัทลักษณ์เอสวีเอ็มทั้งสี่ตัว แสดงด้วยตารางที่ 4.3

ตารางที่ 4.3 ประสิทธิภาพการแบ่งแยกสัทลักษณ์ของลักษณะการออกเสียง
(เปอร์เซ็นต์ความแม่นยำถูกปรับเพื่อให้โอกาสพบข้อมูลที่มีและไม่มีสัทลักษณ์เป็น 50% เท่ากัน)

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

1.3 วิเคราะห์ผลการทดลอง

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

2. การทดลองเพื่อเปรียบเทียบประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียง

ผลลัพธ์ที่ได้จากการจำแนกลักษณะการออกเสียงด้วยซัพพอร์ตเวกเตอร์แมชชีน และจากการรู้จำหน่วยเสียงด้วยเครื่องรู้จำเสียงพูด จะนำมาใช้วัดประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียง โดยมีรายละเอียดเกี่ยวกับการวัดประสิทธิภาพ และผลการทดลองดังต่อไปนี้

2.1 การวัดประสิทธิภาพ

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

2.1.1 ความแม่นยำและความครอบคลุม

การพิจารณาค่าความแม่นยำและค่าความครอบคลุม จะต้องใช้ผลของการตัดสินใจว่าขอบเขตของหน่วยเสียงที่หามาได้นั้นมีความถูกต้องตรงกับขอบเขตของหน่วยเสียงที่เป็นตัวอ้างอิงหรือไม่ ขอบเขตของหน่วยเสียงที่ดีนั้นไม่จำเป็นต้องอยู่ที่ตำแหน่งเดียวกันกับขอบเขตของหน่วยเสียงที่เป็นตัวอ้างอิงก็ได้ โดยอาจยินยอมให้คลาดเคลื่อนกันได้เป็นระยะเวลาสั้นๆในระดับมิลลิวินาที ขึ้นอยู่กับการนำไปใช้งาน ดังนั้นในการทดลองนี้จึงมีการกำหนดระดับความคลาดเคลื่อนยินยอม (Tolerance)

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

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

เมื่อกำหนดให้ D คือจำนวนขอบเขตของหน่วยเสียงทั้งหมดที่ตรวจหามาได้ T คือจำนวนขอบเขตหน่วยเสียงที่เป็นตัวอ้างอิง และ C คือจำนวนขอบเขตของหน่วยเสียงทั้งหมดที่ตรงกับขอบเขตของหน่วยเสียงอ้างอิงโดยยินยอมให้คลาดเคลื่อนไปจากขอบเขตของหน่วยเสียงที่เป็นตัวอ้างอิงได้ไม่เกินระดับความคลาดเคลื่อนยินยอมที่กำหนด

เกณฑ์การตัดสินคุณภาพของขอบเขตของหน่วยเสียงที่หามาได้ว่ายอมรับได้หรือไม่ ส่วนใหญ่แล้วจะพิจารณาโดยดูจากความแม่นยำของขอบเขตที่หามาได้ที่ระดับความคลาดเคลื่อนยินยอมที่แตกต่างกัน ขึ้นอยู่กับความต้องการและประเภทของงานที่นำไปใช้ ในงานวิจัยเกี่ยวกับด้านการ

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

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

2.1.2 เรียลไทม์แฟคเตอร์

การเปรียบเทียบเวลาที่ใช้ในการตรวจหาขอบเขตของหน่วยเสียงจะใช้เรียลไทม์แฟคเตอร์เขียนแทนด้วยสัญลักษณ์ *RTF* ซึ่งเป็นตัววัดความเร็วที่ใช้กันโดยทั่วไปในระบบรู้จำเสียงพูด โดยมีนิยามดังนี้

