บทที่ 2

การสังเคราะห์เสียงพูด

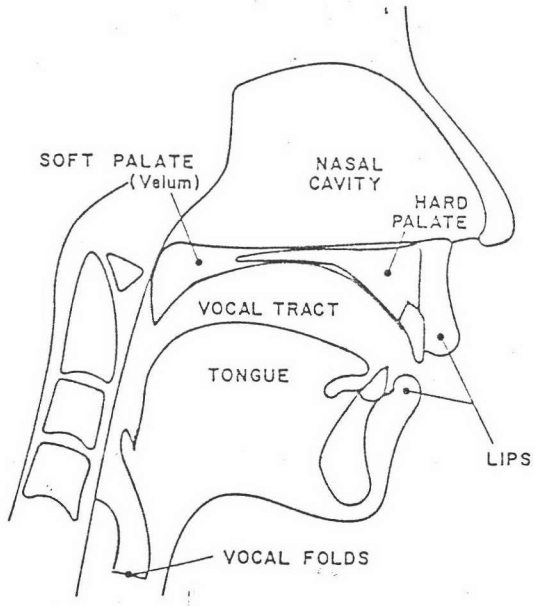
การสังเคราะห์เสียงพูด (Speech Synthesis) มีความหมายกว้างๆ คือ การผลิตเสียงพูดจากอุปกรณ์ใดๆ นอกเหนือจากการผลิตเสียงพูดจากอวัยวะผลิตเสียง ความหมายเฉพาะในปัจจุบันคือ การสร้างเสียงพูดในรูปของสัญญาณไฟฟ้าจากข้อมูลเชิงเลข

2.1 เสียงพูด

ส่วนประกอบที่สำคัญในการผลิตเสียงพูด คือ ต้นกำเนิดเสียง (Sound Source) และทางเดินเสียง (Vocal Tract) ต้นกำเนิดเสียงมีได้ 2 ลักษณะ คือ

1. เสียงก้อง (Voiced Sound) เกิดจากการสั่นของเส้นเสียง (Vocal Cords) ซึ่งเป็นส่วนประกอบของอวัยวะกล่องเสียง (Larynx) เส้นเสียงมีลักษณะเป็นกล้ามเนื้อ 2 แผ่นปิดขวางช่องหลอดลม เมื่อไม่ได้ออกเสียงช่องระหว่างเส้นเสียงจะขยายใหญ่ให้ลมเข้าออกสะดวก เมื่อมีการออกเสียงช่องจะแคบลงจนมีพื้นที่ประมาณ 5 ตารางมิลลิเมตร โดยมีความยาวประมาณ 18 มิลลิเมตร ลมที่ถูกอัดจากปอดเมื่อผ่านช่องแคบ จะทำให้เส้นเสียงสั่นและกำเนิดเสียงซึ่งมีลักษณะเป็นคาบ (Periodic) ความถี่เสียง (Pitch) ควบคุมด้วยกล้ามเนื้อเส้นเสียง ส่วนความดังขึ้นกับแรงของลมที่ผ่านออกมา ตัวอย่างของเสียงก้องได้แก่ เสียงสระต่างๆ และเสียงพยัญชนะ บ,ก ที่เกิดจากการเปล่งเสียงออกทางปาก หรือเสียงพยัญชนะ ม,น,ง เกิดจากการเปล่งเสียงออกทางจมูก หรือที่เรียกว่าเสียงนาสิก (Nasal Sound) เป็นต้น

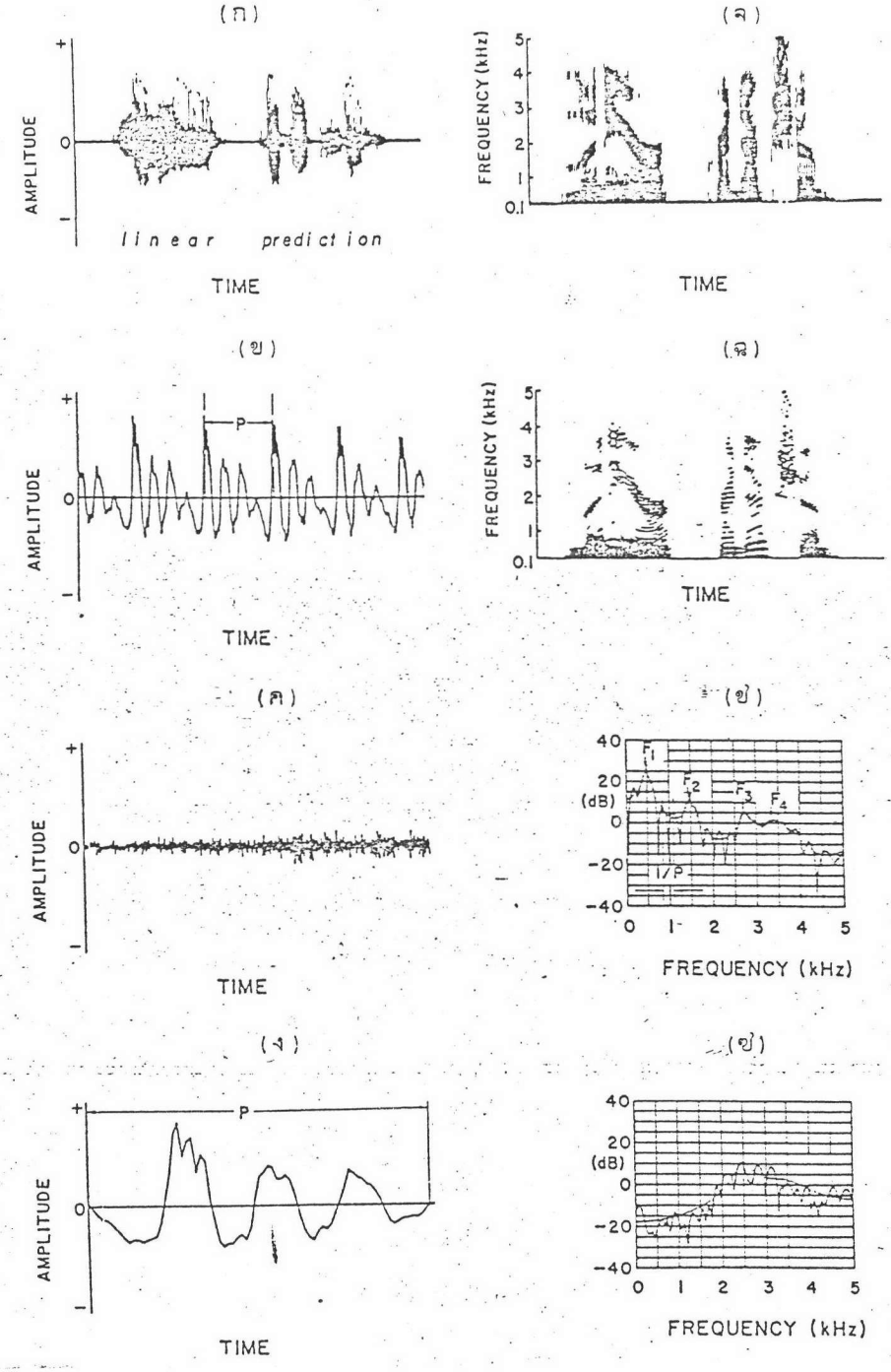
2. เสียงไม่ก้อง (Unvoiced Sound) คือเสียงที่ไม่ได้มีต้นกำเนิดมาจากเส้นเสียง เกิดจากลมที่ผ่านช่องแคบๆ ด้วยความเร็วสูงจนการไหลของอากาศมีลักษณะเป็น Turbulence เสียงที่ได้จะมีลักษณะเป็นเสียงของสัญญาณเสียงรบกวน (Noise) ซึ่งไม่เป็นคาบ ตัวอย่างของเสียงไม่ก้องได้แก่ เสียงพยัญชนะ ฟ,ช,ส ฯลฯ เสียงเหล่านี้มีชื่อเรียกเฉพาะว่าเสียงเสียดแทรก (Fricative-Sound). เสียงกัก (Stop) ที่เกิดจากการกักลมที่ผ่านทางเดินเสียงอย่างกระทันหัน เช่น เสียงพยัญชนะตัวสะกดในคำ นัด, นับ, นัก, เสียงระเบิด (Plosive Sound) หรือการเปล่งเสียงเริ่มแรกของพยัญชนะต้นของคำต่างๆ



รูป 2.1 ส่วนประกอบของอวัยวะผลิตเสียง

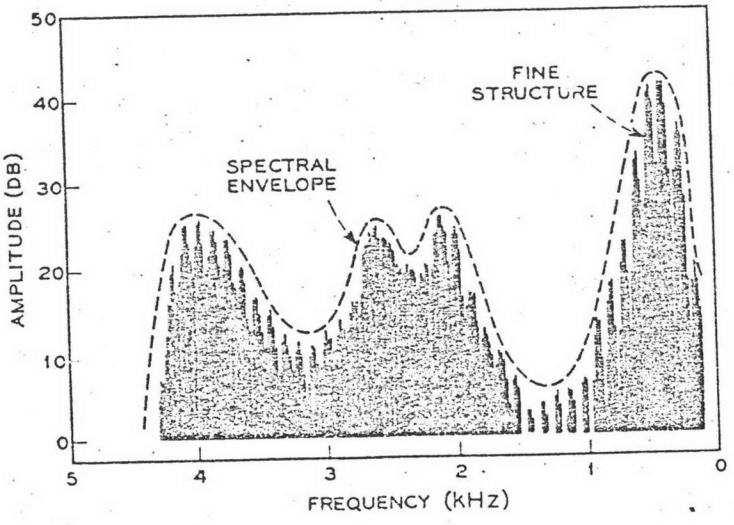
ทางเดินเสียง (Vocal Tract) คือ ช่องที่เสียงจะเดินทางผ่านจากต้นกำเนิดเสียงถึงริมฝีปาก รูป 2.1 แสดงส่วนประกอบของอวัยวะผลิตเสียง ส่วนของทางเดินเสียงคือ ช่องระหว่างลิ้นไก่ (Valum) และเพดานแข็ง (Hard Palate)กับลิ้น (Tongue) เสียงไม่ว่าจะเป็นเสียงก้องหรือไม่ก้อง จะถูกเปลี่ยนแปลงคุณภาพตามรูปร่างของทางเดินเสียงขณะนั้น ในทางวิทยาศาสตร์ สรุปได้ว่า ทางเดินเสียงคือท่อนำเสียงที่มีรูปร่างไม่แน่นอน (Non Uniform Acoustic Tube) มีความยาวประมาณ 17 เซนติเมตร มีพื้นที่หน้าตัดที่เปลี่ยนแปลงได้ระหว่าง 0 ถึง 20 ตารางเซนติเมตร [2]

