บทที่ 3
วิธีดำเนินการวิจัย

ในบทนี้จะกล่าวถึงขั้นตอนวิธีดำเนินการวิจัย ประกอบไปด้วย ขั้นตอนการเลือกใช้คุณลักษณะสำคัญของเสียงที่นำมาจากมาตรฐานการเข้ารหัส G.729 ตัวอย่างเช่น พลังงาน คาบการสั่นของเสียง (pitch period) และคุณลักษณะสำคัญเชิงความถี่ และนำคุณลักษณะเด่นดังกล่าวฝึกฝนระบบการรู้จำคำพูดและทดสอบระบบการรู้จำคำพูด

3.1 การกำหนดชุดคำศัพท์

จุดประสงค์หลักของงานวิจัยนี้ คือ สร้างระบบรู้จำที่ต้องการการคำนวณต่ำ แต่ให้ประสิทธิภาพการรู้จำที่สามารถนำไปใช้งานได้ ดังนั้น จึงกำหนดชุดคำศัพท์ แบ่งเป็น 2 ชุด ได้แก่ ชุดคำศัพท์ตัวเลขศูนย์ถึงเก้าจำนวน 10 คำ และชุดคำศัพท์พยางค์เดียว 20 คำ โดยใช้ชุดคำศัพท์อ้างอิง (วิศรุต, 2539) แสดงในภาคผนวก ก ในงานวิจัยนี้ทำการเก็บตัวอย่างเสียงพูดจำนวน 60 คน แบ่งเป็นชาย 52 คนและหญิง 8 คน โดยแบ่งเป็น 2 ชุด ได้แก่ชุดเสียงเพื่อฝึกฝนระบบจำนวน 40 คน แบ่งเป็นเสียงผู้ชาย 36 เสียงและเสียงผู้หญิง 4 เสียง ชุดเสียงเพื่อทดสอบระบบจำนวน 20 คนแบ่งเป็นเสียงผู้ชาย 16 เสียงและเสียงผู้หญิง 4 เสียง

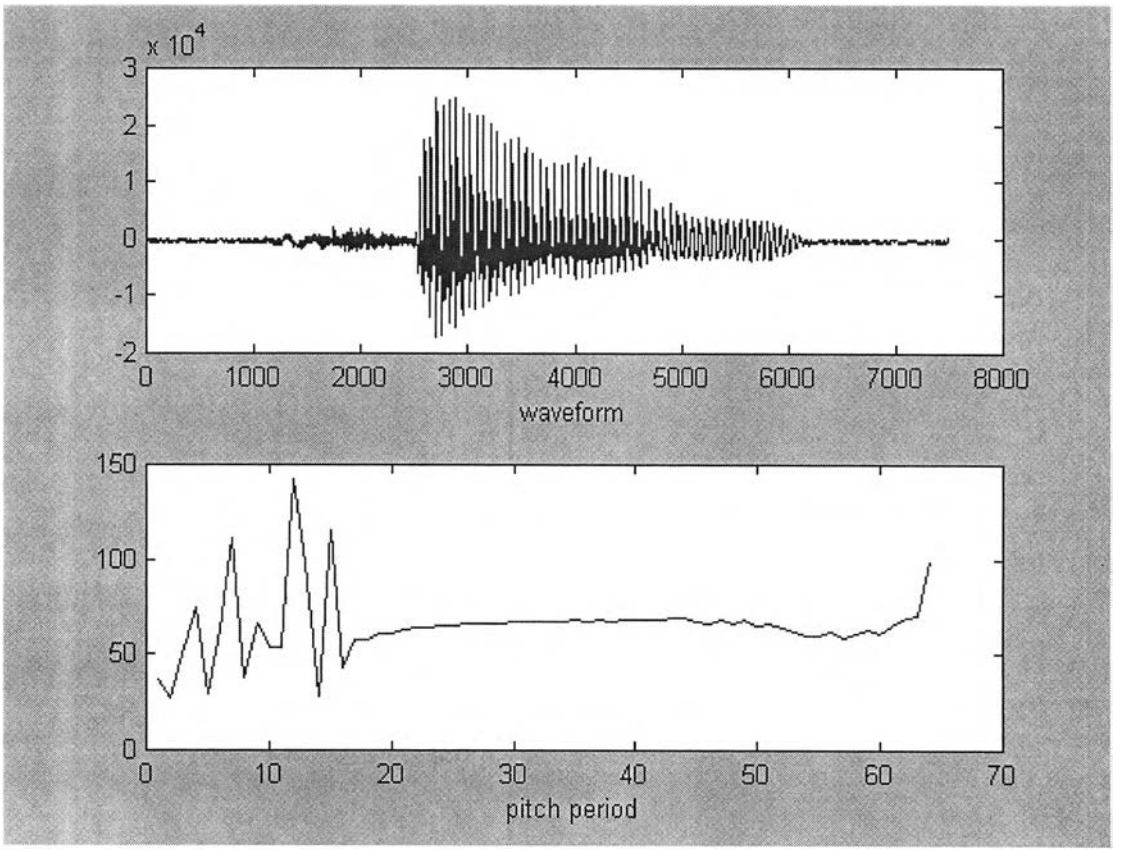
3.2 ลักษณะสำคัญของเสียงที่นำมาใช้

การเลือกลักษณะสำคัญของเสียงเป็นขั้นตอนที่สำคัญ เมื่อเลือกลักษณะที่สำคัญของเสียงที่ดีมีผลทำให้อัตรารู้จำมีค่ามากขึ้น และการคำนวณเพื่อหาลักษณะสำคัญของเสียงทำให้ต้องใช้เวลาในการคำนวณด้วย ถ้าลดความซับซ้อนในการคำนวณส่วนนี้ได้จะทำให้ลดเวลาการคำนวณลงได้

3.2.1 คาบการสั่นของเสียง (Pitch Period)

เป็นพารามิเตอร์ที่บ่งบอกได้ว่าเสียงนั้นเป็น เสียงโฆษะ (voice) หรือ เสียงอโฆษะ (unvoice) จากการทดสอบ เสียงที่เป็น เสียงโฆษะ(voice) จะมีค่า คาบการสั่นของเสียง (Pitch Period) ค่อนข้างคงนี้ในแต่ละช่วงเวลา และเสียงอโฆษะ (unvoice) จะมีค่าคาบการสั่นของเสียงเปลี่ยนแปลงรวดเร็วในแต่ละช่วงเวลา คาบการสั่นของเสียงจึงเป็นตัวบ่งชี้ช่วงเสียงโฆษะและอโฆษะ ซึ่งช่วยในการคำนวณหาจุดสิ้นสุดของเสียง (endpoint detection) เพราะช่วงเวลาใดที่มีพลังงานของเสียงสูงแสดงว่าเป็นช่วงเสียงโฆษะ และในช่วงเวลาใดที่มีพลังงานต่ำอาจจะเป็นช่วงเสียงเงียบ หรือเสียงจากสัญญาณรบกวน หรือเสียงอโฆษะก็ได้ ดังนั้น คาบการสั่นของเสียงจึงเป็น

ตัวบ่งชี้อีกตัวหนึ่งที่ช่วยในการตัดคำ ที่บริเวณช่วงเสียงใดมีการเปลี่ยนแปลงของคาบการสั่นของเสียงให้ลดเกณฑ์พลังงานในการตัดคำของช่วงเสียงลง



รูปที่ 3.1 สัญญาณเสียงและคาบการสั่นของเสียงของคำว่า "สอง"