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

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

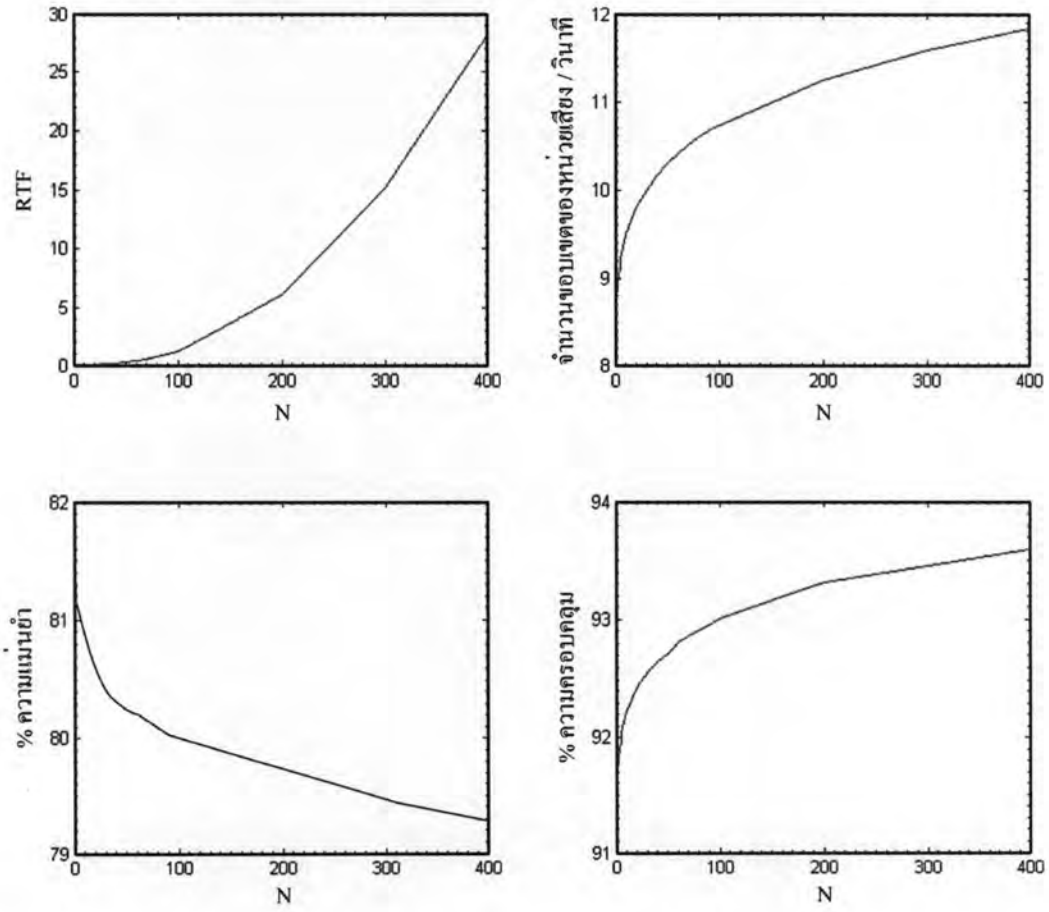
ค่า *RTF* นั้นขึ้นอยู่กับความเร็วของอุปกรณ์ฮาร์ดแวร์ที่ใช้ โดยในอุปกรณ์ฮาร์ดแวร์ที่ใช้ในการทดลองทั้งหมดในวิทยานิพนธ์นี้คือ เครื่องคอมพิวเตอร์พีซีที่มีสถาปัตยกรรมแบบ 32 บิต ความเร็วซีพียู 2,793 เมกะเฮิร์ต ใช้หน่วยความจำหลักขนาด 1 กิกะไบต์ ทำงานบนระบบปฏิบัติการลินุกซ์

2.1.3 จำนวนขอบเขตของหน่วยเสียงที่ตรวจหามาได้

จำนวนขอบเขตของหน่วยเสียงที่ตรวจหามาได้ จะวัดเป็นจำนวนขอบเขตของหน่วยเสียงต่อหนึ่งหน่วยวินาที โดยค่านี้จะสะท้อนให้เห็นขนาดของอินพุตที่จะผ่านเข้าไปยังขั้นตอนการสร้างกราฟของเซกเมนต์ต่อไป โดยหากมีปริมาณมากเกินไปก็จะทำให้กราฟของเซกเมนต์มีขนาดใหญ่ตามไปด้วย แต่ถ้าหากมีน้อยเกินไปก็จะไม่ครอบคลุมขอบเขตของหน่วยเสียงที่ต้องการ

2.2 ผลการทดลองของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูด

ประสิทธิภาพของกระบวนการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูดจะขึ้นอยู่กับตัวแปร N หรือจำนวนของผลการรู้จำ N อันดับที่ดีที่สุด ดังแสดงด้วยกราฟในรูปที่ 4.1



รูปที่ 4.1 กราฟแสดงประสิทธิภาพการตรวจหาขอบเขตหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูด (ที่ระดับความคลาดเคลื่อนยินยอม 30 มิลลิวินาที)

2.3 ผลการทดลองของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีน

ประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนโดยใช้สารสนเทศสวนสัทศาสตร์ ในด้านความเร็วในการทำงานและจำนวนของขอบเขตของหน่วยเสียงที่ตรวจหามาได้แสดงไว้ในตารางที่ 4.4 ส่วนประสิทธิภาพของการตรวจหาขอบเขตของหน่วย

เสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนในแง่ของความแม่นยำและความครอบคลุมแสดงไว้ใน ตารางที่ 4.5 และ 4.6 ตามลำดับ

ตารางที่ 4.4 ความเร็วในการทำงานและจำนวนขอบเขตของหน่วยเสียงที่หามาได้ของกระบวนการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีน แบบเชิงเส้นเทียบกับแบบพหุนาม

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

ตารางที่ 4.5 ความแม่นยำของการตรวจหาขอบเขตของหน่วยเสียง แบบอาศัยซัพพอร์ตเวกเตอร์แมชชีน