กรณีของเสียงนาสิก เสียงจากต้นกำเนิดจะเดินทางผ่านโพรงจมูก (Nasal Cavity) และดังออกทางรูจมูกควบคู่กันไปกับที่ดังออกมาทางปาก โดยลิ้นไก่จะเปิดช่องให้เสียงผ่านเข้าสู่โพรงจมูกซึ่งมีรูปร่างคงที่มีความยาวประมาณ 12 เซนติเมตร ปริมาตรรวม 60 ลูกบาศก์เซนติเมตร [2]



รูป 2.2 ตัวอย่างในการศึกษาคุณสมบัติของเสียงพูด

รูป 2.2 เป็นตัวอย่างในการศึกษาคุณสมบัติของเสียงพูด ในรูป 2.2 ก แสดงขนาดของสัญญาณเสียง (Amplitude) กับเวลาของเสียงพูดคำว่า "Linear Prediction" ซึ่งเมื่อขยายดูรายละเอียดในช่วงเวลาสั้นๆ สามารถแบ่งได้เป็น 2 ส่วนคือ เสียงไม่ก้อง (Unvoiced Sound) รูป 2.2 ค ในกรณีนี้เป็นเสียงเสียดแทรก และเสียงก้อง (Voiced Sound) ซึ่งเป็นสัญญาณที่มีคาบ รูป 2.2 ข และรูป 2.2 ง ในการวิเคราะห์ทางความถี่ เสียงเสียดแทรกจะมีสเปกตรัมความถี่เสียงกระจายไปตลอดย่านดังในรูป 2.2 ช ส่วนเสียงก้องสเปกตรัมจะมียอดหลายยอด ในรูป 2.2 ช ความถี่บริเวณยอดในสเปกตรัมมีชื่อเรียกว่า ความถี่ฟอร์แมนท์ (Formant Frequency) กำกับด้วยอักษร F1-F4 ในรูป 2.2 ช ความถี่ฟอร์แมนท์ เกิดจากการกำทอนคลื่นเสียง (Resonance) ของทางเดินเสียง นอกจากนั้นภายใต้ยอดต่างๆ ของฟอร์แมนท์ ยังมีสเปกตรัมของเสียงจากเส้นเสียง ซึ่งมีความถี่เท่ากับส่วนกลับของคาบ (1/P) รูป 2.2 จ, ฉ แสดงให้เห็นการเปลี่ยนแปลงสเปกตรัมตามเวลา จะเห็นว่าช่วงของเสียงก้อง ความถี่ฟอร์แมนท์จะเปลี่ยนไปตามเวลา ตามการเคลื่อนไหวของอวัยวะที่ประกอบเป็นทางเดินเสียง



รูป 2.3 สเปกตรัมของเสียงสระ "e"

รูป 2.3 แสดงให้เห็นสเปกตรัมของเสียงสระ "e" ในภาษาอังกฤษ ส่วนที่เป็น Spectral Envelope คือผลตอบทางความถี่ของทางเดินเสียง ภายใต้ Spectral Envelope เป็นสเปกตรัมของต้นกำเนิดเสียง ซึ่งยอดเล็กๆ ที่กำกับว่า Fine Structure ในรูป 2.3 คือความถี่ฮาร์มอนิกส์ต่างๆ ของเสียงที่เกิดจากเส้นเสียง

2.2 หลักการสังเคราะห์เสียงพูด

ความพยายามสร้างเสียงเลียนแบบเสียงมนษย์ มีมาตั้งแต่ศตวรรษที่ 18 โดยการเลียนเสียงสระในภาษาอังกฤษด้วยท่ออออร์แกนลมซึ่งสร้างขึ้นเอง จนกระทั่งเข้าสู่หลักการไฟฟ้า โดย Homer Dudley ในปี ค.ศ.1928 [3] หลักการคือการสร้าง Bandpass Filter ต่อขนานกันให้คลุมตลอดย่านความถี่เสียง เพื่อจำลองผลตอบทางความถี่ของทางเดินเสียง ในปัจจุบันการสังเคราะห์เสียงจะสร้างในลักษณะคลื่นไฟฟ้าทั้งหมดซึ่งผลิตจากกระบวนการต่างๆ ทางเชิงเลขทั้งสิ้น

ในการสังเคราะห์เสียงพูด สิ่งจำเป็น คือ การศึกษาวิเคราะห์เสียงพูดจริง เพื่อหาคุณสมบัติหรือตัวแทน (Signal Representations) ตามหลักการต่างๆ ดังนั้นทฤษฎีเกี่ยวกับการสังเคราะห์จึงต้องรวมถึงเรื่องการวิเคราะห์เข้าไปด้วย หลักการวิเคราะห์และสังเคราะห์เสียงพูดที่สำคัญมีอยู่ด้วยกัน 5 วิธี [4] คือ

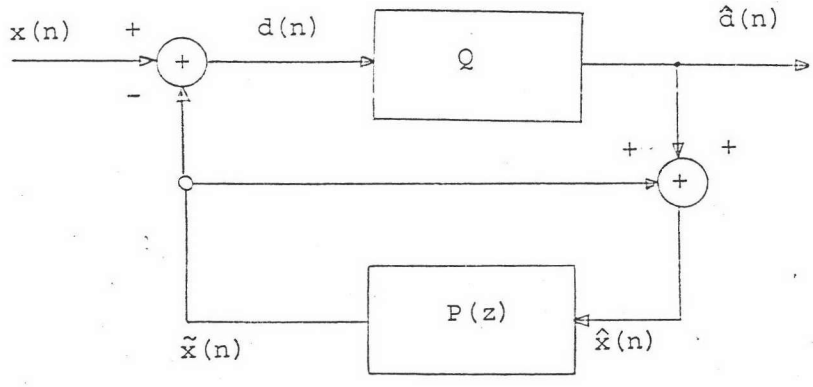
2.2.1 การนำสัญญาณเข้ารหัสเชิงเลข (Digital Waveform Coding)

หลักการคือ แปลงสัญญาณพูดให้เป็นข้อมูลเชิงเลขโดยตรง แบ่งได้เป็น 2 วิธี คือ

1. พัลส์โคดโมดูเลชัน (Pulse Code Modulation) หรือพีซีเอม (PCM) คือ การสุ่มตัวอย่างสัญญาณอนาลอคในช่วงเวลาที่คงที่ ย่านความถี่ (Bandwidth) ที่ได้จะเท่ากับครึ่งหนึ่งของความถี่ในการสุ่ม (Sampling Frequency) จากนั้นจะนำสัญญาณที่สุ่มมาเปลี่ยนเป็นข้อมูลเลขฐานสองด้วยกระบวนการควอนไทเซชัน (Quantization) ความถี่ในการสุ่มสัญญาณอยู่ระหว่าง 6kHz ถึง 20kHz ค่า Signal to Noise Ratio (SNR) จะขึ้นอยู่กับจำนวนบิท และวิธีการควอนไทซ์ (Quantize) ตัวอย่างเช่น การควอนไทซ์แบบเชิงเส้น ความละเอียด 11 บิท จะได้ค่า SNR เท่ากับ 60 db และการเปลี่ยนแปลงความละเอียด 1 บิท จะทำให้ค่า SNR เปลี่ยนไป 6 db การควอนไทซ์ที่ได้ผลดีที่สุดคือ ระบบ Logarithm ซึ่งสามารถลดจำนวนบิท ลงเหลือ 7-8 บิท ที่คุณภาพใช้งานได้ จำนวนข้อมูลของเสียงจะเท่ากับจำนวนบิท ของการควอนไทซ์คูณกับความถี่ในการสุ่ม ที่ใช้งานทั่วไปอยู่ระหว่าง 48,000 bits/s ถึง 220,000 bits/s วิธีการนี้เป็นวิธีที่ง่ายที่สุด แต่ข้อมูลเสียงที่ได้มีมากไม่เหมาะสำหรับการสังเคราะห์เสียง

2. ดิฟเฟอร์เรนเชียลควอนไทเซชัน (Differential Quantization) วิธีการนี้เกิดจากความเป็นจริงที่ว่า ในการสุ่มสัญญาณเสียงพูดตาม Nyquist Rate ค่าสหสัมพันธ์ (Correlation) ของสัญญาณที่สุ่มต่อกันยังคงมีค่าสูง นั่นคือเรามีโอกาสทำนายสัญญาณให้ได้ผลดีโดยไม่ยากนัก เพราะความสามารถในการทำนายได้ของสัญญาณ

จากรูป 2.4 $\tilde{x}(n)$ คือค่าทำนายสัญญาณสุ่ม $x(n) = x(nT)$ ในที่นี้ T คือ เวลาระหว่างการสุ่ม (Sampling Period) ถ้าการทำนายของตัวทำนาย P(z) มีความถูกต้องพอสมควรค่าความแปรปรวน(Variance) ของค่าที่แตกต่างระหว่างสัญญาณจริงกับสัญญาณทำนาย คือ $d(n) = x(n) - \tilde{x}(n)$ จะมีค่าน้อยกว่าค่าวาเรียนซ์ ของสัญญาณ x(n) ดังนั้น วาเรียนซ์ของค่าผิดพลาดจากการควอนไทซ์ของสัญญาณ d(n) ย่อมมีค่าน้อยกว่าวาเรียนซ์ของค่าผิดพลาดจากการควอนไทซ์ของตัวสัญญาณ x(n)



