บทที่ 2

การวิเคราะห์เสียงพูด

ในการศึกษาเรื่องระบบการรู้จำเสียงนั้น สิ่งสำคัญที่เป็นพื้นฐานก็คือเสียงพูด ซึ่งการศึกษาทางด้านภาษาที่ผ่านมาได้มีการศึกษาเสียงพูดของมนุษย์อย่างมีระบบ และมีหลักเกณฑ์ที่แน่นอน โดยเรียกการศึกษาในลักษณะนี้ว่า สัทศาสตร์ (Phonetics) โดยวิชาสัทศาสตร์นี้จะไม่จำกัดอยู่เพียงภาษาใดภาษาหนึ่งเท่านั้น แต่จะใช้ได้ทุกภาษา ซึ่งได้มีการแบ่งวิชาสัทศาสตร์ออกเป็น 3 สาขาวิชาดังนี้ (พิณทิพย์ ทวยเจริญ, 2533)

ก. สรีรสัทศาสตร์ (Articulatory Phonetics) จำเป็นการศึกษาเกี่ยวกับกลไกของกระแสลมที่ใช้ในการเปล่งเสียงพูด อวัยวะที่ใช้ในการออกเสียง กระบวนการออกเสียงต่าง ๆ

ข. กลสัทศาสตร์ (Acoustic Phonetics) เป็นการศึกษาถึงลักษณะทางกายภาพของคำพูดที่เปล่งออกมาว่ามีลักษณะต่าง ๆ เป็นอย่างไร

ค. โสตสัทศาสตร์ (Auditory Phonetics) เป็นการศึกษาเกี่ยวกับสรีระของหูที่มีต่อการรับรู้คลื่นเสียง

โดยที่กระบวนการสำคัญที่ใช้ในการออกเสียงของมนุษย์นั้นจะเกิดจากอวัยวะต่างๆ ที่เกี่ยวกับการออกเสียงที่มีอยู่ เช่น ปอด กล่องเสียง ลิ้น ฟัน ริมฝีปาก เป็นต้น

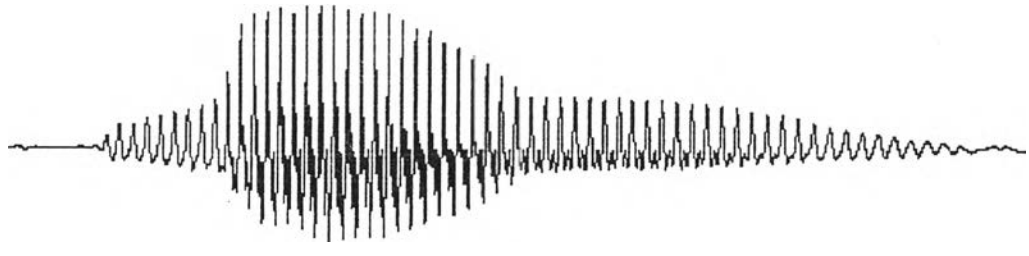
2.1 <u>เสียงที่ใช้ในการวิเคราะห์</u>

เนื่องจากเสียงที่มนุษย์ใช้ในการพูดโดยติดต่อผ่านสื่อในปัจจุบัน เช่น สายโทรศัพท์ เป็นต้น จะมีความถี่อยู่ในช่วง 30 ถึง 3400 เฮริตซ์ ในการวิเคราะห์เสียงในแบบสัญญาณอนาลอกทำได้ยากจึงจะทำการเปลี่ยนสัญญาณเสียงพูด ไปเป็นสัญญาณดิจิตอลแทน โดยในการวิจัยนี้จะใช้การ์ดเสียง (Sound Card) เข้าช่วย โดยที่สัญญาณเสียงพูดที่เป็นอนาลอกนี้จะถูกบันทึกผ่านไมโครโฟน จากนั้นการ์ดเสียงจะเป็นตัวจัดการกับสัญญาณเสียงพูดนี้ให้อยู่ในรูปแบบ (format) ที่เหมาะสม ซึ่งรูปแบบนี้จะถูกกำหนดโดยโปรแกรมการบันทึกเสียงที่ให้มาพร้อมกับการ์ดเสียงซึ่งเป็นข้อมูลเสียงแบบ 8 บิต การ์ดเสียงที่ใช้จะเป็นการ์ดซาว์ดบลาสเตอร์โปร (Sound