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

ตารางที่ 4.6 ความครอบคลุมของการตรวจหาขอบเขตของหน่วยเสียง แบบอาศัยซัพพอร์ตเวกเตอร์แมชชีน

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

2.4 ผลการทดลองของการเปรียบเทียบประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียง

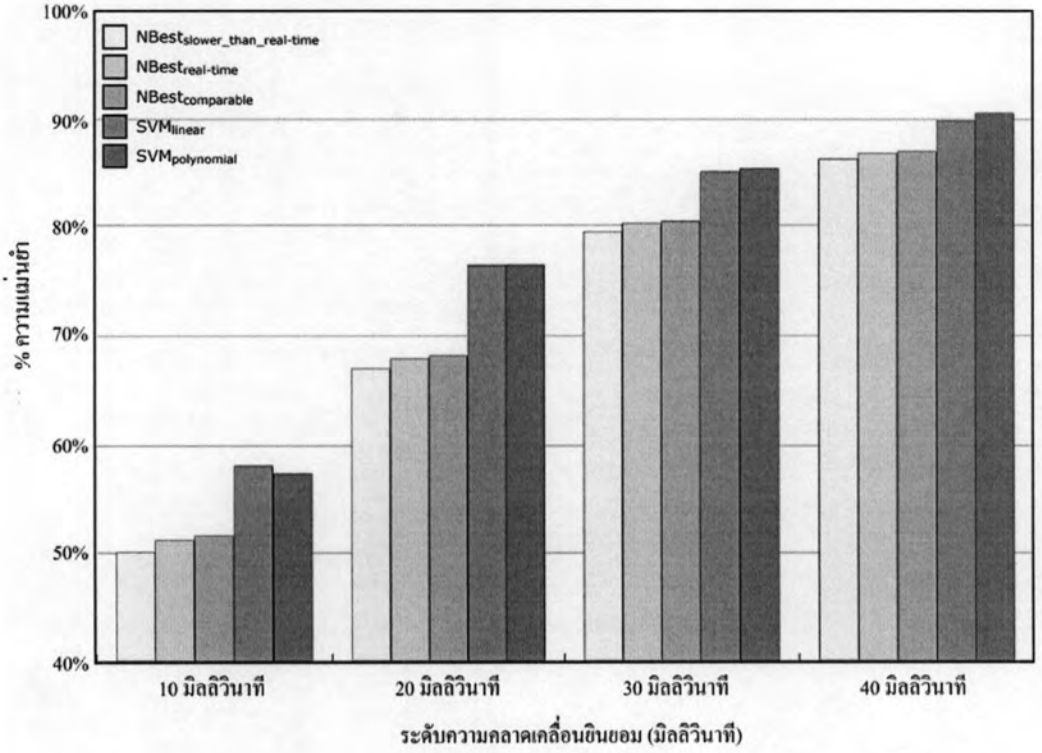
เนื่องจากประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูด สามารถปรับเปลี่ยนได้ตามตัวแปร N โดยในการทดลองเพื่อเปรียบเทียบประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียงระหว่างวิธีที่อาศัยเครื่องรู้จำเสียงพูด กับวิธีที่อาศัยซัพ

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

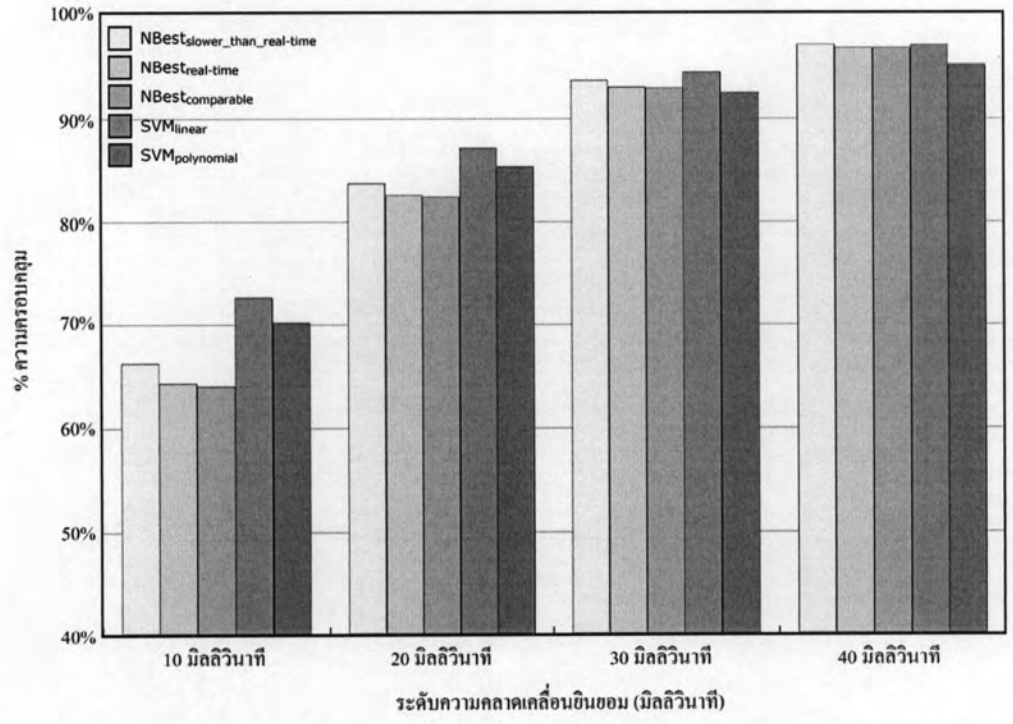
การเปรียบเทียบประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียงของเครื่องตรวจหาขอบเขตของหน่วยเสียงทั้งห้าเครื่องในตารางที่ 4.7 ในด้านความแม่นยำและความครอบคลุมของขอบเขตของหน่วยเสียงที่หามาได้แสดงด้วยกราฟในรูปที่ 4.2 และ 4.3

