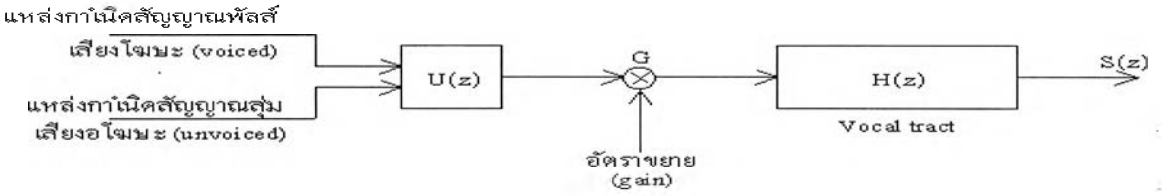
บทที่ 2
แนวคิดและทฤษฎีที่เกี่ยวข้อง

2.1 การรู้จำเสียง(Speech Recognition)

การรู้จำเสียงได้มีการพัฒนามากว่า 50 ปี แล้ว โดยเริ่มต้นครั้งแรกในปี ค.ศ.1950 เมื่อนักวิจัยได้ทำการใช้พื้นฐานความคิดของลักษณะการออกเสียงมาใช้ในการรู้จำเสียง ต่อมาในปี ค.ศ.1952 ที่ห้องทดลองเบลล์ (Bell Laboratories) ได้ทำการสร้างระบบสำหรับรู้จำเสียงของตัวเลขจากผู้พูดเพียงคนเดียว ซึ่งนับเป็นจุดเริ่มต้นงานวิจัยทางด้านนี้

2.1.1 แบบจำลองการสร้างเสียงพูด



รูปที่ 2.1 แบบจำลองการกำเนิดเสียงพูด

กลไกการกำเนิดเสียงจะเริ่มที่การขับเสียงโดยแหล่งกำเนิด (excitation source) คือการบังคับให้อากาศไหลจากปอดผ่านหลอดลมและกล่องเสียง ผ่านช่องปาก ออกมาเป็นเสียง การกระตุ้นสามารถจำแนกออกได้เป็น (Campbell,1997;Rabiner and Juang,1993)

ก. phonation excitation จะเกิดขึ้นเมื่อการไหลของอากาศถูกบังคับโดยช่องเปิดปิดบังคับเสียง เมื่อช่องนี้ปิดอยู่อากาศภายในก็จะสะสมแรงดันจนกระทั่งแรงดันนี้มากพอที่จะทำให้ช่องนี้เปิดออกหลังจากนั้น ช่องเปิดปิดบังคับเสียง ก็จะถูกดึงปิดลงอีกครั้ง โดย แรงดึง ความยืดหยุ่นและปรากฏการณ์แบร์นุลลี ส่งผลให้ส่งอากาศที่เป็นห้วง ๆ ออกมาและอากาศนี้จะเป็นตัวกระตุ้นช่องเสียง ความถี่ของการสั่นนี้จะเรียกว่าความถี่มูลฐาน (fundamental frequency) ความถี่มูลฐานนี้มีค่าเท่าใด จะขึ้นอยู่กับความยาว, แรงดึง และมวลของช่องเปิดปิดบังคับเสียง ดังนั้นความถี่มูลฐานนี้จะแตกต่างกันไปในแต่ละบุคคล

ข. whispered excitation เกิดขึ้นจากการที่อากาศไหลอย่างรวดเร็วผ่านช่องเปิดสามเหลี่ยมเล็กๆ ที่บริเวณปลายของช่องเปิดปิดบังคับเสียงที่เกิบปิด ผลที่เกิดขึ้นคืออากาศจะไหลผ่านแบบไม่มีระเบียบและเสียงที่เกิดขึ้นจะมีลักษณะสมบัติแบบสัญญาณรบกวนแถบกว้าง (wide-band)

ค. Frication excitation ถูกสร้างขึ้นได้โดยการบีบตัวของช่องทางเดินเสียง ขนาด, รูปร่าง, และอัตราของการบีบตัวจะเป็นสิ่งกำหนดรูปร่างของสเปคตรัมของสัญญาณรบกวนแถบกว้าง ถ้าการบีบตัวมากขึ้นจะทำให้ความถี่ที่เกิดขึ้นสูงขึ้น

ง. Compression excitation เป็นผลจากการปล่อยอากาศที่อยู่ในช่องทางเดินเสียงที่ปิดด้วยความดัน ส่งผลให้เกิดเสียงเงียบ (ระหว่างการสะสมแรงดัน) ตามด้วยสัญญาณรบกวนชั่วขณะ

จ. Vibration excitation เกิดจากการที่อากาศถูกบังคับให้ผ่านบริเวณปิดอื่นๆที่ไม่ใช่ช่องบังคับเปิดปิดเสียงใหญ่จะเกิดขึ้นที่ลิ้น เช่นเสียง "ร",ควบกล้ำ

เสียงที่กำเนิดแบบ phonation จะเรียกว่า เสียงโฆษะ (voiced) เสียงที่ถูกสร้างโดย phonation ผสมกับ Frication จะเรียกว่า เสียงโฆษะแบบผสม (mixed voiced) และเสียงที่สร้างโดยวิธีอื่นๆจะเรียกว่า เสียงอโฆษะ (unvoiced) การกระตุ้นเหล่านี้จะผ่านช่องเสียง (vocal tract) ซึ่งจะเปลี่ยนรูปร่างออกไป เป็นผลให้กำหนดลักษณะของเสียง สเปคตรัมของเสียงจะถูกเปลี่ยนไปตามเรโซแนนซ์ของความถี่เรโซแนนซ์ของช่องเสียง (vocal tract) นี้มีชื่อเรียกเฉพาะว่า formant frequency ดังนั้นรูปร่างของช่องเสียงจะสามารถประมาณได้ด้วยรูปร่างของสเปคตรัมของสัญญาณเสียง (เช่น การดูตำแหน่งของ formant และการเบี่ยงเบนของสเปคตรัม)

2.1.2 **การวิเคราะห์และวัดค่าลักษณะสำคัญ** (Feature Measurement)

ในการรู้จำทุกชนิดจำเป็นต้องมีการดึงลักษณะสำคัญก่อน เพื่อใช้ในการรู้จำโดยลักษณะสำคัญที่ดีขึ้นอยู่กับองค์ประกอบ เช่น ประสิทธิภาพในการรู้จำ ความซับซ้อนในการคำนวณ เนื้อที่หน่วยความจำ เป็นต้น คุณลักษณะที่ใช้แทนเสียงพูด มีหลายชนิด เช่น คาบการสั่นของเสียง (Pitch Period) พลังงาน คุณลักษณะเชิงความถี่ เป็นต้น การดึงลักษณะสำคัญ ยังทำหน้าที่ลดปริมาณสัญญาณรบกวน ลดองค์ประกอบของเสียงที่ไม่ต้องการ ลดความหลากหลายของเสียงและลักษณะการพูดของผู้พูดแต่ละคน

ก. พลังงาน เป็นคุณลักษณะสำคัญที่ใช้ในส่วนของการปรับแต่งหัวท้ายคำ (Evangelos and Nikos,1991;Ying, Mitchell and Jamieson,1993) ในเฟรมที่เป็นเสียงพูดจะมีพลังงานของเสียงสูงกว่าในเฟรมที่เป็นเสียงเงียบ ดังนั้นพลังงานจึงเป็นลักษณะสำคัญที่ใช้ในการกำหนดต้นและท้ายคำ

ข. คาบการสั่นของเสียง (Pitch Period) เป็นคุณลักษณะสำคัญที่เสริมให้การปรับแต่งหัวท้ายคำถูกต้องยิ่งขึ้น (Hamada, Takizawa and Norimatsu, 1990) เพราะพลังงานเพียงอย่างเดียว จะให้ค่าที่ผิดพลาดได้ เมื่อเป็นเสียงอโฆษะ (unvoiced) ทั้งนี้เพราะเสียงอโฆษะจะให้พลังงานที่ต่ำ คล้ายกับสัญญาณรบกวน ซึ่งการปรับแต่งหัวท้ายคำอาจผิดพลาดได้ เนื่องจาก

คาบการสั่นของเสียงจะมีค่าค่อนข้างคงที่เมื่อเป็นเสียงโฆษะ (voiced) และจะมีค่าอย่างสุ่มเมื่อเป็นเสียงอโฆษะ (unvoiced) ดังนั้นเสียงในช่วงใดที่คาบการสั่นของเสียงมีค่าอย่างสุ่ม เกณฑ์ของพลังงานที่บ่งบอกว่าเป็นเสียงพูดต้องลดลง

ค. คุณลักษณะสำคัญเชิงความถี่ เป็นลักษณะสำคัญที่ใช้ในการรู้จำ

2.1.3 สัมประสิทธิ์ของการประมาณพันธะเชิงเส้น (Linear Prediction Coefficient)