Blaster Pro) ซึ่งผลิตโดยบริษัท Creative Labs, inc. รูปแบบการจัดเก็บสัญญาณเสียงนั้นจะกล่าวอยู่ในภาคผนวก ค. จากการศึกษาเกี่ยวกับการรู้จำเสียงพูด ความถี่ที่ใช้ในการสุ่ม (Sampling rate) จะอยู่ในช่วง 6 ถึง 16 กิโลเฮริตซ์ (Furui, 1989) โดยในการวิจัยนี้จะใช้ความถี่ในการสุ่ม 8 กิโลเฮริตซ์ ซึ่งเป็นไปตามกฎของ Nyqist rate ที่ระบุไว้ว่าความถี่ในการสุ่มจะต้องมากกว่าความถี่สัญญาณอย่างน้อย 2 เท่าเพื่อป้องกันการเกิด aliasing ของสัญญาณ

2.2 การวิเคราะห์สัญญาณ

เนื่องจากเสียงพูดมีคุณสมบัติเปลี่ยนแปลงตามเวลา (time varying , non-stationary) ดังรูปที่ 2.2.2 ดังนั้นในการวิเคราะห์เสียงพูดจะแบ่งออกเป็นช่วง ๆ (Frame) โดยทั่วไปจะอยู่ในช่วง 10-30 มิลลิวินาที (ไพศาล ธรรมโพธิทอง,2533) ทั้งนี้ในช่วงเวลาดังกล่าว เสียงพูดจะมีการเปลี่ยนแปลงคุณสมบัติน้อยมาก ทำให้ทำการคำนวณได้ง่าย เพราะจะสมมติให้ในแต่ละเฟรมของเสียงจะไม่เปลี่ยนแปลงตามเวลา



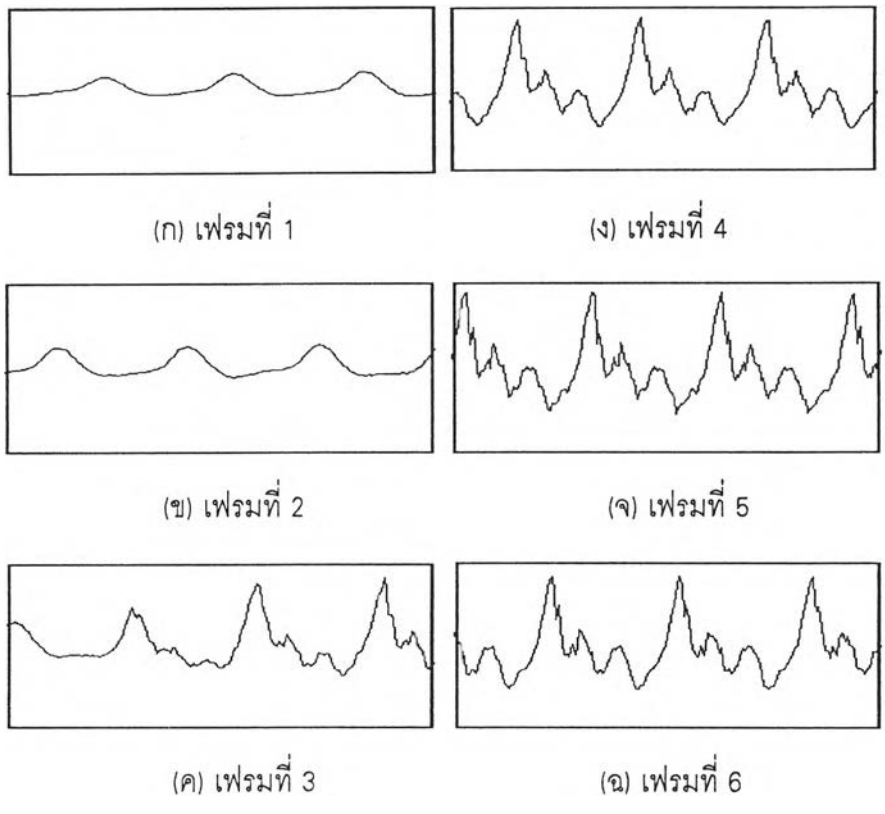
รูปที่ 2.2.1 ตัวอย่างสัญญาณเสียงคำว่า "หนึ่ง"

จากสัญญาณเสียงในรูปที่ 2.2.1 จะเห็นได้ว่าจะสามารถแบ่งลักษณะของเสียงได้ออกเป็น 3 ส่วน ดังนี้คือ

ก. ช่วงที่ยังไม่มีการเปล่งเสียงหรือสภาวะเงียบ (silence) เสียงในช่วงนี้จะค่อนข้างเรียบถ้าไม่มีสัญญาณรบกวนจากภายนอก

ข. ช่วงก่อนที่จะเปล่งเสียงออกมา หรือที่เรียกว่า เสียงอโฆษะ (unvoice speech) ในช่วงนี้แอมปลิจูดของเสียงจะต่ำและจะไม่มีความเป็นคาบ