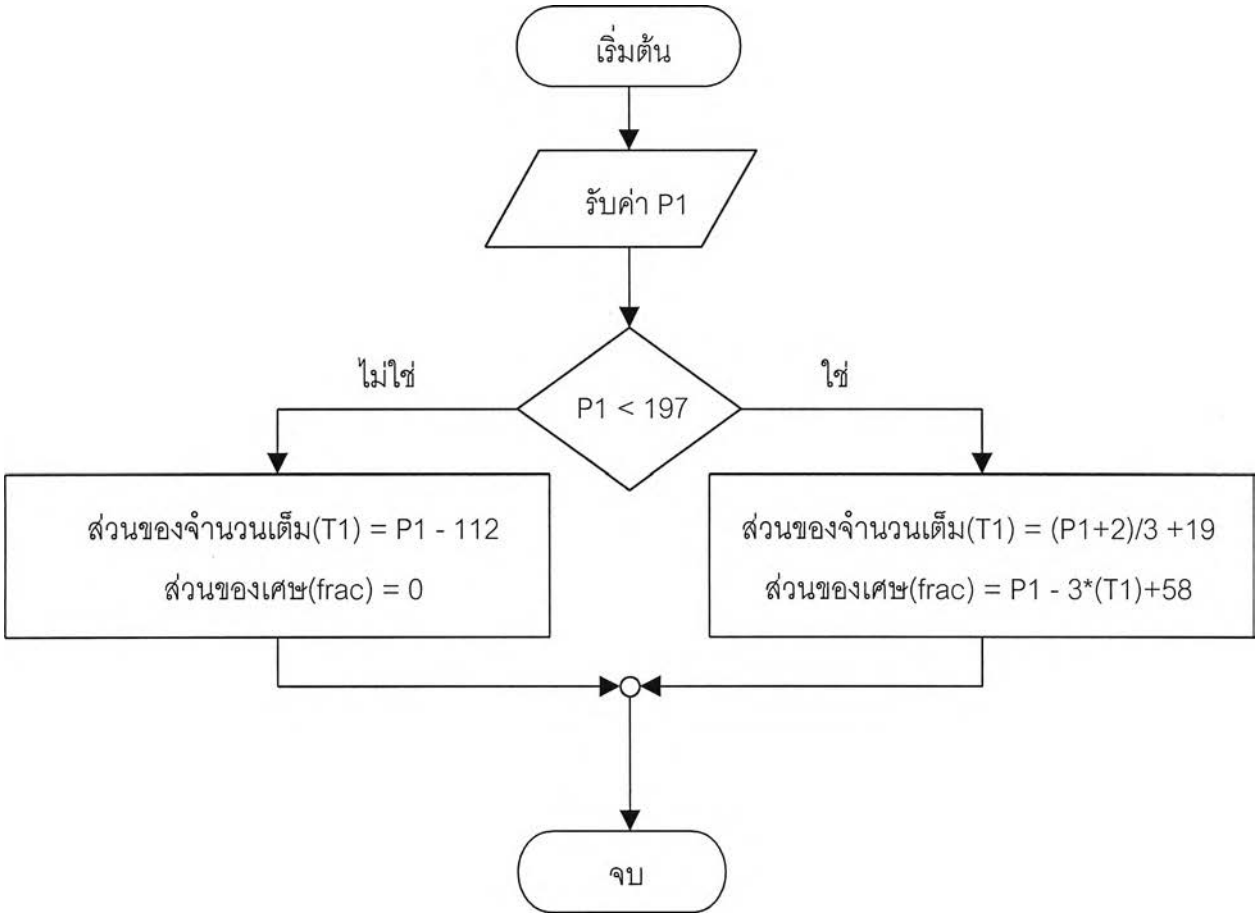
มาตรฐานการเข้ารหัส G.729 ได้ส่งข้อมูลเกี่ยวกับคาบการสั่นของเสียงโดยตรงคือพารามิเตอร์

P0 (Pitch-delay parity) ขนาด 1 บิต

P1 Adaptive-codebook delay ในสับเฟรมแรก ขนาด 8 บิต

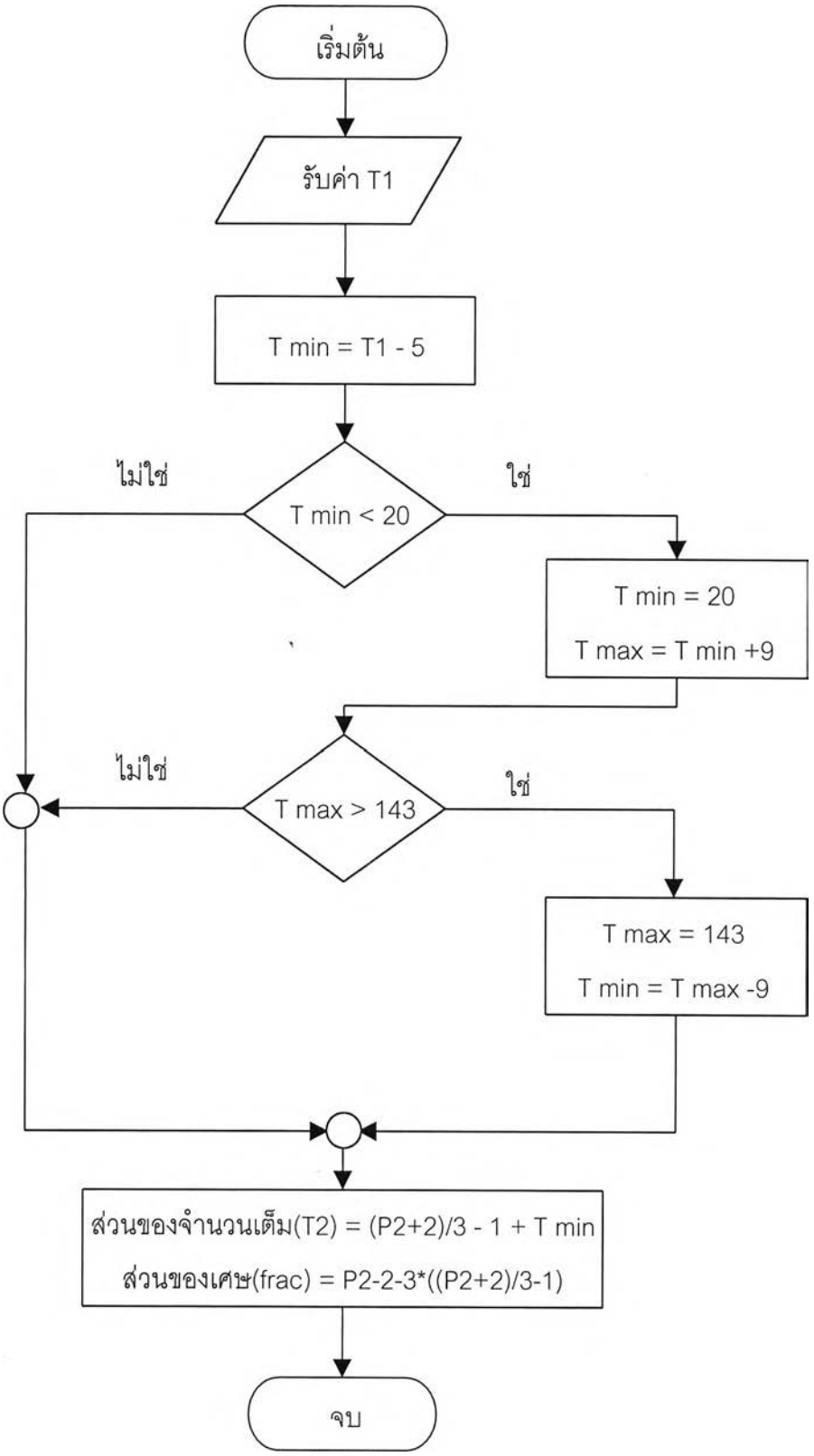
P2 Adaptive-codebook delay ในสับเฟรมหลัง ขนาด 5 บิต

จากข้อมูลที่ได้ทั้ง 1+8+5 = 14 บิตดังกล่าว สามารถนำมาคำนวณค่าคาบการสั่นของเสียงซึ่งแบ่งแต่ละเฟรมเป็น 2 สับเฟรม โดยในสับเฟรมแรก คำนวณตามขั้นตอนดังรูปที่ 3.2