คำว่า "การประมาณพันธะเชิงเส้น" หรือ Linear Prediction ถูกนำเสนอเป็นครั้งแรกโดย N. Weiner ในปี ค.ศ. 1966 โดยเทคนิคนี้ถูกนำมาใช้เป็นครั้งแรกกับการวิเคราะห์และการสังเคราะห์เสียงโดย Itakura กับ Saito และ Atal กับ Schroeder ในปีค.ศ. 1968 (Furui, 1991) ความสำคัญของเทคนิคการประมาณพันธะเชิงเส้นนี้ก็คือ การที่รูปคลื่นและลักษณะสมบัติทางความถี่ของเสียงพูดสามารถแสดงด้วยค่าพารามิเตอร์เพียงไม่กี่ค่าได้อย่างแม่นยำและมีประสิทธิภาพ นอกจากนี้ค่าพารามิเตอร์ดังกล่าวยังสามารถคำนวณได้ง่ายอีกด้วย เหตุผลที่ LPC ได้รับความนิยมในงานประยุกต์เกี่ยวกับเสียงพูดคือ

ก. LPC เป็นแบบจำลองที่ดีสำหรับสัญญาณเสียงพูด โดยเฉพาะในช่วงสถานะกึ่งอยู่ตัว (qrasi-steady state) หรือช่วงเสียงโฆษะ (voiced) นี้ที่แบบจำลอง all-pole ของ LPC จะสามารถประมาณผลตอบของช่องเสียง (vocal tract) ได้เป็นอย่างดี ส่วนในช่วงอโฆษะ (unvoiced) และช่วงเปลี่ยนแปลงจะมีประสิทธิภาพต่ำลงแต่ก็ยังคงเป็นแบบจำลองที่มีประโยชน์ต่องานรู้จำเสียง

ข. วิธีการที่ LPC ใช้ในการวิเคราะห์เสียงนั้นนำไปสู่การแยกแหล่งกำเนิดเสียงและช่องเสียงออกจากกัน ทำให้การนำเสนอลักษณะสมบัติของช่องเสียงซึ่งเกี่ยวเนื่องกับเสียงที่ถูกเปล่งออกมาทำได้ง่าย

ค. LPC เป็นแบบจำลองที่สามารถวิเคราะห์ได้ง่ายโดยกระบวนการทางคณิตศาสตร์และก็ง่ายในการนำไปปฏิบัติทั้งโดยซอฟท์แวร์และฮาร์ดแวร์

ง. แบบจำลอง LPC ทำงานได้ผลลัพธ์ที่ดีในงานรู้จำเสียง

การประมาณพันธะเชิงเส้นเป็นขบวนการทางคณิตศาสตร์ที่ใช้ในการหาเอกลักษณ์ของระบบ โดยพิจารณาว่าเสียงเกิดจากผลรวมเชิงเส้น (linear combination) ของสัญญาณที่ทราบค่าแล้ว โดยใช้วิธี กำลังสองน้อยที่สุด (method of least square) ในการเลือกค่าพารามิเตอร์ของระบบ หลักการประมาณพันธะเชิงเส้นมีวิธีการใหญ่ 2 วิธี คือ วิธีการหาค่าความแปรปรวนร่วมและวิธีอัตสัมพันธ์ วิธี Linear Predictive coding สามารถแสดงคุณสมบัติได้ ใกล้เคียงกับพื้นฐานโมเดลการกำเนิดเสียงของมนุษย์

จากแบบจำลอง เอกลักษณ์ของระบบ H(z)

$$H(z) = \frac{S(z)}{GU(z)} \tag{2.1}$$

เมื่อพิจารณาว่าเสียงเกิดจากผลรวมเชิงเส้นของสัญญาณที่ทราบค่าแล้ว

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p) \tag{2.2}$$

เมื่อ p คือ จำนวนสัญญาณก่อนหน้าที่นำมาประมาณ

สัมประสิทธิ์ $a_1, a_2, \ldots, a_p$ มีค่าคงที่ตลอดทั้งเฟรม เพื่อให้ค่าของ s(n) ถูกต้อง จำเป็นต้องรวมสัญญาณป้อนเข้าในปัจจุบัน u(n) และอัตราขยาย G

$$s(n) = \left[\sum_{k=1}^{p} a_k s(n-k) \right] + Gu(n) \tag{2.3}$$

เขียนใน z-domain จะได้

$$S(z) = \sum_{k=1}^{p} a_k z^{-k} S(z) + GU(z) \tag{2.4}$$

$$\frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2.5}$$

เมื่อ $H(z) = \frac{S(z)}{GU(z)}$ และ $A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}$

ดังนั้น

$$H(z) = \frac{1}{A(z)} \tag{2.6}$$

จากสมการข้างต้น เราสามารถกำหนดเอกลักษณะของระบบได้โดยการระบุสัมประสิทธิ์ $a_1, a_2, \ldots, a_p$ และใช้สัมประสิทธิ์เหล่านี้ในการรู้จำ

2.1.4 Line Spectrum Pair (LSP)

LSP เป็นทางเลือกหนึ่งในการแสดงลักษณะของช่องเสียง (vocal tract) หรือลักษณะเชิงความถี่ (Smith and Schalkwyk , 1988) โดยเริ่มแรกนั้นได้ถูกเสนอโดย Itakura ในชื่อของ line spectral representation (LSR) ซึ่งภายหลังได้ถูกพัฒนาเป็น LSP ซึ่งได้รับการศึกษาอย่างมากในประเทศญี่ปุ่น โดยเฉพาะที่ NTT-ECL ในสหรัฐ Wakita ได้เขียนเอกสารทำการสอนครั้งแรกเกี่ยวกับการสังเคราะห์ LSP ในการวัดคุณภาพเสียงที่ใช้ LSP ในการเข้ารหัสจะดีกว่า LPC มาก (Coetzee and Barnwell, 1989) จากการทดสอบโดยใช้ค่า DRT (diagnostic rhyme test) พบว่า LSP vocoder ที่ 800 b/s มีค่า 87 ในขณะที่มาตรฐาน 2400 b/s LPC vocoder มีค่า 88.4

DRT เป็นวิธีการทดสอบคุณภาพของเสียงที่ทำการเข้ารหัสวิธีหนึ่ง(Deller, Proadis and Hansen, 1993) โดยการบันทึกเสียงด้วยวิธีการเข้ารหัสต่าง ๆ แล้ว ใช้คนทดสอบว่าฟังเสียงเหล่านั้นรู้เรื่องถูกต้องหรือไม่ โดยคำนวณจาก

$$DRT = \frac{N_{correct} - N_{incorrect}}{N_{test}} \times 100\% \quad (2.7)$$

เมื่อ N_{test} คือจำนวนครั้งการทดสอบทั้งหมด

$N_{correct}$ คือจำนวนครั้งที่ตอบถูกต้อง

$N_{incorrect}$ คือจำนวนครั้งที่ตอบไม่ถูกต้อง

LSP มีพื้นฐานจากการประมาณพันธะเชิงเส้น กำหนดพหุนามการประมาณพันธะเชิงเส้น อันดับ m

$$A_m(z) = 1 + a_1z^{-1} + a_2z^{-2} + \ldots + a_mz^{-m} \quad (2.8)$$

เมื่อ a_i สัมประสิทธิ์ของการประมาณพันธะเชิงเส้น

ในการลงรหัสของ LPC โดยทั่วไปจะแปลงสัมประสิทธ์ a_i ไปเป็น สัมประสิทธิ์การสะท้อน (reflection coefficient) ซึ่งมีข้อเสียคือค่าผิดพลาดจากการควอนไทซ์ (quantization error) ที่เกิดกับพารามิเตอร์ตัวใดตัวหนึ่งจะมีผลกระทบกับสเปคตรัมทั้งหมด แต่มีวิธีการอื่นที่มีประสิทธิภาพดีกว่าคือการลงรหัสความถี่ฟอร์แมนท์ (formant frequencies) และ แบนด์วิดท์ ซึ่งค่าผิดพลาดจากการควอนไทซ์จะถูกจำกัดในบริเวณหนึ่งเท่านั้น และที่ความถี่สูงขึ้นจะถูกควอนไทซ์หยาบกว่าที่ความถี่ต่ำ ทำให้สอดคล้องกับความสามารถในการฟังของมนุษย์ แต่การคำนวณจะยุ่งยาก ทำให้สัมประสิทธิ์การสะท้อนยังนิยมใช้มากกว่า

เพื่อให้สอดคล้องกับแบบจำลองของหลอดเสียง (acoustic tube) ที่มีปลายเปิดอนันต์และปิดสนิทตามลำดับ เนื่องจากสภาวะขอบเขตทั้งสองนี้ ทำให้ไม่มีพลังงานสะท้อนถูกดูดกลืนทางด้านปลาย และจะเป็นระบบที่ไม่มีการสูญเสีย จึงนิยามเงื่อนไขของเขตใหม่อีก 2 เงื่อนไข(Smith and Schalkwyk, 1988) คือ

ก. สอดคล้องกับแบบจำลองของหลอดเสียงที่มีปลายปิดสนิท (สัมประสิทธิ์การสะท้อนเป็น +1) จะได้

$$P(z) = A_m(z) - z^{-(m+1)}A_m(z^{-1}) = A_m(z)[1-R(z)] \quad (2.9)$$

ข. สอดคล้องกับแบบจำลองของหลอดเสียงที่มีปลายเปิดอนันต์ (สัมประสิทธิ์การสะท้อนเป็น -1) จะได้

$$Q(z) = A_m(z) + z^{-(m+1)}A_m(z^{-1}) = A_m(z)[1+R(z)] \quad (2.10)$$

เมื่อ $R(z) = \frac{z^{-(m+1)}A(z^{-1})}{A(z)}$

R(z) มีคุณสมบัติเป็นตัวกรองผ่านทุกความถี่ (all pass filter) และเรียกว่า “Ratio Filter” จากการตรวจสอบพหุนาม LSP ทั้งสอง พบว่ามีคุณสมบัติที่น่าสนใจดังต่อไปนี้