ค. ช่วงที่เป็นคำพูด หรือที่เรียกว่าเสียงโฆษะ (voice speech) ในช่วงนี้เสียงพูดจะมีลักษณะเป็นคาบจะมีแอมปลิจูดสูง



รูปที่ 2.2.2 แสดงรูปคลื่นในแต่ละเฟรมของคำว่า "หนึ่ง" ขนาดของเฟรมเท่ากับ 25 มิลลิวินาที

ในการวิเคราะห์สัญญาณเสียงนี้จะแบ่งออกเป็น 2 ส่วน ดังนี้คือ

2.2.1 <u>การวิเคราะห์สัญญาณในเชิงเวลา</u> (Time Domain Analysis)

ในส่วนนี้จะเป็นการวิเคราะห์ในเชิงเวลาโดยใช้ autocorrelation เพื่อหาระดับของพลังงานของเสียงพูด ถ้ากำหนดให้อนุกรมของข้อมูลเสียงพูดเป็น

$$x[1], x[2], x[3], \ldots, x[n], \ldots, x[K] \qquad \ldots..2.2.1$$

โดยที่ K จะเป็นจำนวนของข้อมูลเสียงทั้งหมด เราสามารถเขียนสมการของ autocorrelation ได้เป็น

$$R_l(m) = \sum_{n=0}^{N1-1-|m|} x[m]x[n+m] \quad ,m = 0,1,\ldots\ p \qquad \ldots..2.2.2$$

โดยที่ l แทนลำดับของเฟรมข้อมูลเสียง, l = 0,1,2,...,L

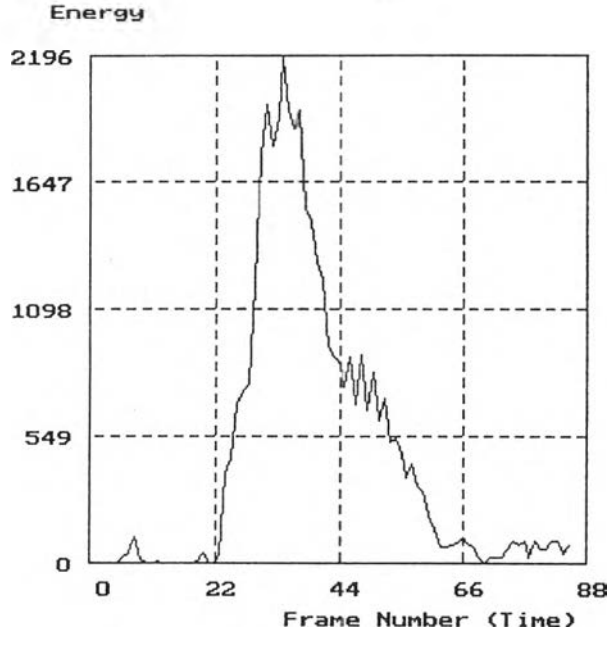
$N1$ แทนจำนวนข้อมูลเสียงในแต่ละเฟรม

p จะเป็น order ของ ระบบ ซึ่งในที่นี้จะกำหนดให้ p = 0 จะได้เป็น

$$R_l(m) = \sum_{n=0}^{N1-1} x^2\, n \qquad \text{.... 2.2.3}$$

เพื่อให้การคำนวนเร็วขึ้น จึงเปลี่ยนมาใช้ค่าสัมบูรณ์แทนดังสมการ (ชัยศรี เอี่ยมอำไพ, ม.ป.ป)

$$M_l(m) = \sum_{n=0}^{N1-1} |x[n]| \qquad \text{....2.2.4}$$



รูปที่ 2.2.1.1 แสดงพลังงานของสัญญาณเสียงของคำว่า " หนึ่ง" ตามสมการที่ 2.2.4 โดย N1 เท่ากับ 100

จากสมการ 2.2.4 ถ้าแบ่งข้อมูลเสียงออกเป็นเฟรม ได้ L เฟรม และที่ m เท่ากับ 0 จะสามารถเขียนได้เป็น

$$M_1(0), M_2(0), \ldots, M_L(0) \qquad \text{.... 2.2.5}$$

ซึ่งในรูปที่ 2.2.1.1 จะแสดงตัวอย่างของฟังก์ชันที่ได้จากสมการที่ 2.2.4,2.2.5 โดยที่ค่าของ N1 มีค่าเท่ากับ 100 โดยที่ค่าของข้อมูลไม่มีการเหลื่อม (overlap) กัน ซึ่งในส่วนนี้จะนำไปใช้ในส่วนของการหาส่วนเริ่มต้นและสิ้นสุดสัญญาณ (End Point Detection) และการหาขอบเขตของคำ (Word Segmentation) ต่อไป