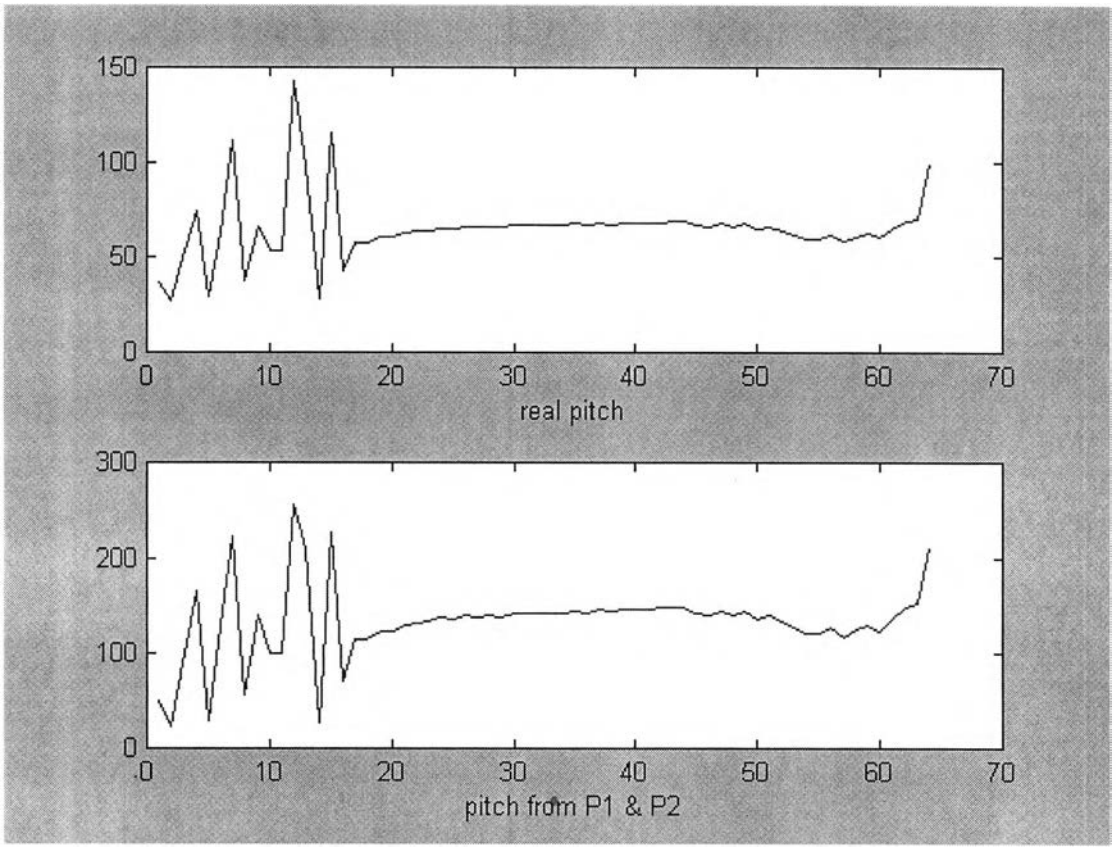
รูปที่ 3.2 ขั้นตอนในการหาคาบการสั่นของเสียงในสับเฟรมแรก

และคำนวณคาบการสั่นของเสียงในสับเฟรมหลังดังรูปที่ 3.3



รูปที่ 3.3 ขั้นตอนในการหาคาบการสั่นของเสียงในสับเฟรมหลัง

จากการศึกษาพบว่าคาบการสั่นของเสียงที่คำนวณได้จากขั้นตอนดังกล่าวมีความละเอียดมาก กล่าวคือมีความละเอียดถึงระดับ $\frac{1}{3}$ ของตัวอย่าง ซึ่งไม่จำเป็นสำหรับการนำมาตรวจสอบว่าช่วงเสียงใดเป็นเสียงโฆษะหรือเสียงอโฆษะ



รูปที่ 3.4 คาบการสั่นของเสียงจริงเทียบกับค่าที่ได้จากข้อมูล P1 และ P2 โดยตรง

การดึงข้อมูลจากบิต P1 และ P2 จากมาตรฐาน G.729 โดยตรง ต้องอาศัยการตรวจสอบจากบิต P0 ซึ่งเป็นบิตตรวจสอบ ถ้าผลลัพธ์ของ P0 บ่งชี้ว่าเกิดข้อผิดพลาด ตั้งค่า P1 เท่ากับค่า P2 ของเฟรมก่อนหน้านี้ และถ้าผลลัพธ์ของ P0 บ่งชี้ว่าไม่มีข้อผิดพลาดให้ใช้ค่า P1 ได้ตามปกติ

จากรูปที่ 3.4 คาบการสั่นของเสียงจริงเทียบกับค่าที่ได้จากข้อมูลจากมาตรฐาน G.729 โดยตรงมีลักษณะคล้ายกันมาก สามารถใช้แทนกันได้

3.2.2 พลังงานของเสียง

เป็นลักษณะที่สำคัญในการรู้จำเสียงพูด โดยเฉพาะในส่วนของการตัดคำ แต่มาตรฐานนี้ไม่ได้ส่งข้อมูลเกี่ยวกับพลังงานโดยตรง จึงต้องทำการคำนวณเอง โดยการคำนวณนี้ มิได้สร้างสัญญาณเสียงขึ้นใหม่ทั้งหมดแล้วจึงคำนวณพลังงาน เพราะเป็นการเพิ่มภาระในการคำนวณ แต่ทำการคำนวณโดยตรงจากข้อมูลที่ได้จากการเข้ารหัส ซึ่งต้องทำการคำนวณให้มีความซับซ้อนน้อยที่สุด

3.2.2.1 การหาพลังงานโดยเก็บค่าพลังงานในเฟรมอดีต แทนการเก็บเวกเตอร์ชุดรหัสที่เปลี่ยนแปลงได้ (Adaptive codebook vector)

เป็นแนวความคิดที่เก็บค่าพลังงานในอดีตในแต่ละสับเฟรมแทนที่จะใช้หน่วยความจำปริมาณมากในการเก็บการกระตุ้นในอดีตของแต่ละเฟรม

เนื่องจากคาบการสั่นของเสียงจะอยู่ระหว่าง 20 ถึง 143 ทำให้ต้องเก็บพลังงานย้อนหลังไป 2 เฟรม(หรือ 4 สับเฟรม) จาก

$$u(n) = \hat{g}_p\, v(n) + \hat{g}_c\, c(n) \qquad (3.1)$$

ถ้า $\hat{g}_p\, v(n)$ กับ $\hat{g}_c\, c(n)$ ไม่มีเฟสตรงกัน จะได้

$$|u(n)|^2 = |\hat{g}_p\, v(n)|^2 + |\hat{g}_c\, c(n)|^2$$

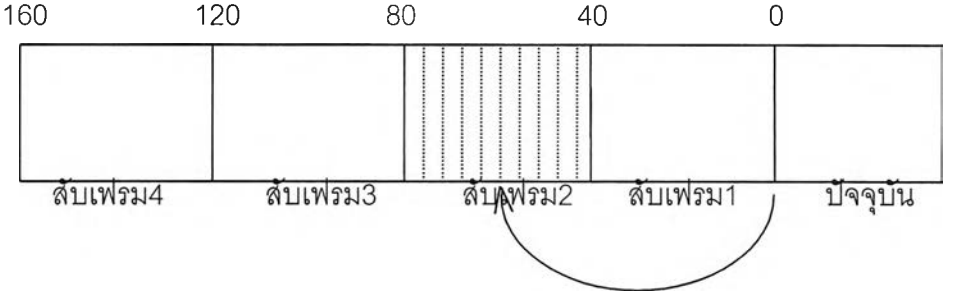
$$= \hat{g}_p{}^2 v(n)^2 + 4\, \hat{g}_c{}^2 \qquad (3.2)$$

$v(n)^2$ ประมาณด้วยค่า antilog ของพลังงานที่เก็บไว้ในแต่ละสับเฟรม

$4\, \hat{g}_c{}^2$ คือ antilog ของพลังงานในสับเฟรมปัจจุบัน

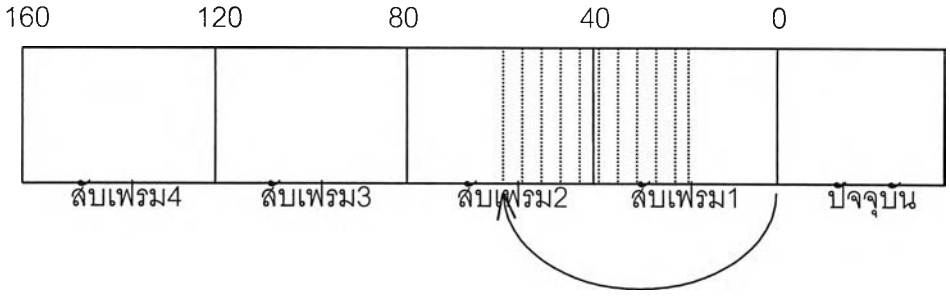
โดยมีแนวความคิด อยู่ 3 แบบ ในการคำนวณพลังงาน