1. ศูนย์(Zero)ทั้งหมดของพหุนาม LSP วางตัวบนวงกลมหนึ่งหน่วย

2. ศูนย์(Zero)ของ P(z) และ Q(z) สลับกันไปมาอย่างมีลำดับ
3. คุณสมบัติเฟสต่ำสุด (minimum phase) ของ A(z) ยังเป็นจริง ถ้าคุณสมบัติทั้งสองข้อแรกเป็นจริงหลังจากควอนไทซ์

พหุนาม LSP ทั้งสองจะมีอันดับ m+1 ดังนั้นจึงมีศูนย์ m+1 ค่า ถ้า m เป็นเลขคู่ P(z) จะมีศูนย์ ที่ z = 1 และ Q(z) จะมีซีโรที่ z = -1 ถ้า m เป็นเลขคี่ ทั้ง P(z) และ Q(z) จะมีศูนย์ที่ z = -1 และ z = 1 อีกทั้งสัมประสิทธิ์ของการทำนายเป็นจำนวนจริง ดังนั้นซีโรที่เกิดขึ้นจะเป็นคู่สังยุคกัน เนื่องจากสมบัติของความเป็นคู่ ครึ่งล่างของระนาบ z (z-plane) จะเกินความจำเป็นเพราะสามารถหาได้จากครึ่งบนของวงกลม และ LSP ทุกตัวอยู่บนวงกลม 1 หน่วย ทำให้เราแทนสัมประสิทธิ์ LSP แต่ละตัวด้วยมุมก็เพียงพอต่อการบอกตำแหน่งของสัมประสิทธิ์ LSP แล้ว และเรียกว่า Line Spectral Frequency (LSF) และมีค่าอยู่ในช่วง 0 ถึง π

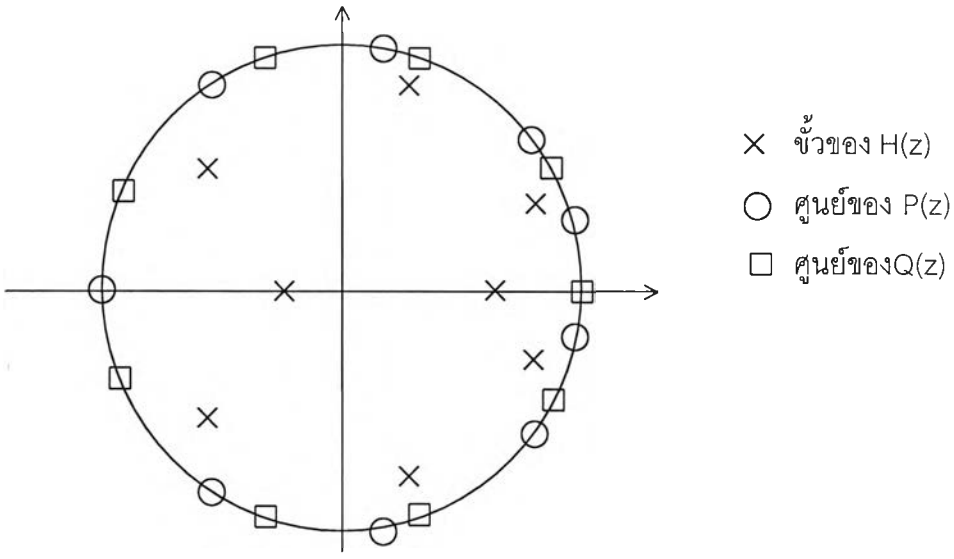
โดยการเฉลี่ยระหว่าง P(z) และ Q(z) ทำให้เราสามารถหาสัมประสิทธิ์การประมาณพันธะเชิงเส้นได้จาก LSP ดังนี้

$$A(z) = \frac{1}{2}\left(P(z) + Q(z)\right) \tag{2.11}$$

ตัวอย่างการแปลง LPC ที่มีอันดับ 8 ไปเป็น LSP โดยเป็นเสียง /U/ ในคำว่า "Foot" (Campbell, 1997)

กำลังของ z	0	-1	-2	-3	-4	-5	-6	-7	-8
LPC	1	-2.346	1.657	-0.006	0.323	-1.482	1.155	-0.190	-0.059

ซึ่งมี ขั้วของ H(z) ทั้งหมด 8 ค่าดังรูปที่ 2.2



รูปที่ 2.2 สัมประสิทธิ์ LP และสัมประสิทธิ์ LSP บนระนาบ Z ของเสียงคำว่า /U/

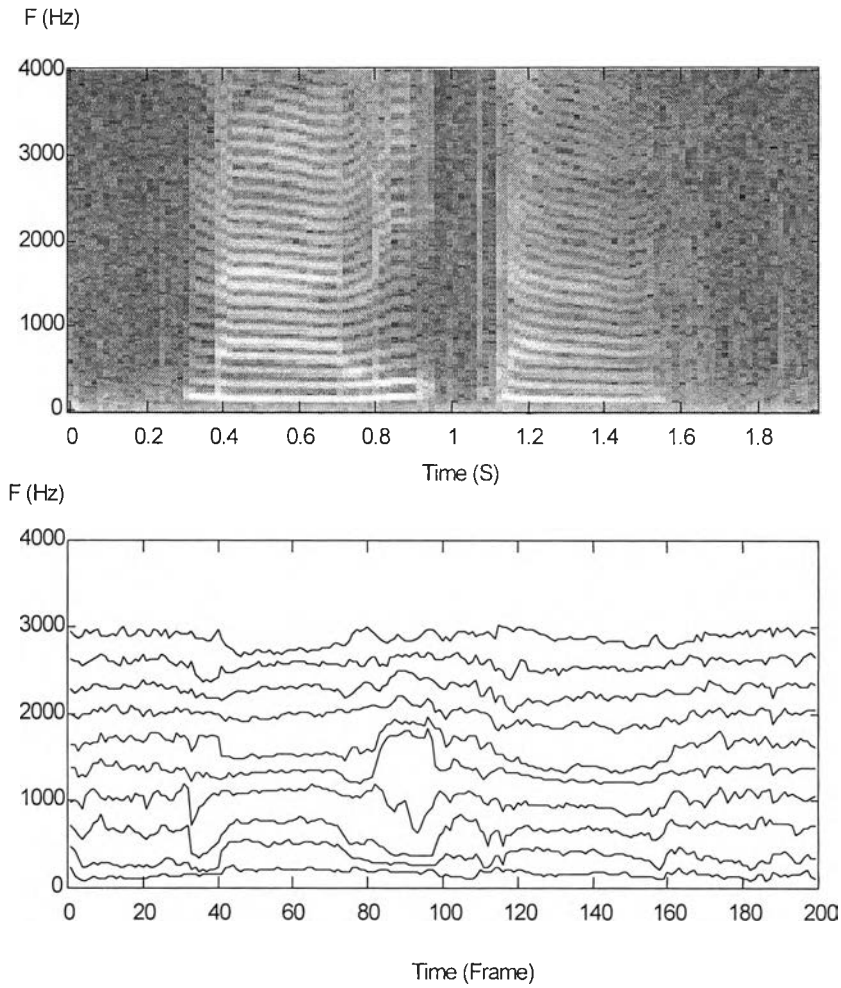
นอกจาก LSP จะมีความสัมพันธ์โดยตรงกับการประมาณพันธะเชิงเส้น แล้วยังมีลักษณะทางกายภาพที่บอกถึง ความถี่ฟอร์แมนท์

เนื่องจากเอกลักษณ์ของระบบ $H(z)=\frac{1}{A(z)}$

ถ้า A(z) มีค่าน้อยย่อมทำให้เกิด ความถี่ฟอร์แมนท์ และ $A(z)=\frac{1}{2}\left(P(z)+Q(z)\right)$

เมื่อ ซีโรของ P(z) และ Q(z) ใกล้กันมากเท่าใด ตำแหน่งระหว่างกลางของ P(z) กับ Q(z) จะทำให้ได้ค่า A(z) ต่ำ จึงเกิดเป็น ความถี่ฟอร์แมนท์ ดังนั้นตำแหน่งที่เกิด ความถี่ฟอร์แมนท์ คือ ตำแหน่งที่ศูนย์ของ P(z) กับ Q(z) อยู่ชิดกันมาก ๆ

เมื่อวาดความถี่ LSP กับ สเปกโตรแกรมจะได้ดังรูปที่ 2.3 จะเห็นได้ว่าลักษณะของรูปทั้งสองมีความสอดคล้องกัน



รูปที่ 2.3 เปรียบเทียบสเปคโตรแกรมและความถี่ LSP ของคำว่า 'นาฬิกา'

2.1.5 การทดสอบความคล้ายคลึงกันของรูปแบบ (Pattern Similarity Testing)