2.2.2 <u>การวิเคราะห์สัญญาณในเชิงความถี่</u>

ในที่นี้จะใช้ดีสครีตฮาร์ตเลย์ทรานส์ฟอร์ม(Discrete Hartley Transform , DHT) มาใช้ในการวิเคราะห์หาพารามิเตอร์ (parameter) .เพื่อใช้ในการรู้จำเสียงพูดตัวเลขไทย ซึ่งจะขอกล่าวโดยสรุปดังนี้ (สุนิสา จันทวีกูล, 2536)

ฮาร์ตเลย์ทรานส์ฟอร์มเป็นอินทิกรัลทรานส์ฟอร์มที่เสนอโดย R.V.L Hartley โดยมีรูปแบบดังนี้คือ

$$H(w) = (2\pi)^{-1} \int_{-\infty}^{\infty} x(t)\cos(wt) + \sin(wt)\, dt \qquad \text{.... 2.2.6}$$

$$x(t) = (2\pi)^{-1} \int_{-\infty}^{\infty} H(w)\cos(wt) + \sin(wt)\, dw \qquad \text{.... 2.2.7}$$

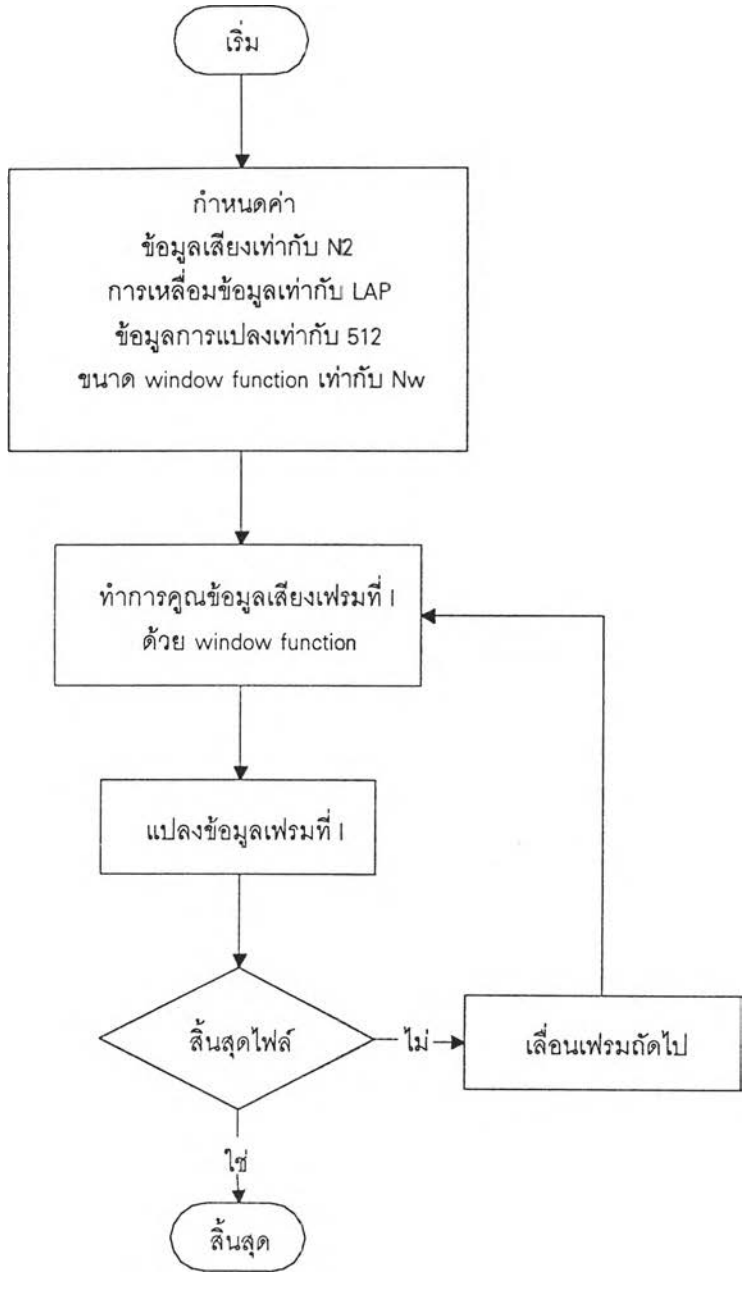
และรูปแบบฮาร์ตเลย์ทรานส์ฟอร์มในรูปของดีสครีตทรานส์ฟอร์มเป็น

$$H[k] = \sum_{n=0}^{N-1} x[n]cas(2\pi nk / N) \quad , \; 0 \le n \le N-1 \qquad \text{.... 2.2.8}$$

$$x[n] = N^{-1} \sum_{k=0}^{N-1} H[k]cas(2\pi nk / N) \quad , \; 0 \le n \le N-1 \qquad \text{.... 2.2.9}$$

