บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

ในบทนี้เป็นการอธิบายทฤษฎีต่าง ๆ ที่เกี่ยวข้องกับงานวิจัย โดยจะกล่าวถึงทฤษฎีทางด้านภาษาศาสตร์ ทฤษฎีทางด้านการวิเคราะห์สัญญาณเสียง และทฤษฎีปรับปรุงคุณลักษณะสัญญาณเสียงพูดและปรับปรุงขอบเขตพยางค์

ทฤษฎีทางด้านภาษาศาสตร์ ประกอบด้วยหน่วยเสียงพยัญชนะและหน่วยเสียงสระในภาษาไทย การอธิบายพยางค์ โครงสร้างพยางค์ทางสัทวิทยา ทฤษฎีทางด้านการวิเคราะห์สัญญาณเสียงประกอบด้วย การวิเคราะห์สัญญาณเสียงพูดเบื้องต้น ซึ่งมีกรรมวิธีการปรับบรรทัดฐานแอมปลิจูด (Amplitude Normalization) กรรมวิธีการเน้นล่วงหน้า (Preemphasis) กรรมวิธีการวางกรอบหน้าต่าง (Windowing) การวิเคราะห์คุณลักษณะของสัญญาณเสียงพูด ซึ่งเป็นคุณลักษณะที่ใช้ในการพิจารณาหาขอบเขตพยางค์ ส่วนทฤษฎีการปรับปรุงคุณลักษณะและปรับปรุงขอบเขตพยางค์ ประกอบด้วยวิธีการปรับเรียบ (Smoothing) ซึ่งเป็นวิธีที่นำมาใช้เพื่อปรับให้แผนภูมิเส้นของคุณลักษณะของสัญญาณเสียงพูดเรียบขึ้น ช่วยให้การตัดสินใจในการเลือกขอบเขตพยางค์ถูกต้องมากยิ่งขึ้น และใช้กรรมวิธีปรับปรุงขอบเขตพยางค์ในขั้นตอนสุดท้ายของกรรมวิธีการหาขอบเขตพยางค์

2.1 ทฤษฎีทางด้านภาษาศาสตร์ (Linguistics Theory)

2.1.1 หน่วยเสียงพยัญชนะและสระ (Consonant and Vowel Phonemes)

พยัญชนะในภาษาไทยมี 21 หน่วยเสียง แบ่งเป็น พยัญชนะกัก (Stop Consonants) 11 หน่วยเสียง และพยัญชนะไม่กัก (Non-Stop Consonants) 10 หน่วยเสียง ดังแสดงในตารางที่ 2.1 พยัญชนะทั้ง 21 หน่วยเสียง สามารถอยู่ในตำแหน่งต้นพยางค์ได้ทุกหน่วยเสียง แต่มีเพียง 9 หน่วยเสียงเท่านั้นที่สามารถอยู่ในตำแหน่งท้ายพยางค์ได้ คือ เสียงกักไม่พ่นลม 4 หน่วยเสียง [p, t, k, ?] เสียงนาสิก 3 หน่วยเสียง [m, n, ŋ] และเสียงกึ่งสระ 2 หน่วยเสียง [w, j]

ตารางที่ 2.1 เสียงพยัญชนะไทย (ณัฐกร ทับทอง, 2538)

ฐานที่เกิดเสียง (Points of Articulation) / ลักษณะของเสียง (Manners of Articulation)		ริมฝีปาก (Bilabial)	ปุ่มเหงือก (Alveolar)	เพดานแข็ง (Palatal)	เพดานอ่อน (Velar)	เส้นเสียง (Gottal)
พยัญชนะกัก (Stop Consonants)	ไม่พ่นลม (Unaspirated)	p*	t*	c	k*	?*
	พ่นลม (Aspirated)	ph	th	ch	kh	
	ก้อง (Voiced)	b	d			
พยัญชนะไม่กัก (Non Stop Consonant)	นาสิก (Nasal)	m*	n*		ŋ*	
	เสียดแทรก (Fricative)	f	s			h
	กระทบ (Flapped)		r			
	ข้างลิ้น (Lateral)		l			
	กึ่งสระ (Semi-Vowel)	w*		j*		

<u>หมายเหตุ</u> หน่วยเสียงพยัญชนะที่มีเครื่องหมายดอกจันทร์ (*) คือ หน่วยเสียงที่สามารถปรากฏในตำแหน่งท้ายพยางค์ได้

เสียงสระในภาษาไทยมี 24 หน่วยเสียง แบ่งตามลักษณะการเคลื่อนไหวของลิ้นได้เป็น 2 ชนิดคือ สระเดี่ยว 18 หน่วยเสียง และสระประสม 6 หน่วยเสียง ในการบรรยายลักษณะของสระจะพิจารณาระดับของลิ้น ตำแหน่งของลิ้น และลักษณะของริมฝีปาก ดังตารางที่ 2.2

ตารางที่ 2.2 เสียงสระไทย (อุดม วิโรตม์สิกขดิตถ์, 2521)

ประเภท (Vowel Type)	ตำแหน่งของลิ้น (Tongue Position) / ระดับของลิ้น (Tongue Level)	หน้า (Front)		กลางค่อนไปทางหลัง(Central)		หลัง (Back)	
		สั้น	ยาว	สั้น	ยาว	สั้น	ยาว
สระเดี่ยว (Single Vowel)	สูง (High)	i	ii	ɯ	ɯɯ	u	uu
	กลาง (Mid)	e	ee	ɤ	ɤɤ	o	oo
	ต่ำ (Low)	ɛ	ɛɛ	a	aa	ɔ	ɔɔ
สระประสม (Diphthongs)		ia	iia	ɯa	ɯɯa	ua	uua

2.1.2 การอธิบายพยางค์

การอธิบายและให้คำจำกัดความพยางค์ แบ่งออกเป็น

1. การอธิบายและให้คำจำกัดความพยางค์ในทางสัทศาสตร์ (Phonetics)
2. การอธิบายและให้คำจำกัดความพยางค์ในทางสัทวิทยา (Phonology)

1. การอธิบายและให้คำจำกัดความพยางค์ในทางสัทศาสตร์ แบ่งออกเป็น 2 ลักษณะ (อุมาพร ศรีรักษา, 2538) ดังนี้คือ

- การอธิบายพยางค์ในทางสรีรศาสตร์ (Articulatory Phonetics)

เป็นการอธิบายพยางค์โดยพิจารณาในแง่การเปล่งเสียงพูด (Activities of the Speakers) เช่น การพิจารณาปริมาณความกว้างของอวัยวะที่ใช้ในการเปล่งเสียง การพิจารณาแรงดันลมจากปอด เป็นต้น