มีหน้าที่แบ่งชนิดของเสียงต่าง ๆ เพื่อสอดคล้องกับจุดประสงค์ของการรู้จำเสียง เทคนิคที่ใช้ในปัจจุบันแบ่งออกได้เป็น 4 ประเภทใหญ่ ๆ(Rabiner and Juang,1993) ดังนี้

1. หน่วยเสียงย่อย (Acoustic-Phonetic) โดยการใช้ทฤษฎี กฎ คุณสมบัติต่าง ๆ ของเสียงมาใช้ในการรู้จำเสียง

2. การเข้าคู่ต้นแบบ (Template Matching) เทคนิคนี้ใช้การเปรียบเทียบ รูปแบบ(pattern) ของเสียงเป็นสำคัญ

3. แบบจำลองฮิดเดน มาร์คอฟ (Hidden Markov Models ,HMM) เป็นการรู้จำเสียงโดยอาศัยความรู้ทางสถิติ เป็นเทคนิคที่ได้รับความนิยมสูง

4. เครือข่ายประสาท (Neural Networks) เป็นระบบจำลองระบบประสาทของสิ่งมีชีวิต

เนื่องจากในงานวิจัยนี้ จะดึงคุณลักษณะเด่นของเสียงพูดจากการเข้ารหัส G.729 ดังนั้นการใช้วิธีหน่วยเสียงย่อย และการเข้าคู่ต้นแบบจึงไม่เหมาะสม อีกทั้งคุณลักษณะเด่นเหล่านี้ยังเหมาะสมที่จะใช้แบบจำลองฮิดเดนมาร์คอฟ วิธีการเราเลือกในงานวิจัยนี้คือ แบบจำลองฮิดเดนมาร์คอฟ ดังนั้นจะกล่าวเฉพาะแบบจำลองฮิดเดนมาร์คอฟเท่านั้น

2.1.6 **แบบจำลองฮิดเดน มาร์คอฟ** (Hidden Markov Models, HMM)

แบบจำลองฮิดเดน มาร์คอฟ ถูกนำเสนอในปลายทศวรรษ 1960 ถึงต้นทศวรรษ 1970 และได้รับความนิยมในการใช้งานเพิ่มขึ้นเรื่อย ๆ (Rabiner,1989) ด้วยเหตุผล 2 ประการ คือ 1. แบบจำลองนี้อาศัยโครงสร้างทางคณิตศาสตร์และสามารถเปลี่ยนแปลงทฤษฎีพื้นฐานเพื่อประยุกต์ใช้งานได้อย่างกว้างขวาง และ 2. แบบจำลองนี้สามารถทำงานได้เป็นอย่างดีเมื่อเลือกประยุกต์ใช้งานอย่างเหมาะสม แบบจำลองฮิดเดน มาร์คอฟ เป็นวิธีจำแนกรูปแบบโดยอาศัยวิธีการทางสถิติ ซึ่งได้เปรียบกว่าวิธีเข้าคู่ต้นแบบ คือ สามารถเก็บข้อมูลรายละเอียดในทางสถิติเกี่ยวกับเสียงพูดไว้ได้มากกว่าวิธีการเข้าคู่ต้นแบบ อีกทั้งขั้นตอนวิธีการนี้ยังอาศัยโปรแกรมแบบพลวัต (Dynamic Programming) ทำให้มีความรวดเร็วในการประมวลผลมากยิ่งขึ้น

ประเภทของแบบจำลองสัญญาณสามารถแบ่งได้เป็น 2 ประเภท ได้แก่ แบบจำลองที่กำหนดถาวร (Deterministic Model) และแบบจำลองทางสถิติ (Statistical Models) แบบจำลองที่กำหนดถาวรจะบอกถึงคุณสมบัติเฉพาะของสัญญาณ โดยอาศัยเพียงการประมาณค่าพารามิเตอร์ที่จำเป็นให้แก่แบบจำลองสัญญาณเท่านั้น เช่น แอมปลิจูด, ความถี่, เฟสของสัญญาณไซน์ (sine) เป็นต้น แบบจำลองทางสถิติจะอาศัยคุณสมบัติทางสถิติของสัญญาณในการบอกคุณสมบัติของสัญญาณ เช่น กระบวนการเการ์เซียน (Gaussian processes), กระบวนการปัวซอง (Paisson processes), กระบวนการมาร์คอฟ (Markov processes), กระบวนการฮิดเดน มาร์คอฟ (Hidden Markov processes) เป็นต้น

แบบจำลองสัญญาณทั้ง 2 แบบข้างต้นใช้งานได้ดีกับงานประยุกต์ทั้งคู่ ส่วนแบบจำลองฮิดเดน มาร์คอฟจัดอยู่ในแบบจำลองสัญญาณทางสถิติ เพื่อออกแบบแก้ปัญหาพื้นฐาน 3 ข้อคือ

ปัญหาที่1 การคำนวณค่าความน่าจะเป็นของลำดับค่าสังเกตเมื่อกำหนดแบบจำลองให้แล้ว

ปัญหาที่2 การหาลำดับที่ดีที่สุดในแต่ละสถานะของแบบจำลอง

ปัญหาที่3 การปรับค่าพารามิเตอร์เพื่อให้ได้ค่าที่เหมาะสมที่สุดกับค่าที่สังเกต

2.1.6.1 **องค์ประกอบของแบบจำลองฮิดเดน มาร์คอฟ**

แบบจำลองฮิดเดน มาร์คอบ เป็นวิธีการใช้ทฤษฎีความน่าจะเป็น มาอธิบายการเกิดของ ลำดับ 2 ตัว คือ ลำดับของสถานะและลำดับของค่าสังเกต โดยผู้สังเกตจะเห็นเพียงผลลัพธ์ของแต่ละสถานะ (ค่าสังเกต) แต่จะไม่ทราบแน่ชัดว่าอยู่ที่สถานะใด ประกอบไปด้วยพารามิเตอร์ต่าง ๆ ดังนี้

ก. จำนวนสถานะในแบบจำลองแทนด้วย N แต่ละสถานะสามารถเชื่อมกันได้ด้วยค่าความน่าจะเป็นค่าหนึ่ง ๆ แต่ละสถานะแทนด้วย $S = \{S_1, S_2, S_3, \ldots, S_N\}$ และขณะที่เวลา t แสดงได้ด้วย q_t

ข. จำนวนสัญลักษณ์ของค่าสังเกตต่อหนึ่งสถานะแทนด้วย M ซึ่งสัญลักษณ์ของค่าสังเกตจะสัมพันธ์กับผลลัพธ์ขาออกของระบบที่ถูกจำลอง แต่ละสัญลักษณ์สามารถแสดงได้ด้วย $V = \{V_1, V_2, V_3, \ldots, V_M\}$

ค. การกระจายความน่าจะเป็นในการเปลี่ยนสถานะจากสถานะที่ i เป็นสถานะที่ j (state Transition Probability Distribution) แทนด้วย $A = \{a_{ij}\}$ เมื่อ

$$a_{ij} = P\left[q_{t+1} = S_j \middle| q_t = S_i\right], \qquad 1 \le i, j \le N \tag{2.12}$$

ง. การกระจายความน่าจะเป็นของสัญลักษณ์ของค่าสังเกต k ที่สถานะ j (Observation Symbol Probability Distribution) แทนด้วย $B = \{b_j(k)\}$ เมื่อ

$$b_j(k) = P\left[v_k \quad \text{ที่เวลา} \quad t \middle| q_t = S_j\right], \quad 1 \le j \le N, 1 \le k \le M \tag{2.13}$$

ฉ. การกระจายของสภาวะเริ่มต้น (Initial State Distribution) แทนด้วย $\pi = \{\pi_i$ เมื่อ

$$\pi_i = P\left[q_1 = S_i\right], \quad 1 \le i \le N \tag{2.14}$$

ตัวอย่างเช่น กำหนดให้

1) แบบจำลองมี 5 สถานะ

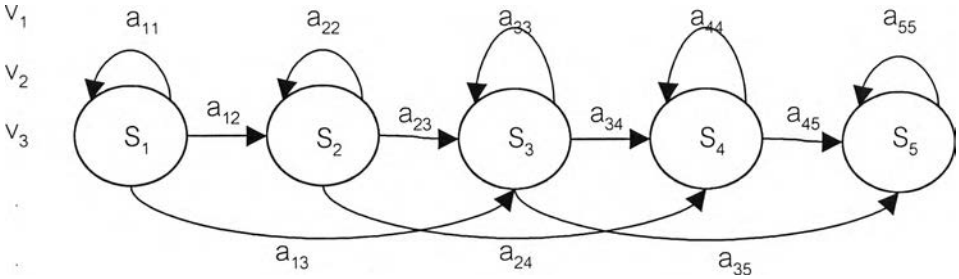
$N = 5$ ดังนั้น $S = \{S_1, S_2, S_3, S_4, S_5\}$

2) แบบจำลองมี 10 สัญลักษณ์ต่อค่าสังเกต

$M = 10$ ดังนั้น $V = \{V_1, V_2, V_3, \ldots, V_{10}\}$

3) แบบจำลองมีการกระจายของความน่าจะเป็นในการเปลี่ยนสถานะดังนี้

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 & 0 \\ 0 & a_{22} & a_{23} & a_{24} & 0 \\ 0 & 0 & a_{33} & a_{34} & a_{35} \\ 0 & 0 & 0 & a_{44} & a_{45} \\ 0 & 0 & 0 & 0 & a_{55} \end{bmatrix}$$