รูป 2.4 แผนภาพหลักการทำงานของดิฟเฟอร์เรนเชียลควอนไทเซชัน

ให้ ค่าผิดพลาดจากการควอนไทซ์ ของ d(n) เท่ากับ e(n) หรือ

$$e(n) = d(n) - \hat{d}(n) \qquad (2.1)$$

เนื่องจาก $d(n) = x(n) - \tilde{x}(n)$ และ $\hat{d}(n) = \hat{x}(n) - \tilde{x}(n)$

ดังนั้น $$e(n) = x(n) - \hat{x}(n) \qquad (2.2)$$

นั่นคือ ค่าผิดพลาดจากการควอนไทซ์ของสัญญาณที่ได้ (x(n)) จะมีค่าเท่ากับค่าผิดพลาดจากการควอนไทซ์ของค่าที่แตกต่างระหว่างสัญญาณจริงกับสัญญาณทำนาย ผลก็คือ ในจำนวนบิทเท่ากันดิฟเฟอร์เรนเชียลควอนไทเซชันจะได้ SNR สูงกว่าพีซีเอม ตัวอย่างการใช้งานด้วย หลักการนี้ก็มี 1) การทำควอนไทเซชันแบบเชิงเส้นกับสัญญาณ d(n) เรียกว่า Linear Delta Modulation (LDM) 2) ดิฟเฟอร์เรนเชียลพีซีเอม (DPCM) ซึ่งใช้ Quantization เพียง 1 บิท ข้อเสียของ

วิธีนี้ คือ การเกิด Granular Distortion ในสัญญาณความถี่ต่ำๆ และจะเกิด Slope Overload ในสัญญาณความถี่สูงที่มี Amplitude สูง

วิธีการแก้ปัญหาของ DPCM คือเพิ่มข้อมูลเกี่ยวกับ Step Size เข้าไป คือเป็นดิฟเฟอร์เรนเชียลควอนไทเซชันที่เปลี่ยนแปลง Step Size ได้ทำให้คุณภาพเสียงดีขึ้น วิธีการดังกล่าวเรียกว่า Adaptive DPCM (ADPCM) ซึ่งจำนวนข้อมูลลดลงมากพอสมควร เมื่อเทียบกับพีซีเอม ที่ใช้งานทั่วไปมีอัตราข้อมูลอยู่ประมาณ 24k bits/s

2.2.2 การวิเคราะห์ในโดเมนเวลา (Time Domain Analysis)

หลักการนี้เป็นการวิเคราะห์หาคุณสมบัติของเสียงพูดใน โดเมนเวลา เนื่องจากเสียงพูดมีการเปลี่ยนแปลงคุณสมบัติตามเวลา ในการวิเคราะห์จึงต้องแบ่งเสียงพูดออกเป็นช่วงๆ (Frame, Segment) โดยมีช่วงเวลาอยู่ระหว่าง 10-30 ms ในช่วงเวลาดังกล่าวเสียงจะมีการเปลี่ยนแปลงคุณสมบัติน้อยมาก ดังนั้นในแต่ละเฟรมจึงสมมุติให้คุณสมบัติของเสียงไม่เปลี่ยนแปลงตามเวลา ทำให้การวิเคราะห์ทำได้ง่ายยิ่งขึ้น การแบ่งเสียงพูดเป็นช่วงๆ นี้ยังใช้กับการวิเคราะห์เสียงพูดด้วยหลักการอื่นๆ อีกมาก ดังจะกล่าวต่อไป

ตัวอย่างการวิเคราะห์เสียงพูดด้วย การวิเคราะห์ในโดเมนเวลา มีดังนี้

1. การวัดพลังงานของสัญญาณ (Energy Measurement) พลังงานของเสียงใน 1 เฟรม ซึ่งมีจำนวน N แซมเปิล มีค่าเท่ากับ

$$E(n) = \sum_{m=0}^{N-1} \{w(m)x(n-m)\}^2 \qquad (2.3)$$

ในที่นี้ w(m) คือ Window Function

สมการนี้สามารถนำไปใช้ได้อย่างกว้างขวาง มีความเกี่ยวข้องโดยตรงกับความดังของเสียง และค่าพลังงานก็เป็นพารามิเตอร์ที่สำคัญของการสังเคราะห์เสียงพูดด้วยวิธีต่างๆ รวมทั้งวิธีการแอลพีซี

2. Short-Time Autocorrelation Analysis เป็นวิธีการสำคัญในการวิเคราะห์ความเป็นคาบของสัญญาณเสียงพูด สมการ Short-Time Autocorrelation คือ

$$\phi_{\ell}(m) = \frac{1}{N}\sum_{n=0}^{N-m} x_{\ell}(n)x_{\ell}(n+m), \quad 0 \leq m \leq M_0 - 1 \qquad (2.4)$$

ในที่นี้ ℓ เป็นตัวบอกถึงเฟรมใดๆ N คือ จำนวนแซมเปิล ใน 1 เฟรม

$$x_{\ell}(n) = x(n+\ell), \quad 0 \leq n \leq N-1 \qquad (2.5)$$

ถ้าสัญญาณ $x(n)$ เป็นสัญญาณที่มีคาบเท่ากับ P นั่นคือ $x(n+p) = x(n)$ ที่ n ใดๆ ถ้าให้ $M_0 >> P$ เราจะได้

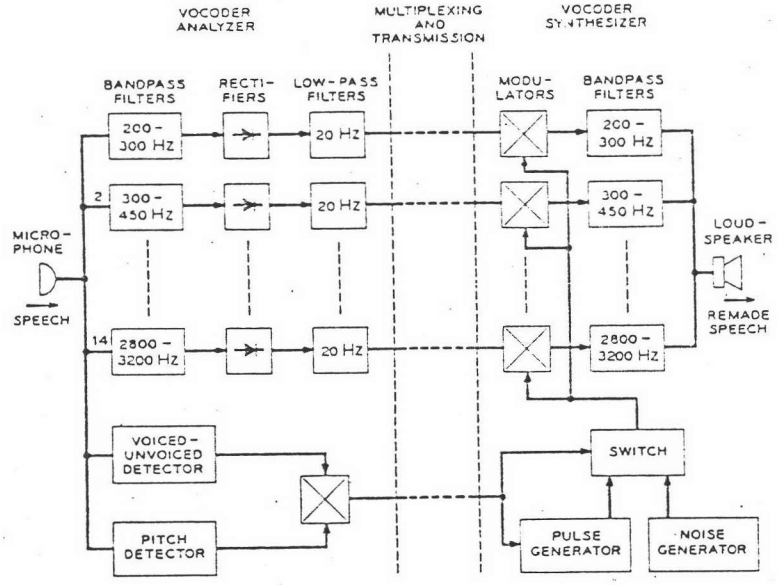
$$\phi(m) \simeq \phi(m+P) \qquad (2.6)$$

นั่นคือ ฟังก์ชัน อัตสหสัมพันธ์ ที่ได้จะเป็นคาบด้วย นอกจากนั้นการลดลงของค่าอัตสหสัมพันธ์ จากจุดยอดที่ m=o กับค่าที่ m เพิ่มขึ้นเรื่อยๆ จะบ่งบอกถึงความสามารถในการทำนายของสัญญาณ นอกจากนั้นวิธีการนี้ยังถูกนำไปใช้ในการสังเคราะห์เสียงที่เรียกว่า Autocorrelation Vocoder [3]

2.2.3 การวิเคราะห์และสังเคราะห์ทางสเปกตรัม (SPECTRAL ANALYSIS AND SYNTHESIS) [3]

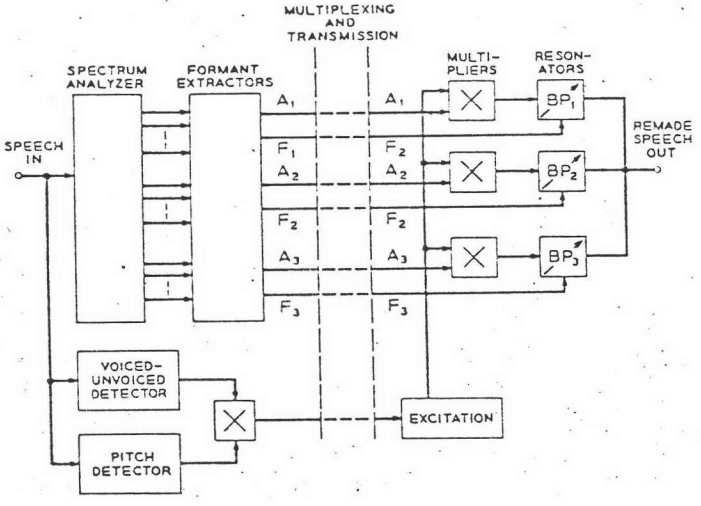
หลักการนี้อาศัยการวิเคราะห์ส่วนประกอบทางความถี่ (Spectrum) ของคลื่นเสียง มี 2 วิธีที่สำคัญ คือ

1. Channel Vocoder เป็นวิธีการแรกในประวัติการสังเคราะห์เสียงพูดโดยการผลิตรูปคลื่นในรูปของสัญญาณไฟฟ้า Channel Vocoder ประกอบด้วย 2 ภาค คือ ภาควิเคราะห์ (Analyzer) และภาคสังเคราะห์ (Synthesizer) ดังในรูป 2.5 ภาควิเคราะห์ประกอบด้วยวงจร Bandpass Filter คลุมย่านความถี่ระหว่าง 200-3200 Hz จำนวน 14 ช่อง ในแต่ละช่องจะมีวงจร Rectifier เพื่อดึงเอาเฉพาะ Envelope ของสัญญาณที่ผ่าน Bandpass Filter



