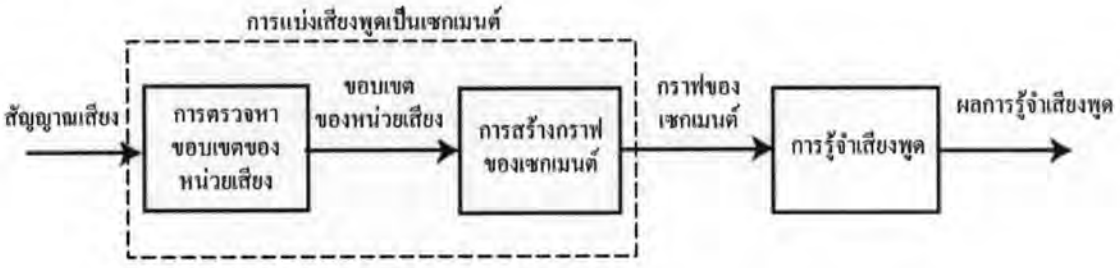
บทที่ 3
การแบ่งเสียงพูดเป็นเซกเมนต์โดยใช้สารสนเทศสวนสัทศาสตร์

บทนี้จะนำเสนอเกี่ยวกับการแบ่งเสียงพูดเป็นเซกเมนต์ โดยจะเริ่มจากภาพรวมของการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ การตรวจหาขอบเขตของหน่วยเสียง การสร้างกราฟของเซกเมนต์ และการให้คะแนนขอบเขตของหน่วยเสียง ตามลำดับ

ภาพรวมของการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์

การแบ่งเสียงพูดเป็นเซกเมนต์ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์นั้นประกอบไปด้วยขั้นตอนย่อยสองขั้นตอนคือ การตรวจหาขอบเขตของหน่วยเสียง และการสร้างกราฟของเซกเมนต์ ดังแสดงด้วยแผนภาพในรูปที่ 3.1 โดยการทำงานจะเริ่มต้นจากการป้อนสัญญาณเสียงพูดเข้าไปเป็นข้อมูลอินพุตของการตรวจหาขอบเขตของหน่วยเสียง ได้เอาท์พุตออกมาเป็นเซตหรือลำดับของขอบเขตของหน่วยเสียงที่เป็นขอบเขตของหน่วยเสียง แล้วผ่านเอาท์พุตนี้ไปเป็นอินพุตของการสร้างกราฟของเซกเมนต์ต่อไป สรุปสุดท้ายได้ผลลัพธ์ออกมาเป็นกราฟของเซกเมนต์สำหรับนำไปใช้ต่อในขั้นตอนการรู้จำเสียงพูดต่อไป



รูปที่ 3.1 แผนภาพแสดงส่วนประกอบของกระบวนการแบ่งเสียงพูดเป็นเซกเมนต์

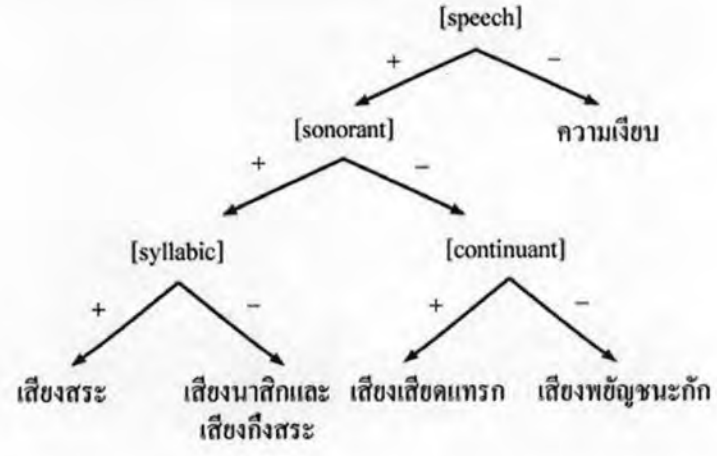
ทั้งการตรวจหาขอบเขตของหน่วยเสียง และการสร้างกราฟของเซกเมนต์นั้นสามารถทำได้หลายวิธี ซึ่งคุณภาพของขอบเขตของหน่วยเสียงและกราฟของเซกเมนต์ที่ได้ จะส่งผลโดยตรงต่อประสิทธิภาพในการรู้จำเสียงพูด งานวิจัยนี้จึงพัฒนาวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ โดยนำสารสนเทศสวนสัทศาตร์มาช่วยเพิ่มประสิทธิภาพในขั้นตอนการตรวจหาขอบเขตของหน่วยเสียง รวมถึงเสนอวิธีการสร้างกราฟของเซกเมนต์แบบต่างๆ เพื่อให้ได้ขอบเขตของหน่วยเสียงและเซกเมนต์กราฟที่มีคุณภาพมากขึ้น อีกทั้งยังสามารถทำงานได้รวดเร็วแบบทันกาลอยู่ในระดับที่นำไปใช้ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ได้

การตรวจหาขอบเขตของหน่วยเสียง

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

การวิเคราะห์และอธิบายคุณสมบัติของเสียงพูดได้จะอาศัยสัทลักษณ์ ดังที่กล่าวไว้ในบทที่ 2 จากความสัมพันธ์ระหว่างสัทลักษณ์ลักษณะการออกเสียง กับประเภทของหน่วยเสียงในตารางที่ 2.3 เราสามารถแสดงให้อยู่ในรูปของโครงสร้างของสัทลักษณ์ลักษณะการออกเสียงเป็นลำดับชั้นได้ดังรูปที่ 3.2

ด้วยโครงสร้างแบบนี้ เราสามารถจำแนกเสียงพูดในแต่ละกรอบเวลาตามลักษณะการออกเสียงได้โดยการแบ่งแยกเสียงพูดที่มีและไม่มีสัทลักษณ์ลักษณะการออกเสียงแต่ละแบบออกจากกันเป็นลำดับชั้น เพื่อระบุว่าสัญญาณเสียงนั้นมีลักษณะการออกเสียงเป็นแบบใด ต่อจากนั้นจึงค่อยพิจารณาตำแหน่งที่มีการเปลี่ยนแปลงลักษณะการออกเสียงให้เป็นขอบเขตของหน่วยเสียงต่อไป ซึ่งวิธีการนี้มีข้อได้เปรียบตรงที่สามารถเลือกใช้ค่าลักษณะสำคัญที่เหมาะสมต่อการจำแนกสัทลักษณ์แต่ละแบบแตกต่างกันได้