รูปที่ 2.4 แบบจำลองฮิดเดน มาร์คอฟที่มี N=5

ดังนั้น การกำหนดแบบจำลองฮิดเดน มาร์คอฟตัวหนึ่ง ๆ ต้องระบุทั้ง N, M และ $\lambda=(A,B,\pi)$

2.1.6.2 ปัญหาพื้นฐาน 3 ข้อของแบบจำลองฮิดเดน มาร์คอฟ

ในการประยุกต์ใช้งานแบบจำลองฮิดเดน มาร์คอฟ ในทางปฏิบัติ จะพบปัญหา 3 ข้อ ซึ่งต้องใช้ขั้นตอนวิธีต่าง ๆ ในการแก้ปัญหา

ก. ปัญหาข้อที่1 คือ ปัญหาในการคำนวณค่าความน่าจะเป็นของลำดับค่าสังเกตที่สร้างจากแบบจำลอง เมื่อกำหนดค่าสังเกตที่สร้างจากแบบจำลอง และแบบจำลองหนึ่ง กล่าวคือ ถ้ามีลำดับของค่าสังเกต $O=O_1O_2O_3...O_T$ และมีแบบจำลอง $\lambda=(A,B,\pi)$ จะทดสอบได้อย่างไรว่าลำดับนี้มีความเข้าคู่กันได้มากน้อยเพียงใดกับแบบจำลอง ซึ่งก็คือการหาค่า $P(O|\lambda)$ นั่นเอง ปัจจุบันแก้ปัญหานี้ด้วยเทคนิคกระบวนการไปหน้า (Forward procedure) และกระบวนการย้อนกลับ (Backward procedure)

ข. ปัญหาข้อที่ 2 คือ การหาลำดับที่ดีที่สุดในแต่ละสถานะของแบบจำลอง กล่าวคือ ถ้ามีลำดับของค่าสังเกต $O=O_1O_2O_3...O_T$ และมีแบบจำลอง $\lambda=(A,B,\pi)$ จะทำการเลือกลำดับสถานะที่สัมพันธ์กับ $Q=q_1q_2q_3...q_T$ ที่มีความเหมาะสมที่สุดกับแบบจำลองที่กำหนดให้ได้อย่างไร ปัจจุบันแก้ปัญหานี้ด้วยเทคนิคขั้นตอนวิธีการ Viterbi (Viterbi Algorithm)

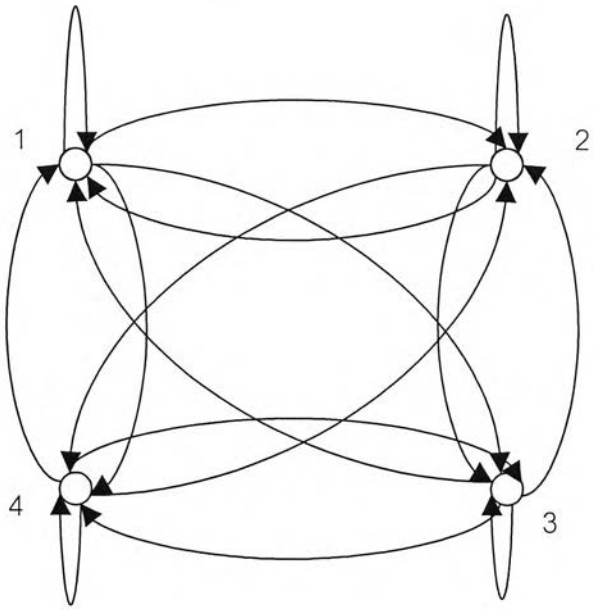
ค. ปัญหาข้อที่ 3 คือ การปรับค่าพารามิเตอร์เพื่อให้ได้ค่าที่เหมาะสมที่สุดกับค่าที่สังเกต กล่าวคือ ทำการปรับค่าพารามิเตอร์ของแบบจำลอง $\lambda=(A,B,\pi)$ เพื่อให้ได้ค่าความน่าจะเป็นของลำดับค่าสังเกต $P(O|\lambda)$ มีค่ามากที่สุด หรือการทำให้พารามิเตอร์ของแบบจำลองมีประสิทธิภาพมากที่สุด เพื่อที่จะอธิบายลำดับค่าสังเกตได้ดีที่สุด ซึ่งเป็นพื้นฐานที่สำคัญมากในกระบวนการฝึกของแบบจำลองฮิดเดน มาร์คอฟ ปัจจุบันการปรับพารามิเตอร์ใช้กระบวนการประมาณค่าซ้ำของ Baum-Welch (Baum-Welch Reestimation Procedure)

2.1.6.3 ประเภทของแบบจำลองฮิดเดน มาร์คอฟ

มีอยู่ด้วยกัน 3 แบบ หลัก(Rabiner,1989) คือ

ก. แบบจำลองที่ไม่มีเงื่อนไข (unconstrained model) แบบจำลองนี้ทุกสถานะสามารถติดต่อกับสถานะอื่น ๆ ได้ทุกสถานะ

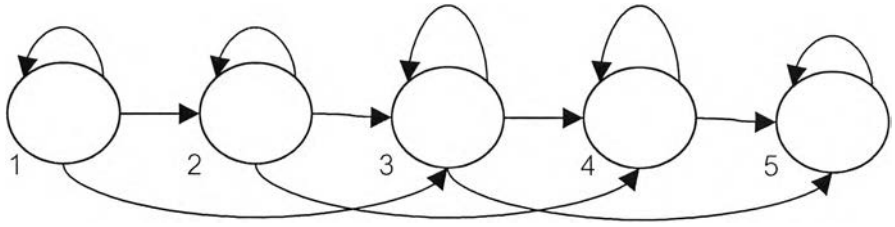
$$a_{ij} \neq 0 \quad \text{ทุกค่า i และ j}$$



รูปที่ 2.5 แบบจำลองHMM แบบเออร์กอดิกที่มี 4 สถานะ

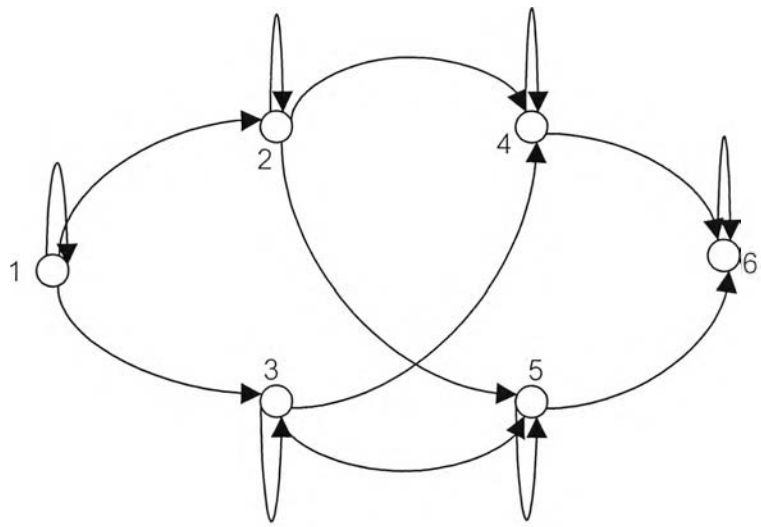
ข. แบบจำลองแบบอนุกรม (constrained serial model) แบบจำลองนี้จะเลื่อนจากสถานะหนึ่งไปอีกสถานะหนึ่ง โดยไม่มีการย้อนกลับมาที่สถานะเดิม

$$a_{ij} = 0 \quad \text{เมื่อ } i > j$$



รูปที่ 2.6 แบบจำลองHMMจากซ้ายไปขวาที่มี 5 สถานะ

ค. แบบจำลองแบบขนาน (constrained parallel model) แบบจำลองนี้จะมีคุณสมบัติเช่นเดียวกับแบบจำลองแบบอนุกรม แต่มีความซับซ้อนกว่า



รูปที่ 2.7 แบบจำลองHMM แบบขนานจากซ้ายไปขวา ที่มี 6 สถานะ

2.2 มาตรฐานการเข้ารหัส G.729

มาตรฐาน G.729 เป็นมาตรฐานการเข้ารหัสเสียง ซึ่งเป็นมาตรฐานในระบบการส่งข้อมูลและตัวกลาง (ITU-T RECOMMENDATION G.729, 1996) มาตรฐานนี้ทำการสุ่มสัญญาณเสียงที่อัตรา 8000 Hz แบ่งการคำนวณเป็นเฟรม เฟรมละ 80 ตัวอย่าง (ดังนั้นใน 1 เฟรม จะยาว 10 มิลลิวินาที) แต่การคำนวณบางส่วนเพื่อความละเอียด จะแบ่ง 1 เฟรมเป็น 2 สับเฟรม (subframe) ขนาดเท่า ๆ กัน (ดังนั้นใน 1 สับเฟรม จะยาว 5 มิลลิวินาที หรือ 40 ตัวอย่าง) ในการเข้ารหัสจะอาศัยการจำลองการสร้างเสียงของมนุษย์ ทำให้ได้เสียงที่มีคุณภาพสูงที่อัตราการเข้ารหัสต่ำ