ตารางที่ 4.7 เครื่องตรวจหาขอบเขตของหน่วยเสียงที่นำมาใช้เปรียบเทียบประสิทธิภาพ

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



รูปที่ 4.2 กราฟแสดงเปอร์เซ็นต์ความแม่นยำในการตรวจหาขอบเขตของหน่วยเสียง



รูปที่ 4.3 กราฟแสดงเปอร์เซ็นต์ความครอบคลุมในการตรวจหาขอบเขตของหน่วยเสียง

2.5 วิเคราะห์ผลการทดลอง

จากการทดลองเปรียบเทียบประสิทธิภาพในการตรวจหาขอบเขตของหน่วยเสียง โดยเริ่มต้นจากการพิจารณาประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูดจากกราฟในรูปที่ 4.1 จะพบว่าเมื่อเราปรับให้ N มีค่ามากขึ้น ก็จะทำให้ได้จำนวนขอบเขตของหน่วยเสียงเพิ่มมากขึ้น ส่งผลให้ขอบเขตของหน่วยเสียงที่หามาได้ครอบคลุมขอบเขตของหน่วยเสียงที่เป็นตัวอ้างอิงมากยิ่งขึ้นด้วย แต่ทั้งนี้ก็ต้องแลกเปลี่ยนกับเปอร์เซ็นต์ความแม่นยำที่ลดลงและเวลาในการทำงานที่เพิ่มมากขึ้นตามไปด้วย

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

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

ฟังก์ชันเคอร์เนลแบบพหุนามมีค่าสูงถึง 4.83 ในระหว่างที่ *RTF* ของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลแบบเชิงเส้นมีค่าเพียง 0.57 ด้วยเหตุนี้จึงสามารถสรุปได้ว่า วิธีการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลแบบเชิงเส้นนั้นมีประสิทธิภาพดีกว่าวิธีการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลแบบพหุนาม

เนื่องจากเราต้องการวิธีการตรวจหาขอบเขตของหน่วยเสียงที่สามารถทำงานได้รวดเร็ว ดังนั้นตัวเลือกที่เหมาะสมที่สุดในที่นี้ก็คือ การตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลเชิงเส้นเนื่องจากสามารถทำงานได้รวดเร็ว อีกทั้งยังสามารถตรวจหาขอบเขตของหน่วยเสียงที่มีความแม่นยำและความครอบคลุมใกล้เคียงกับการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลพหุนาม ด้วยเหตุนี้ในการทดลองเพื่อเปรียบเทียบประสิทธิภาพการสร้างกราฟของเซกเมนต์ในหัวข้อถัดไปจะนำขอบเขตของหน่วยเสียง ที่ได้จากวิธีตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลเชิงเส้นไปใช้เท่านั้น

3. การทดลองเพื่อเปรียบเทียบประสิทธิภาพการสร้างกราฟของเซกเมนต์

การทดลองในที่นี้จะเป็นการทดลองวัดประสิทธิภาพในการสร้างกราฟของเซกเมนต์ โดยจะเปรียบเทียบวิธีการสร้างกราฟทั้งหมดสามวิธีตามที่เสนอไว้ในบทที่ 2 คือ วิธีการสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง วิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม และวิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับ โดยอาศัยขอบเขตของหน่วยเสียงที่ได้จากการตรวจหาขอบเขตของหน่วยเสียงทั้งแบบอาศัยเครื่องรู้จำเสียงพูดและแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีน แล้วนำกราฟของเซกเมนต์ที่ได้มาวัดประสิทธิภาพเปรียบเทียบกัน ซึ่งมีรายละเอียดการวัดประสิทธิภาพและผลการทดลองดังต่อไปนี้