เดวิด อเบอร์ครอมบี (Abercrombie, อ้างถึงใน อุมาพร ศรีรักษา, 2538) ได้อธิบายว่า พยางค์เกิดจากการหดตัวของกล้ามเนื้อในช่องอกทำให้เกิดแรงดันลม ดันลมจากปอดเป็นช่วง ๆ ต่อกัน ซึ่งแรงดันลมที่ออกมาเป็นช่วง ๆ เกิดจากการที่กล้ามเนื้อที่ใช้ในการหายใจ (Respiratory Muscles) หดตัวและคลายตัวสลับกัน ลมดังกล่าวเคลื่อนที่ผ่านช่องทางเดินเสียง (Vocal Tract) และอวัยวะกำเนิดเสียงในช่องทางเดินเสียงคือ ช่องปาก (Oral Cavity) ช่องคอ (Pharyngeal Cavity) ช่องจมูก (Nasal Cavity) ลิ้น ริมฝีปาก เพดานอ่อน และเส้นเสียง ที่เปลี่ยนรูปไปในลักษณะต่าง ๆ ทำให้เกิดเสียงที่มีคุณสมบัติต่าง ๆ กัน เสียงสระถือเป็นแกนกลางพยางค์โดยมีเสียงพยัญชนะเป็นขอบพยางค์ พยัญชนะที่ต้นพยางค์เกิดจากการระบายลมที่มีการหดรัดช่องทางเดินเสียง การหดรัดเกิดขึ้นที่อวัยวะกำเนิดเสียง (Articulatory Organ) จึงให้กำเนิดแรงดันลม แต่ถูกกักกั้นไว้ชั่วขณะหนึ่ง และเมื่อลมนั้นถูกปล่อยออกมาจะเป็นเสียงสระ ส่วนพยัญชนะที่อยู่ท้ายพยางค์เกิดจากการ

เคลื่อนไหวอวัยวะกำเนิดเสียงหดรัดช่องทางเดินเสียง เมื่อช่องทางลมถูกปิดกั้นจึงทำให้พยางค์สิ้นสุดลง

- การอธิบายพยางค์ในทางโสตศาสตร์ (Auditory Phonetics)

เป็นการอธิบายพยางค์โดยพิจารณาในแง่คุณสมบัติของเสียง (Properties of Sound) ที่ได้จากการฟัง เช่น ระดับความก้องของเสียง (Sonority) ความดังเด่นของเสียง (Prominence) เป็นต้น

ออตโต เจสเปอร์เซน (Jesperson , อ้างถึงใน อุมาพร ศรีรักษา, 2538) ได้อธิบายว่า เสียงมักจะจับกลุ่มกันตามคุณสมบัติของความก้องประจำเสียง (Sonority) หน่วยเสียงจะจับกลุ่มรอบหน่วยเสียงที่มีความก้องสูงสุด เจสเปอร์เซน ได้จัดกลุ่มหน่วยเสียงที่มีความก้องมากที่สุดไปหาน้อยที่สุดดังนี้

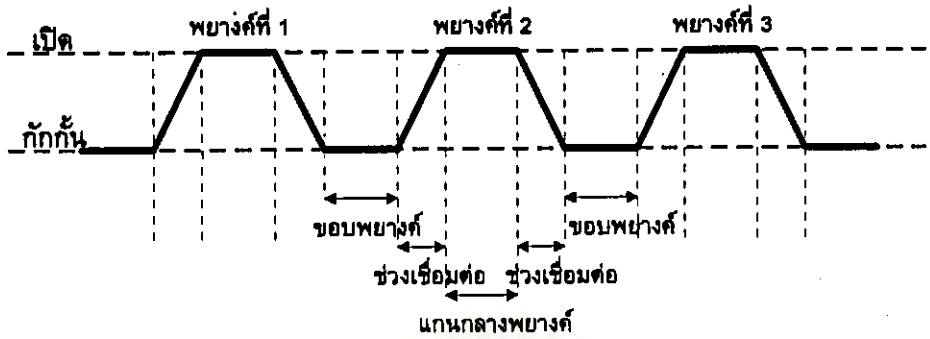
1. เสียงสระเปิด [a,...]
2. เสียงสระกึ่งเปิด [e,ø,ɤ,ɐ]
3. เสียงสระปิด [ɪ,y,u]
4. เสียงรัว [r]
5. เสียงนาสิกและข้างลิ้น [m,n,l,...]
6. เสียงเสียดแทรกก้อง [v,z,...]
7. เสียงกักก้อง [b,d,g]
8. เสียงเสียดแทรกไม่ก้อง [f,s]
9. เสียงกักไม่ก้อง [p,t,k]

2. การอธิบายและให้คำจำกัดความพยางค์ในทางสัทวิทยา

ทฤษฎีทางสัทวิทยา อธิบายพยางค์ในแง่โครงสร้างและหน้าที่ของพยางค์ในภาษา ดังนั้นการอธิบายจึงมุ่งไปที่การรวมหน่วยเสียงเข้าเป็นพยางค์ (Ernst Pulgram, 1970 อ้างถึงใน อุมาพร ศรีรักษา, 2538)

โรเจอร์ แลส (Roger Lass, อ้างถึงใน อุมาพร ศรีรักษา, 2538) อธิบายว่า พยางค์เป็นหน่วยโครงสร้างทางเสียงที่เล็กที่สุดหน่วยหนึ่ง ที่มีสระเป็นแกนกลางพยางค์ และมีพยัญชนะเดี่ยวหรือควบกล้ำอยู่รอบ ๆ แกนกลางพยางค์

รูปที่ 2.1 แสดงให้เห็นว่าแต่ละพยางค์ประกอบด้วยส่วนขอบพยางค์ และแกนกลางพยางค์ ช่วงที่อยู่ระหว่างส่วนขอบพยางค์และใจกลางพยางค์ เรียกว่า "ช่วงเชื่อมต่อ" (Formant Transition, F-Trans) ถ้า F-Trans อยู่หน้าพยัญชนะเรียกว่า Pre-Consonantal Formant Transition (Pre-C-F-Trans) ส่วน F-Trans ที่อยู่หลังพยัญชนะเรียกว่า Post Consonantal Formant Transition (Post-C-F-Trans)



รูปที่ 2.1 ส่วนประกอบของพยางค์

2.1.3 สัทสัมพันธ์ (Suprasegmentals)

ลักษณะสำคัญที่ปรากฏร่วมกับพยัญชนะและสระในพยางค์หรือถ้อยความดลอดเวลา คือ สัทสัมพันธ์ ซึ่งได้แก่ การเน้น (Stress) ความดัง (Loudness) จังหวะการพูด (Rhythm) ความเร็วในการพูด (Tempo) ความยาว (Length) และระดับเสียง (Pitch)