รูปที่ 3.2 โครงสร้างลำดับชั้นของสัทลักษณ์ลักษณะการออกเสียง

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

ในหัวข้อนี้จึงจะนำเสนอเกี่ยวกับการสกัดลักษณะสำคัญของเสียง กระบวนการเรียนรู้การจำแนกลักษณะการออกเสียง และกระบวนการตรวจหาขอบเขตของหน่วยเสียงจากผลการจำแนกลักษณะการออกเสียง โดยมีรายละเอียดดังต่อไปนี้

1. การสกัดลักษณะสำคัญของเสียง

ลักษณะสำคัญของเสียงเพื่อการเรียนรู้ในที่นี้แบ่งออกเป็นสองส่วนคือ 1) สัมประสิทธิ์เซปสตรัมบนสเกลเมล 2) ลักษณะสำคัญอื่นๆที่ได้จากการใช้สารสนเทศสวนสัทศาสตร์ ซึ่งมีรายละเอียดดังต่อไปนี้

1.1 สัมประสิทธิ์เซปสตรัมบนสเกลเมล

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

1.2 ลักษณะสำคัญที่ได้จากการใช้สารสนเทศสวนสัทศาสตร์

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

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

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

ตารางที่ 3.1 สารสนเทศสวนสัทศาสตร์และค่าลักษณะสำคัญที่ใช้จำแนกสัทลักษณ์แต่ละประเภท

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

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

การสกัดค่าลักษณะสำคัญที่ได้จากการใช้สารสนเทศสวนสัทศาสตร์ในที่นี้ จะต้องอาศัยค่าพลังงาน อัตสหสัมพันธ์ และอัตราการตัดศูนย์มาคำนวณเป็นระดับความก้องของเสียง และระดับความไม่เป็นคาบของสัญญาณเสียง ในหัวข้อต่อไปนี้จะนำเสนอเกี่ยวกับการหาค่าลักษณะสำคัญดังกล่าวโดยมีรายละเอียดดังต่อไปนี้

1.2.1 ค่าพลังงาน

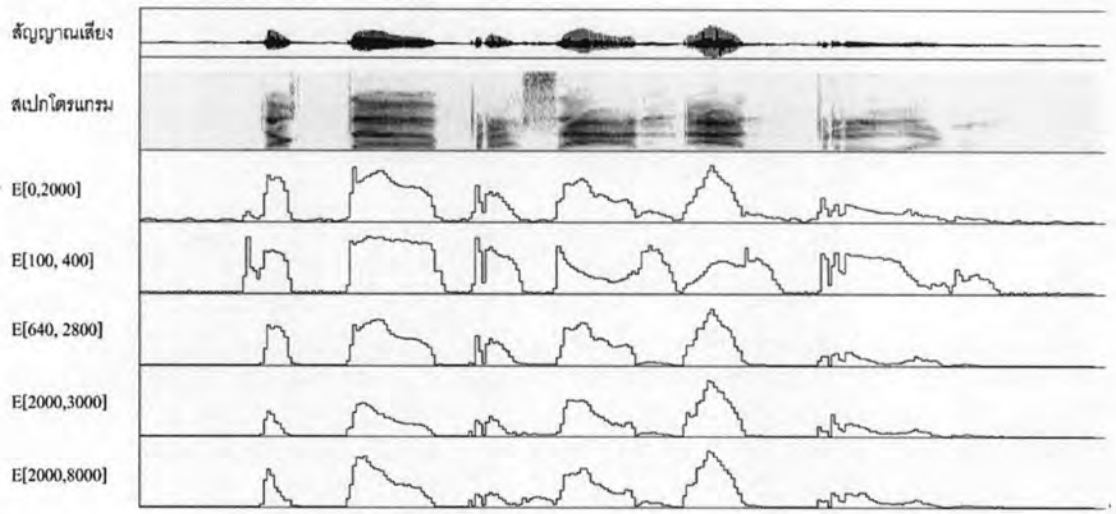
พลังงานของสัญญาณเป็นค่าลักษณะสำคัญที่นิยมนำมาใช้วิเคราะห์สัญญาณเสียง โดยค่าพลังงานของสัญญาณ $s(n)$ ใดๆที่แปรตามเวลาสามารถนิยามได้ว่า

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

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

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

และในการหาค่าพลังงานในช่วงความถี่ต่างๆ จะทำการกรองความถี่ของสัญญาณเสียงโดยใช้ตัวกรองความถี่ชนิดแถบความถี่ผ่าน (Band-pass filter) ซึ่งจะยอมให้สัญญาณที่มีความถี่ในขอบเขตที่กำหนดผ่านไปได้และจะกรองสัญญาณที่ความถี่ส่วนอื่นๆทิ้ง แล้วจึงนำสัญญาณที่กรองได้นี้ไปคำนวณค่าพลังงานของสัญญาณเสียงในช่วงความถี่ที่ต้องการดังแสดงได้ด้วยรูปที่ 3.3



รูปที่ 3.3 ค่าพลังงานของสัญญาณเสียงบนช่วงความถี่ต่างๆ

1.2.2 อัตราการตัดศูนย์

ค่าอัตราการตัดศูนย์ของสัญญาณเสียง คือจำนวนครั้งของการเปลี่ยนแปลงเครื่องหมายจากบวกเป็นลบหรือจากลบเป็นบวกของแอมพลิจูดของสัญญาณเสียงต่อหนึ่งหน่วยเวลา ค่าลักษณะสำคัญนี้นิยมใช้กันมากในระบบรู้จำเสียงพูด โดยสามารถนำมาใช้วัดระดับเสียงรบกวนหรือนำมาไปประยุกต์วัดระดับความก้องของสัญญาณเสียงได้ อัตราการตัดศูนย์สามารถคำนวณได้จากสมการดังต่อไปนี้

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

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

1.2.3 อัตสหสัมพันธ์ (Autocorrelation coefficients)