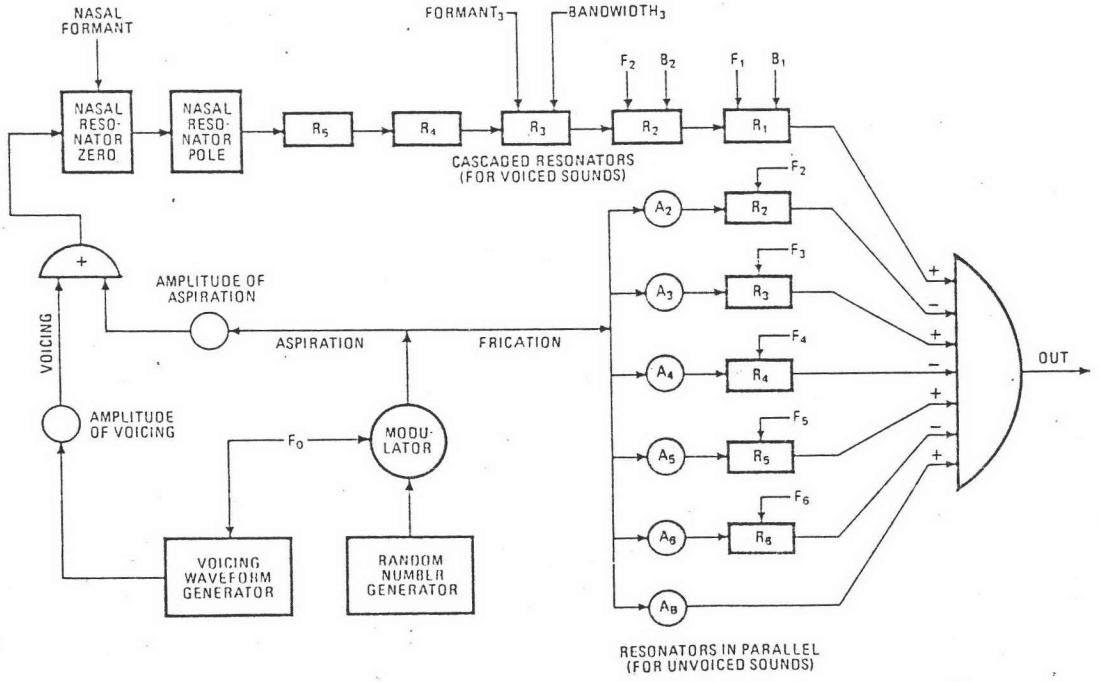
รูป 2.5 ส่วนประกอบของ Channel Vocoder

ตัวตรวจสอบเสียงก้อง (Voiced-Unvoiced Detector) จะทำหน้าที่ตัดสินว่าเสียงพูดขณะนั้นเป็นเสียงก้อง หรือเสียงไม่ก้อง ตัวตรวจสอบดาบ (Pitch Detector) จะทำหน้าที่หาความถี่หลักมูล (Fundamental Frequency)ของเสียงขณะนั้น ภาคสังเคราะห์ประกอบด้วย Modulator ทำหน้าที่คูณหรือปรับขนาดของสัญญาณจากตัวกำเนิดพัลส์ หรือ สัญญาณรบกวน เพื่อป้อนเข้าสู่ Bandpass Filter ในแต่ละช่อง ตัวคูณ คือค่าที่ส่งมาจากภาควิเคราะห์ เสียงพูดจะได้จากสัญญาณที่ผ่าน Bandpass Filter ทุกช่องนำมารวมกัน วิธีการนี้ถูกนำไปสร้างเป็นเครื่องวิเคราะห์-สังเคราะห์เสียงพูดสำเร็จเป็นครั้งแรกเมื่อ ปี 1939 โดย Homer W. Dudley ไม่นานมานี้ได้มีการประยุกต์วิธีการโดยหันมาใช้ Short-time Fast Fourier Transform ทำการคำนวนด้วยเครื่องคอมพิวเตอร์ [5] ซึ่งสามารถใช้งานได้ทั้งวิเคราะห์และสังเคราะห์ แต่มีข้อเสียคือต้องมีการคำนวณเป็นจำนวนมากและข้อมูลเสียงยังมีจำนวนมาก ปัจจุบันนิยมใช้วิธีการนี้สำหรับวิเคราะห์เสียงพูด เช่น ใช้ในเครื่อง Sono Graph หรือ Spectrum Analyzer



รูป 2.6 ส่วนประกอบของ Formant Vocoder ยุคแรก

2. Formant Synthesis วิธีการนี้คล้ายคลึงกับ Channel Vocoder แต่แทนที่จะใช้ Bandpass Filter ซึ่งความถี่กลาง (Center Frequency) คงที่หลายๆ ช่อง ก็หันไปใช้ Bandpass Filter ที่สามารถเปลี่ยนแปลงความถี่กลางได้ รูป 2.6 แสดงแผนภาพของ Formant Vocoder ยุคแรกๆ ที่ใช้ Filter ต่อขนานกัน 3 ช่อง ภาควิเคราะห์ประกอบด้วย Spectrum Analyzer ทำหน้าที่วิเคราะห์ความถี่ของเสียง Formant Extractors ทำหน้าที่หาความถี่ฟอร์แมนท์ (Formant Frequency) (F1,F2,F3) หรืออีกนัยหนึ่ง คือ ความถี่กำทอนของทางเดินเสียง และขนาดของยอดฟอร์แมนท์ Formant (A1,A2,A3) ภาคสังเคราะห์ประกอบด้วย Multiplier ทำหน้าที่ควบคุมความแรงของสัญญาณเข้าสู่ Filter และ Bandpass Filter ซึ่งสามารถเปลี่ยนแปลงความถี่กลางได้ 3 ชุดต่อแบบขนานกัน วิธีการนี้มีความคล้ายคลึงกับการผลิตเสียงของมนุษย์มาก และเป็นวิธีการสังเคราะห์เสียงวิธีหนึ่งซึ่งแพร่หลายในปัจจุบัน



รูป 2.7 Formant Synthesizer ยุคปัจจุบัน

รูป 2.7 [6] แสดงส่วนประกอบของ Formant Synthesizer สมัยใหม่ซึ่งมีความซับซ้อนมากขึ้น การทำงานอาศัยการคำนวณด้วยเครื่องคอมพิวเตอร์ หรือใช้ Digital Signal Processor จากรูป 2.7 ส่วนบนเป็น Digital Filter ทำหน้าที่เป็นตัวกำทอน(Resonator) R1 ถึง R5 ซึ่งต่อกันแบบ Cascade จำลองทางเดินเสียง และมี Filter จำลองผลตอบของโพรงจมูก ส่วนบนทั้งหมดทำหน้าที่ผลิตเสียงในกรณีที่เสียงเป็น Voiced Sound ส่วนล่างเป็นตัวกำทอนที่ต่อกันอย่างขนาน ทำหน้าที่จำลองผลตอบความถี่ของทางเดินเสียงในกรณีที่เสียงเป็นเสียงเสียดแทรก (Fricative Sound)

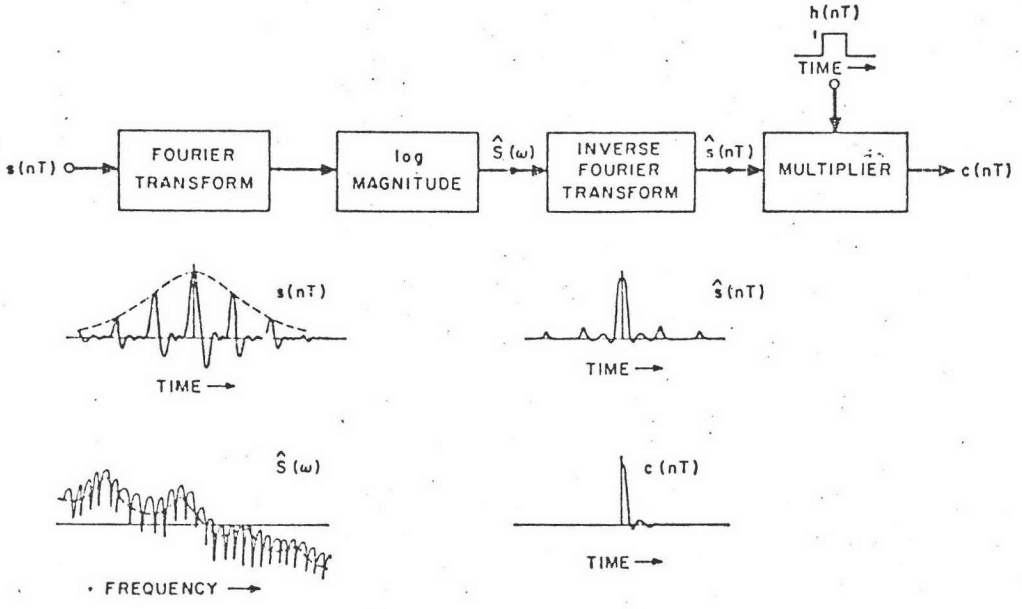
Formant Synthesis เป็นวิธีการลดทอนข้อมูลได้มาก คือข้อมูลเสียงจะอยู่ประมาณ 0.6 kb/s [7] หรือมากกว่า ขึ้นอยู่กับความซับซ้อนของระบบผลิตเสียง ข้อเสียของวิธีการนี้ คือ การวิเคราะห์เพื่อหาความถี่ฟอร์แมนท์ ยุ่งยากมาก นอกจากนี้คุณภาพเสียงจะขึ้นกับการวิเคราะห์และต้นแบบของเสียงที่พูดเป็นอย่างมาก [8]

2.2.4 การกรองแบบโฮโมมอร์ฟิค (HOMOMORPHIC FILTERING) [9][10]