$cas(\theta) \triangleq \cos(\theta) + \sin(\theta)$

โดย $H[k]$ ในสมการจะเป็นดีสครีตฮาร์ตเลย์ทรานส์ฟอร์ม ส่วน x[n] เป็นอินเวอร์สดีสครีตฮาร์ตเลย์ทรานส์ฟอร์ม N แทนจำนวนข้อมูลที่จะนำมาวิเคราะห์ k จะเป็นลำดับของดีสครีตฮาร์ตเลย์ทรานส์ฟอร์ม



รูปที่ 2.2.2.1 ขั้นตอนการแปลงข้อมูลเสียง

ต่อมาได้มีการเสนออัลกอริทึมฟาสต์ฮาร์ตเลย์ทรานส์ฟอร์มขึ้น (Fast Hartley Transform) ซึ่งเป็นส่วนสำคัญที่นำมาใช้ในงานวิจัยนี้ โดยจะมีรูปสมการเป็น

$$H[k] = H_{2n}[k] + H_{2n+1}[k]\cos(2\pi k / N) + H_{2n+1}[k]\sin(2\pi k / N) \quad \text{..... 2.2.10}$$

$$0 \le k \le N-1$$

H[k] จะมีจำนวนเท่ากับ N ส่วน $H_{2n}[k]$ และ $H_{2n+1}[k]$ เป็นค่าของ H[k] โดยที่ k เป็นเลขคู่และเลขคี่ตามลำดับ ซึ่งจะมีจำนวนเท่ากันคือ $N/2$

ในการนำดีสครีตฮาร์ตเลย์ทรานส์ฟอร์มมาใช้พิจารณาจากข้อมูลเสียง x[n] ดังในสมการที่ 2.2.1 โดยจะทำการแบ่งข้อมูลเสียงออกเป็นช่วง ๆ (Frames) โดยให้มีขนาดเท่ากับ N2 และจะมีการเหลื่อมกัน (overlap) ของข้อมูลขนาดเท่ากับ LAP ซึ่งมีค่าเป็น $N2/2$ จากนั้นจะทำการคูณข้อมูลในแต่ละช่วงของข้อมูลด้วย hamming window function ขนาด N_w จุด แล้วจึงทำ ดีสครีตฮาร์ตเลย์ทรานส์ฟอร์มขนาด 512 จุด ดังแสดงในรูปที่ 2.2.2.1 ซึ่งผลของการทรานส์ฟอร์มนี้จะใช้หาค่าของพารามิเตอร์ของระบบการรู้จำต่อไป

2.3 ฟังก์ชั่นหน้าต่าง

ในการวิเคราะห์เสียงก่อนที่จะทำการทรานส์ฟอร์มข้อมูล จะทำการคูณด้วยค่าของ Hamming window function กับข้อมูลเสียง x[n] แล้วจึงทำการทรานส์ฟอร์มข้อมูลดังในสมการที่ 2.3.1 ซึ่งในการคูณนี้จะมีผล 2 อย่างคือ (Furui, 1989)

ก. ป้องกันการเปลี่ยนแปลงอย่างฉับพลันของการลดทอนของแอมปิจูดที่ปลายทั้งสองข้างของข้อมูลที่ตัดออกมาคำนวณ

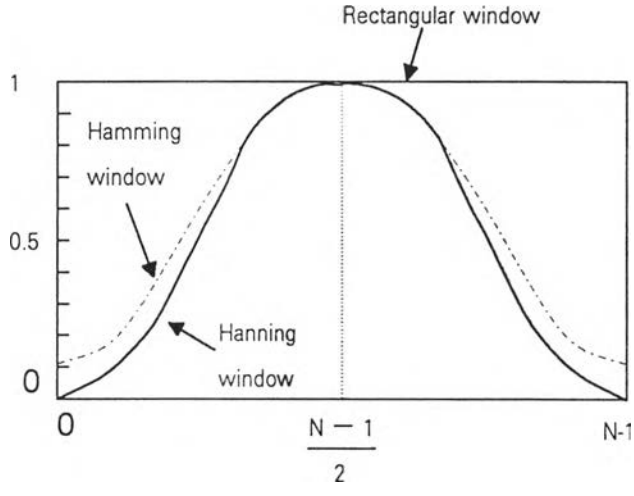
ข. เป็น weighted เฉลี่ย ในเชิงความถี่ เป็นผลให้มีการลด spectral distortion ที่เกิดจากการใส่ rectangular window โดยที่ความถี่ที่สูง robe ข้าง ๆ จะมีขนาดเล็ก นั้นคือมีการลดทอนมากที่ robe ด้านข้างทั้งสอง ดังในรูปที่ 2.3.1 (ข)