ก. นำพลังงานของสับเฟรมปัจจุบันบวกกับพลังงานของสับเฟรมอดีตที่มีค่าคาบการสั่นของเสียงชี้อยู่ ไม่ว่าคาบการสั่นของเสียงจะชี้ที่ตำแหน่งใดของสับเฟรมนั้นก็ตาม



เช่น ค่าคาบการสั่นของเสียงอยู่ระหว่าง 40 ถึง 80 จะนำค่าพลังงานในสับเฟรมที่ 2 มาบวกกับ สับเฟรมปัจจุบัน ซึ่งจะเห็นว่าเป็นการคำนวณค่าประมาณอย่างหยาบ ๆ

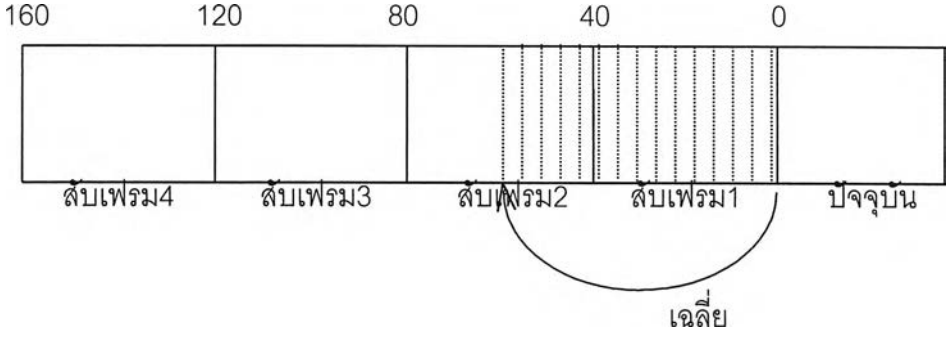
ข. เพื่อความละเอียดยิ่งขึ้นจะประมาณว่าในแต่ละสับเฟรม จะมีการกระจายของพลังงานสม่ำเสมอและประมาณพลังงานจากค่าคาบการสั่นของเสียง แบบเชิงเส้น



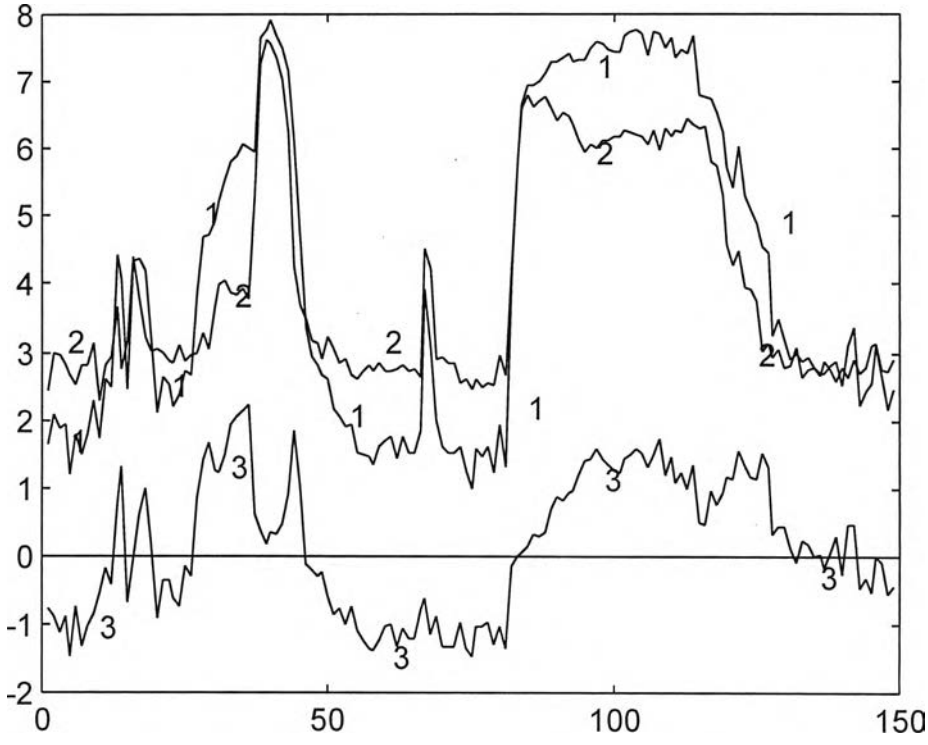
เช่น ค่าคาบการสั่นของเสียงเป็น 70 จะนำพลังงานของ สับเฟรมที่1 กับ สับเฟรมที่ 2 มาประมาณเป็นเชิงเส้น โดย

$$\text{พลังงานรวม} = \text{พลังงานของสับเฟรมปัจจุบัน รวมกับ} \left[\frac{10}{40}(\text{subframe1}) + \frac{30}{40}(\text{subframe2}) \right] \quad (3.3)$$

ค. เพื่อความละเอียดยิ่งขึ้น จะทำการคำนวณพลังงานจากการเฉลี่ยสับเฟรมปัจจุบันถึงสับเฟรม ที่คาบการสั่นของเสียงชี้อยู่



ได้ผลดังรูปที่ 3.5



รูปที่ 3.5 พลังงานเสียงจริง(1)เทียบกับประมาณจากการบวกแบบ เฉลี่ยทุกสับเฟรม (2) และอัตราส่วนระหว่างพลังงานเสียงจริงกับการประมาณจากการบวกแบบเฉลี่ยทุกสับเฟรม(3)

ซึ่งในแบบ ค. นี้จะได้ระดับพลังงานที่เรียบและเหมือนพลังงานเสียงจริงที่สุด จึงเลือกแบบ ค. นี้ในการปรับปรุงการหาพลังงานเสียงต่อไป

3.2.2.2 การประมาณโดยอาศัยผลตอบเชิงความถี่ร่วมด้วย

ขั้นต่อไปจะทำการปรับปรุงให้ค่าพลังงานที่คำนวณใกล้เคียงกับพลังงานจริงมากที่สุดจากรูปที่ 3.5 ตัวอย่างเสียงจะมี 2 ช่วงที่ค่าพลังงานต่างกันอย่างเห็นได้ชัด ถ้าต้องการพลังงานเสียงที่ใกล้เคียงกับพลังงานเสียงจริงจึงต้องทำความเข้าใจกับส่วนของฟิลเตอร์ จาก

$$s(n) = u(n) - \left[a_1 s(n-1) + a_2 s(n-2) + \ldots + a_{10} s(n-10)\right] \qquad (3.4)$$

หาผลการแปลง Z จะได้

$$S(Z) = U(Z) - \left[a_1 Z^{-1} S(Z) + a_2 Z^{-1} S(Z) + \ldots + a_{10} Z^{-10} S(Z)\right] \qquad (3.5)$$

$$S(Z) = \frac{U(Z)}{1 + a_1 Z^{-1} + a_2 Z^{-2} + ... + a_{10} Z^{-10}} \quad (3.6)$$

$$|S(Z)|^2 = \left| \frac{U(Z)}{1 + a_1 Z^{-1} + a_2 Z^{-2} + ... + a_{10} Z^{-10}} \right|^2 \quad (3.7)$$