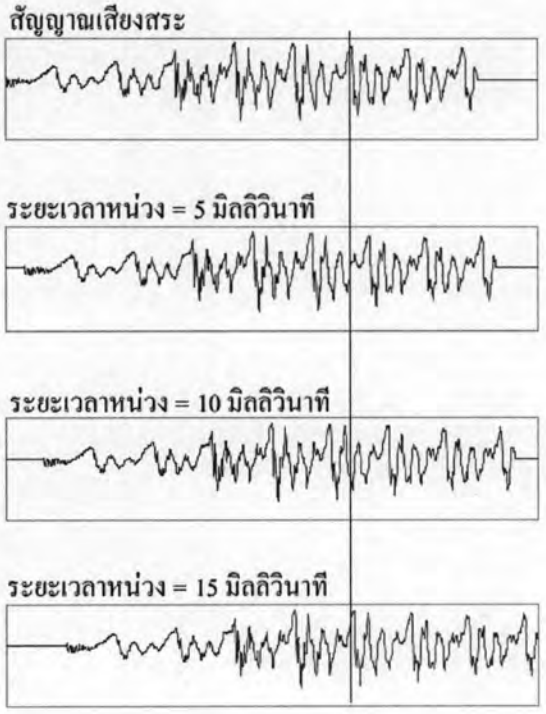
ค่าอัตสหสัมพันธ์เป็นค่าลักษณะสำคัญที่สามารถนำมาใช้วัดระดับความเป็นคาบของสัญญาณเสียง โดยพื้นฐานแล้วค่าอัตสหสัมพันธ์สามารถคำนวณได้จากการเปรียบเทียบสัญญาณเสียงที่เวลาหนึ่งกับสัญญาณเสียงเดียวกันที่ตำแหน่งซึ่งเยื้องหรือหน่วงเวลาออกไปดังแสดงด้วยรูปที่ 3.4 จากกราฟด้านบนของรูปจะเป็นสัญญาณเสียงสระซึ่งมีลักษณะเป็นคาบ ซึ่งการหา

ค่าอัตสหสัมพันธ์จะพิจารณาสัญญาณเสียงที่ตำแหน่งที่ต้องการหาค่าอัตสหสัมพันธ์ว่ามีความคล้ายกันกับสัญญาณเสียงที่ตำแหน่งที่มีการหน่วงเวลาไปมากน้อยแค่ไหน

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

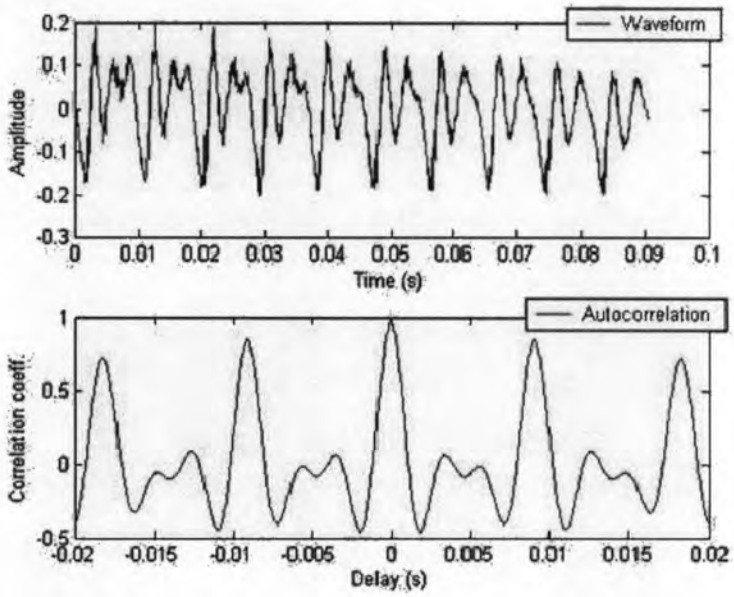
$$autocorr(m,k) = \frac{1}{N\sigma^2}\sum_{n=1}^{N}[s(n)w(m-n)-\mu][s(n+k)w(m-n+k)-\mu]$$

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



รูปที่ 3.4 สัญญาณเสียงสระที่ระยะเวลาหน่วงต่างๆ

สัญญาณเสียงที่มีลักษณะเป็นคาบจะมีค่าอัตสหสัมพันธ์สูงเมื่อหน่วงเวลาไปเป็นจำนวนเท่าของคาบของสัญญาณเสียงนั้น ดังรูปที่ 3.5 จากกราฟในรูปจะสังเกตเห็นว่าค่าอัตสหสัมพันธ์จะมีค่าสูงสุดเมื่อไม่มีการหน่วงเวลา และค่าอัตสหสัมพันธ์สูงสุดอันดับต่อๆไปจะอยู่ที่ระยะหน่วงเวลาซึ่งเป็นจำนวนเท่าของคาบของสัญญาณเสียงนั้น



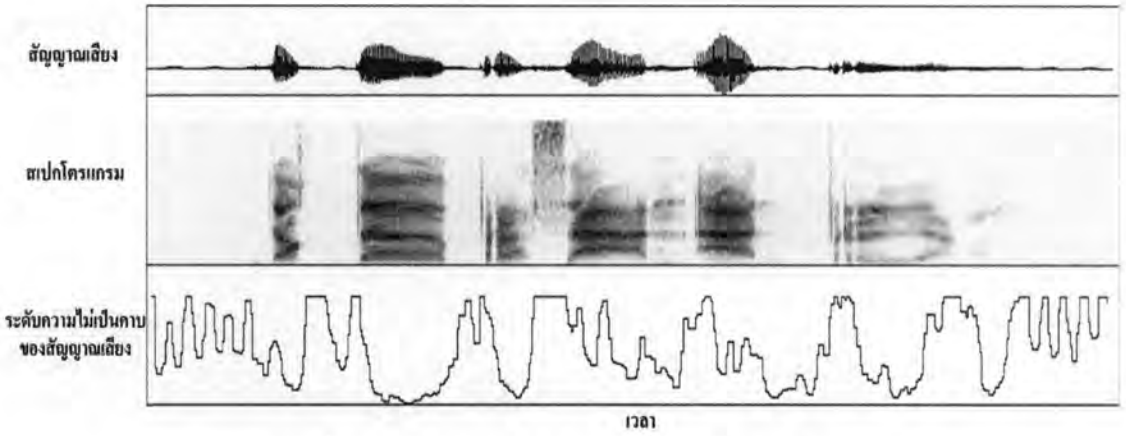
รูปที่ 3.5 สัญญาณเสียงที่มีลักษณะเป็นคาบ และค่าอัตสหสัมพันธ์ที่ระยะเวลาหน่วงต่างๆ