จากในรูปที่ 2.3.1 (ก) Hamming window จะมีสมการดังนี้

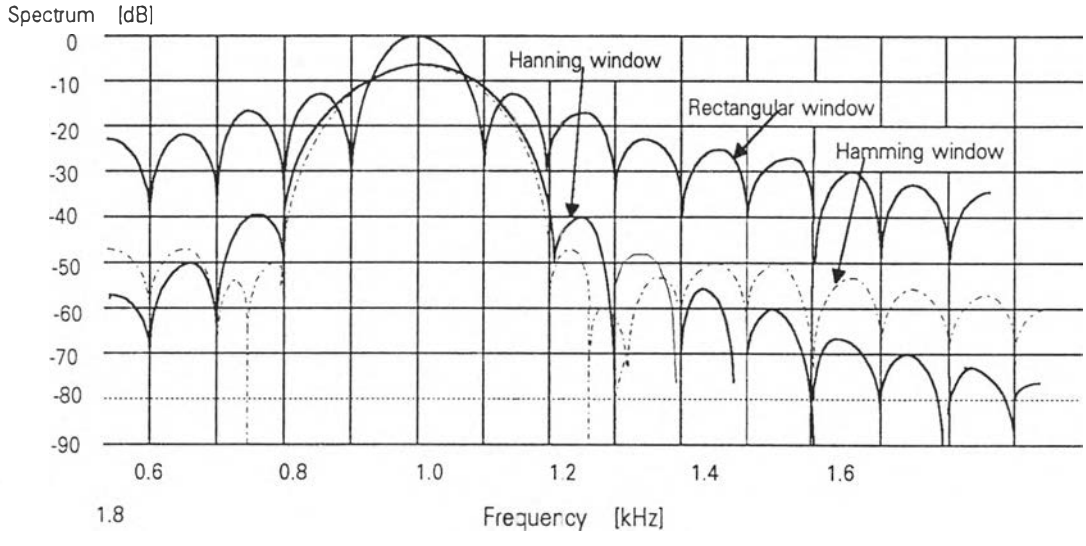
$$W_H(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{N_W - 1}\right) \qquad \text{....2.3.1}$$

ในขณะที่ Rectangular window มีสมการเป็น

$$W_R(n) = 1, (0 \le n \le N_W - 1) \qquad \text{....2.3.2}$$



(ก)



(ข)

รูปที่ 2.3.1 (ก) แสดงรูปของ window function แบบต่าง ๆ
(ข) แสดงถึง spectrum ของสัญญาณ ที่ใช้ window function แบบต่าง ๆ ใน (ก)
(Furui,1989)

2.4 การตัดคำ (Word Segmentation)

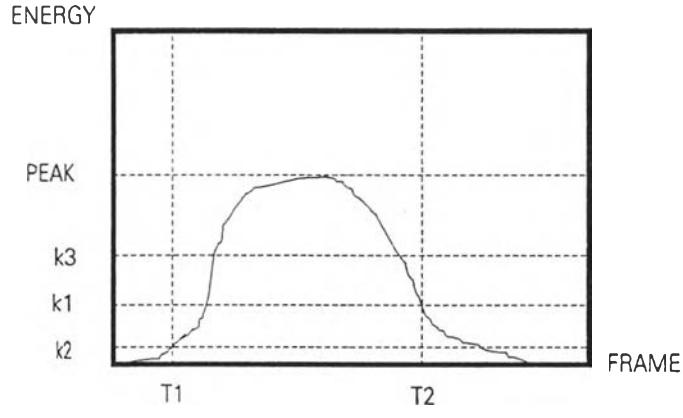
ในระบบการวิเคราะห์เสียงแบบ isolated word recognition สิ่งสำคัญคือการตัดคำเพื่อที่จะหาขอบเขตของคำ ในกรณีที่ input เป็นกลุ่มของคำที่ต่อเนื่องกัน เราจำเป็นต้องแยกคำแต่ละคำออกมา ซึ่งการแยกคำที่ถูกต้องจะให้ความเชื่อถือสูงและการคำนวณข้อมูลเสียงใช้เวลาน้อย (Lamel et al.,1981) อย่างไรก็ดี ในงานวิทยานิพนธ์นี้ จะเป็นการออกเสียงเป็นแต่ละพยางค์โดด ๆ