1. การเน้นพยางค์

การเน้นพยางค์ หมายถึง การใช้ปริมาณพลังกระแสลมที่ออกมาจากปอดมากในขณะพูดพยางค์ใดมีการเน้น (Stressed Syllable) จะมีพลังกระแสลมสูงกว่าพยางค์ที่ไม่มีการเน้น (Unstressed Syllable) พลังกระแสลมจากปอดจะสัมพันธ์กับความดัง ดังนั้นผู้ฟังจะมีความรู้สึกว่า พยางค์ที่มีการเน้น จะมีเสียงดังกว่าพยางค์ที่ไม่มีการเน้น

2. ความดัง

ความดังของเสียงพูดขึ้นอยู่กับกำลังลมดันออกจากปอด หรือคือพลังกระแสลมจากปอด เมื่อใดที่พลังกระแสลมจากปอดสูงจะมีความดังมาก และเมื่อใดที่พลังกระแสดลมจากปอดต่ำจะมีความดังน้อย พลังกระแสลมจากปอดมีความสัมพันธ์กับการเน้นพยางค์ และจังหวะการพูด

3. จังหวะการพูด

จังหวะการพูด หมายถึง พลังกระแสลมที่เกิดขึ้นอย่างสม่ำเสมอในช่วงเวลาหนึ่ง ๆ การกำหนดจังหวะที่เกิดขึ้นในภาษาต่าง ๆ มี 2 แบบ คือ การกำหนดจังหวะโดยอาศัยช่วงเวลาการเกิดพยางค์ (Syllable-Timed Rhythm) คือจังหวะจะตกบนพยางค์สุดท้ายของกลุ่ม ซึ่งแต่ละกลุ่มจะมีจำนวนพยางค์ต่าง ๆ กัน และการกำหนดจังหวะโดยอาศัยช่วงเวลาการเน้นพยางค์ (Stress-Timed Rhythm) คือจังหวะจะตกบนพยางค์ที่มีการเน้นเสมอ

4. ความเร็วในการพูด

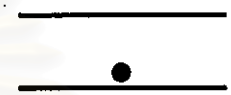
ความเร็วในการพูดสามารถพิจารณาได้จากความต่อเนื่องของพยางค์ในถ้อยความหนึ่ง ๆ ถ้าความต่อเนื่องของพยางค์มีมากจะพูดเร็ว ความชัดเจนของสัทลักษณ์อาจเปลี่ยนแปลงไป แต่ถ้าความต่อเนื่องของพยางค์มีน้อย การพูดจะช้าลง และมีความชัดถ้อยชัดคำ

5. ความยาวของพยางค์

ความสั้นยาวของพยางค์เกิดจากการยืดเสียงในช่วงเวลาหนึ่งออกไปมากหรือน้อย ความสั้นยาวของพยางค์จะสัมพันธ์กับความเร็วช้าในการพูดด้วย

6. ระดับเสียง

ระดับเสียง หมายถึง ความถี่ในการสั่นของเส้นเสียงในการพูด ถ้าเส้นเสียงมีความถี่ในการสั่นสะเทือนสูง ระดับเสียงจะสูง แต่ถ้าเส้นเสียงมีความถี่ในการสั่นสะเทือนต่ำ ก็จะทำให้ระดับเสียงต่ำ ระดับเสียงในระดับต่าง ๆ มีชื่อเรียกดังนี้

ระดับเสียง	รูปแบบของระดับเสียง
ระดับต่ำ (Low)	
ระดับกลาง (Mid)	
ระดับสูง (High)	
ระดับสูง-ตก (Falling)	
ระดับต่ำ-ขึ้น (Rising)	
ระดับตก-ขึ้น (Fall-Rise)	
ระดับขึ้น-ตก (Rise-Fall)	

ระดับเสียงมีบทบาทสำคัญ 2 ประการ คือ ระดับเสียงที่สัมพันธ์กับพยางค์ เมื่อมีการเปลี่ยนระดับเสียงไป ทำให้ความหมายของคำนั้นเปลี่ยนไป เรียกว่า "วรรณยุกต์" (Tone) และระดับเสียงที่สัมพันธ์กับโครงสร้างของถ้อยความหรือประโยค ทำให้ความหมายทางไวยากรณ์ หรือทางทัศนคติเปลี่ยนไป เรียกว่า "ทำนองเสียง" (Intonation)

2.1.4 โครงสร้างพยางค์ทางสัทวิทยา

พยางค์ในภาษาไทย สามารถเขียนรูปของโครงสร้าง ได้ดังนี้

$$S = C(C)V(C)T\#$$

โดยที่ S = พยางค์

C = พยัญชนะ

V = สระ

T = วรรณยุกต์

\# = ขอบเขตพยางค์

( ) = มีหรือไม่มีก็ได้

โครงสร้างพยางค์ทางสัทวิทยา สามารถใช้กำหนดขอบเขตพยางค์ในภาษาไทยได้ ในกรณีที่พยางค์มีพยัญชนะตัวสะกดเป็นเสียงกักตามด้วยพยางค์ที่มีพยัญชนะต้นเป็นเสียงเสียดแทรก นาสิก กัก และเสียงกึ่งสระ แต่ในกรณีที่เป็นปัญหาไม่สามารถใช้โครงสร้างพยางค์ทางสัทวิทยากำหนดขอบเขตพยางค์ได้มี 3 กรณีคือ

1. ปัญหาจากเสียงพยัญชนะ
2. ปัญหาจากเสียงพยัญชนะเรียง
3. ปัญหาจากเสียงสระเรียง

1. ปัญหาจากเสียงพยัญชนะ

เมื่อพยัญชนะปรากฏอยู่ระหว่างสระ 2 เสียง และเป็นพยัญชนะ [p, t, k, m, n, ŋ, w, j] ซึ่งพยัญชนะนี้อาจทำหน้าที่เป็นพยัญชนะต้นของพยางค์ที่สองหรือพยัญชนะท้ายของพยางค์แรกก็ได้ ดังโครงสร้าง A และ B

A: $CVC\#V$

B: $CV\#CV$

2. ปัญหาจากเสียงพยัญชนะเรียง

เมื่อมีพยัญชนะ 2 เสียงปรากฏเรียงกันอยู่ระหว่างสระ 2 เสียง และพยัญชนะ C_1 ได้แก่ [p, t, k] และพยัญชนะ C_2 ได้แก่ [r, l, w] ในกรณีนี้พยัญชนะ C_1 สามารถทำหน้าที่เป็นพยัญชนะท้ายของพยางค์แรก หรือทำหน้าที่เป็นพยัญชนะต้นควบกล้ำกับพยัญชนะ C_2 ของพยางค์หลังก็ได้ ดังโครงสร้าง C และ D