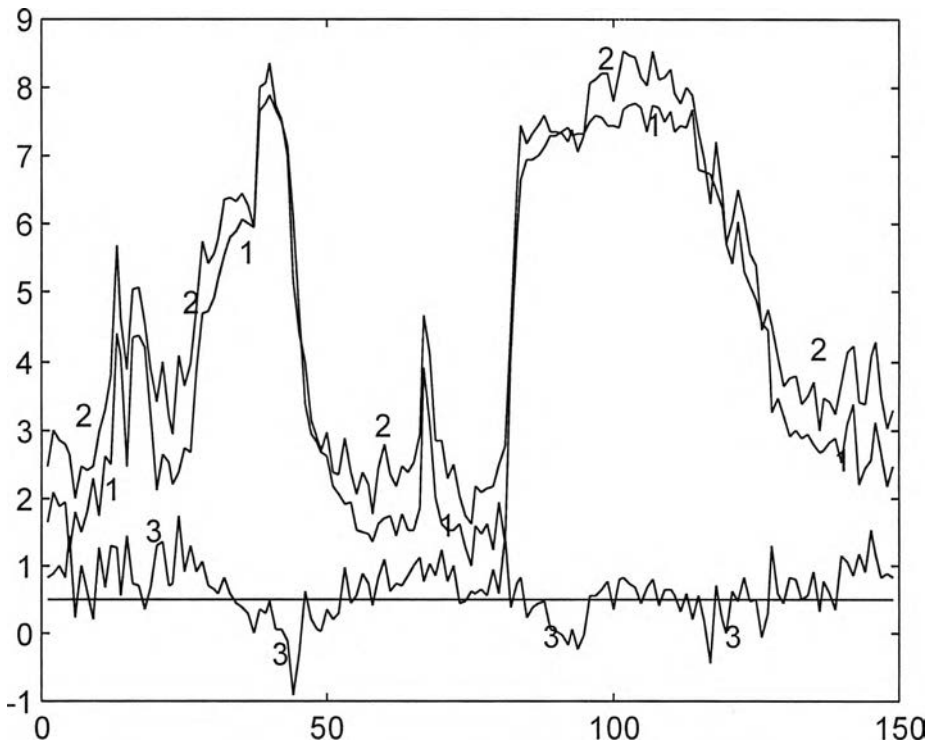
$$|S(Z)|^2 = \frac{|U(Z)|^2}{\left|1 + a_1 Z^{-1} + a_2 Z^{-2} + ... + a_{10} Z^{-10}\right|^2} \quad (3.8)$$

จากสมการ ถ้าต้องการหาพลังงานของเสียงจริง นอกจากต้องหาพลังงานของ excitation แล้วยังต้องคำนวณ $\left|1 + a_1 Z^{-1} + a_2 Z^{-2} + ... + a_{10} Z^{-10}\right|^2$

เนื่องจากพัลส์ของการกระตุ้นจะอยู่ในช่วงความถี่ต่ำ จึงสมมติว่าใกล้เคียงกับความถี่ 0 หรือ ค่า z=1 จึงได้

$$\left|1 + a_1 Z^{-1} + a_2 Z^{-2} + ... + a_{10} Z^{-10}\right|^2 = \left|1 + a_1 + a_2 + ... + a_{10}\right|^2 \quad (3.9)$$

เมื่อทำการคำนวณขนาดกำลังสองของผลบวกของ LPC ในแต่ละสับเฟรมแล้วนำไปหารกับพลังงานของการกระตุ้นจะได้ดังรูปที่ 3.6



รูปที่ 3.6 พลังงานเสียงจริง(1)เทียบกับประมาณจากการบวกแบบ เฉลี่ยและคิด LPC(2) และอัตราส่วนระหว่างพลังงานเสียงจริงกับการประมาณจากการบวกแบบเฉลี่ยและคิด LPC (3)

ซึ่งมีความใกล้เคียงกับพลังงานจริงมาก

จากขั้นตอนดังกล่าวทำให้มีการคำนวณใกล้เคียงกับพลังงานเสียงจริงในระดับที่น่าพอใจ อีกทั้งยังลดการคำนวณค่อนข้างมาก

3.2.2.3 การคำนวณพลังงาน โดยใช้ค่า LSF โดยตรง

เครื่องเข้ารหัส G.729 จะเข้ารหัส 80 บิตต่อเฟรม ในการคำนวณ LPC ต้องทำการถอดรหัส เป็น LSF ก่อนและคำนวณ LSP แล้วจึงทำการคำนวณ LPC ดังนั้นจึงน่าจะใช้ LSF แทน LPC ในการคำนวณแทน

จาก $1+a_1Z^{-1}+a_2Z^{-2}+...+a_{10}Z^{-10} = (Z-Z_1)(Z-Z_2)...(Z-Z_{10}) = \prod_{i=1}^{10}(1-Z_i)$ (3.10)

ซึ่งคือผลคูณของระยะจาก Z=1 ถึง โพลของพหุนามของการประมาณพันธะเชิงเส้น (linear prediction) และจะมีค่าใกล้เคียงกับระยะจาก Z=1 ถึงศูนย์ของ $F_1(Z)$ หรือระยะจาก Z=1 ถึงศูนย์ของ $F_2(Z)$

เมื่อ $F_1(Z)$, $F_2(Z)$ คือสัมประสิทธิ์ LSP โดย

$$F_1(Z) = \frac{A(z)+Z^{-11}A(z^{-1})}{1+Z^{-1}} \quad (3.11)$$

$$F_2(Z) = \frac{A(z)-Z^{-11}A(z^{-1})}{1-Z^{-1}} \quad (3.12)$$

และ $F_1(Z)$ คือ รากคำตอบในตำแหน่งคู่ และ $F_2(Z)$ คือรากคำตอบในตำแหน่งคี่ อีกทั้งในการเข้ารหัสเราจะเปลี่ยน สัมประสิทธิ์ LSP เป็นสัมประสิทธิ์ LSF แทน ดังนั้นในขั้นตอนการถอดรหัสเราจะได้สัมประสิทธิ์ LSF ก่อน ซึ่งสัมประสิทธิ์ LSF ก็คือค่ามุมบนวงกลมหนึ่งหน่วย

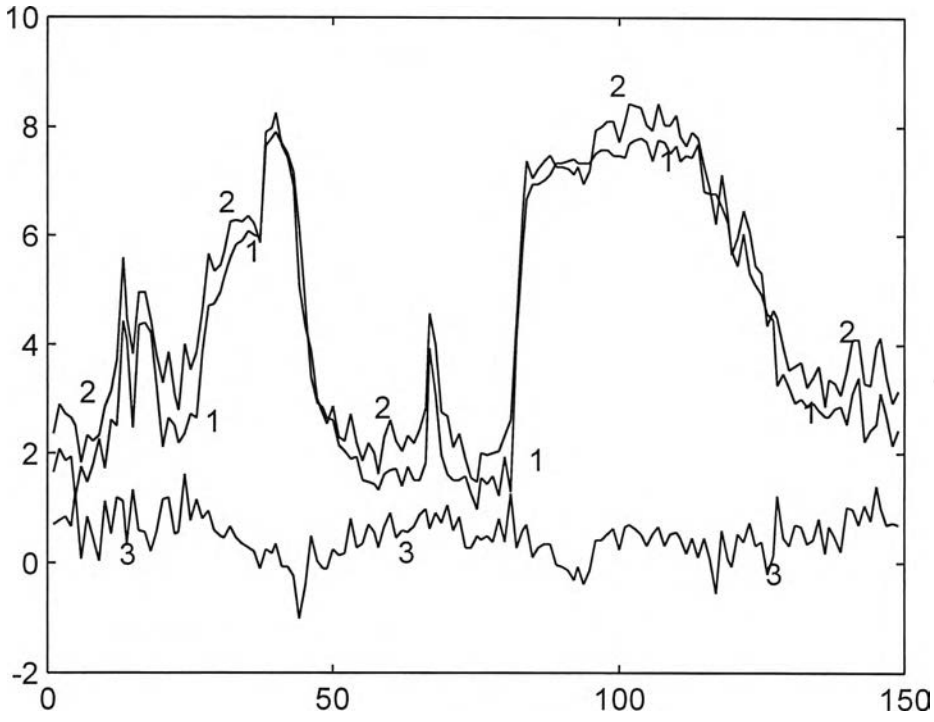
$$\omega_i = \arccos(q_i) \quad (3.13)$$

ระยะจาก Z=1 ถึง $F_1(Z)$ หรือ $F_2(Z)$

$$= \sqrt{(1-\cos\omega_i)^2+(\sin\omega_i)^2}$$
$$= \sqrt{1-2\cos\omega_i+\cos^2\omega_i+\sin^2\omega_i}$$
$$= \sqrt{2-2\cos\omega_i} \quad (3.14)$$