ในการวิเคราะห์เพื่อหาขอบเขตของคำนี้ เราจะใช้พารามิเตอร์ในสมการที่ 2.2.4 โดยจะกำหนดระดับพลังงานอ้างอิง (Energy Thresholds) 3 ค่า คือ k1, k2, k3 และมีพารามิเตอร์ช่วยอีก 3 ตัว คือ PEAK และ ค่า G1, G2

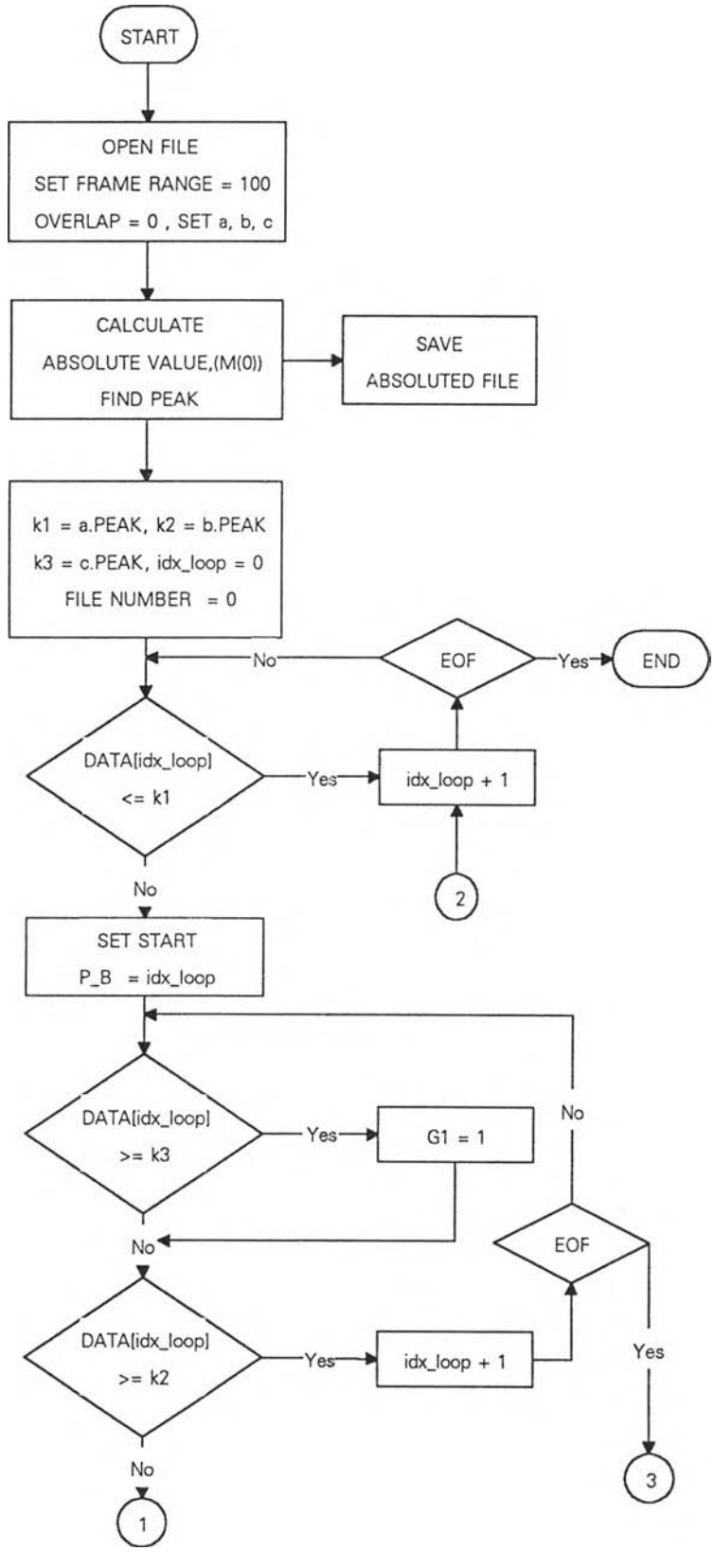
เนื่องจากเราได้ทำการบันทึกเสียงในสภาวะเสียงที่มีความแตกต่างกัน บางครั้งมีเสียงรบกวนค่อนข้างมาก อันเนื่องจากสภาวะภายนอกรวมทั้งเกิดจากอุปกรณ์ที่ใช้ในการบันทึกเสียง โดยที่ พารามิเตอร์ PEAK จะแทน ค่าสูงสุด (peak) ของสัญญาณเสียงที่วิเคราะห์ ซึ่งจะนำค่านี้ไปกำหนดค่าของระดับพลังงานอ้างอิงดังนี้คือ

$$\begin{aligned} k1 &= a.PEAK \\ k2 &= b.PEAK \\ k3 &= c.PEAK \end{aligned} \qquad \text{...2.4.1}$$