1.2.4 ค่าระดับความไม่เป็นคาบของสัญญาณเสียง

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

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

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



รูปที่ 3.6 ระดับความไม่เป็นคาบของสัญญาณเสียง

1.2.5 ค่าระดับความก้องของเสียง

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

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

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

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

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

ค่าคะแนนที่ได้จากค่าพลังงานของสัญญาณเสียงจะคำนวณจากสมการดังนี้

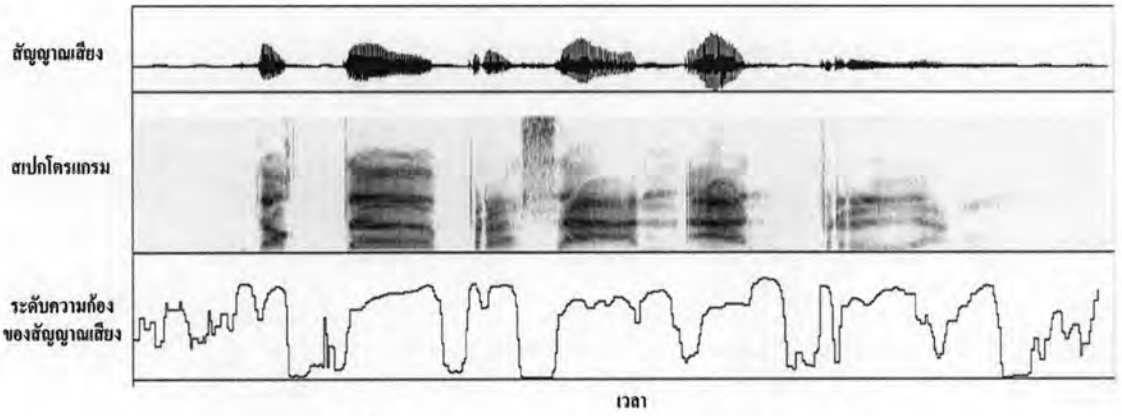
$$ep(m) = \frac{1}{1 + e^{-\left(\frac{E(m) - maxenergy}{5}\right)}}$$

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

ค่าคะแนนที่ได้จากอัตราการตัดศูนย์จะคำนวณจากสมการต่อไปนี้

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

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



รูปที่ 3.7 ระดับความก้องของสัญญาณเสียง

2. การเรียนรู้การจำแนกลักษณะการออกเสียง

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

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

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

ตารางที่ 3.2 หน่วยเสียงที่มีและไม่มีคุณสมบัติตามสัทลักษณ์แบ่งตามลักษณะการออกเสียง

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

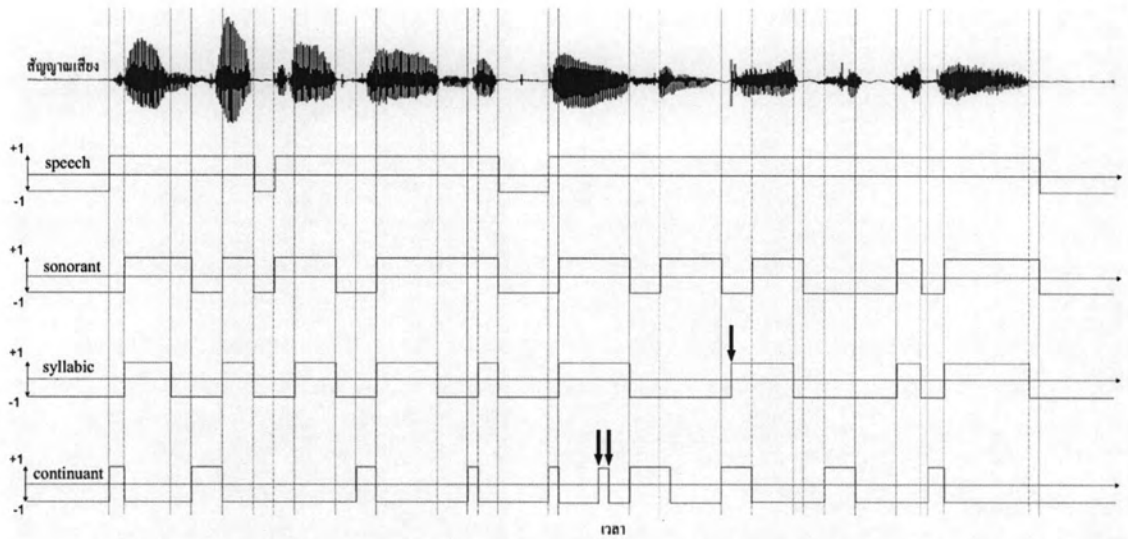
[1]ใช้หน่วยเสียงตามสัทอักษรสากล (International Phonetic Alphabet – IPA)

3. การตรวจหาขอบเขตของหน่วยเสียงจากผลการจำแนกลักษณะการออกเสียง

การตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนนี้ จะอาศัยผลที่ได้จากการจำแนกลักษณะการออกเสียง โดยพิจารณาตำแหน่งที่มีการเปลี่ยนแปลงลักษณะการออกเสียงออกมาเป็นขอบเขตของหน่วยเสียง ดังแสดงด้วยรูปที่ 3.8 ซึ่งขอบเขตของหน่วยเสียงนี้จะมีความละเอียดอยู่ในระดับเดียวกันกับระยะห่างของแต่ละกรอบเวลาที่กำหนดเป็นพารามิเตอร์ไว้ในขั้นตอนการสกัดลักษณะสำคัญเพื่อการเรียนรู้และการจำแนก โดยในที่นี้ความละเอียดจะอยู่ที่ 10 มิลลิวินาที

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

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



รูปที่ 3.8 ผลการแบ่งแยกสัทลักษณ์ลักษณะการออกเสียงด้วยซัพพอร์ตเวกเตอร์แมชชีน ตำแหน่งที่ลูกศรชี้คือขอบเขตของหน่วยเสียงที่ถูกกรองออก