3.1 การวัดประสิทธิภาพ

การแบ่งเสียงพูดเป็นเซกเมนต์ที่ดีนั้นจะต้องสามารถสร้างกราฟของเซกเมนต์ที่มีคุณภาพรวมถึงทำงานได้อย่างรวดเร็ว ความหมายคุณภาพในที่นี้คือกราฟของเซกเมนต์ที่มีขนาดเล็ก และครอบคลุมเซกเมนต์ของหน่วยเสียงอ้างอิง ในส่วนของความเร็วในการแบ่งเสียงพูดเป็นเซกเมนต์เนื่องจากเวลาที่ในการทำงานเฉพาะขั้นตอนในการสร้างกราฟของเซกเมนต์นั้นอยู่ในระดับที่ยอมกันได้ โดยใช้เวลาในการทำงานสั้นมากเมื่อเทียบกับขั้นตอนการตรวจหาขอบเขตของหน่วยเสียง ดังนั้นการเปรียบเทียบความเร็วในการแบ่งเสียงพูดเป็นเซกเมนต์จึงจะพิจารณาจากความเร็วในการตรวจหาขอบเขตของหน่วยเสียงเป็นหลัก

3.1.1 ขนาดของกราฟของเซกเมนต์

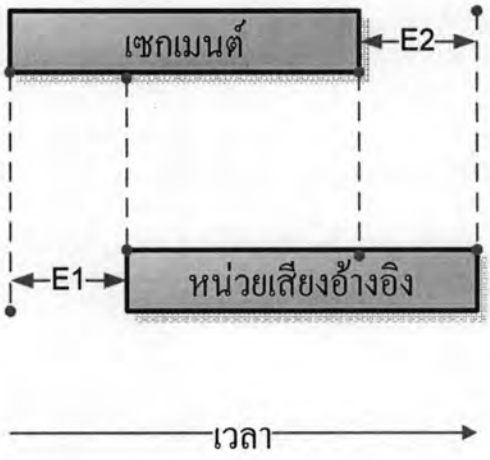
การวัดขนาดของกราฟของเซกเมนต์จะวัดจากจำนวนเซกเมนต์ที่พบในกราฟต่อหนึ่งหน่วยวินาที โดยค่านี้จะสะท้อนให้เห็นขนาดของกราฟของเซกเมนต์ที่จะนำไปใช้ต่อในกระบวนการค้นหาและให้คะแนนเพื่อการรู้จำเสียงพูดในขอบข่ายงานของระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ต่อไป โดยหากมีปริมาณมากเกินไปก็จะทำให้กราฟของเซกเมนต์มีขนาดใหญ่ แต่ถ้าหากมีน้อยเกินไปก็อาจจะได้จำนวนเซกเมนต์ที่ไม่ครอบคลุมหน่วยเสียงในระดับที่ต้องการ

3.1.2 ความครอบคลุม

การพิจารณาความครอบคลุมของกราฟของเซกเมนต์ จะต้องใช้ผลของการตัดสินใจว่าเซกเมนต์ที่หามาได้นั้นมีความถูกต้องตรงกับหน่วยเสียงที่เป็นตัวอ้างอิงหรือไม่ โดยเซกเมนต์ที่ดี

นั้นไม่จำเป็นต้องอยู่ที่ตำแหน่งเดียวกันกับหน่วยเสียงที่เป็นตัวอ้างอิงก็ได้ โดยอาจยินยอมให้คลาดเคลื่อนกันได้เป็นระยะเวลาสั้นๆในระดับมิลลิวินาที ค่าความคลาดเคลื่อนของเซกเมนต์จะคิดจากผลรวมระหว่างค่าความคลาดเคลื่อนของขอบเขตทางด้านซ้ายของเซกเมนต์เทียบกับขอบเขตทางด้านซ้ายของหน่วยเสียงอ้างอิง และค่าความคลาดเคลื่อนของขอบเขตทางด้านขวาของเซกเมนต์เทียบกับขอบเขตทางด้านขวาของหน่วยเสียงอ้างอิง โดยแสดงได้ด้วยรูปที่ 4.4

ค่าความคลาดเคลื่อน = E1 + E2



รูปที่ 4.4 การวัดความคลาดเคลื่อนของเซกเมนต์

เปอร์เซ็นต์ความครอบคลุมเขียนแทนด้วยสัญลักษณ์ $\%Re$ สามารถคำนวณได้จากสมการต่อไปนี้

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

เมื่อกำหนดให้ T คือจำนวนหน่วยเสียงที่เป็นตัวอ้างอิง และ C คือจำนวนเซกเมนต์ทั้งหมดที่ถูกต้องอยู่ตรงกับหน่วยเสียงที่เป็นตัวอ้างอิงโดยยินยอมให้มีค่าคลาดเคลื่อนไปจากหน่วยเสียงที่เป็นตัวอ้างอิงได้ไม่เกินระดับความคลาดเคลื่อนยินยอมที่ 10, 20, 30 และ 40 มิลลิวินาทีตามลำดับ