เป็นวิธีการวิเคราะห์สัญญาณเสียงพูดเพื่อแยกผลตอบอิมพัลส์ (Impulse Response) ของทางเดินเสียงกับฟังก์ชันของต้นกำเนิดเสียง (Excitation Function) ออกจากกัน คำว่าโฮโมมอร์ฟิค (Homomorphic) เป็นชื่อเรียกของระบบ Non-Linear ซึ่งมีคุณสมบัติคล้องจองกับ Generalized Superposition [11] จากสมมุติฐานว่าเสียงได้จากการทำ Convolution ระหว่างฟังก์ชันต้นกำเนิดเสียงกับผลตอบอิมพัลส์ของทางเดินเสียง ดังในสมการ (2.7)

$$s(nT) = p(nT) * v(nT) \qquad (2.7)$$

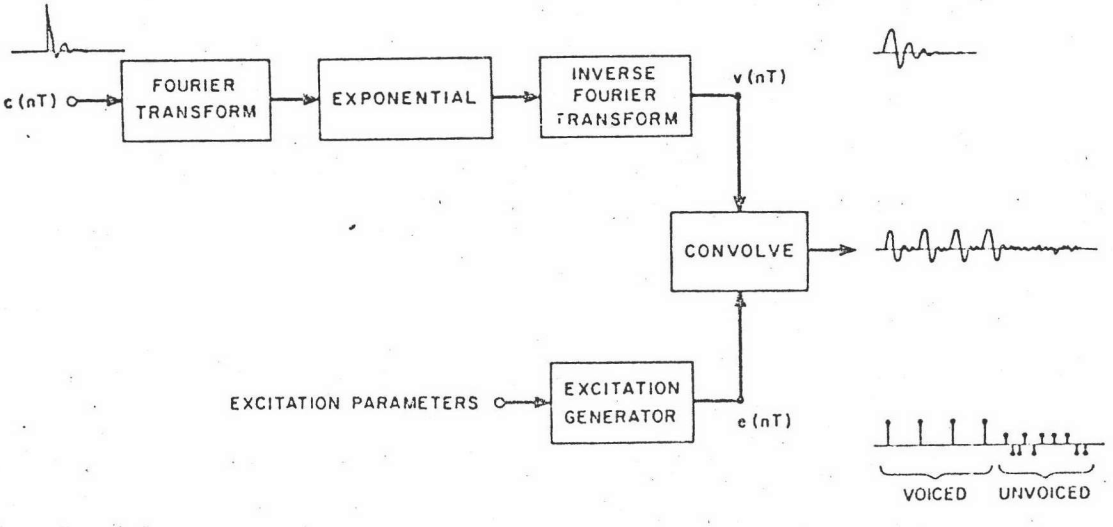
T คือช่วงเวลาระหว่างการสุ่มสัญญาณ s(nT) คือสัญญาณเสียงพูด p(nT) คือ ฟังก์ชันของต้นกำเนิดเสียง และ v(nT) คือ ผลตอบอิมพัลส์ของทางเดินเสียง การวิเคราะห์ด้วยการกรองแบบโฮโมมอร์ฟิค คือการเปลี่ยนจากการทำ Convolution มาเป็นการบวกกันโดยอาศัยการแปลงทางคณิตศาสตร์ ซึ่งทำให้สามารถแยกส่วนประกอบของเสียง หรือประมวลผลสัญญาณใน



รูป 2.8 ส่วนประกอบของภาควิเคราะห์เสียงพูดด้วยวิธีการกรองแบบโฮโมมอร์ฟิค

ลักษณะต่าง ๆ ได้ง่าย รูป 2.8 แสดงถึงส่วนประกอบของภาควิเคราะห์เสียงพูดด้วยการกรองแบบโฮโมมอร์ฟิค จากซ้ายสัญญาณเสียงพูดที่ผ่าน Hamming Window s(nT) จะผ่านการคำนวณ Discrete Fourier Transform (DFT) เพื่อเปลี่ยนจากการทำ Convolution ใน Time Domain มาเป็นการคูณกันใน Frequency Domain นำค่าที่ได้มาเปลี่ยนเป็นค่า Logarithm เพื่อเปลี่ยนจากการคูณกันเป็นการบวกกัน จากรูป 2.8 $\hat{S}(\omega)$ คือ Spectrum ที่ได้ จะเห็นว่ามีส่วนประกอบ 2 ส่วน คือ Spectrum Envelope ซึ่งเกิดจากผลตอบความถี่ของทางเดินเสียง และยอดเล็ก ๆ จำนวนมากซึ่งเป็นส่วนประกอบความถี่ของต้นกำเนิดเสียง จากนั้นสัญญาณ $\hat{S}(\omega)$ จะผ่านการคำนวณ Inverse DFT ได้เป็นสัญญาณ $\hat{s}(nT)$ ซึ่งมีชื่อเรียกเฉพาะว่า Cepstrum จากรูป $\hat{s}(nT)$ ประกอบด้วยยอดสูงที่ n ใกล้ๆ o เป็นข้อมูลสำคัญซึ่งเป็นผลมาจากผลตอบความถี่ของทางเดินเสียง ส่วนยอดเล็ก ๆ ซึ่งจะปรากฏบริเวณเวลาเท่ากับจำนวนเท่าของคาบ (Pitch Period) คือข้อมูลของสัญญาณต้นกำเนิดเสียง จากนั้นนำไปผ่านการกรองด้วย Cepstrum Window h(nT) ซึ่งมีความกว้างเกือบเท่ากับคาบ c(nT) ที่ได้ จะเป็นข้อมูลของทางเดินเสียงล้วน ๆ

ในการสังเคราะห์ก็นำ Cepstrum c(nT) ไปสร้างผลตอบอิมพัลส์ของทางเดินเสียง v(nT) โดยผ่านการคำนวณ Inverse Characteristic System คือ คำนวณ DFT นำค่าที่ได้เปลี่ยนเป็นค่ายกกำลัง (Exponential) แล้วผ่าน Inverse DFT ดังในรูป 2.9



รูป 2.9 ส่วนประกอบของภาคสังเคราะห์เสียงพูดด้วยวิธีการกรองแบบโฮโมมอร์ฟิค

$v(nT)$ ที่ได้คือ ผลตอบอิมพัลส์ของทางเดินเสียง ในการสร้างเสียงกลับคืนมา ให้นำ $v(nT)$ ไปผ่านกระบวนการ Convolution กับสัญญาณ Excitation ซึ่งเป็น Pulse Train ที่มีคาบเท่ากับ Pitch Period ในกรณีที่เสียงเป็นเสียงก้อง หรือ Random Pulse ในกรณีที่เสียงเป็นเสียงไม่ก้อง ส่วน Excitation Parameter เป็นข้อมูลบอกความถี่หลักมูล (หรือ Pitch Period) ขนาดของสัญญาณและตัวเลือกสัญญาณกรณีที่เป็นเสียงก้อง หรือเสียงไม่ก้อง ในการทดลองโดย A.V. Oppenheim [10] จากเสียงที่สุ่มด้วยความถี่ 10 kHz ผ่านการวิเคราะห์ด้วย FFT เฟรมละ 512 จุด คำนวนทุกๆ 20 ms และในการสังเคราะห์ใช้ข้อมูล Cepstrum จำนวน 26 จุด ผ่านการ Quantization ละเอียด 6 บิท จะได้ข้อมูลเสียงประมาณ 7,800 bits/s ที่คุณภาพเสียงใช้งานได้

2.2.5 การเข้ารหัสแบบลิเนียร์พรีดิกทีฟ (LINEAR PREDICTIVE CODING)

การเข้ารหัสแบบลิเนียร์พรีดิกทีฟ หรือ แอลพีซี เป็นวิธีการวิเคราะห์-สังเคราะห์เสียงพูดที่แพร่หลายมากที่สุดวิธีหนึ่งในปัจจุบัน หลักการเบื้องต้นของแอลพีซี คือ สัญญาณเสียงพูดสามารถประมาณด้วยการนำมารวมกันแบบเชิงเส้นระหว่างสัญญาณที่สุ่มมาในอดีต $\tilde{s}(n)$ คือ สัญญาณจากการทำนายโดยวิธีแอลพีซี a_i คือสัมประสิทธิ์การทำนาย p คือ จำนวนสัญญาณที่สุ่มในอดีตที่ใช้ในการทำนาย หรือเรียกว่า ออร์เดอร์ (Order) ของตัวทำนาย $s(n)$ คือสัญญาณจริงที่สุ่มไว้ สมการเบื้องต้นของการทำนายคือ

$$\tilde{s}(n) = -\sum_{i=1}^{p} a_i s(n-i) \qquad (2.8)$$

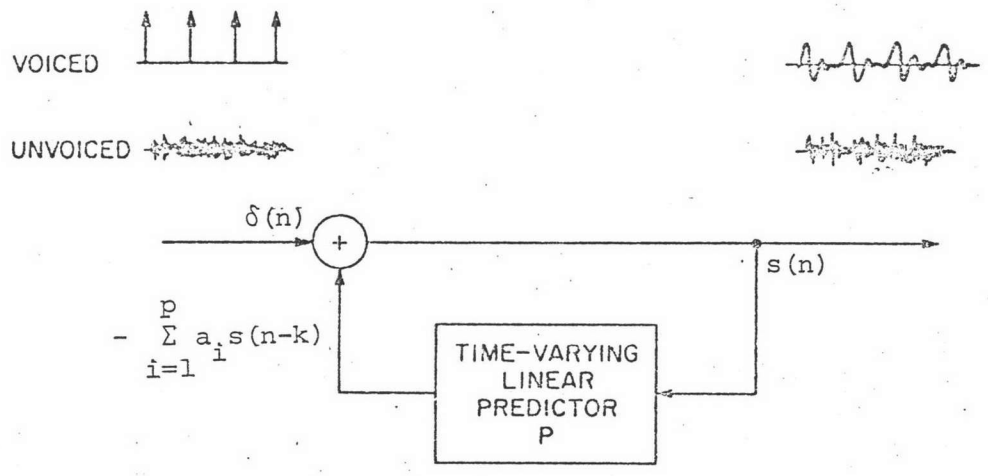
และค่าผิดพลาดของการทำนายเทียบกับสัญญาณจริง คือ

$$\begin{aligned} e(n) &= s(n) - \tilde{s}(n) \\ &= s(n) + \sum_{i=1}^{p} a_i s(n-k) \end{aligned} \qquad (2.9)$$