รูปที่ 2.8 แบบจำลองการเกิดเสียงอย่างหยาบ

จากแบบจำลองการเกิดเสียง มาตรฐานการเข้ารหัส G.729 สามารถแยกการทำงาน เป็น 2 ส่วนสำคัญ

2.2.1 **ในส่วนของฟิลเตอร์** (filter)

ในส่วนของฟิลเตอร์ ถ้าเปรียบเทียบกับการกำเนิดเสียงพูดของมนุษย์ คือ ช่องเสียง (vocal tract) เป็นตัวสร้างความถี่ของเสียงที่ต้องการเปล่งออกมา ในมาตรฐาน G.729 จะใช้การจำลองโดยอาศัยการวิเคราะห์การประมาณพันธะเชิงเส้น (Linear prediction analysis) โดยนิยาม

$$\frac{1}{A(z)} = \frac{1}{1+\sum_{i=1}^{10} \hat{a}_i z^{-i}} \tag{2.15}$$

เมื่อ $\hat{a}_i$ เป็นสัมประสิทธิ์การประมาณพันธะเชิงเส้น (linear prediction coefficients, LPC) ที่ผ่านการควอนไตซ์แล้ว โดย i = 1, 2, 3, ... ,10 ในการคำนวณหา LPC จะทำการเลือกช่วงเสียงที่จะนำมาคำนวณ โดยใช้วางกรอบขนาดสัญญาณ (window)

$$\omega_{lp}(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{399}\right), n = 0,...,199 \\ \cos\left(\frac{2\pi(n-200)}{159}\right), n = 200,...,239 \end{cases} \tag{2.16}$$

กำหนดให้ s(n) คือสัญญาณเสียง

จะได้

$$s'(n) = \omega_{lp}(n)s(n) \quad \text{เมื่อ } n = 0,1, \ldots, 239 \tag{2.17}$$

เมื่อ $s'(n)$ คือ สัญญาณเสียงที่ผ่านการวางกรอบขนาดสัญญาณแล้ว

เมื่อได้สัญญาณเสียงที่ผ่านการวางกรอบขนาดสัญญาณ แล้ว ซึ่งเป็นสัญญาณเสียงที่ตัดเป็นท่อน ๆ แล้ว จึงนำมาคำนวณ สัมประสิทธิ์อัตสหสัมพันธ์ (autocorrelation coefficient)

$$r(k) = \sum_{n=k}^{299} s'(n)s'(n-k) \quad \text{เมื่อ } k = 0, \ldots, 10 \tag{2.18}$$

เพื่อหลีกเลี่ยงปัญหาในการคำนวณที่สัญญาณต่ำจึงมีการคูณด้วยตัวประกอบเพื่อให้ได้ค่าที่เหมาะสม ซึ่งเป็น สัมประสิทธิ์อัตสหสัมพันธ์ ที่ปรับปรุงได้ดังนี้

$$\begin{aligned} r'(0) &= 1.0001r(0) \\ r'(k) &= \omega_{lag}(k)r(k) \end{aligned} \tag{2.19}$$

โดยที่

$$\omega_{lag}(k) = \exp\left[-\frac{1}{2}\left(\frac{2\pi f_0 k}{f_s}\right)^2\right]$$

และ k = 1, 2, ... ,10

หลังจากที่คำนวณได้ค่า สัมประสิทธิ์อัตสหสัมพันธ์ เป็นที่เรียบร้อยแล้ว จึงทำการคำนวณสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น โดยอาศัยความสัมพันธ์ดังสมการ

$$\sum_{i=1}^{10} a_i r'(|i-k|) = -r'(k) \quad \text{เมื่อ } k = 1, 2, .. ,10 \tag{2.20}$$

ซึ่งสามารถแก้สมการนี้ได้โดยขั้นตอนวิธีการวนซ้ำของ Levinson-Durbin ถึงขั้นตอนนี้เราจะได้ LPC ได้แก่ $a_1, a_2, ..., a_{10}$

ส่วนในการลงรหัสของ LPC โดยทั่วไปจะแปลงสัมประสิทธิ์ a_i ไปเป็น สัมประสิทธิ์การสะท้อน (reflection coefficient) ซึ่งมีข้อเสียคือความผิดพลาดที่เกิดจากการควอนไทซ์ที่เกิดกับพารามิเตอร์ตัวใดตัวหนึ่งจะมีผลกระทบกับสเปคตรัมทั้งหมด ในเครื่องเข้ารหัส G.729 เมื่อทำการคำนวณได้สัมประสิทธิ์ a_i แล้วจะทำการแปลงเป็น line spectrum pair (LSP) เพราะสามารถกำจัดปัญหาดังกล่าวได้ อีกทั้งคุณภาพเสียงที่ใช้ LSP ในการเข้ารหัสจะดีกว่า LPC มาก

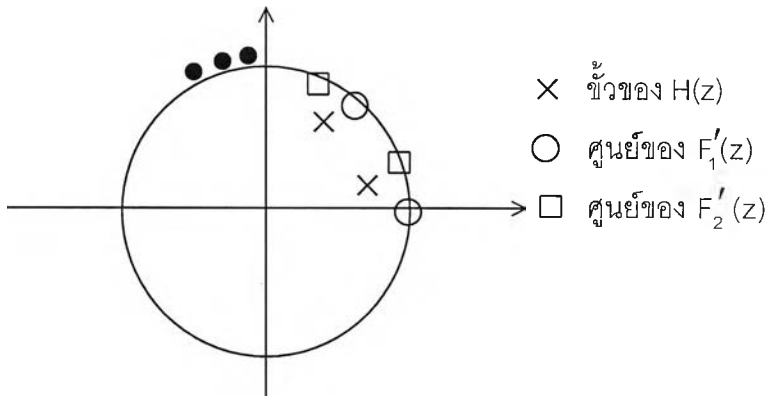
สมบัติอีกประการหนึ่งที่สำคัญ คือ LSP สามารถทำการประมาณค่าในช่วง (interpolation) ได้ เรานิยามสัมประสิทธิ์ LSP คือรากของผลบวกและผลต่าง ต่อไปนี้

$$\begin{aligned} F_1'(z) &= A(z) + z^{-11}A(z^{-1}) \\ F_2'(z) &= A(z) - z^{-11}A(z^{-1}) \end{aligned} \tag{2.21}$$

ดังนั้น

$$A(z) = \frac{1}{2}\left[F_1'(z) + F_2'(z)\right] \tag{2.22}$$

จากสมการ ศูนย์ทั้ง 10 ค่าของ A(z) ซึ่งอยู่ในวงกลม 1 หน่วยบนระนาบ Z จะถูกเปลี่ยนให้วางตัวอยู่บนวงกลม 1 หน่วยซึ่งเป็นรากคำตอบของ $F_1'(z)$ กับ $F_2'(z)$ อีกทั้งศูนย์ของ $F_1'(z)$ กับ $F_2'(z)$ จะสลับไปมาอย่างมีลำดับ ดังรูปที่ 2.9



รูปที่ 2.9 การวางตัวของศูนย์ของ $F_1'(z)$ กับ $F_2'(z)$

ด้วยเหตุผลที่ว่า a_i เป็นจำนวนจริง ดังนั้นรากของ $F_1'(z)$ กับ $F_2'(z)$ จะเป็นคู่สังยุคกันเสมอ จึงไม่มีความจำเป็นที่ต้องเข้ารหัสรากที่อยู่ครึ่งล่างของระนาบ Z อีกทั้งมีคำตอบที่แน่นอน 2 คำตอบ คือ $z = -1(\omega=\pi)$ กับ $z = 1$ $(\omega=0)$ ซึ่งถูกกำจัดทิ้งโดย

$$F_1(z) = \frac{F_1'(z)}{(1+z^{-1})}$$

$$F_2(z) = \frac{F_2'(z)}{(1-z^{-1})} \qquad (2.23)$$

โดยศูนย์ 1 ค่าของ A(z) จะเปลี่ยนไปเป็นรากของ $F_1'(z)$ 1 ค่ากับรากของ $F_2'(z)$ อีก 1 ค่า แต่รากที่อยู่ครึ่งล่างของระนาบ Z ทำให้มีรากของ LSP ที่สนใจอยู่ทั้งหมด 10 ค่า