เนื่องจาก ω_i จะมีทั้งหมด 10 ค่า ซึ่งมาจากสัมประสิทธิ์ LSP พจน์คู่และพจน์คี่ ถ้าทำการประมาณที่ความถี่ 0 เช่นเดียวกับกรณีของ LPC ทำให้ไม่ต้องคิดพจน์ที่เป็นเลขคู่เพราะที่ตำแหน่งความถี่ 0 จะมีพจน์ที่เป็นเลขคู่อยู่ 1 ตัวทำให้ผลคูณของระยะของพจน์ที่เป็นเลขคู่เป็น 0 จึงคิดระยะของพจน์ที่เป็นเลขคี่เท่านั้น

$$\text{พลังงานเสียง(คิด LSFที่เป็นพจน์คี่)} = 10\log\left[\sum_{n=0}^{79}\frac{u^2(n)}{\prod_{i=1,3,5,7,9}2-2\cos\omega_i(n)}\right] \quad (3.15)$$



รูปที่ 3.7 พลังงานเสียงจริง(1)เทียบกับประมาณจากการบวกแบบเฉลี่ยทุกสับเฟรมและคิด LSF พจน์คี่(2) และอัตราส่วนระหว่างพลังงานเสียงจริงกับการประมาณจากการบวกแบบเฉลี่ยทุกสับเฟรมและคิด LSF พจน์คี่(3)

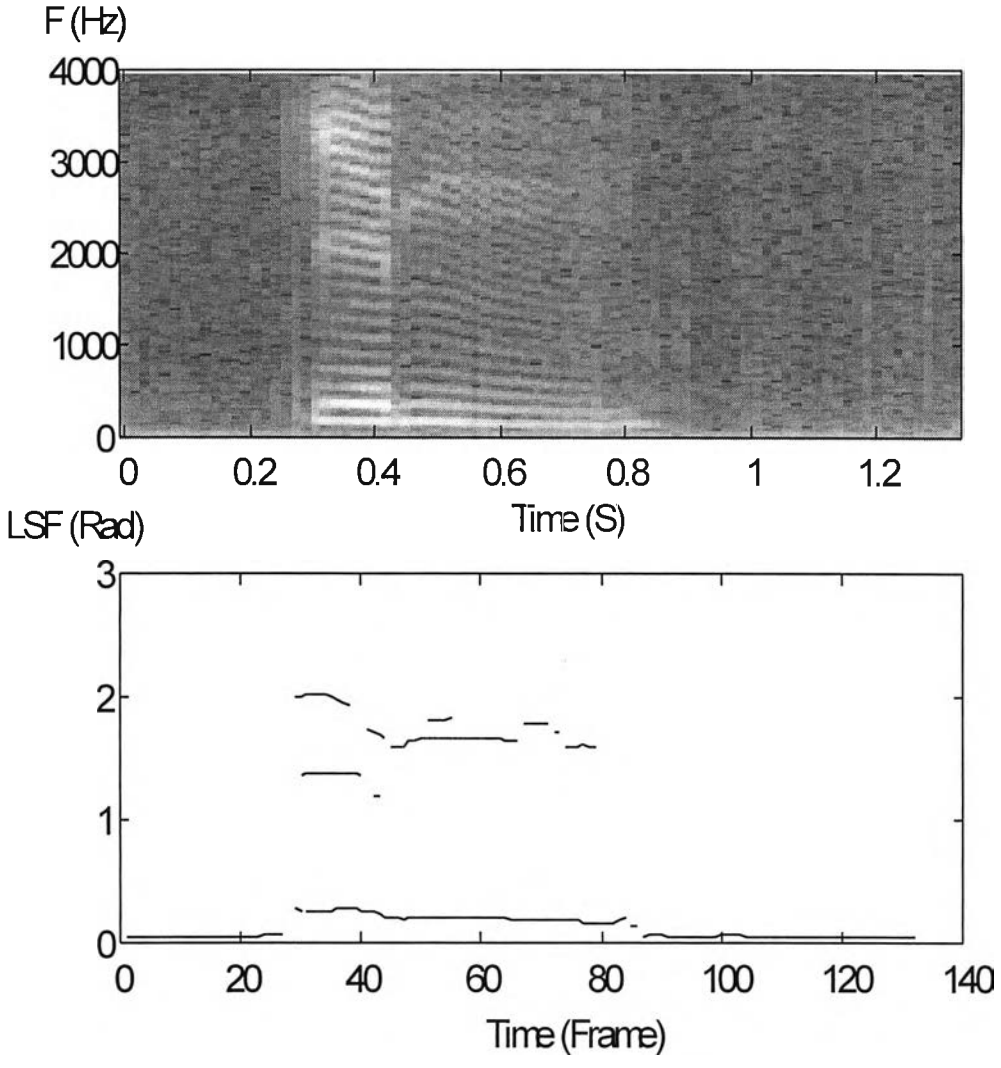
3.2.3 ลักษณะสำคัญด้านความถี่

ลักษณะสำคัญด้านความถี่ที่นิยมใช้งานอย่างแพร่หลาย เช่น สัมประสิทธิ์การประมาณพันธะเชิงเส้น (Linear Predictive Coefficient, LPC) เป็นต้น เนื่องจากมาตรฐานการเข้ารหัส G.729 ทำการเข้ารหัสของช่องเสียง (vocal tract) โดยใช้ LSP(Line Spectral Pair) ซึ่งเป็นตัวแทนของ LPC ทำให้ได้ลักษณะสำคัญด้านความถี่โดยไม่ต้องทำการคำนวณซ้ำอีก จากการศึกษาสมบัติของตัวประมวลผลเบื้องต้นของระบบรู้จำเสียงที่ดี ทำให้ทราบว่าจุดเด่นของเสียงควรจะสามารถถูกแยกออกมาได้ง่ายจากตัวประมวลผลเบื้องต้น นอกจากนี้ตัวประมวลผลเบื้องต้นควรจะสามารถลดปริมาณข้อมูลที่ไม่จำเป็นในการรู้จำลงให้เหลือแต่ข้อมูลที่สำคัญเท่านั้น จากการศึกษาคุณสมบัติของ LSP พบว่าบริเวณที่มีความถี่ LSP กระจุกตัวกันอย่างหนาแน่นจะเป็นบริเวณที่เป็นความถี่ฟอร์แมนท์ของมนุษย์ หรือความถี่มูลฐานที่ช่องเสียง (vocal tract) ของมนุษย์สั่นพ้องซึ่งเป็นลักษณะเฉพาะของเสียงและสามารถใช้ในการรู้จำเสียงได้ โดยการดูตำแหน่งหรือการเปลี่ยนแปลงของความถี่ฟอร์แมนท์

การใกล้ชิดกันของ Line Spectral Frequency (LSF) ที่ทำให้เกิดเป็นความถี่ฟอร์แมนท์นั้นสามารถประมาณได้อย่างง่ายสองแบบ คือ การใกล้ชิดกันของ LSF 2 ตัว และแบบ 3 ตัว หลักการตรวจสอบความใกล้กันนั้นใช้หลัก 2 ประการคือ

อัตราส่วนของความใกล้ชิดของคู่ LSF ที่น่าจะเกิดความถี่ฟอร์แมนท์ต่อความใกล้ชิดของคู่ LSF ที่อยู่ด้านข้าง