ในการวิเคราะห์ด้วยวิธีแอลพีซี เราจะแบ่งสัญญาณเสียงออกเป็นเฟรมในช่วงเวลา 10-30 ms ในแต่ละเฟรมเราจะสมมุติให้คุณสมบัติของเสียงไม่เปลี่ยนแปลงตามเวลา ดังนั้นถ้าให้ $\tilde{s}(n)$ ในสมการ (2.8) แทนการทำนายสัญญาณในเฟรมหนึ่งๆ ชุดของ a_i ; $i=1,2,..,p$ ในแต่ละเฟรมจะเป็นค่าคงที่ เมื่อเรากำหนดเงื่อนไขบางอย่างในการทำนาย คือ ให้ค่าผิดพลาดยกกำลังสองรวมในหนึ่งเฟรมมีค่าน้อยที่สุด หรือ α_ℓ มีค่าน้อยที่สุด

$$\alpha_{\ell} = \sum_{n=0}^{N-1} \{s_{\ell}(n) - \tilde{s}_{\ell}(n)\}^2 \qquad (2.10)$$

เมื่อ $s_{\ell}(n) = s(n+\ell)$



รูป 2.10 แบบจำลองการผลิตเสียงพูดด้วยวิธีแอลพีซี

ในที่นี้ ℓ คือตัวกำหนดจุดเริ่มต้นของเฟรม N คือ จำนวนแซมเปิล ใน 1 เฟรม ร่วมกับข้อมูลเสียงที่สุ่มมา เราสามารถสร้างสมการเพื่อหาค่าของ a_i ได้ ค่า a_i นี้จะเป็นตัวแทนของสัญญาณ s(n) และสามารถนำไปใช้ในการสังเคราะห์ s(n) กลับคืนมาได้ รูป 2.10 แสดงแบบจำลองการผลิตเสียงพูดด้วยวิธีแอลพีซี $\delta(n)$คือ สัญญาณ Excitation ซึ่งจะเป็น Pulse Train มีคาบเท่ากับ Pitch Period กรณีที่เสียงเป็นเสียงก้อง หรือเป็น White Noise กรณีที่เสียงเป็นเสียงไม่ก้อง ตัวทำนายมี Transfer Function เท่ากับ

$$P(z) = 1 + \sum_{i=1}^{p} a_i z^{-i} \qquad (2.11)$$

เสียงสัญญาณที่ได้ คือ

$$s(n) = - \sum_{i=1}^{p} a_i s(n-k) + \delta(n) \qquad (2.12)$$

เทียบกับสมการ (2.9) ถ้าการทำนายโดยชุดของ a_i ได้ผลถูกต้องที่สุด สัญญาณ $\delta(n)$ จะเท่ากับค่าผิดพลาด $e(n)$ พอดี และชุดของ a_i จะเป็นตัวแทนของสัญญาณเสียงพูดในเฟรมนั้นๆ ถ้านำวิธีการมาเปรียบเทียบกับการผลิตเสียงพูดของอวัยวะผลิตเสียง LPC Predictor Filter ทั้งหมดตามรูป 2.12 ซึ่งมี Transter Function เท่ากับ

$$H(z) = \frac{A}{1 + \sum_{i=1}^{p} a_i z^{-i}} \qquad (2.13)$$

จะทำหน้าที่จำลองการทำงานของทางเดินเสียงนั่นเอง และสัญญาณ Excitation $\delta(n)$ คือเสียงจากต้นกำเนิดเสียง

การวิเคราะห์หาพารามิเตอร์ของเสียงพูดตามแบบจำลองแอลพีซี มีได้หลายวิธี อาทิเช่น

1. Covariance Method [12]
2. Autocorrelation Formulation [13][14]
3. PARCOR Method by Inverse Filter Formulation [14]
4. PARCOR Method by Maximum Likelihood Formulation [15][16]

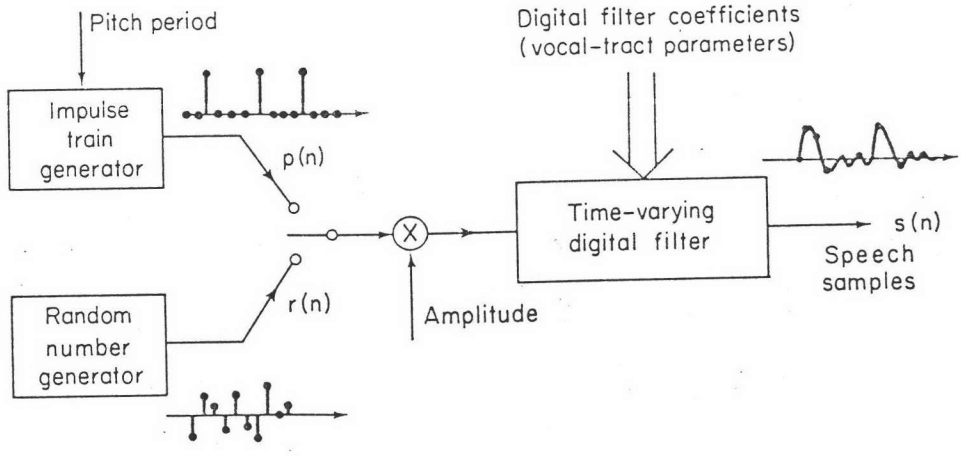
สำหรับวิทยานิพนธ์นี้ ใช้วิธีการของ J.D. Merkel และ A.H. Gray [14] ตามวิธีข้อ 3 ที่กล่าวข้างต้น รายละเอียดของเรื่องนี้อยู่ในหัวข้อ 2.3

ในการสังเคราะห์เสียงด้วยวิธีแอลพีซี ข้อมูลเสียงที่ใช้ในการสังเคราะห์จะอยู่ระหว่าง 2,400 bits/s ถึง 9,600 bits/s โดยข้อมูลแอลพีซีต้องผ่าน Optimal Quantization เพื่อลดจำนวน bit rate [12][16] ตัวอย่างของอุปกรณ์ที่ใช้วิธีแอลพีซีในการสังเคราะห์เสียง คือ เครื่อง Speak and Spell ของบริษัท Taxas Instruments ใช้วิธีแอลพีซีที่มีออร์เดอร์ของตัวทำนายเท่ากับ 10 และใช้ Optimal Quantization ร่วมกับการ Pack ข้อมูลสามารถลดจำนวนข้อมูลเสียงลงเหลือ 600 bits/s ถึง 2,400 bits/s [17]

2.2.6 สรุปหลักการสังเคราะห์เสียงพูด

เราสามารถแบ่งวิธีการผลิตเสียงพูดได้เป็น 2 พวกใหญ่ [7] คือ

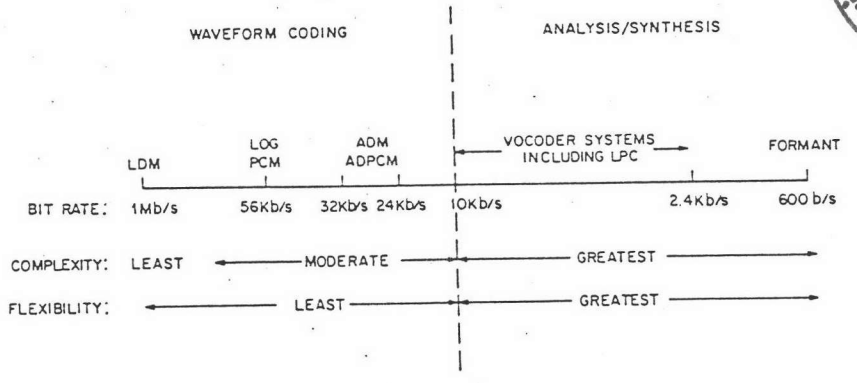
1. Waveform Representation คือ หลักการแปลงรูปคลื่นเสียงให้เป็นข้อมูลเชิงเลขโดยตรง เรียกอีกอย่างหนึ่งว่า การเข้ารหัสรูปคลื่น (Waveform Coding) เช่น LDM, PCM, DPCM และ ADPCM ดังที่ได้กล่าวไว้ในหัวข้อ 2.2.1



รูป 2.11 แบบจำลองของการผลิตเสียงพูด

2. Parametric Representation คือ หลักการที่อาศัยสมมุติฐานเกี่ยวกับการผลิตเสียงพูด หรือมีการสร้างแบบจำลอง (Model) ของการผลิตเสียงพูด และจะต้องมีกระบวนการวิเคราะห์ (Analysis) คู่กับกระบวนการสังเคราะห์ (Synthesis) เสมอ หลักการนี้ได้แก่ Spectral Analysis and Synthesis, Homomorphic Filtering, และ LPC ดังที่กล่าวไว้ในหัวข้อ 2.2.3, 2.2.4 และ 2.2.5 ตามลำดับ จะเห็นว่าทั้งสามวิธีมีการสมมุติแบบจำลองของการผลิตเสียงพูดซึ่งมีลักษณะคล้ายคลึงกันมาก คือ ประกอบด้วยต้นกำเนิดเสียงซึ่งจะผลิต Pulse Train กรณีที่เป็นเสียงก้อง หรือ Random Pulse กรณีที่เป็นเสียงไม่ก้อง และ Digital Filter ทำหน้าที่จำลองการทำงานของทางเดินเสียงดังแสดงในรูป 2.11