และคำนวณจาก

$$f_1(i+1) = a_{i+1} + a_{10-i} - f_1 \text{ ,I=0,...,4} \qquad (2.24)$$

$$f_2(i+1) = a_{i+1} + a_{10-i} - f_2 \text{ ,I=0,...,4} \qquad (2.25)$$

เมื่อ $f_1(0) = f_2(0) = 1.0$

เมื่อได้สัมประสิทธิ์ LSP เราต้องทำการเปลี่ยนเป็น line spectrum frequency(LSF) เพื่อให้ง่ายต่อการควอนไทซ์ เพราะสัมประสิทธิ์ LSP ทุกตัวจะอยู่บนวงกลม 1 หน่วย ทำให้เราแทนสัมประสิทธิ์ LSP แต่ละตัวด้วยมุมก็เพียงพอต่อการบอกตำแหน่งของสัมประสิทธิ์ LSP แล้ว โดย

$$\omega_i = \arccos(q_i) \text{ , i = 1,2,...,10} \qquad (2.26)$$

เมื่อได้ ω_i แล้วจะนำมาทำการควอนไทซ์โดยเลือก การควอนไทซ์แบบเวกเตอร์ 2 ขั้น ในขั้นแรกเป็นการควอนไทซ์เพื่อให้ได้ค่า LSF อย่างคร่าว ๆ ดังนั้นจึงเป็นการควอนไทซ์ค่า LSF ค่าหลัก ซึ่งจะควอนไทซ์ค่า LSF ทั้ง 10 ค่า ดังนั้นต้องใช้ การควอนไทซ์แบบเวกเตอร์ ที่มีขนาด 10 มิติ (แทน LSF ทั้ง 10ค่า) อาศัยชุดรหัส $\ell 1$ ที่มีขนาด 128 ชุดรหัส (ใช้ 7 บิตในการเก็บข้อมูล) ขั้นที่สองเป็นการควอนไทซ์เพื่อให้ได้ค่า LSF ที่ละเอียดขึ้น โดยแบ่ง LSF 10 ค่าเป็น 2 ชุด ชุดแรกคือ LSF 5 ค่าแรก ใช้ชุดรหัส $\ell 2$ ซึ่งเป็นการควอนไทซ์แบบเวกเตอร์ ที่มีขนาด 5 มิติ (แทน LSF 5 ค่าแรก) และชุดสองคือ LSF 5 ค่าหลัง ใช้ชุดรหัส $\ell 3$ ซึ่งเป็นการควอนไทซ์แบบเวกเตอร์ ที่มีขนาด 5 มิติ (แทน LSF 5 ค่าหลัง) และชุดรหัส $\ell 2$, $\ell 3$ จะมีขนาด 32 ชุดรหัส (ใช้ชนิดละ 5 บิตในการเก็บข้อมูล)

ดังนั้นจะได้ LSP ที่ผ่านการควอนไทซ์เป็น

$$\hat{l}_i = \begin{cases} \ell 1_i(L1) + \ell 2_i(L2), i = 1,...,5 \\ \ell 1_i(L1) + \ell 3_{i-5}(L3), i = 6,...,10 \end{cases} \qquad (2.27)$$

เมื่อ L1, L2, L3 เป็นดัชนีที่ชี้ชุดรหัส

เพื่อป้องกันการเกิดความถี่ของเสียงบางเสียงมีขนาดสูงผิดปกติ ถ้าค่า $\hat{l}_i$ มีค่าใกล้กันเกินไป ดังนั้นต้องทำการเรียง $\hat{l}_i$ ใหม่โดยให้ $\hat{l}_i$ มีค่าใกล้กันที่สุดเป็น J ซึ่งทำตามขั้นตอนต่อไปนี้

```
for i = 2, ... 10
    if( l̂_{i-1} > l̂_i − J )
        l̂_{i-1} = (l̂_i + l̂_{i-1} − J)/2
        l̂_i = (l̂_i + l̂_{i-1} + J)/2
    end
end
```

ซึ่งขั้นตอนดังกล่าว จะทำ 2 ครั้ง โดยในครั้งแรกกำหนดให้ J = 0.0012 และครั้งที่สองกำหนดให้ J = 0.0006 หลังจากที่ทำการเรียงค่า $\hat{l}_i$ ใหม่แล้ว จำทำการคำนวณ LSF ที่ผ่านการควอนไทซ์ $\hat{\omega}_i^{(m)}$ ในเฟรมปัจจุบัน ที่คำนวณจากการให้น้ำหนักของ $\hat{l}_i^{(m-k)}$ ในเฟรมอดีต และ $\hat{l}_i^{(m)}$ ในเฟรมปัจจุบัน ดังสมการ

$$\hat{\omega}_i^{(m)} = \left(1 - \sum_{k=1}^{4} \hat{p}_{i,k}\right) \hat{l}_i^{(m)} + \sum_{k=1}^{4} \hat{p}_{i,k} \hat{l}_i^{(m-k)} \quad , i = 1, 2, ..., 10 \qquad (2.28)$$

เมื่อ $\hat{p}_{i,k}$ คือ สัมประสิทธิ์ของ Switched MA predictor มีหน้าที่ให้น้ำหนักของ $\hat{l}_i^{(m-k)}$ ในเฟรมอดีต และ $\hat{l}_i^{(m)}$ ในเฟรมปัจจุบันและสัมประสิทธิ์นี้จะมี 2 ชุด ซึ่งต้องเก็บ L0 อีก 1 บิตเพื่อเลือกว่าจะใช้สัมประสิทธิ์ชุดใด ดังนั้น การเก็บ LSP จะใช้ทั้งหมด 1+7+5+5 = 18 บิต

2.2.2 **ในส่วนของการกระตุ้น** (excitation)

จากที่กล่าวในตอนต้นแล้วว่า การกำเนิดเสียงต้องประกอบไปด้วย การขับเสียงโดยแหล่งกำเนิด (excitation source) ซึ่งคือการบังคับให้อากาศไหลจากปอดผ่านหลอดลมและกล่องเสียงผ่านช่องปาก ออกมาเป็นเสียง ในการสร้างการกระตุ้น จะประกอบไปด้วย 2 ส่วนสำคัญ คือ

ก. พัลส์ที่คำนวณจากสับเฟรมปัจจุบัน

เป็นพัลส์ค่าหลัก กล่าวคือ เป็นการกระตุ้นค่าหลัก แต่ยังเป็นการกระตุ้นที่ผ่านการคำนวณอย่างหยาบ ยังมีค่าแตกต่างจากการกระตุ้นที่แท้จริงอยู่มาก พัลส์ที่คำนวณจากสับเฟรมปัจจุบัน ซึ่งจะคำนวณจากชุดรหัสคงที่ (fixed codebook) คือ ตำแหน่งของพัลส์ที่มีขนาด +1 หรือ -1 ในแต่ละสับเฟรมที่คำนวณได้ ซึ่งในแต่ละสับเฟรมจะทำการคำนวณพัลส์ 4 ลูก และนับว่าเป็นพัลส์หลักในแต่ละสับเฟรม นอกจากนี้ยังมีอัตราการขยายของชุดรหัสคงที่ (fixed codebook gain) ซึ่งนำมาคูณกับพัลส์ที่ได้จากชุดรหัสคงที่ เพื่อให้ได้พัลส์ที่มีขนาดเหมือนจริงมากที่สุด

ข. การกระตุ้นที่เกิดจากเฟรมในอดีต

ทำหน้าที่ปรับค่าการกระตุ้นให้เหมือนจริงยิ่งขึ้น คำนวณมาจากชุดรหัสที่ปรับเปลี่ยนได้ (adaptive codebook) ที่มีลักษณะคล้ายกับการกระตุ้นในเฟรมอดีตซึ่งจะนำมาคูณกับอัตราขยายของชุดรหัสที่ปรับเปลี่ยนได้ (adaptive gain) แล้วจึงนำไปรวมกับพัลส์ที่คำนวณได้จากเฟรมปัจจุบัน เพื่อให้เป็นการกระตุ้นที่สมบูรณ์แบบยิ่งขึ้น

การคำนวณและการเข้ารหัสของการกระตุ้น

2.2.2.1 **ชุดรหัสคงที่** (fixed codebook)

ในการเข้ารหัสของมาตรฐาน G.729 จะเข้ารหัสโดยมีเฟรมขนาด 80 ตัวอย่าง แต่การสร้างตำแหน่งของการกระตุ้นจะกระทำทีละสับเฟรมโดยใน 1 สับเฟรม ประกอบไปด้วย 40 ตัวอย่าง ในแต่ละสับเฟรมจะมีการสร้างตำแหน่ง พัลส์ขึ้นมาใหม่เสมอโดยมีพัลส์เพิ่มขึ้น 4 พัลส์ขนาด +1 หรือ -1 โดยมีตำแหน่งดังตารางที่ 2.1

ตารางที่ 2.1 เครื่องหมายและตำแหน่งของพัลส์

พัลส์	เครื่องหมายของพัลส์	ตำแหน่งของพัลส์
i_0	s_0 : ± 1	M_0 : 0,5,10,15,20,25,30,35
i_1	s_0 : ± 1	M_1 : 1,6,11,16,21,26,31,36
i_2	s_0 : ± 1	M_2 : 2,7,12,17,22,27,32,37
i_3	s_0 : ± 1	M_3 : 3,8,13,18,23,28,33,38 4,9,14,19,24,29,34,39

และมีเวกเตอร์ของชุดรหัส (codebook vector) c(n) เป็นดังนี้

$$c(n) = s_0\delta(n-m_0)+ s_1\delta(n-m_1)+ s_2\delta(n-m_2)+ s_3\delta(n-m_3) \quad ,n=0,\ldots,39 \qquad (2.29)$$