ค่าความใกล้ชิดของคู่ LSF นั้น

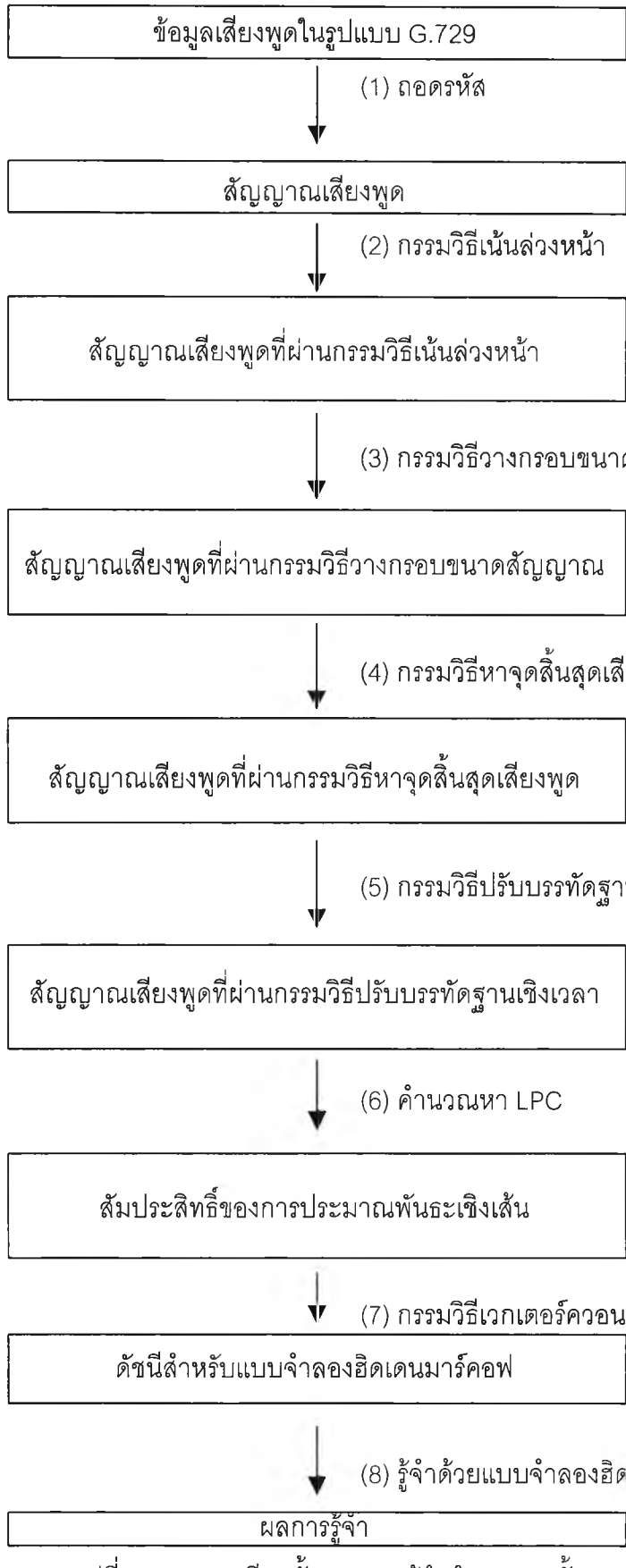


รูปที่ 3.8 การประมาณความถี่ฟอร์แมนท์ของคำว่า 'กิน' เมื่อเทียบกับสเปกโตรแกรม

ในการคำนวณค่า LSF จากมาตรฐาน G.729 นั้น เริ่มจากรับบิต L0, L1 และ L2 ตามลำดับ และนำมาคำนวณตามสมการ (2.27) และ สมการ (2.28) ตามลำดับ

3.3 ขั้นตอนการรู้จำคำพูด

ในขั้นตอนนี้ นำคุณลักษณะเด่นที่ได้คำนวณจากข้างต้นเข้าสู่ขั้นตอนการรู้จำคำพูดแบ่งออกได้เป็น 3 ขั้นตอน ได้แก่ ขั้นตอนการประมวลผลสัญญาณเบื้องต้น (Signal Preprocessing) ขั้นตอนการฝึกฝนระบบรู้จำคำพูด และขั้นตอนการทดสอบระบบรู้จำคำพูด โดยขั้นตอนการฝึกฝนระบบรู้จำคำพูด แบ่งออกเป็น 2 ขั้นตอนย่อยคือ ขั้นตอนการสร้างและฝึกฝนชุดรหัส (Codebook Training Procedure) และขั้นตอนการสร้างและฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ (Hidden Markov Model Parameters Training)



รูปที่ 3.9 รายละเอียดขั้นตอนการรู้จำคำพูดตามขั้นตอนทั่วไป

3.3.1 ขั้นตอนการประมวลผลสัญญาณเบื้องต้น (Signal Preprocessing)

โดยปกติ ขั้นตอนการประมาลผลสัญญาณเบื้องต้นเป็นการเตรียมข้อมูลเสียงที่ได้จากการบันทึกเสียง เพื่อใช้ในการรู้จำต่อไป ประกอบไปด้วยวิธีย่อย ดังนี้ กรรมวิธีเน้นล่วงหน้า (Preemphasis) กรรมวิธีวางกรอบขนาดสัญญาณ (Smoothing Window) กรรมวิธีหาจุดสิ้นสุดเสียงพูด (Endpoint Detection) กรรมวิธีปรับบรรทัดฐานเชิงเวลา (Time Normalization) แต่การรู้จำเสียงพูดโดยตรงจากการเข้ารหัส G.729 นั้นอาศัยรหัสที่ผ่านการเข้ารหัส G.729 แล้ว ซึ่งผ่านกรรมวิธีประมวลผลสัญญาณเบื้องต้นแล้ว จำเป็นเพียงใช้กรรมวิธีหาจุดสิ้นสุดเสียงพูดเท่านั้น อีกทั้งคำศัพท์ที่ใช้เป็นคำพูดพยางค์เดียว จึงเลือกใช้พลังงาน (คำนวณได้จากหัวข้อ 3.3.2) และ คาบการสั่นของเสียง (คำนวณได้จากหัวข้อ 3.3.1) เป็นพารามิเตอร์ที่ใช้หาจุดสิ้นสุดของเสียง กล่าวคือ ที่เฟรมข้อมูลที่มีการเปลี่ยนแปลงค่า คาบการสั่นของเสียงมาก แสดงว่าบริเวณดังกล่าวอาจเป็นเสียงเงียบ หรือ ช่วงเสียงอโฆษะ (unvoice) ให้ลดระดับพลังงานที่เป็นตัวบ่งบอกว่ามีเสียงหรือไม่ที่ระดับร้อยละ 5 ของพลังงานเสียงสูงสุด แต่ที่เฟรมข้อมูลที่มีการเปลี่ยนแปลงค่า คาบการสั่นของเสียงน้อย แสดงว่าบริเวณดังกล่าวเป็นเสียงโฆษะ (voice) ระดับพลังงานที่เป็นตัวบ่งบอกว่ามีเสียงหรือไม่อยู่ที่ระดับร้อยละ 10 ของพลังงานเสียงสูงสุด

3.3.2 ขั้นตอนการฝึกฝนระบบรู้จำคำพูด

เป็นขั้นตอนที่สำคัญเพราะเป็นขั้นตอนที่เลือก คุณลักษณะเด่นของเสียงเพื่อสร้างชุดรหัส

3.3.2.1 การสร้างและฝึกฝนชุดรหัส (Codebook Training Procedure)

ขั้นตอนการสร้างและฝึกฝนชุดรหัสนี้ ใช้วิธีการแบ่งเฉลี่ย K ส่วน (Deller,J., Proakis, J., Hansen,J.,1993) ของข้อมูลเสียงพูดเพื่อใช้ในการควอนไทซ์แบบเวกเตอร์ โดยการสุ่มจากชุดตัวอย่างเสียงพูดทั้งหมด ซึ่งต้องเลือกคุณสมบัติสำคัญของเสียงพูด

3.3.2.1.1 วิธีที่ 1 ทำการคำนวณค่า line spectral frequency (lsf) โดยตรง

ในส่วนของการลดการคำนวณ เมื่อใช้ค่า lsf จากการคำนวณโดยตรง สามารถลดการคำนวณในขั้นตอนที่ (1) ถึง (6) จากรูปที่ 3.9 ในการคำนวณค่า LSF จากมาตรฐาน G.729 นั้น เริ่มจากรับบิต L0, L1 และ L2 ตามลำดับ และนำมาคำนวณตามสมการ (2.27) และ สมการ (2.28) ตามลำดับ และแก้ไขค่า LSF ที่ทำให้ระบบไม่มีเสถียรภาพ ตามขั้นตอนดังนี้

ก. เรียงลำดับ LSF ($\hat{\omega}_i$) จากน้อยไปหามาก

ข. ถ้า $\hat{\omega}_1 < 0.005$ แล้ว $\hat{\omega}_1 = 0.005$

ค. ถ้า $\hat{\omega}_{i+1} - \hat{\omega}_i < 0.0391$ แล้ว $\hat{\omega}_{i+1} = \hat{\omega}_i + 0.0391$,i = 1,...,9