ส่วนของต้นกำเนิดเสียงทั้งสามวิธีจะเหมือนกันทุกประการ และข้อมูลที่ใช้ในการผลิตเสียงจะประกอบด้วย 1) ตัวกำหนดเสียงก้องและเสียงไม่ก้อง 2) คาบของเสียง (Pitch Period) กรณีที่เป็นเสียงก้อง 3) ขนาดความแรงของสัญญาณที่จะป้อนเข้าสู่ Digital Filter ส่วนของ Digital Filter ในแต่ละวิธีจะแตกต่างกัน ข้อมูลที่ใช้ผลิตเสียงจึงแตกต่างกันไป เช่น วิธี Formant Synthesis จะเป็นข้อมูล Formant Frequency,Gain, Bandwidth วิธี Homomorphic จะเป็น Ceptrum ที่ใช้ในการสร้าง Impulse Responseของทางเดินเสียง และวิธีแอลพีซีจะเป็นสัมประสิทธิ์ของตัวทำนาย อย่างไรก็ตาม ทั้งหมดที่กล่าวมาทำหน้าที่จำลองการทำงานของทางเดินเสียง



รูป 2.12 การเปรียบเทียบประสิทธิภาพของวิธีผลิตเสียงพูด

ในแง่ของประสิทธิภาพในการผลิตเสียงพูด จุดใหญ่จะขึ้นอยู่กับจำนวนของข้อมูลเสียง ถ้าข้อมูลเสียงยิ่งน้อยประสิทธิภาพยิ่งสูง เพราะใช้หน่วยความจำในการเก็บเสียงพูดน้อยกว่า หรือในการส่งข้อมูลเสียงก็มีอัตราการส่งข้อมูลที่ต่ำกว่า นอกจากนั้นสิ่งที่ต้องพิจารณาอีกก็คือคุณภาพของเสียงพูด อย่างไรก็ตาม คุณภาพเสียงพูดจะเกี่ยวโยงกับจำนวนข้อมูลเสียง ในวิธีการใดๆ ถ้าต้องการคุณภาพเสียง ข้อมูลเสียงจะเพิ่มขึ้น รูป 2.12 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการผลิตเสียงพูดต่างๆ อัตราข้อมูลเสียงมีหน่วยเป็นจำนวนบิทของข้อมูลเสียงพูดที่มีความยาว 1 วินาที จากรูปจะเห็นว่า หลักการ Waveform Coding จะได้ข้อมูลเสียงที่มีจำนวนมาก ความยืดหยุ่นในการใช้งานจำกัด แต่เป็นวิธีที่มีความซับซ้อนน้อย ด้วยหลักการที่อาศัยแบบจำลองการผลิตเสียง ข้อมูลที่ได้จะน้อย มีความยืดหยุ่นในการใช้งานสูง แต่วิธีการผลิตเสียงมีความซับซ้อนมาก

2.3 การหาฟิลเตอร์พารามิเตอร์ของวิธีแอลพีซี

ดังที่กล่าวไว้ในหัวข้อ 2.2.5 การหาฟิลเตอร์พารามิเตอร์สามารถทำได้หลายวิธี ในส่วนของวิทยานิพนธ์นี้ใช้วิธีการที่พัฒนาโดย J.D. Markel และ A.H. Gray jr.[14] วิธีการดังกล่าวมีชื่อว่า อินเวิร์สฟิลเตอร์ (Inverse Filter) การพิสูจน์สมการเพื่อใช้ในการหาคำตอบ อาศัยคุณสมบัติออโธโกนัลลิตี (Orthogonality) ซึ่งผลลัพธ์ที่ได้ตรงกับวิธีพาร์คอร์ (PARCOR) ซึ่งพัฒนาโดย Saito และ Itakura.

2.3.1 แบบจำลองการผลิตเสียงพูดตามวิธีแอลพีซี

หลักการเบื้องต้น คือ สัญญาณปัจจุบันสามารถประมาณได้ด้วยผลรวมเชิงเส้นของสัญญาณในอดีต เราเรียกกระบวนการดังกล่าวว่าการทำนายเชิงเส้น ให้ $s(n)$ คือสัญญาณเสียงสุ่มที่ n ใดๆ $\tilde{s}(n)$ คือ สัญญาณที่ n ที่ได้จากการทำนาย a_i คือสัมประสิทธิ์ของการทำนาย โดยใช้สัญญาณในอดีตทั้งหมด M ตัว หรือการทำนายมีออร์เดอร์ (Order) เท่ากับ M สมการของการทำนายเชิงเส้น คือ

$$\tilde{s}(n) = -\sum_{i=1}^{M} a_i s(n-i) \qquad (2.14)$$

Z-Transform ของสมการ (2.14) คือ

$$\tilde{S}(z) = F(z)S(z) \qquad (2.15)$$

โดยให้ $$F(z) = -\sum_{i=1}^{M} a_i z^{-i} \qquad (2.16)$$

$F(z)$ มีชื่อว่าฟิลเตอร์ทำนาย (Predictor Filter) ถ้าเอาสัญญาณจริงลบด้วยสัญญาณที่ได้จากการทำนาย จะได้ค่าผิดพลาด $e(n)$ เท่ากับ

$$e(n) = s(n) - \tilde{s}(n) \qquad (2.17)$$

$$= s(n) + \sum_{i=1}^{M} a_i s(n-i)$$

$$= \sum_{i=0}^{M} a_i s(n-i) \qquad ; a_0 = 1 \qquad (2.18)$$

Z-Transform ของ (2.18) คือ

$$E(z) = S(z)\{1 - F(z)\} \qquad (2.19)$$

1-F(z) ในสมการ (2.19) มีชื่อว่าอินเวิร์สฟิลเตอร์ (Inverse Filter) ใช้สัญลักษณ์ A(z)

$$A(z) = 1 - F(z) = 1 + \sum_{i=1}^{M} a_i z^{-i} \qquad (2.20)$$

ดังนั้น $$E(z) = S(z)A(z) \qquad (2.21)$$

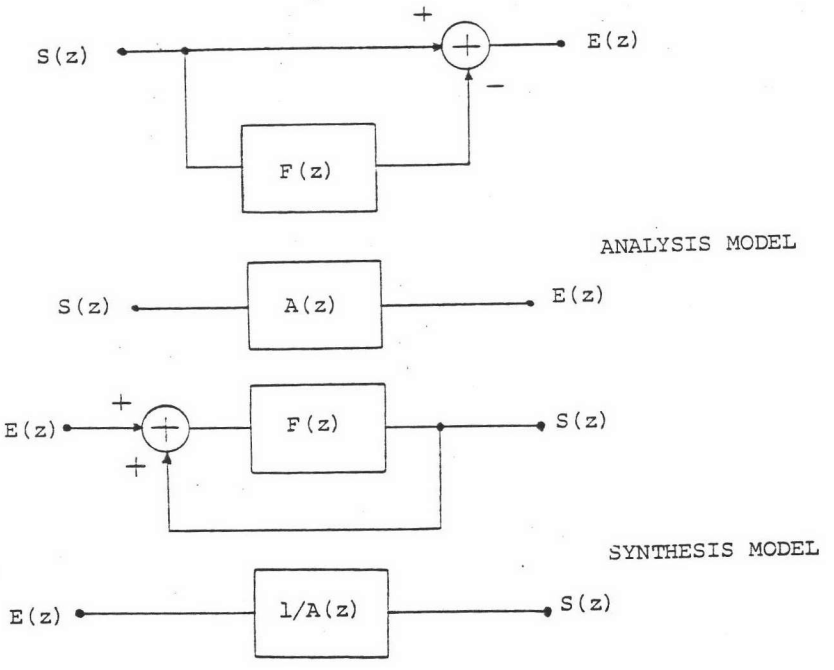
สมการ (2.21) คือส่วนของการวิเคราะห์ เมื่อเราทราบค่าชุดของ a_i ก็สามารถหาค่าผิดพลาดของการทำนายได้ในการสังเคราะห์ เราสามารถหาค่า S(z) ได้ถ้าทราบชุดของ a_i และ E(z) จากความสัมพันธ์

$$S(z) = \frac{E(z)}{1 - F(z)}$$

$$= \frac{E(z)}{A(z)} \qquad (2.22)$$

หรือ $$s(n) = e(n) - \sum_{i=1}^{M} a_i s(n-1) \qquad (2.23)$$

การสร้างสัญญาณตามสมการ (2.22) มีชื่อเรียกเฉพาะว่า ALL Pole Model [13] เพราะฟิลเตอร์ 1/A(z) มีแต่ Pole ล้วนๆ ซึ่งทำให้การคำนวณง่ายขึ้น รูป 2.13 แสดงแบบจำลองของการวิเคราะห์และสังเคราะห์ตามสมการ (2.21) และ (2.22) ในการวิเคราะห์เราจะแบ่งเสียงพูดเป็นส่วนย่อยๆ เรียกว่าเฟรม (frame) แต่ละเฟรมกินเวลาประมาณ 10-20 ms ซึ่งช่วงเวลาดังกล่าวเราสมมุติให้คุณสมบัติของเสียงไม่เปลี่ยนแปลงตามเวลา a_i จะเป็นค่าคงที่



รูป 2.13 แบบจำลองการวิเคราะห์และสังเคราะห์เสียงพูดด้วยวิธีแอลพีซี

การหาพารามิเตอร์เพื่อใช้ในการสังเคราะห์เสียง เรากำหนดเงื่อนไขค่าผิดพลาดยกกำลังสองรวมในหนึ่งเฟรมมีค่าน้อยที่สุด สมมุติให้เฟรมที่ทำการวิเคราะห์อยู่ระหว่าง $n = n_0$ ถึง $n = n_1$ ค่าผิดพลาดยกกำลังสองรวม มีค่าเท่ากับ

$$\alpha = \sum_{n=n_0}^{n_1} e^2(n) \qquad (2.24)$$

แทนค่า e(n) จากสมการ (2.19) ได้

$$\alpha = \sum_{n=n_0}^{n_1} \{ \sum_{i=0}^{M} a_i s(n-i) \}^2$$

$$= \sum_{n=n_0}^{n_1} \sum_{i=0}^{M} \sum_{j=0}^{M} a_i s(n-i) s(n-j) a_j \qquad (2.25)$$