การสร้างกราฟของเซกเมนต์

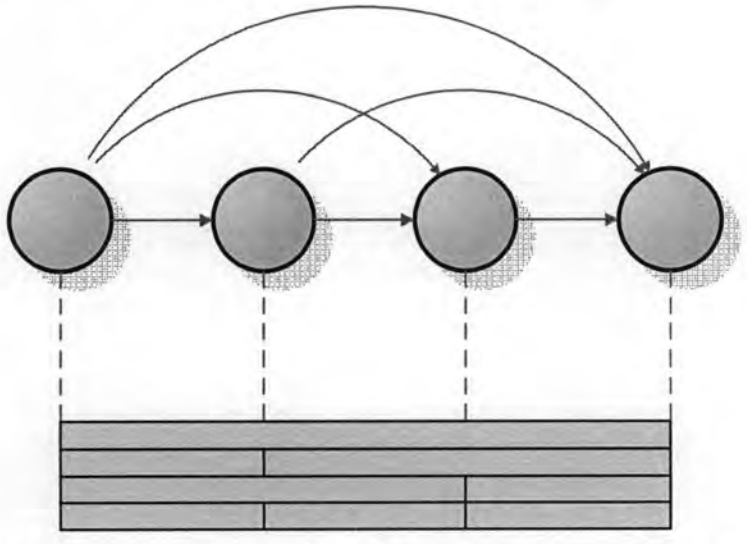
การสร้างกราฟของเซกเมนต์ เป็นขั้นตอนที่อาศัยลำดับของขอบเขตของหน่วยเสียงที่ได้จากกระบวนการตรวจหาขอบเขตของหน่วยเสียงมาสร้างกราฟของเซกเมนต์ เพื่อนำมาใช้ในการรู้จำเสียงพูดในขอบข่ายงานของระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ โดยในหัวข้อนี้จะเริ่มต้นด้วยการนำเสนอเกี่ยวกับวิธีในการสร้างกราฟของเซกเมนต์สามวิธีได้แก่

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

ต่อจากนั้นจะนำเสนอเกี่ยวกับ การทดลองเปรียบเทียบประสิทธิภาพในการสร้างกราฟของเซกเมนต์ของทั้งสามวิธี และวิเคราะห์ผลการทดลอง

1. การสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง

การสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง เป็นวิธีการสร้างกราฟของเซกเมนต์แบบที่ง่ายที่สุด โดยจะอาศัยลำดับของขอบเขตของหน่วยเสียงที่ได้จากขั้นตอนการตรวจหาขอบเขตของหน่วยเสียงมาเป็นอินพุต แล้วพิจารณาสร้างเซกเมนต์ขึ้นมาด้วยการจับคู่ขอบเขตของหน่วยเสียงทุกๆคู่ขึ้นมาเป็นขอบเขตของเซกเมนต์ ให้ครบทุกกรณี ดังแสดงด้วยภาพตัวอย่างในรูปที่ 3.9 วงกลมแต่ละวงแทนขอบเขตของหน่วยเสียงแต่ละตำแหน่ง เมื่อเชื่อมต่อครบทุกขอบเขตของหน่วยเสียงก็จะได้เซกเมนต์ซึ่งแสดงด้วยสี่เหลี่ยมผืนผ้า

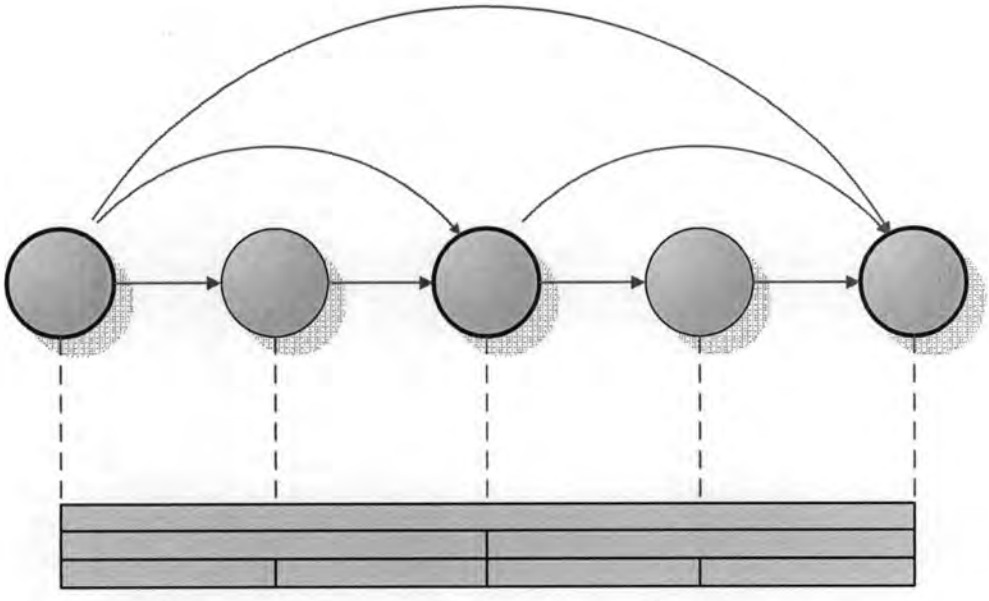


รูปที่ 3.9 กราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง

2. การสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม

การสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม เป็นวิธีการสร้างกราฟของเซกเมนต์แบบหนึ่งที่อาศัยการเปลี่ยนแปลงสเปกตรัมมาตัดบางเซกเมนต์ออกไป เพื่อให้ได้กราฟที่มีขนาดเล็กลงกว่าวิธีการแบบเชื่อมต่อทุกขอบเขตของหน่วยเสียงเข้ากันหมด โดยใช้แนวคิดที่ว่าตำแหน่งของสัญญาณเสียงที่มีการเปลี่ยนแปลงสเปกตรัมสูงมักจะเป็นขอบเขตของหน่วยเสียง โดยจะวัดค่าการเปลี่ยนแปลงสเปกตรัมที่ทุกๆขอบเขตของหน่วยเสียงเอาไว้ และพิจารณาขอบเขตของหน่วยเสียงออกเป็นสองพวกคือ ขอบเขตของหน่วยเสียงที่มีการเปลี่ยนสเปกตรัมสูง และขอบเขตของหน่วยเสียงที่มีการเปลี่ยนแปลงสเปกตรัมต่ำ โดยจะใช้ค่าเฉลี่ยของการเปลี่ยนแปลงสเปกตรัมจากทุกๆขอบเขตของหน่วยเสียงในเสียงพูดนั้นมากำหนดระดับของการแบ่งแยก