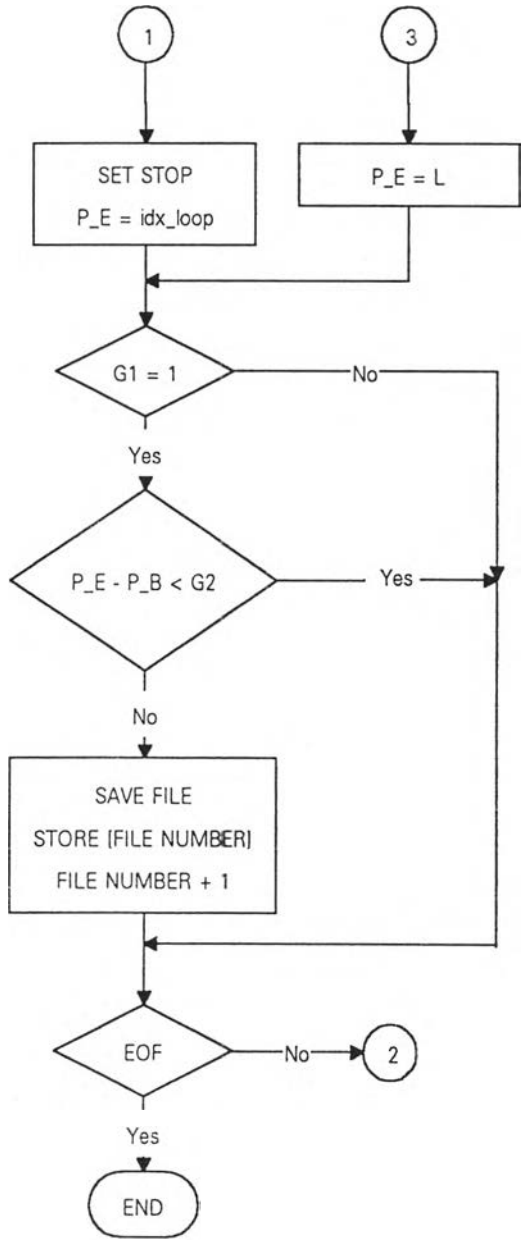
ค่าของ a, b, c จะเป็นตัวกำหนดค่าของระดับพลังงาน k1,k2,k3 ตามลำดับ จะเห็นได้ว่าค่าของระดับพลังงานอ้างอิงจะเปลี่ยนแปลงตามค่าของพารามิเตอร์ PEAK ที่ได้จากการคำนวณก่อนหน้าที่จะกำหนดระดับพลังงานอ้างอิงดังในรูปที่ 2.4.1 โดยค่าระดับพลังงาน k1 จะเป็นตำแหน่งเริ่มต้นของเสียง, k2 จะเป็นระดับพลังงานสิ้นสุดของเสียงที่ตัดคำได้ เมื่อทำการตรวจสอบได้ค่าเริ่มต้นและสิ้นสุดสัญญาณเสียงแล้วจะทำการตรวจสอบว่า ระดับพลังงานอ้างอิง k3 อยู่ภายในช่วงดังกล่าวหรือไม่ ถ้าไม่อยู่ภายในช่วงดังกล่าวจะทำการหาจุดเริ่มต้นของเสียงใหม่ แต่ถ้าค่าระดับพลังงานอ้างอิง k3 นี้อยู่ภายใน จะทำการตรวจสอบอีกครั้งว่าช่วงดังกล่าวนี้มีจำนวนเฟรมของสัญญาณน้อยกว่า ค่า G2 หรือไม่ ถ้ามีค่าน้อยกว่าจะไม่ถือว่าเป็นเสียงของคำ แต่ถ้าจำนวนเฟรมมีค่าไม่น้อยกว่า G2 แล้ว จะถือว่าเป็นส่วนของคำที่ตรวจสอบได้ ขั้นตอนการตัดคำแสดงอยู่ในรูปที่ 2.4.2 โดยจะทำการตรวจสอบจนหมดความยาวของข้อมูลเสียง ผลที่ได้ของคำแต่ละคำที่ผ่านขั้นตอนของการตัดคำนี้จะแยกเก็บเป็นคำ ๆ โดยจะบันทึกเก็บลงไฟล์



รูปที่ 2.4.1 แสดงจุดอ้างอิงเพื่อหาจุดเริ่มต้นและจุดสิ้นสุดของรูปคลื่นพลังงาน



รูปที่ 2.4.2 แสดงขั้นตอนการตัดคำ



รูปที่ 2.4.2 แสดงขั้นตอนการตัดคำ (ต่อ)