C: $CVC_1\#C_2V$

D: $CV\#C_1C_2V$

3. ปัญหาจากเสียงสระเรียง

เมื่อเสียงสระ [i, ii, u, uu, ɯ, ɯɯ] ปรากฏเรียงกับเสียงสระ [a] เสียงสระทั้ง 2 เสียงนี้อาจเป็นสระเดี่ยวทั้งคู่ใน 2 พยางค์ หรือเป็นสระผสมใน 1 พยางค์ ดังโครงสร้าง X และ Y

X : $CVVC\#$

Y : $CV\#VC\#$

จากปัญหาเหล่านี้สามารถสร้างคำพูดที่มีความกำกวมของจุดแปงพยางค์ในภาษาไทยได้ 18 คู่ 36 ประโยค (ดูรายละเอียดในบทที่ 3)

2.2 ทฤษฎีการวิเคราะห์สัญญาณเสียงพูดเบื้องต้น

โดยธรรมชาติของสัญญาณเสียงพูดจะไม่เสถียร และเปลี่ยนแปลงตามเวลา (Non-stationary) ดังนั้น เมื่อต้องการนำสัญญาณเสียงพูดมาประมวลผลสัญญาณดิจิทัล (Digital Signal Processing) จึงจำเป็นต้องแบ่งสัญญาณเสียงพูดออกเป็นช่วงเวลาสั้น ๆ (Short Time) เพื่อให้สัญญาณเสียงมีความเสถียรและไม่เปลี่ยนแปลงตามเวลา (Stationary) จากนั้นจึงจะสามารถนำสัญญาณเสียงไปประมวลผลต่อไปได้ กรอบเสียงพูด (Speech Frame) ความยาวประมาณ 20-30 มิลลิวินาที ทำให้สัญญาณเสียงพูดในแต่ละกรอบเสียงพูดเป็นสัญญาณที่มีความเสถียรและไม่เปลี่ยนแปลงตามเวลา การเหลื่อมกรอบเสียงพูด (Frame Overlap) จะทำให้รอยต่อของลักษณะสำคัญของเสียงพูด จากกรอบเสียงพูดหนึ่งไปยังอีกกรอบเสียงพูดหนึ่งเรียบ (Smooth) ขึ้น กรรมวิธีการวิเคราะห์สัญญาณเสียงพูดเบื้องต้น ประกอบด้วย

2.2.1กรรมวิธีการปรับบรรทัดฐานแอมปลิจูด (Amplitude Normalization)

สัญญาณเสียงพูดเป็นสัญญาณต่อเนื่อง (Continuous Signal) จะต้องทำการสุ่มตัวอย่าง (Sampling) สัญญาณเสียงพูด โดยอัตราการสุ่มตัวอย่าง (Sampling Rate) เท่ากับ อย่างน้อย 2 เท่าของความถี่ของสัญญาณเสียงพูด (Carlson · A.B., 1975 อ้างถึงใน ไพศาล ธรรมโพธิทอง, 2533) ซึ่งจะมีตัวอย่างของสัญญาณเสียงพูด (Speech Sample) อย่างน้อย 8,000–10,000 ตัวอย่างต่อวินาที ตัวอย่างสัญญาณเสียงจะถูกนำมาปรับระดับสัญญาณ (Quantization) ให้เป็นสัญญาณดิจิทัล (Digital Signal) กรรมวิธีการปรับบรรทัดฐานแอมปลิจูดของสัญญาณเสียงพูดเป็นการเพิ่มหรือลดขนาดของสัญญาณเสียงพูด เพื่อให้ขนาดของสัญญาณเสียงพูดมีความเหมาะสม เนื่องจากสัญญาณเสียงพูดของแต่ละบุคคลมีขนาดไม่เท่ากัน จึงจำเป็นต้องปรับให้ขนาดของสัญญาณเสียงพูดอยู่ในบรรทัดฐานเดียวกัน เพื่อง่ายต่อการวัดค่าคุณลักษณะและเปรียบเทียบสัญญาณเสียงกับค่ากำหนด (Threshold) การปรับบรรทัดฐานแอมปลิจูดแสดงดังสมการที่ (2.1)

$$\widetilde{S}[i] = \frac{S[i]}{2^{N-1}} \quad \text{.....(2.1)}$$

เมื่อ $\widetilde{S}[i]$ คือสัญญาณเสียงพูดที่ปรับบรรทัดฐานแอมปลิจูดแล้ว

$S[i]$ คือสัญญาณเสียงพูดดิจิทัล

N คือจำนวนบิต (Bit) ที่ใช้แทนค่าของสัญญาณเสียงพูด

2.2.2 กรรมวิธีการเน้นล่วงหน้า (Preemphasis)

กรรมวิธีการเน้นล่วงหน้าจะทำให้ความลาดเอียงในเชิงความถี่แบนราบลง (วิศรุต อาชุบุตร, 2539) และทำให้อัตราส่วนของสัญญาณเสียงพูดต่อสัญญาณรบกวน (Signal to Noise Ratio) มีค่าสูงขึ้น โดยนำสัญญาณเสียงพูด มาผ่านตัวกรองดิจิทัลลำดับหนึ่ง (First Order Digital Filter) ดังแสดงในสมการที่ (2.2)

$$S'[i] = S[i] - aS[i-1] \quad \text{.....(2.2)}$$

เมื่อ $S'[i]$ คือค่าของสัญญาณเสียงพูดที่ผ่านกรรมวิธีการเน้นล่วงหน้าที่ i

$S[i]$ คือสัญญาณเสียงพูดที่ i

$S[i-1]$ คือสัญญาณเสียงพูดที่ $i-1$ และ

a คือสัมประสิทธิ์ของตัวกรอง (Preemphasis Factor)

โดยทั่วไปในระบบรู้จำเสียงพูดจะกำหนดให้สัมประสิทธิ์ของตัวกรองเท่ากับ 0.95 (Ganapathiraju, 1996)

2.2.3 กรรมวิธีการวางกรอบหน้าต่าง (Windowing)