การเชื่อมต่อขอบเขตของหน่วยเสียงด้วยวิธีการนี้ ขอบเขตของหน่วยเสียงที่มีการเปลี่ยนแปลงสเปกตรัมต่ำทั้งหมดที่อยู่ระหว่างขอบเขตของหน่วยเสียงที่มีการเปลี่ยนสเปกตรัมสูงจะเชื่อมต่อกันหมด และขอบเขตของหน่วยเสียงที่มีการเปลี่ยนสเปกตรัมสูงทั้งหมดจะเชื่อมต่อกันเอง โดยจะไม่ยินยอมให้ขอบเขตของหน่วยเสียงที่มีการเปลี่ยนแปลงสเปกตรัมต่ำเชื่อมต่อข้ามขอบเขตของหน่วยเสียงที่มีการเปลี่ยนสเปกตรัมสูง ดังแสดงด้วยภาพตัวอย่างในรูปที่ 3.10 วงกลมที่มีเส้นรอบวงหนาจะใช้แทนขอบเขตของหน่วยเสียงที่มีการเปลี่ยนสเปกตรัมสูง วงกลมที่มีเส้นรอบวงบางจะแทนขอบเขตของหน่วยเสียงที่มีการเปลี่ยนแปลงสเปกตรัมต่ำ



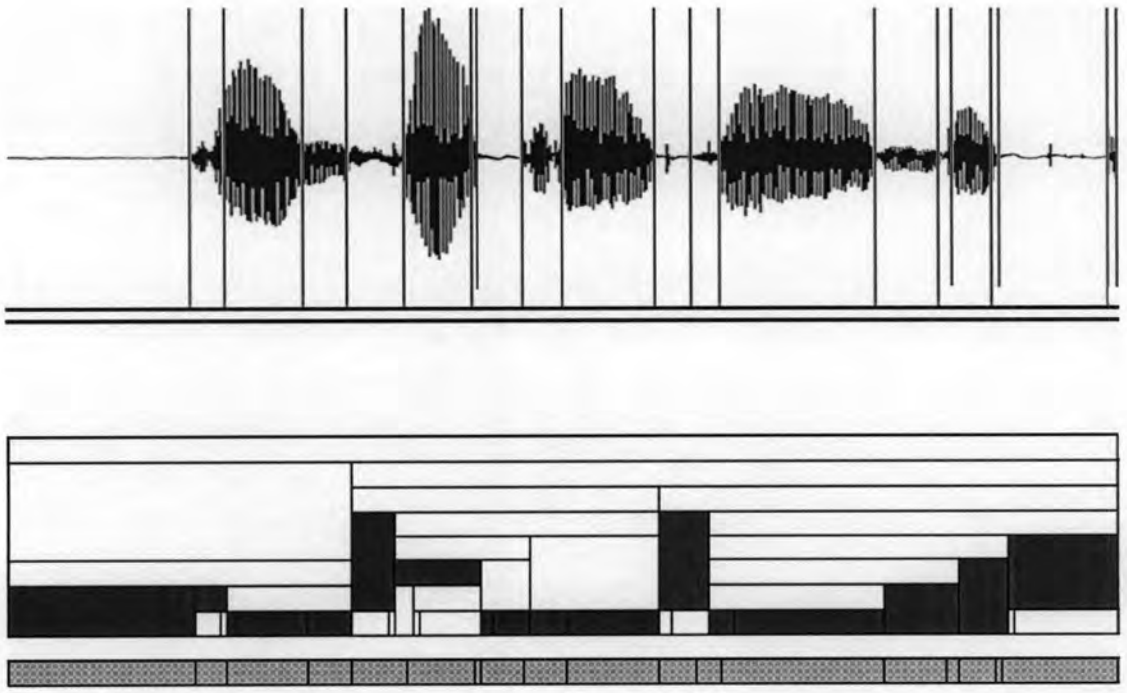
รูปที่ 3.10 กราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม

3. การสร้างกราฟของเซกเมนต์แบบหลายระดับ

วิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับ เป็นการสร้างกราฟของเซกเมนต์ให้มีลักษณะเหมือนเดนโดแกรมดังรูปที่ 3.11

โดยอาศัยหลักการที่พิจารณาให้เสียงพูดเป็นลำดับของของเซกเมนต์ที่เป็นหน่วยเสียง และสัญญาณเสียงที่อยู่ในเซกเมนต์เดียวกันจะมีความคล้ายคลึงกันมากกว่าสัญญาณเสียงที่อยู่ในเซกเมนต์อื่นซึ่งอยู่ติดกัน จากหลักการนี้ลักษณะของปัญหาการสร้างกราฟของเซกเมนต์สามารถเปลี่ยนไปเป็นปัญหาการรวมกลุ่ม (Clustering) โดยอาศัยการวัดค่าความคล้ายคลึงกันของสัญญาณเสียง แล้วค่อยๆรวมเซกเมนต์ที่มีความคล้ายกันเข้าด้วยกัน อัลกอริทึมการสร้างกราฟของเซกเมนต์แบบหลายระดับแสดงด้วยอัลกอริทึมดังตารางที่ 3.3

ในที่นี้เราใช้เวกเตอร์ลักษณะสำคัญของค่าสัมประสิทธิ์สเปกตรัมบนสเกลเมลที่ได้จากขั้นตอนการสกัดลักษณะสำคัญมาเป็นตัวเปรียบเทียบความคล้ายคลึงกันระหว่างสัญญาณเสียง การพิจารณาหาความคล้ายกันระหว่างสัญญาณเสียง จะวัดและแสดงอยู่ในรูปผลต่างของการกระจัดแบบยูคลิเดียน (Euclidean Distance) ระหว่างสัญญาณ โดยหากผลต่างของการกระจัดจากการเปรียบเทียบกันระหว่างหน่วยเสียงมีค่าน้อยเข้าใกล้ศูนย์ ก็จะสรุปได้ว่าสัญญาณเสียงคู่นั้นมีความคล้ายคลึงกันมาก