เมื่อ $\delta(0)$ เป็นพัลส์ขนาด +1

s_0 , s_1, s_2, s_3 เป็นเครื่องหมายของแต่ละพัลส์ว่าเป็นบวกหรือลบ

เราสามารถนำมาเข้ารหัสขนาด 4 บิต สำหรับบ่งบอกว่าพัลส์เป็นชนิดบวกหรือลบ โดยให้ s =1 หมายถึง พัลส์ขนาด +1 และ s =0 หมายถึง พัลส์ขนาด -1

$$s = s_0+ 2s_1+4s_2+ 8s_3 \qquad (2.30)$$

และนำมาเข้ารหัสขนาด 13 บิต สำหรับการบอกตำแหน่งของชุดรหัสคงที่ (fixed codebook)

$$c = (m_0/5)+ 8(m_1/5)+ 64(m_2/5)+ 512(2(m_3/5+jx) \qquad (2.31)$$

โดย jx = 0 ถ้า m_3 = 3,8,...,38 และ jx = 1 ถ้า m_3 ,4,9,...,39

รวมตำแหน่งของพัลส์และเครื่องหมายของพัลส์เป็น 4+13= 17 บิต

2.2.2.2 **อัตราขยายของชุดรหัสคงที่** (Fixed codebook gain) $\hat{g}_c$

คำนวณค่าของอัตราขยายของชุดรหัสคงที่เพื่อนำมาคูณกับค่าพัลส์ที่คำนวณได้จากพัลส์ของชุดรหัสคงที่ และคำนวณหาจากค่าผิดพลาดกำลังสองถ่วงน้ำหนัก (mean-squared weighted error) ระหว่างสัญญาณเดิมกับสัญญาณเสียงที่สร้างขึ้นให้มีค่าน้อยที่สุด และสามารถถอดรหัสได้ดังนี้

$$\hat{g}_c = g'_c\,(GA_2(GA)+GB_2(GB)) \qquad (2.32)$$

โดย *GA* เป็น ชุดรหัส ที่มีขนาด 2 มิติ และแตกต่างกัน 8 แบบ (ใช้ 3 บิตในการเก็บข้อมูล) และ *GB* เป็น ชุดรหัส ที่มีขนาด 2 มิติ และแตกต่างกัน 16 แบบ (ใช้ 4 บิตในการเก็บข้อมูล)

2.2.2.3 **ชุดรหัสที่ปรับเปลี่ยนได้** (adaptive codebook)

คำนวณค่าที่ได้จากชุดรหัสที่ปรับเปลี่ยนได้ เพื่อรวมกับผลคูณของพัลส์ที่คำนวณจากชุดรหัสคงที่กับอัตราขยายของชุดรหัสคงที่ เพื่อให้ได้ค่าใกล้เคียงกับการกระตุ้นมากที่สุด ในการคำนวณหาชุดรหัสที่ปรับเปลี่ยนได้ต้องคำนวณค่าต่าง ๆ ตามขั้นตอนดังนี้

2.2.2.3.1 **อัตราขยายของชุดรหัสที่ปรับเปลี่ยนได้** (adaptive codebook gain หรือ Pitch gain) $\hat{g}_p$

เป็นค่าที่คำนวณเพื่อไปคูณกับค่าที่ได้จากชุดรหัสที่ปรับเปลี่ยนได้ และรวมกับผลคูณของพัลส์ที่ได้จากชุดรหัสคงที่กับอัตราขยายชุดรหัสคงที่ สามารถถอดรหัสดังนี้

$$\hat{g}_p = GA_1(\text{GA}) + GB_1(\text{GB}) \qquad (2.33)$$

2.2.2.3.2 **คาบการสั่นของเสียง** (Pitch Period)

เมื่อได้ทั้งตำแหน่งและขนาดของการกระตุ้นในแต่ละเฟรม แล้วยังเป็น การกระตุ้นที่ไม่คล้ายของจริงมากนัก สำหรับเสียงโฆษะ (voiced) ซึ่งเป็นเสียงที่มีการสั่นเป็นคาบ เราจึงต้องคำนวณคาบการสั่นของเสียง (Pitch Period) เพื่อนำ excitation ของเฟรมอดีตที่ถัดจากเฟรมปัจจุบันเท่ากับคาบการสั่นของเสียง มาทำการบวกเข้าเพื่อสร้างการกระตุ้นที่คล้ายของจริงยิ่งขึ้น

ในการคำนวณคาบการสั่นของเสียงจะใช้การหาความคล้ายคลึงกัน (correlation) ของเสียงในเฟรมปัจจุบันกับเฟรมอดีตมีค่ามากที่สุด นั่นคือในบริเวณใดที่มีความคล้ายคลึงกับเสียงในเฟรมปัจจุบันมากที่สุด จำนวนตัวอย่างที่ย้อนกลับไปถึงเฟรมอดีตดังกล่าวคือคาบการสั่นของเสียง โดยทำการคำนวณเป็น 2 ขั้นตอนย่อย คือ

ก. **การวิเคราะห์พิตช์แบบวงรอบเปิด** (Open-loop pitch analysis)

เป็นการลดความซับซ้อนในการคำนวณหาคาบการสั่นของเสียง ที่เหมาะสมที่สุดโดยตรง โดยในการคำนวณค่าคาบการสั่นของเสียงแท้จริงจะถูกจำกัดช่วงโดยค่าพิตช์แบบวงรอบเปิด (Open-loop pitch) นี้ วิธีการคำนวณพิตช์แบบวงรอบเปิดจะใช้การหาความคล้ายคลึงกันกับเฟรมอดีตที่มีค่ามากที่สุด ใน 3 ช่วงอดีต คือ ช่วงย้อนหลังตัวอย่างที่ 20 ถึง 39 , ช่วงย้อนหลังตัวอย่างที่ 40 ถึง 79 และช่วงย้อนหลังตัวอย่างที่ 80 ถึง 143

ข. คาบการสั่นของเสียงอย่างละเอียด

เมื่อได้พิตช์แบบวงรอบเปิดจะทำการคำนวณคาบการสั่นของเสียง โดยจำกัดค่าในการคำนวณถึงพิตช์แบบวงรอบเปิดเท่านั้น เป็นการลดการคำนวณ ไม่ต้องทำการคำนวณทั้งหมด

2.2.2.3.3 **เวกเตอร์ของชุดรหัสที่ปรับเปลี่ยนได้** (adaptive codebook vector)

เมื่อได้ค่าคาบของสั่นของเสียงนี้แล้วทำให้ทราบช่วงของเฟรมในอดีตที่มีลักษณะคล้ายกับเฟรมปัจจุบัน โดยเลือกตำแหน่งก่อนและหลัง ค่าคาบการสั่นของเสียง 10 ตัวอย่างมาทำการประมาณค่าในช่วง โดยเลือกการวางกรอบขนาดสัญญาณแบบแฮมมิง (Hamming window) เป็นฟังก์ชันในการประมาณค่าในช่วง อีกทั้งคาบการสั่นอาจมีค่าละเอียดกว่าอัตราการสุ่มตัวอย่าง ดังนั้นจึงคำนวณค่าคาบการสั่นของเสียงให้ละเอียดในระดับ 1 ใน 3 ของอัตราการสุ่มและใช้การวางกรอบขนาดสัญญาณแบบแฮมมิง ในการประมาณค่าในช่วง โดยคำนึงถึงความละเอียดของค่าคาบการสั่นของเสียงนี้ซึ่งการประมาณค่าในช่วงนี้คือการคำนวณชุดรหัสที่ปรับเปลี่ยนได้นั่นเอง

นำคาบการสั่นของเสียงมาคำนวณ เวกเตอร์ของชุดรหัสที่ปรับเปลี่ยนได้ v(n) ดังนี้

$$v(n)=\sum_{i=0}^{9}u(n-k-i)b_{30}(t+i\cdot 3)+\sum_{i=0}^{9}u(n-k+1+i)b_{30}(3-t+i\cdot 3) \qquad (2.34)$$

เมื่อ n = 0, ... ,39 และ t = 0,1,2

u(n) คือ การกระตุ้นของแต่ละลับเฟรม

b_{30} คือ การวางกรอบขนาดสัญญาณแบบแฮมมิงที่มีการตัดช่วงที่ ± 30 ($b_{30}(30)=0$)

เมื่อได้ พัลส์จากชุดรหัสคงที่ (fixed codebook) c(n)

เวกเตอร์จากชุดรหัสที่ปรับเปลี่ยนได้ (adaptive codebook vector) v(n)

อัตราขยายของชุดรหัสคงที่ (fixed codebook gain) $\hat{g}_c$

อัตราขยายของชุดรหัสที่ปรับเปลี่ยนได้ (adaptive codebook gain) $\hat{g}_p$

ดังนั้นจะได้ excitation ในแต่ละลับเฟรม คือ

$$u(n) = \hat{g}_p\, v(n) + \hat{g}_c\, c(n) \qquad (2.35)$$

ในการเก็บพารามิเตอร์ต่าง ๆ ของมาตรฐาน G.729 ซึ่งมีทั้งหมด 80 บิตต่อเฟรม เป็นไปดังตารางที่ 2.2

ตารางที่ 2.2 จำนวนบิตที่ใช้ในการเก็บพารามิเตอร์ต่าง ๆ ของมาตรฐาน G.729

พารามิเตอร์	สัญลักษณ์	Subframe 1	Subframe 2	รวม
Line Spectrum Pairs	L0,L1,L2,L3			18
Adaptive-codebook delay	P1,P2	8	5	13
Pitch-delay parity	P0	1		1
Fixed-codebook index	C1,C2	13	13	26
Fixed-codebook sign	S1,S2	4	4	8
Codebook gains(stage 1)	GA1,GA2	3	3	6
Codebook gains(stage 2)	GB1,GB2	4	4	8
รวม				80