กรรมวิธีการวางรูปแบบกรอบหน้าต่าง เป็นการเตรียมข้อมูลในแต่ละกรอบเสียงพูด เพื่อนำไปวิเคราะห์ในขั้นตอนการวัดค่าคุณลักษณะของสัญญาณเสียง โดยสัญญาณเสียงพูดจะถูกคูณด้วยฟังก์ชันกรอบหน้าต่าง (Window Function) ซึ่งฟังก์ชันกรอบนั้นมีหลายประเภท เช่น Rectangular Window, Hamming Window, Hanning Window, Blackman Window, Kaiser เป็นต้น (Alan V. Oppenheim, 1989) รูปที่ 2.1 แสดงรูปแบบหน้าต่างประเภทต่าง ๆ ในงานวิจัยนี้เลือกใช้ฟังก์ชันกรอบแบบ Hamming Window (วิศรุต อาชุบุตร, 2539) ดังแสดงในสมการที่ (2.3) เนื่องจาก Hamming Window จะทำให้ขนาดของสัญญาณลดลงอย่างช้า ๆ ที่บริเวณปลายของกรอบหน้าต่าง จึงเป็นการช่วยป้องกันการเปลี่ยนแปลงอย่างกระทันหันที่บริเวณปลายของกรอบเสียงพูด สมการการวางกรอบหน้าต่างทับสัญญาณเสียงพูด แสดงในสมการที่ (2.4)

$$W_n = 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right] \quad \text{.....(2.3)}$$

$$\hat{S}_n[i] = S_n[i] * W_n \quad \text{.....(2.4)}$$

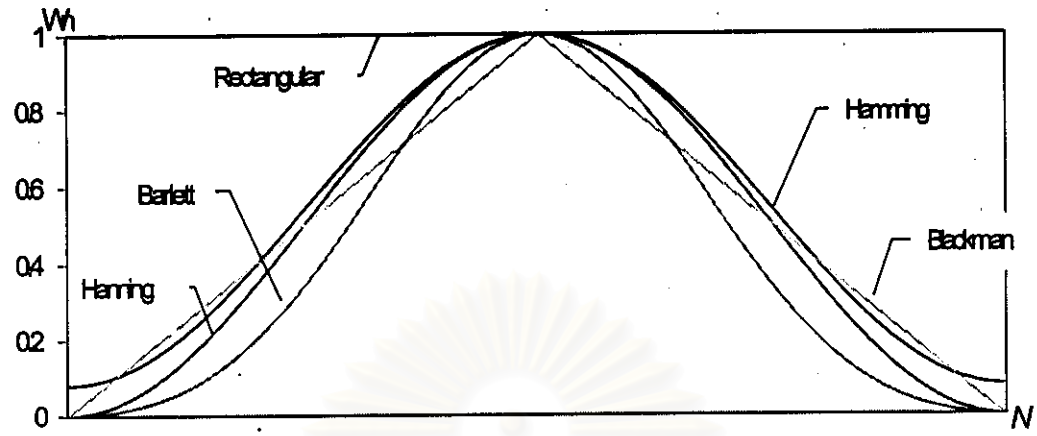
โดย $n = 0,1,\ldots,N-1$ และ $i = 0,1,\ldots,K-1$

เมื่อ $\hat{S}_n[i]$ คือสัญญาณเสียงพูด เมื่อวางรูปแบบหน้าต่างแล้ว

$S_n[i]$ คือสัญญาณเสียงพูดที่ i ในกรอบเสียงพูด n

W_n คือฟังชันหน้าต่างแบบ Hamming Window ของกรอบเสียงพูด n

N คือจำนวนตัวอย่างเสียงพูดภายในกรอบเสียงพูด



รูปที่ 2.2 รูปแบบของฟังก์ชันกรอบแบบต่าง ๆ

2.3 ทฤษฎีการวิเคราะห์คุณสมบัติของสัญญาณเสียงพูด

คุณลักษณะของสัญญาณเสียงพูดที่สามารถนำมาหาค่าและวิเคราะห์หาขอบเขตพยางค์ได้มีดังนี้คือ

2.3.1 พลังงานของสัญญาณเสียงพูด

พลังงานของสัญญาณเสียงพูด ถูกนำมาใช้ในการวิเคราะห์เสียงพูดตั้งแต่ ปี ค.ศ. 1970 และเป็นคุณลักษณะที่นิยมนำมาใช้กันอย่างแพร่หลาย เนื่องจากเป็นวิธีที่คำนวณง่าย และรวดเร็ว พลังงานของสัญญาณเสียงพูดเป็นคุณสมบัติที่แสดงให้เห็นว่ามีสัญญาณเสียง (รวมทั้งสัญญาณรบกวน) เกิดขึ้น ณ เวลานั้นหรือไม่ การคำนวณหาค่าพลังงาน จะทำทีละกรอบเสียงพูด โดย E_n คือ ค่าพลังงานของกรอบเสียงพูดที่ n $S_n[i]$ คือสัญญาณเสียงพูดที่ i ในกรอบเสียงพูด n และในแต่ละกรอบเสียงพูดจะมีสัญญาณเสียงพูดจำนวน K ซึ่งวิธีการคำนวณหาค่าพลังงานของสัญญาณเสียงพูด มีดังนี้ คือ

1) พลังงานสัมบูรณ์ (Absolute Energy)

เป็นการหาผลรวมของสัญญาณเสียงพูดสัมบูรณ์ ในแต่ละกรอบเสียงพูด ดังสมการที่ (2.5)

$$E_n = \sum_{i=1}^{K} |S_n[i]| \quad \text{.....(2.5)}$$

2) พลังงานเฉลี่ย (Root Mean Square Energy)

เป็นการพลังงานของสัญญาณเสียงพูดจากรากที่สองของผลรวมกำลังสองเฉลี่ย ดังสมการที่ (2.6)

$$E_n = \left[\frac{1}{K} \sum_{i=1}^{K} S_n^2[i] \right]^{1/2} \quad \text{.....(2.6)}$$

3) พลังงานความถี่และเวลา (Frequency-Time Energy)

พลังงานความถี่และเวลา คือ พลังงานแถบความถี่ (Band Frequency Energy)รวมกับพลังงานในเชิงเวลา พลังงานแถบความถี่เฉลี่ย คือพลังงานจากการแปลงสัญญาณเสียงพูดในเชิงเวลา (Time Domain) ให้อยู่ในเชิงความถี่ (Frequency Domain) โดยใช้การแปลงฟูเรียร์แบบเร็ว (Fast Fourier Transform) แล้วนำขนาดของความถี่ในช่วงความถี่ 250 –3500 Hz มาหาค่าพลังงาน ส่วนพลังงานในเชิงเวลา คือ ค่า log ของพลังงานเฉลี่ย

4) พลังงานกำลังสอง (Square Energy)

เป็นการวัดค่าพลังงานจากสัญญาณเสียงพูดยกกำลังสอง ทำให้ค่าพลังงานมีความไวต่อสัญญาณที่มีขนาดใหญ่ การหาค่าพลังงานกำลังสอง แสดงดังสมการที่ (2.7)

$$E_n = \sum_{i=1}^{K} S_n^2[i] \quad \text{.....(2.7)}$$

5) พลังงานของ Teager (Teager Energy)