รูปที่ 3.11 กราฟของเซกเมนต์แบบหลายระดับ [26]

ตารางที่ 3.3 อัลกอริทึมการแบ่งเสียงพูดเป็นเซกเมนต์แบบหลายระดับ [26]

Find boundaries

$$\{b_i, 0 < i < N\},\ t_i < t_j,\ \forall i < j$$

Create initial region set

$$R = \{r_0(i), 0 < i < N\},\ r_0(i) \equiv r(i, i+1)$$

Create initial distance set

$$D_0 = \{d_0(i), 0 < i < N\},\ d_0(i) \equiv d(r_0(i), r_0(i+1))$$

Until $R = \{r_N(0)\} \equiv r(0, N)$

For any k such that:

$$d_j(k-1) > d_j(k) < d_j(k+1)$$

(a) $r_{j+1}(i) = r_j(i),\ 0 \le i < k$

(b) $r_{j+1}(k) = merge(r_j(k),\ r_j(k+1))$

(c) $r_{j+1}(i) = r_j(i+1),\ k < i < N - j - 1$

(d) $R_{j+1} = \{r_{j+1}(i),\ 0 \le i < N - j - 1$

(e) $d_{j+1}(i) = d_j(i),\ 0 \le i < k - 1$

(f) $d_{j+1}(k-1) = \max(d_j(k-1), d(r_j(k-1), r_j(k)))$

(g) $d_{j+1}(k) = \max(d_j(k+1), d(r_{j+1}(k), r_j(k+1)))$

(h) $d_{j+1}(i) = d_j(i+1),\ k \le i < N - j - 1$

(i) $D_{j+1} = \{d_{j+1}(i),\ 0 \le i < N - j - 1\}$

เมื่อกำหนดให้

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

การให้คะแนนขอบเขตของหน่วยเสียง

ในการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับนำไปใช้ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์นั้น กราฟของเซกเมนต์ที่สร้างออกมาจะนำไปใช้ต่อในขั้นตอนการค้นหาและให้คะแนนเพื่อการรู้จำเสียงพูด งานวิจัยนี้ได้เสนอให้มีการคิดคะแนนขอบเขตของหน่วยเสียงให้กับกราฟของเซกเมนต์ที่สร้างออกมา โดยเป็นคะแนนแสดงความมั่นใจว่าขอบเขตของหน่วยเสียงนั้นเป็นขอบเขตของหน่วยเสียงจริง ซึ่งจะเป็นประโยชน์ในขั้นตอนการรู้จำเสียงพูดต่อไป

โดยอาศัยแนวคิดที่ว่าตำแหน่งของสัญญาณเสียงที่เป็นขอบเขตของหน่วยเสียงโดยส่วนใหญ่แล้วจะมีการเปลี่ยนแปลงสเปกตรัมสูง [27] ดังนั้นค่าคะแนนของขอบเขตของหน่วยเสียงในที่นี้จะอาศัยค่าการเปลี่ยนแปลงสเปกตรัม มาคำนวณออกมาเป็นความน่าจะเป็นที่ตำแหน่งทางเวลานั้นจะเป็นขอบเขตของหน่วยเสียง ซึ่งการเปลี่ยนแปลงสเปกตรัมนิยามให้เป็นไปตามสมการต่อไปนี้

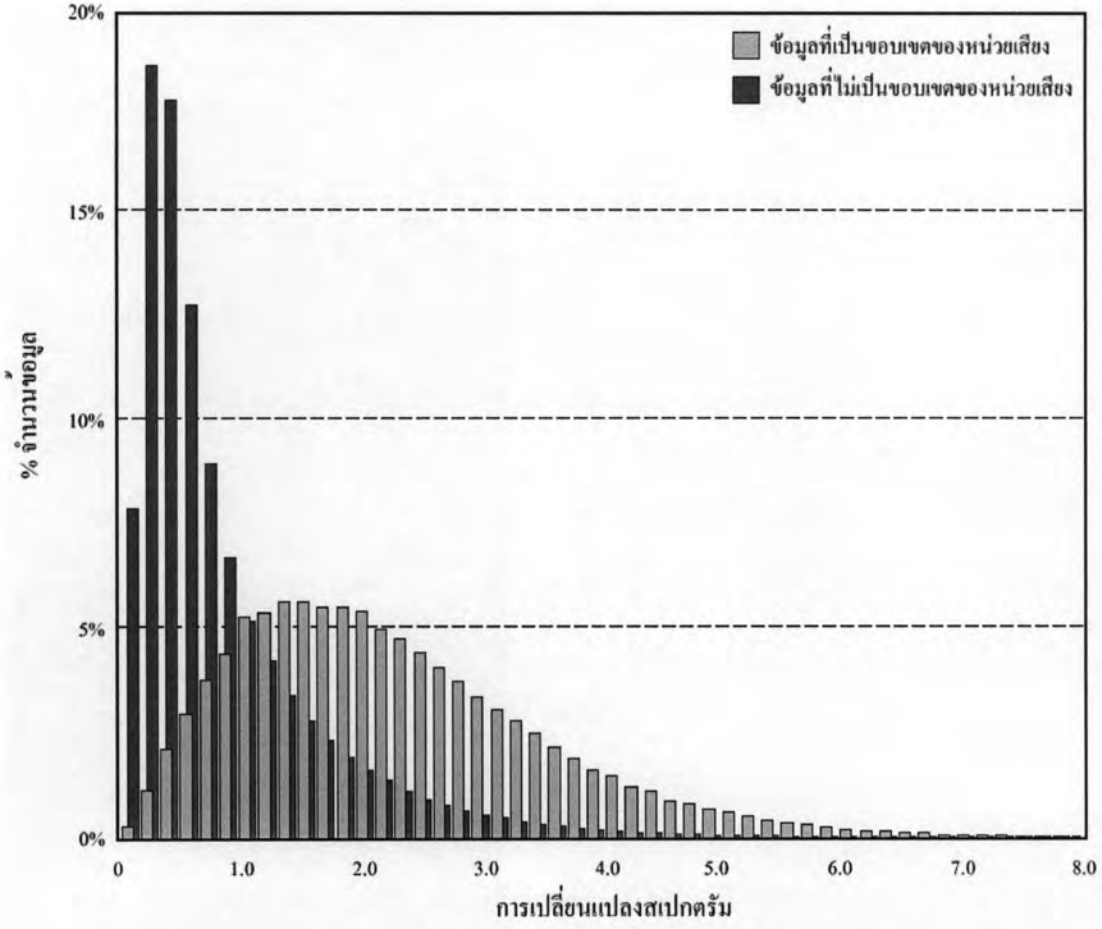
$$STM(m) = \frac{\sum_{i=1}^{D} a_i^2(m)}{D}$$