3.2 ผลการทดลอง

ประสิทธิภาพของการสร้างกราฟของเซกเมนต์ ในด้านความครอบคลุมและขนาดของกราฟแสดงไว้ในตารางที่ 4.8 4.9 และ 4.10

ตารางที่ 4.8 คุณภาพกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง

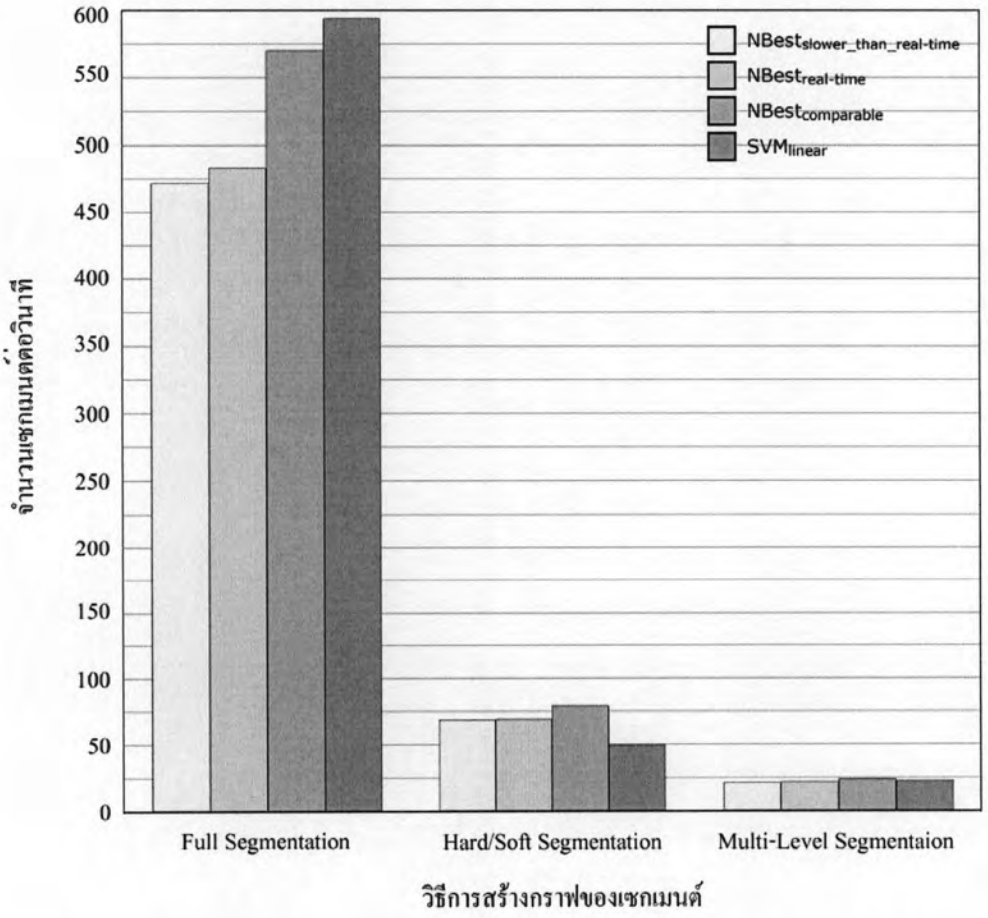
	เปอร์เซ็นต์ความครอบคลุมของเซกเมนต์ที่ระดับความคลาดเคลื่อน (มิลลิวินาที)				จำนวนเซกเมนต์ต่อวินาที
	10	20	30	40	
$NBest_{comparable}$	40.01	63.34	82.83	91.28	471.4
$NBest_{real-time}$	40.40	64.71	83.10	91.41	483.1
$NBest_{slower\ than\ real-time}$	43.10	67.09	84.72	92.39	570.0
SVM_{linear}	50.95	74.80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