เมื่อพูดถึงพลังงานของสัญญาณ มักจะหมายถึงค่าผลรวมเฉลี่ยกำลังสองของขนาดสัญญาณ เมื่ออยู่ในเชิงเวลา หรือ ทำการแปลงฟูเรียร์แบบไม่ต่อเนื่อง (Discrete Fourier Transform, DFT) แล้วยกกำลังขนาดของความถี่ของสัญญาณเสียงพูด ซึ่งเป็นการหาค่าพลังงานในเชิงความถี่ ตัวอย่างของสัญญาณความถี่ 10 Hz และสัญญาณความถี่ 1000 Hz ไม่ว่าจะหาค่าพลังงานในเชิงเวลา หรือในเชิงความถี่ พลังงานของสัญญาณ 1000 Hz จะมีค่ามากกว่าสัญญาณความถี่ 10 Hz เสมอ พลังงานของสัญญาณรายคาบ (Sinusoidal) เกิดจากการคูณกันของขนาดกำลังสองกับความถี่กำลังสอง สัญญาณเสียงมีลักษณะเป็นรายคาบในช่วงเวลาสั้น ๆ ซึ่งได้กำหนดขึ้นเป็นกรอบเสียงพูด โดยมีสมการการเคลื่อนที่ของสัญญาณเสียงดัง สมการที่ (2.8)

$$S_i = A\cos(\Omega i + \phi) \quad \text{.....(2.8)}$$

โดยที่ S_i คือตัวอย่างของสัญญาณเสียงพูด

A คือ ขนาดของสัญญาณเสียงพูด

Ω คือ ความถี่ดิจิทัล $\Omega = 2\pi f / f_s$ ซึ่ง f คือความถี่แอนาลอก (Analog Frequency) และ f_s คืออัตราการสุ่มตัวอย่าง (Sampling Rate)

ϕ คือ เฟสเริ่มต้น

ดังที่ได้กล่าวแล้วว่า พลังงานของสัญญาณรายคาบ จะเท่ากับขนาดของสัญญาณกำลังสองคูณกับความถี่กำลังสอง ในสมการที่ (2.9) แสดงให้เห็นว่าพลังงานของสัญญาณสามารถหาได้จากสัญญาณของตัวเอง ลบด้วยผลคูณของสัญญาณข้างหน้ากับสัญญาณข้างหลัง

$$S_i^2 - S_{i+1}S_{i-1} = A^2\Omega^2 \quad \text{.....(2.9)}$$

ค่าพลังงานที่ได้เป็นเพียงค่าพลังงานของสัญญาณที่เวลาใดๆ เราเรียกว่าพลังงานชั่วขณะ (Instantaneous Energy, E_i) และพลังงานของแต่ละกรอบเสียงพูดสามารถหาได้จากสมการที่ (2.10)

$$E_n = \sum_{i=1}^{K} E_i \quad \text{.....(2.10)}$$

6) พลังงานจากการแปลงแบบ Walsh (Walsh Transform Energy)

รูปแบบของฟังก์ชันการแปลงเชิงเส้น (Linear Transformation) สามารถเขียนเป็นสมการได้ดังสมการที่ (2.11) และ (2.12)

$$F(n) = \sum_{k=0}^{N-1} f(k)a(k,n) \quad \text{.....(2.11)}$$

$$f(k) = \sum_{n=0}^{N-1} F(n)b(k,n) \quad \text{.....(2.12)}$$

ซึ่ง $a(k,n)$ คือฟังก์ชันการแปลงไปข้างหน้า (Forward Transformation Kernel) และ $b(k,n)$ คือฟังก์ชันการแปลงกลับ (Inverse Transformation Kernel) ในกรณีของการแปลงฟูเรียร์แบบไม่ต่อเนื่อง ฟังก์ชันของ $a(k,n)$ และ $b(k,n)$ คือ $\exp[-j2\pi kn/N]$ และ $\exp[j2\pi kn/N]/N$ ตามลำดับ การหาค่าพลังงานในเชิงความถี่ โดยใช้การแปลงฟูเรียร์แบบไม่ต่อเนื่อง หรือแม้แต่ใช้การแปลงฟูเรียร์อย่างเร็วแล้วก็ตาม การคำนวณก็ยังใช้เวลานาน ดังนั้นจึงนำสัญญาณเสียงมาทำการแปลง Walsh (Walsh Transform) ซึ่งเป็นการบวกหรือลบกันของค่าจริง (Real Value) เท่านั้น แล้วนำขนาดของสเปคตราแบบ Walsh (Walsh Spectra) มาหาค่าพลังงานของสัญญาณเสียงพูด ฟังก์ชันการแปลง Walsh แสดงในสมการที่ (2.13)

$$F(n) = \sum_{k=0}^{N-1} f(k)(-1)^{\sum_{r=0}^{r-1} p_r(k)q_r(n)} \quad \text{.....(2.13)}$$

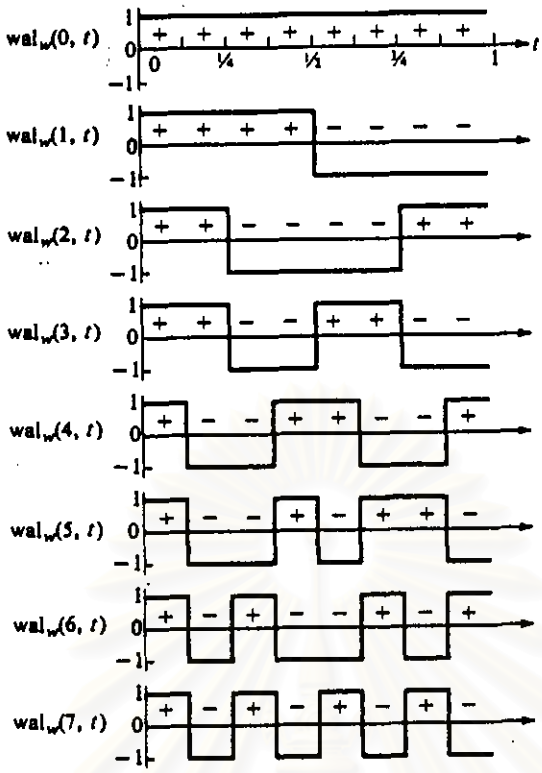
ดังนั้น

$$a(k,n) = (-1)\sum_{r=0}^{r-1} p_r(k)q_r(n) \quad \text{.....(2.14)}$$

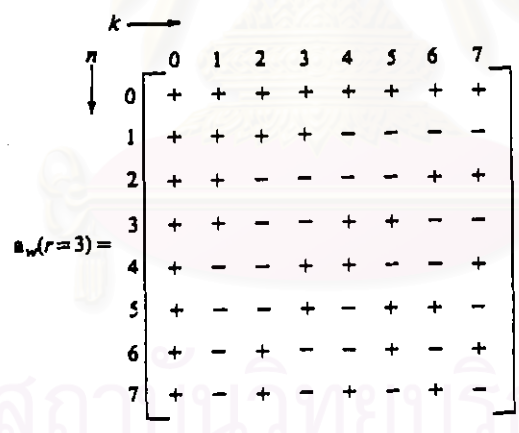
ซึ่ง

$$\begin{aligned} q_0(n) &= p_{r-1}(n) \\ q_1(n) &= p_{r-1}(n) + p_{r-2}(n) \\ q_2(n) &= p_{r-2}(n) + p_{r-3}(n) \quad \text{.....(2.15)} \\ &\vdots \\ q_{r-1}(n) &= p_1(n) + p_0(n) \end{aligned}$$