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

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

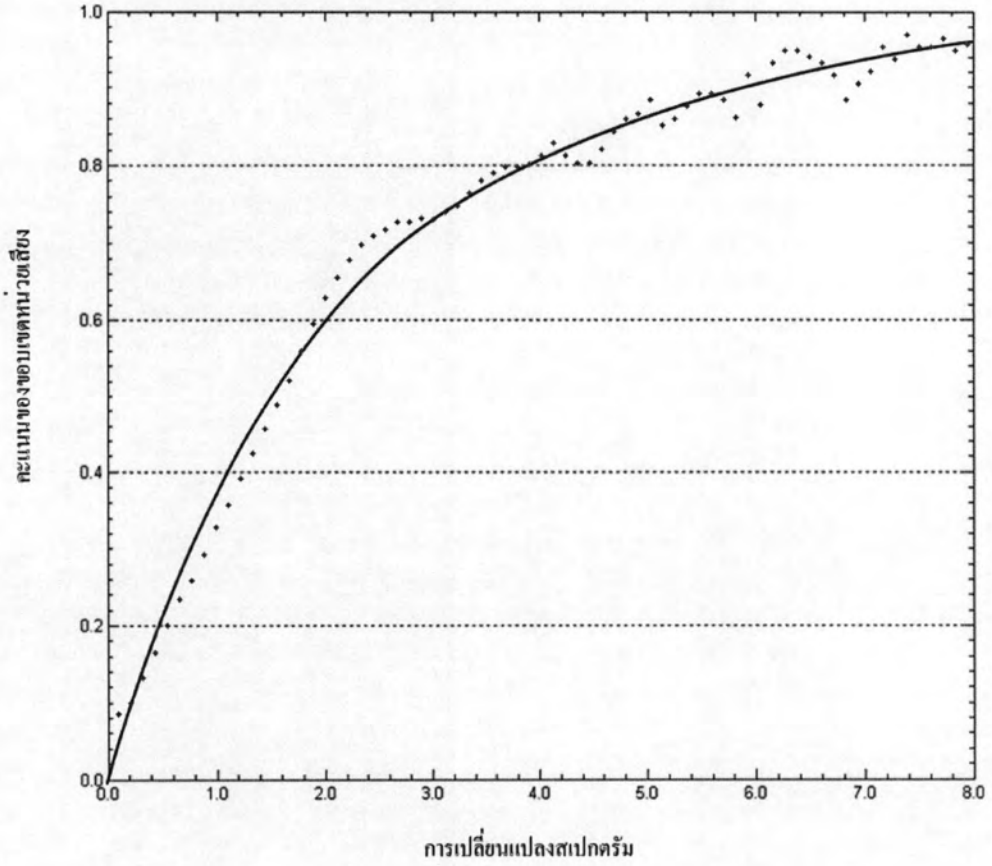
โดยที่ n แทนดัชนีของกรอบเวลา และ I แทนจำนวนของกรอบเวลาสำหรับใช้คำนวณอัตราการเปลี่ยนแปลงสเปกตรัม ในที่นี้กำหนดให้ $I = 2$ ดังนั้นจะเป็นการนำกรอบเวลาที่อยู่ติดกันสองกรอบเวลาก่อนหน้าและสองกรอบเวลาตามหลังมาคำนวณหาการเปลี่ยนแปลงสเปกตรัมที่ตำแหน่งกึ่งกลางตรงกลาง

เพื่อตรวจสอบความถูกต้องและความเป็นไปได้ของการนำแนวคิดนี้มาใช้ งานวิจัยนี้จึงทดลองวัดการกระจายค่าการเปลี่ยนแปลงสเปกตรัมของข้อมูลที่เป็นและไม่เป็นขอบเขตของหน่วยเสียงให้อยู่ในรูปของฮิสโตแกรม เพื่อนำมาอธิบายการแจกแจงความน่าจะเป็นในการเป็นขอบเขตของหน่วยเสียงที่การเปลี่ยนแปลงสเปกตรัมต่างๆ ดังรูปที่ 3.12



รูปที่ 3.12 ฮิสโตแกรมของข้อมูลที่เป็นและไม่เป็นขอบเขตของหน่วยเสียง

จากฮิสโตแกรม พบว่าตำแหน่งของสัญญาณเสียงพูดที่เป็นขอบเขตของหน่วยเสียงส่วนใหญ่จะมีระดับการเปลี่ยนแปลงสเปกตรัมสูงกว่าตำแหน่งของสัญญาณเสียงพูดที่ไม่ได้เป็นขอบเขตของหน่วยเสียง โดยค่าความน่าจะเป็นของการที่ตำแหน่งของสัญญาณเสียงนั้นเป็นขอบเขตของเสียงพูดที่ระดับการเปลี่ยนแปลงสเปกตรัมใดๆจะคำนวณมาจากอัตราส่วนระหว่างจำนวนข้อมูลที่เป็นขอบเขตของหน่วยเสียง ต่อจำนวนข้อมูลทั้งหมดที่มีระดับการเปลี่ยนแปลงสเปกตรัมนั้นๆ แทนด้วยจุดต่างๆ และสามารถประมาณเส้นโค้งของกราฟความสัมพันธ์ระหว่างการเปลี่ยนแปลงสเปกตรัมและคะแนนขอบเขตของหน่วยเสียงได้ดังรูปที่ 3.13



รูปที่ 3.13 ความสัมพันธ์ระหว่างการเปลี่ยนแปลงสเปกตรัมและคะแนนของขอบเขตหน่วยเสียง