ตารางที่ 4.9 คุณภาพกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม

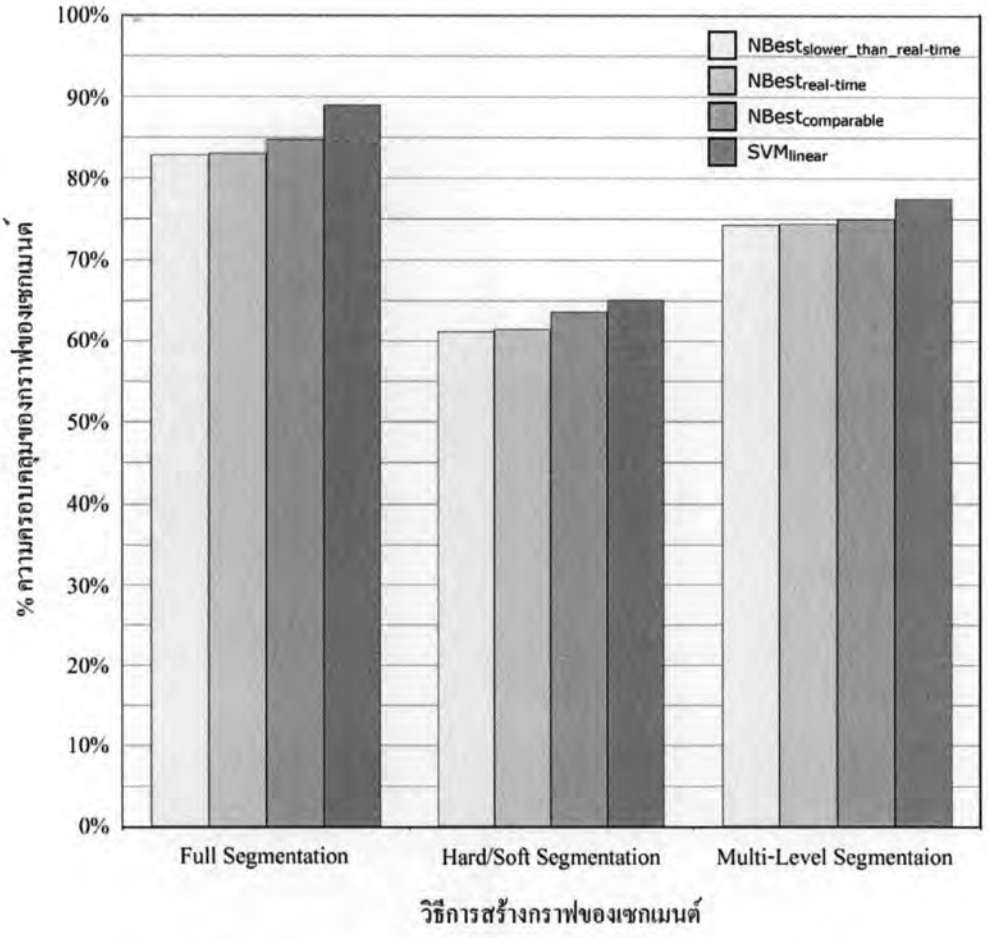
	เปอร์เซ็นต์ความครอบคลุมของเซกเมนต์ที่ระดับความคลาดเคลื่อน (มิลลิวินาที)				จำนวนเซกเมนต์ต่อวินาที
	10	20	30	40	
$NBest_{comparable}$	27.02	45.72	61.13	68.76	68.0
$NBest_{real-time}$	27.30	46.02	61.40	69.00	68.7
$NBest_{slower\ than\ real-time}$	29.32	48.43	63.65	71.00	79.1
SVM_{linear}	32.90	51.85	65.02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

ตารางที่ 4.10 คุณภาพกราฟของเซกเมนต์แบบหลายระดับ

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



รูปที่ 4.5 กราฟเปรียบเทียบขนาดของกราฟของเซกเมนต์ที่สร้างด้วยวิธีต่างๆ



รูปที่ 4.6 กราฟเปรียบเทียบเปอร์เซ็นต์ครอบคลุมของกราฟของเซกเมนต์ที่ระดับความคลาดเคลื่อน 30 มิลลิวินาที

3.3 วิเคราะห์ผลการทดลอง

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

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

ด้วยเหตุนี้จึงสรุปได้ว่า วิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับนั้นสามารถนำมาใช้สร้างกราฟของเซกเมนต์ที่มีคุณภาพดีกว่าวิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม และเมื่อพิจารณาความต้องการที่จะนำการแบ่งเสียงพูดเป็นเซกเมนต์นี้ไปใช้ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่จะต้องทำงานได้รวดเร็วแบบทันกาล วิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับจึงมีความเหมาะสมกว่าวิธีการสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง

สรุปผลการทดลอง

การทดลองการแบ่งเสียงพูดเป็นเซกเมนต์ก็เพื่อเป็นการเปรียบเทียบประสิทธิภาพของขั้นตอนย่อยของการแบ่งเสียงพูดเป็นเซกเมนต์ซึ่งประกอบไปด้วยสองขั้นตอน คือ ขั้นตอนการตรวจหาขอบเขตของหน่วยเสียงและขั้นตอนการสร้างกราฟของเซกเมนต์ โดยสามารถสรุปการเปรียบเทียบวิธีการที่ใช้ในขั้นตอนต่างๆได้ดังแผนภาพในรูปที่ 4.7

รูปที่ 4.7 แผนภาพเปรียบเทียบวิธีการที่ใช้ในแต่ละขั้นตอนของการแบ่งเสียงพูดเป็นเซกเมนต์