รูปที่ 2.3 แสดงฟังก์ชันแบบ Walsh อันดับ 8 (N =8, $N = 2^r$) และแทนลงในเมตริกขนาด 8 x 8 ได้ดังรูปที่ 2.4



รูปที่ 2.3 ฟังก์ชัน Walsh ลำดับ 8



รูปที่ 2.4 ค่าลำดับการแปลง Walsh $r = 3$ (เครื่องหมาย + และ – แทน +1 และ –1)

2.3.2 อัตราการตัดผ่านศูนย์ (Zero Crossing Rate)

อัตราการตัดผ่านศูนย์ เป็นการวัดจำนวนครั้งการตัดผ่านแกนเวลาที่ระดับศูนย์ ของสัญญาณเสียงพูด (จำนวนครั้งของการเปลี่ยนเครื่องหมายสัญลักษณ์ทางคณิตศาสตร์) อัตราการตัดผ่านศูนย์ เป็นเครื่องมืออย่างง่าย ที่สามารถนำมาใช้อธิบายได้ว่าสัญญาณเสียงพูดนั้นเป็นเสียงก้อง (Voiced) หรือเสียงไม่ก้องเสียดแทรก (Unvoiced Fricative) สัญญาณเสียงพูดที่มีอัตราการตัดผ่านศูนย์ต่ำ จะเป็นเสียงก้องส่วนสัญญาณเสียงพูดที่มีอัตราการตัดผ่านศูนย์สูงจะเป็นเสียงไม่ก้องเสียดแทรก อัตราการตัดผ่านศูนย์สามารถหาได้จากสมการที่ (2.16)

$$Z_n = \frac{1}{K}\sum_{i=1}^{K}\frac{|\text{sgn}\{S_n[i]\} - \text{sgn}\{S_n[i-1]\}|}{2} \quad \text{.....(2.16)}$$

เมื่อ $\text{sgn}\{S[i]\} = \begin{cases} +1, & S[i] \geq 0 \\ -1, & S[i] < 0 \end{cases}$

โดยที่ Z_n คืออัตราการตัดผ่านศูนย์ ที่กรอบเสียงพูด n

$S_n[i]$ คือสัญญาณเสียงพูดที่ i ในกรอบเสียงพูด n

K คือความกว้างของกรอบเสียงพูด (จำนวนตัวอย่างเสียงพูดในแต่ละกรอบเสียงพูด)

ได้มีการปรับปรุงอัตราการตัดผ่านศูนย์เป็น<u>อัตราการตัดผ่านระดับกำหนด</u> (Band Crossing Rate) ดังแสดงในสมการที่ (2.17)

$$B_n = \sum_{i=1}^{K}|\text{sgn}\{S_n[i]\} - \text{sgn}\{S_n[i-1]\}| \quad \text{.....(2.17)}$$

เมื่อ $\text{sgn}\{S[i]\} = \begin{cases} +1 & if \quad S[i] \geq L \\ \text{sgn}\{S[i-1]\} & if \quad -L \leq S[i] < L \\ -1 & if \quad S[i] < -L \end{cases}$

ซึ่ง L คือความกว้างของระดับที่พิจารณา

2.3.3 ความถี่มูลฐาน (Fundamental Frequency)

เสียงพูด (Voicing) เกิดจากลมหายใจที่ถูกดัดแปลงไป (Modified Breathing) โดยอาศัยการทำงานร่วมกันของอวัยวะในการออกเสียงต่าง ๆ (Articulators) เมื่อกระแสอากาศจากแหล่งพลังงานต่าง ๆ ซึ่งส่วนใหญ่หมายถึงปอด เคลื่อนที่มาสู่กล่องเสียงและถูกดัดแปลงให้เป็นเสียงแบบต่างๆ ตามรูปแบบการทำงานของเส้นเสียง รูปแบบของการสั่นของเส้นเสียงทำให้เกิดเป็นเสียงแบบต่าง ๆ การสั่นของเส้นเสียงแบบธรรมดา (Normal Vibration) ทำให้เกิดเสียงก้อง (Voiced Sound) ส่วนเสียงไม่ก้อง (Voiceless Sound) จะไม่มีการสั่นสะเทือนของเส้นเสียง ต่อมาอากาศก็จะเดินทางเข้าสู่ช่องปาก ซึ่งประกอบด้วยอวัยวะแปรเสียงหรือฐานกรณ์ (Articulators) มากมาย เพื่อทำหน้าที่ในการกล่อมเกลาเสียงให้ออกมามีคุณลักษณะแตกต่างกัน สระซึ่งเป็นแกนกลางพยางค์มีลักษณะเป็นเสียงก้อง ซึ่งเสียงก้องจะมีความเป็นรายคาบของระดับเสียง (Pitch Period) และส่วนกลับของระยะเวลารายคาบของสัญญาณเสียงคือความถี่มูลฐาน (Fundamental Frequency) วิธีการตรวจหาระดับเสียงที่เป็นรายคาบมีดังนี้

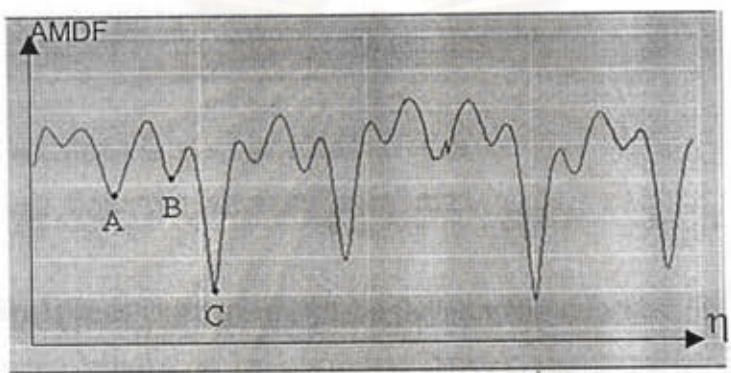
1. วิธี Cepstrum (CEP) (John R. Deller, et al., 1991)
2. วิธี Simplified Inverse Filtering Technique (SIFT) (John R. Deller, et al., 1991)
3. วิธี Modified Autocorrelation Using Clipping (AUTOC) (John R. Deller, et al., 1991)

4. วิธี Average Magnitude Difference Function (AMDF) (John R. Deller, et al., 1991)