ง. ถ้า $\hat{\omega}_{10} > 3.135$ แล้ว $\hat{\omega}_{10} = 3.135$

ซึ่งจะได้ $\hat{\omega}_i$ เมื่อ i = 1, ..., 10 และ นำสัมประสิทธิ์ดังกล่าว เรียกว่าเวกเตอร์ฝึกฝน มาทำการฝึกฝนเพื่อสร้างชุดรหัสอ้างอิงด้วยขั้นตอนวิธีการแบ่งเฉลี่ย K ส่วน (K-Means Clustering Algorithm) ซึ่งมีประสิทธิภาพใกล้เคียงกับการแบ่งกลุ่มทวิภาค เมื่อจำนวนข้อมูลเวกเตอร์ฝึกฝนมีมากพอ และทุกครั้งที่ต้องการรู้จำ ต้องทำการคำนวณ LSF ดังขั้นตอนข้างต้น และทำการควอนไทซ์แบบเวกเตอร์ โดยอาศัยชุดรหัสอ้างอิงข้างต้น

3.3.2.1.2 วิธีที่ 2 ทำการคำนวณค่า lsf โดยประมาณ

ในการคำนวณค่า LSF จากมาตรฐาน G.729 จะต้องรับบิต L0, L1, L2 และ L3 ตามลำดับ จากการศึกษาพบว่าชุดรหัสที่มีบิต L2 และ L3 เป็นดัชนีมีผลต่อการคำนวณค่า LSF น้อย (ประมาณ 10% ของ LSF ที่คำนวณจาก L1) จึงทำการทดลองฟังเสียงจากการเข้ารหัสและถอดรหัสของมาตรฐาน G.729 เมื่อคำนึงถึง L2, L3 และไม่คำนึงถึง L2, L3 ซึ่งให้ผลการฟังเสียงใกล้เคียงกันมาก ดังนั้น LSF ที่คำนวณได้จากมาตรฐาน G.729 ที่ละ L2, L3 น่าจะใช้ในการรู้จำได้ ทำให้ LSF ที่ได้จากการรู้จำเปลี่ยนจาก สมการ (2.27) เป็น

$$\hat{l}_i = \ell 1_i(L1) \quad i = 1,...,10 \tag{3.16}$$

เมื่อ L1 เป็นดัชนีที่ชี้ชุดรหัส

ในขั้นตอนอื่นให้ทำเหมือนกับหัวข้อ 3.3.2.1.1

3.3.2.1.3 วิธีที่ 3 ทำการคำนวณค่า lsf แล้วทำการผ่านตัวกรองมัธยฐาน (Median filter)

เนื่องจาก ความกระจุกตัวของ LSF คือตัวบ่งชี้ ความถี่ฟอร์แมน ดังนั้นบริเวณที่มีความกระจุกตัวของ LSF มาก ย่อมทำให้ บริเวณนั้นมีความถี่ฟอร์แมนชัดเจน จากการศึกษาเส้น LSF พบว่ามีการแกว่งของเส้น LSF แต่ละตัว ถ้าต้องการดูแนวโน้มของแต่ละเส้นให้ชัดเจนยิ่งขึ้น ควรทำให้เส้น LSF เหล่านั้นดูราบเรียบขึ้น โดยคงลักษณะการกระจุกตัวเช่นเดิม จึงเลือกตัวกรองมัธยฐานเข้าช่วย และคาดหวังว่าจะให้ผลการรู้จำดีขึ้น เพราะไม่มีการแกว่งตัวของ LSF ในการใช้ตัวกรองมัธยฐานนี้เลือกใช้ขนาดกรอบ คือ 5 แล้วจึงทำตามขั้นตอนเหมือนหัวข้อ 3.3.2.1.1

3.3.2.1.4 วิธีที่ 4 อาศัยชุดรหัสจากมาตรฐานการเข้ารหัส G.729 โดยตรง

จากวิธีทั้งหมดที่ผ่านมา ต้องคำนวณหาลักษณะสำคัญของเสียงก่อน แล้วจึงทำการสร้างและฝึกฝนชุดรหัส โดยลักษณะสำคัญที่ใช้มากที่สุด คือ LSF ซึ่งคำนวณมาจากบิต L0,

L1, L2, L3 และบิตที่มีผลต่อการคำนวณที่สุดคือ บิต L1 จากการศึกษาพบว่าลักษณะการเก็บข้อมูลของบิต L1 มีลักษณะเป็นชุดรหัสที่ผ่านการควอนไทต์แบบเวกเตอร์ ซึ่งมีขนาด 128 ชุดรหัส หรือ 7 บิต และเป็นขนาดของชุดรหัสที่มากเพียงพอสำหรับทำการรู้จำ จึงเลือกบิต L1 เป็นพารามิเตอร์สำหรับทำการรู้จำโดยตรง ไม่ต้องสร้างและฝึกฝนชุดรหัสอีก ทำให้ลดการคำนวณได้สูงมาก กล่าวคือไม่ต้องคำนวณในขั้นตอนที่ (1) ถึง (7) ในรูปที่ 3.9

3.3.2.2 **การสร้างและฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ** (Hidden Markov Model Parameters Training)

ผลลัพธ์จากหัวข้อ 3.3.2.1 ที่ผ่านการควอนไทซ์แบบเวกเตอร์ จะเป็นข้อมูลขาเข้าของการสร้างและฝึกฝนชุดพารามิเตอร์ เพื่อทำการสร้างชุดรูปร่องต้นแบบอ้างอิงของเสียงพูดแต่ละคำ (Word Reference Template) โดยทำการปรับเปลี่ยนค่าพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ $\lambda = (A,B,\pi)$ ให้มีค่าดีที่สุดตามเสียงพูดแต่ละคำ โดยแบบจำลองที่ใช้เป็นแบบจำลองซ้าย-ขวา (Left-Right Model) ที่มี 3, 5 และ 15 สถานะตามลำดับ และการเชื่อมโยงระหว่างสถานะไม่จำเป็นต้องเชื่อมโยงกันทั้งหมด ดังนั้นแบบจำลองที่ใช้จึงมีเพียงการเชื่อมโยงในสถานะ การเชื่อมโยง ระหว่างสถานะ และการเชื่อมโยงข้ามสถานะเท่านั้น (ซึ่งอาศัยการแก้ไขปัญหาพื้นฐานข้อที่ 1 และปัญหาข้อที่ 3 ของแบบจำลองฮิดเดน มาร์คอฟ

3.3.3 ขั้นตอนการทดสอบระบบรู้จำคำพูด

ขั้นตอนการทดสอบระบบรู้จำคำพูดโดยใช้แบบจำลองฮิดเดน มาร์คอฟประกอบด้วย 3 ขั้นตอน ได้แก่ ขั้นตอนการประมวลผลสัญญาณเบื้องต้น (ทำเช่นเดียวกับหัวข้อ 3.3.1) ขั้นตอนการวิเคราะห์และวัดค่าลักษณะสำคัญ (ทำเช่นเดียวกับหัวข้อ 3.3.2) เพียงแต่ไม่ต้องสร้างชุดรหัสใหม่ แต่ใช้ชุดรหัสต้นแบบอ้างอิง (Codebook Reference Templates) ที่คำนวณได้ เพื่อหาดัชนีที่เป็นตัวแทนเสียงในแต่ละเฟรม ขั้นตอนการจำแนกรูปแบบร่วมกับขั้นตอนวิธีการตัดสินใจ เป็นการทดสอบเสียงตัวอย่าง ร่วมกับรูปรองต้นแบบอ้างอิงของเสียงพูดแต่ละคำ ซึ่งได้จากขั้นตอนการฝึกฝนชุดพารามิเตอร์ของแบบจำลองฮิดเดน มาร์คอฟ เพื่อตรวจสอบว่าเสียงทดสอบมีความคล้ายคลังกับรูปร่องต้นแบบอ้างอิงของเสียงพูดใดมากที่สุด สำหรับในงานวิจัยนี้ ขั้นตอนวิธีการตัดสินใจจะอาศัยแบบจำลองฮิดเดน มาร์คอฟ โดยการแก้ไขปัญหาพื้นฐานข้อที่ 2 ด้วยขั้นตอนวิธีการ Viterbi (Viterbi Algorithm) ผลลัพธ์ที่ได้จากขั้นตอนนี้จะเป็นชุดของเสียงพูดที่รู้จำได้ซึ่งมีความน่าจะเป็นสูงสุด