ในขั้นตอนการตรวจหาขอบเขตของหน่วยเสียง งานวิจัยนี้เสนอวิธีการตรวจหาขอบเขตของเสียงพูดแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนโดยใช้สารสนเทศสวนสัทศาสตร์ มาเปรียบเทียบกับวิธีตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูด จากการวิเคราะห์ผลการทดลองเปรียบเทียบประสิทธิภาพของทั้งสองวิธีพบว่าวิธีการตรวจหาขอบเขตของเสียงพูดแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนโดยใช้สารสนเทศสวนสัทศาสตร์ มีประสิทธิภาพดีกว่าทั้งในด้านคุณภาพของขอบเขตของหน่วยเสียงที่ได้ และความเร็วในการทำงาน

ในขั้นตอนการสร้างกราฟของเซกเมนต์ งานวิจัยนี้เสนอวิธีการสร้างกราฟของเซกเมนต์สามวิธีได้แก่ วิธีการสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียงซึ่งเป็นวิธีดั้งเดิมที่ใช้ในการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด วิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม และวิธีการสร้างกราฟแบบหลายระดับ จากการวิเคราะห์ผลการทดลองพบว่า วิธีการสร้างกราฟของเซกเมนต์แบบจับคู่ทุกขอบเขตของหน่วยเสียงสามารถสร้างกราฟของเซกเมนต์ที่ได้ความครอบคลุมมากที่สุดแต่กราฟนั้นจะมีขนาดใหญ่มากไม่

เหมาะกับระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ต้องการความรวดเร็วในการทำงาน และจากการวิเคราะห์ผลการทดลองเปรียบเทียบประสิทธิภาพของวิธีการสร้างกราฟอีกสองวิธีที่เหลือ พบว่าสามารถลดขนาดของกราฟลงได้มาก แต่วิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับนั้นมีประสิทธิภาพดีกว่าวิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัมเนื่องจากสามารถลดขนาดของกราฟลงได้มากกว่าอีกทั้งยังสามารถรักษาระดับความครอบคลุมไว้ได้มากกว่าด้วยเหตุนี้เมื่อพิจารณาความจำเป็นของระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่จะต้องทำงานได้แบบทันกาล วิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับจะมีความเหมาะสมที่สุด

ด้วยเหตุนี้เมื่อพิจารณาผลการทดลองเปรียบเทียบประสิทธิภาพของวิธีการต่างๆที่นำมาใช้ในแต่ละขั้นตอน จะสามารถสรุปได้ว่าวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ที่ใช้วิธีการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนโดยใช้สารสนเทศสวนสัทศาสตร์ และใช้วิธีการสร้างกราฟแบบหลายระดับ มีประสิทธิภาพดีที่สุด ซึ่งสามารถสร้างกราฟของเซกเมนต์ที่มีคุณภาพดีกว่า อีกทั้งยังทำงานได้รวดเร็วกว่าวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด

ด้วยข้อจำกัดของวิธีการตรวจหาขอบเขตของหน่วยเสียง ในบางครั้งขอบเขตของหน่วยเสียงที่ตรวจหามาได้จะอยู่ติดกันมาก โดยมีระยะห่างระหว่างขอบเขตของหน่วยเสียงที่สั้นประมาณ 10 มิลลิวินาที ทำให้เซกเมนต์ที่สร้างจากขอบเขตของหน่วยเสียงนี้มีขนาดเล็กเกินกว่าจะเป็นหน่วยเสียงได้ ซึ่งหากนำเอาขอบเขตของหน่วยเสียงนี้ไปใช้ต่อก็จะส่งผลต่อคุณภาพกราฟของเซกเมนต์โดยตรง

การทดลองจำแนกลักษณะการออกเสียงในงานวิจัยนี้ ทำการทดลองกับเฉพาะซัพพอร์ตเวกเตอร์แมชชีนที่มีฟังก์ชันเคอร์เนลแบบเชิงเส้นและแบบพหุนามเท่านั้น ไม่ได้ทดลองกับซัพพอร์ตเวกเตอร์แมชชีนที่มีฟังก์ชันเคอร์เนลแบบอื่นๆอย่างครบถ้วน และด้วยข้อจำกัดทางด้านทรัพยากรเกี่ยวกับข้อมูลเสียงที่นำมาใช้ในการทดลอง ซึ่งจะต้องมีการกำกับขอบเขตของหน่วยเสียงไว้ก่อนแล้ว ทำให้ไม่สามารถนำข้อมูลเสียงพูดในฐานข้อมูลเสียงโลตัสมาใช้ได้ทั้งหมด ทำให้เลือกใช้ได้เฉพาะข้อมูลเสียงชุดหน่วยเสียงสมดุล ซึ่งเป็นชุดข้อมูลเสียงเพียงชุดเดียวในฐานข้อมูลเสียงโลตัสที่มีการกำกับขอบเขตของหน่วยเสียงไว้