ในงานวิจัยนี้เลือกใช้วิธี Average Magnitude Difference Function (AMDF) เพราะเป็นวิธีที่มีการคำนวณไม่ซับซ้อนมาก ดังสมการที่ (2.18)

$$AMDF_n(\eta) = \frac{1}{K}\sum_{i=1}^{K}|S[i] - S[i-\eta]| \quad \text{.....(2.18)}$$

ซึ่ง η คือค่าความแตกต่างของเวลา $\eta = 0,1,\ldots,\eta_{max}$ รูปที่ 2.5 แสดงระดับเสียงที่เป็นรายคาบ เมื่อใช้วิธี AMDF ตรวจหา แกน X คือ η และแกน Y คือค่าของ AMDF



รูปที่ 2.5 ลักษณะของสัญญาณเสียงที่เป็นรายคาบ

การคำนวณหาค่า AMDF จะกระทำในช่วงประมาณ 2.5 – 20 มิลลิวินาที (John R. Deller, et al., 1991) เพื่อแก้ปัญหาที่อาจเกิดขึ้นเนื่องจาก η เท่ากับศูนย์ ที่ตำแหน่งของ η ที่มีค่า AMDF ต่ำที่สุดเป็นจุดแรก คือค่าระยะเวลาที่สัญญาณเสียงครบคาบพอดี ดังนั้นก็จะสามารถคำนวณหาความถี่มูลฐานได้จากสมการที่ (2.19)

$$F_0(n) = f_s/\eta \quad \text{.....(2.19)}$$

ซึ่ง $F_0(n)$ คือความถี่มูลฐานของกรอบเสียงพูด n

f_s คืออัตราการสุ่มตัวอย่าง

2.4 ทฤษฎีการปรับปรุงคุณลักษณะของสัญญาณเสียงพูดและการปรับปรุงขอบเขตพยางค์

2.4.1 วิธีการปรับเรียบ (Smoothing)

เนื่องจากการหาขอบเขตพยางค์อาศัยแผนภูมิเส้นระดับพลังงาน อัตราการตัดผ่านระดับที่กำหนด ความถี่มูลฐานของสัญญาณเสียงพูด จึงจำเป็นต้องปรับให้แผนภูมิเส้นคุณลักษณะของสัญญาณเสียงเรียบขึ้น เพื่อปรับปรุงคุณลักษณะของสัญญาณเสียงให้เหมาะสมก่อนนำไปวิเคราะห์หาขอบเขตพยางค์ต่อไป กรรมวิธีการปรับเรียบที่น่าสนใจมี 2 วิธี คือ

1. วิธีการปรับเรียบโดยค่ากลาง (Median Smoothing)

ณ จุดข้อมูลที่เราสนใจ ทำการกำหนดขนาดหน้าต่างขึ้น โดยจุดข้อมูลที่เราสนใจจะเป็นจุดกึ่งกลางของหน้าต่าง ทำการหาค่ากลาง (Middle Value) ของข้อมูลภายในกรอบหน้าต่างที่กำหนดขึ้น จากนั้นแทนค่ากลางที่หาได้ลงที่จุดกึ่งกลางหน้าต่าง กรรมวิธีนี้ทำให้ข้อมูลไม่เกิดการกระโดด (Excite Value)

2. วิธีการปรับเรียบโดยค่าเฉลี่ยเคลื่อนไหว (Moving Average Smoothing)

ณ จุดข้อมูลที่เราสนใจ ทำการกำหนดขนาดหน้าต่างขึ้น โดยจุดข้อมูลที่เราสนใจจะเป็นจุดกึ่งกลางของหน้าต่าง ทำการหาค่าเฉลี่ยของข้อมูลภายในกรอบหน้าต่าง แล้วแทนค่าลงในจุดกึ่งกลางหน้าต่าง ดังแสดงในสมการที่ (2.20)

$$E'_n = \frac{1}{m1+m2+1}\sum_{n=n-m1}^{n+m2} E_n \qquad \text{.....(2.20)}$$

โดยที่ $m1$ คือความกว้างของหน้าต่างครึ่งซ้าย

$m2$ คือความกว้างของหน้าต่างครึ่งขวา

2.4.2 วิธีการปรับปรุงขอบเขตพยางค์ (Adjust Syllable Boundary)

จุดต้นและจุดปลายพยางค์ที่หาได้จากกรรมวิธีการหาขอบเขตพยางค์วิธีต่างๆ อาจยังมีตำแหน่งไม่ถูกต้อง จึงต้องนำมาผ่านขั้นตอนการปรับตำแหน่งอีกครั้งหนึ่ง โดยอาศัยอัตราส่วนของพลังงานเป็นพื้นฐาน ตำแหน่งของจุดปลายพยางค์ที่ได้รับการปรับปรุง จะเป็นตำแหน่งที่มีอัตราส่วนพลังงานด้านขวาต่อด้านซ้ายสูงที่สุดในช่วงที่กำหนด คือจากบริเวณ R1 ถึง R2 ดังสมการที่ 2.32 ส่วนตำแหน่งของจุดต้นพยางค์ที่ได้รับการปรับปรุง จะเป็นตำแหน่งที่มีอัตราส่วนพลังงานด้านซ้ายต่อด้านขวาสูงที่สุดในช่วงที่กำหนด คือจากบริเวณ R2 ถึง R1 ดังสมการที่ 2.33

$$EP1 = \underset{n}{\arg\max}\left(\frac{\sum_{i=n+1}^{n+K} S[i]^2}{\sum_{i=n-K+1}^{n} S[i]^2}\right), \quad n \in [R1, R2] \qquad \text{.....(2.32)}$$

$$SP2 = \underset{n}{\arg\max}\left(\frac{\sum_{i=n-K+1}^{n} S[i]^2}{\sum_{i=n+1}^{n+K} S[i]^2}\right), \quad n \in [R2, R1] \qquad \text{.....(2.33)}$$

โดยที่

$EP1$ คือ จุดปลายของพยางค์ที่หนึ่ง เป็นจุดที่มีอัตราส่วนพลังงานมากที่สุดภายในบริเวณ $R1$ ถึง $R2$

$SP2$ คือ จุดต้นของพยางค์ที่สอง เป็นจุดที่มีอัตราส่วนพลังงานมากที่สุดภายในบริเวณ $R2$ ถึง $R1$

$R1$ คือ จุดที่เลื่อนไปทางซ้ายของจุดที่ต้องการปรับเป็นระยะครึ่งหนึ่งของระยะเวลาพยางค์ (ระยะเวลาพยางค์ ,Syllable Duration คือ การกำหนดจำนวนกรอบเสียงพูดใน 1 พยางค์)

$R2$ คือ จุดที่เลื่อนไปทางขวาของจุดที่ต้องการปรับเป็นระยะครึ่งหนึ่งของระยะเวลาพยางค์

K คือ จำนวนตัวอย่างเสียงพูดในแต่ละกรอบเสียงพูด