ถ้ากำหนดให้

$$C(i,j) = \sum_{n=n_0}^{n_1} s(n-i)s(n-j) \qquad (2.26)$$

สมการ (2.25) ได้เป็น

$$\alpha = \sum_{i=0}^{M} \sum_{j=0}^{M} a_i C(i,j) a_j \qquad (2.27)$$

กำหนดเงื่อนไขค่าผิดพลาดยกกำลังสองรวมในเฟรมหนึ่งๆ มีค่าน้อยที่สุด ค่าอนุพันธ์เทียบกับ a_k จะเท่ากับ 0

$$\frac{\partial \alpha}{\partial a_k} = 0 = 2 \sum_{i=0}^{M} C(i,k) a_k \ , \quad a_0 = 1 \ ; \quad k = 1,2,\ldots,M \qquad (2.28)$$

ซึ่งจะได้ชุดของสมการเชิงเส้นจำนวน M สมการ คือ

$$\sum_{i=1}^{M} a_i C(i,k) = - C(0,k) \qquad ; \ k = 1,2,\ldots,M \qquad (2.29)$$

ค่าของ a_i ที่ได้จากการแก้สมการ (2.29) สามารถนำไปหาค่าผิดพลาด e(n) ได้ด้วยสมการ (2.18) ซึ่ง e(n) และชุดของ a_i ที่ได้สามารถนำไปผลิตเสียงพูดกลับคืนมาได้ด้วยสมการ (2.23) ในทางปฏิบัติเราจะสมมุติให้ e(n) เป็น Pulse Train หรือ Noise ตามแบบจำลองการผลิตเสียงพูดดังที่กล่าวมา จึงไม่จำเป็นต้องคำนวนหาค่า e(n) แต่จะคำนวณเกี่ยวกับคาบของเสียงแทน

การแก้สมการเพื่อหาค่า a_i แบ่งได้เป็น 3 วิธีที่สำคัญ คือ

1) วิธีโควาเรียนซ์ (Covariance Method)
2) วิธีออโตโครีเลชัน (Autocorrelation Method)
3) วิธีพาร์คอร์ (PARCOR)

2.3.2 การหาคำตอบด้วยวิธีโควาเรียนซ์

จากสมการ (2.24) ถ้ากำหนดเงื่อนไขค่าผิดพลาดยกกำลังสองรวมในช่วง $n_0 = M$ ถึง $n_1 = N-1$ จากสมการ (2.26) เราต้องทราบค่าของสัญญาณจริงทั้งหมด N ตัว คือ {s(n)}={s(0), s(1),s(2),.....,s(N-1)} จึงจะคำนวณค่า C(i,j) ; i=0,1,2,..,M ; j=1,2,..,M

ชุดของสมการจะได้เป็น

$$\sum_{i=1}^{M} a_i C(i,j) = -C(0,j) \qquad ;j = 1,2,\ldots,M \qquad (2.30)$$

โดย

$$C(i,j) = \sum_{n=M}^{N-1} s(n-i)s(n-j) \qquad (2.31)$$

$$;i = 0,1,2,\ldots,M; \; j = 1,2,\ldots,M$$

หรือเขียนอยู่ในรูปเมตริกซ์ ได้เป็น

$$\begin{bmatrix} C(1,1) & C(1,2) & \cdots & C(1,M) \\ C(2,1) & C(2,2) & \cdots & C(2,M) \\ \vdots & \vdots & & \vdots \\ C(M,1) & C(M,2) & \cdots & C(M,M) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_M \end{bmatrix} = - \begin{bmatrix} C(0,1) \\ C(0,2) \\ \vdots \\ C(0,M) \end{bmatrix} \qquad (2.32)$$

ซึ่งค่า $C(i,j)$ มีความคล้ายคลึงกับฟังค์ชันโควาเรียนซ์ของสัญญาณ $s(n)$ จึงเรียกเมตริกซ์ของ $C(i,j)$ ว่าโควาเรียนซ์เมตริกซ์ (Covariance Matrix) จากสมการ (2.31) จะพบว่า $C(i,j) = C(j,i)$ ดังนั้นเมตริกซ์จึงมีคุณสมบัติสมมาตรและค่าในเส้น ทแยง มีความสัมพันธ์ คือ

$$C(i+1,j+1) = C(i,j) + s(-i-1)s(-j-1) - s(n-1-i)s(n-1-j) \qquad (2.33)$$

การหาคำตอบ a_i โดยสมการ (2.32) เรียกว่าวิธีโควาเรียนซ์ (Covariance Method) ซึ่งสามารถทำได้อย่างมีประสิทธิภาพด้วย Cholesky Decomposition Method. [7,pp.407-411]

2.3.3 การหาคำตอบด้วยวิธีอัตสหสัมพันธ์

สมการ (2.26) ถ้ากำหนดให้ $n = -\infty$ และ $n = \infty$ และให้ $s(n)=0$ เมื่อ $n<0$ และ $n>=N$ สมการ ของ $C(i,j)$ จะกลายเป็น

$$C(i,j) = \sum_{n=-\infty}^{\infty} s(n-i)s(n-j)$$

$$= \sum_{n=-\infty}^{\infty} s(n)s(n+|i-j|)$$

$$= \sum_{n=0}^{N-1-|i-j|} s(n)s(n+|i-j|)$$

$$= r(|i-j|) \quad ; i = 0,1,2,\ldots,M \qquad (2.34)$$

$$j = 1,2,\ldots,M$$

ซึ่งกรณีนี้ $C(i,j)$ จะตรงกับฟังก์ชัน Short-time Autocorrelation $r(|i-j|)$ กำหนดเงื่อนไขค่าผิดพลาดยกกำลังสองรวมน้อยที่สุดในช่วง $n = -\infty$ ถึง $n = \infty$ หรือให้ผลจริงในช่วง $n=0$ ถึง $n=N+M-1$ (เนื่องจาก $s(n)$ มีค่าเท่ากับศูนย์นอกช่วงดังกล่าว) จะได้ชุดของสมการเชิงเส้นจำนวน M สมการ คือ

$$\sum_{i=1}^{M} a_i r(|i-j|) = -r(j) \quad ; j = 1,2,\ldots,M \qquad (2.35)$$

เมื่อเขียนให้อยู่ในรูปเมตริกซ์จะได้

$$\begin{bmatrix} r(0) & r(1) & r(3) & \cdots & r(M-1) \\ r(1) & r(0) & r(1) & \cdots & r(M-2) \\ r(2) & r(1) & r(0) & \cdots & r(M-3) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ r(M-1) & r(M-2) & r(M-3) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ \vdots \\ a_M \end{bmatrix} = - \begin{bmatrix} r(1) \\ r(2) \\ r(3) \\ \vdots \\ \vdots \\ r(M) \end{bmatrix} \qquad (2.36)$$

ในที่นี้ [r(j)] คือ เมตริกซ์ของค่า อัตสหสัมพันธ์ มีลักษณะเป็น Toeplitz Matrix คือมีคุณสมบัติสมมาตรและ Element ในเส้นทแยง มีค่าเท่ากัน วิธีหาคำตอบอย่างมีประสิทธิภาพสามารถทำได้ด้วยวิธีของ Durbin [7] ข้อดีของวิธี อัตสหสัมพันธ์ คือ การคำนวณเพื่อแก้สมการมีน้อยกว่าวิธีโควาเรียนซ์ และมีความแน่นอนด้านเสถียรภาพ [7,p.148] แต่สัญญาณจริงที่ใช้คำนวณ r(j) ต้องผ่าน Smoothing Window เพื่อลดความผิดเพี้ยนทางองค์ประกอบความถี่ นอกจากนั้นการใช้ Rectangular Window ยังทำให้ค่าผิดพลาด e(n) มีค่ามากในช่วงท้ายของแต่ละเฟรม สัญญาณ s(n) ที่ใช้ในการคำนวณ r(j) จึงมีค่าเท่ากับ $s(n) = s_i(n)w(n)$ ซึ่ง $s_i(n)$ คือ สัญญาณสุ่มของเสียง w(n) คือฟังก์ชันของ window โดยทั่วไปใช้ Hammimg Window ซึ่งมีฟังก์ชันเท่ากับ

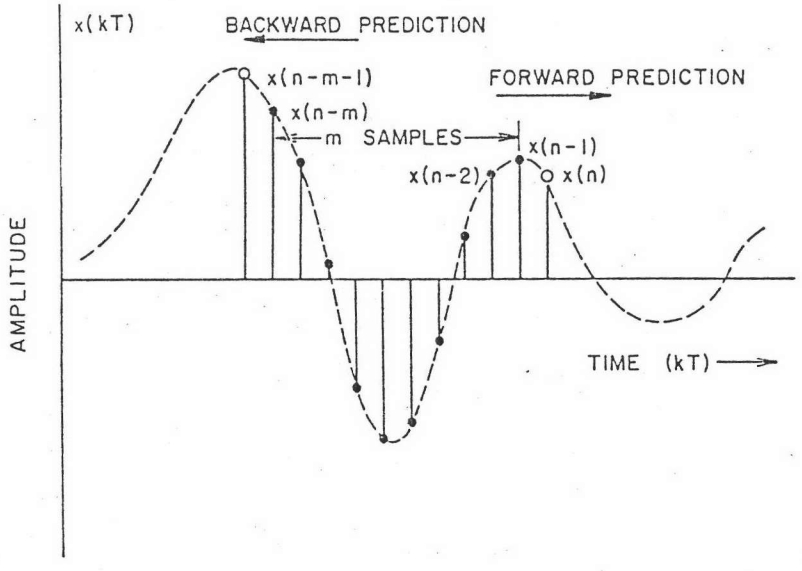
$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \qquad 0 \leq n \leq N-1 \quad (2.37)$$

2.3.4 วิธีพาร์คอร์ (PARCOR)

หลักการเบื้องต้นของวิธีพาร์คอร์ (PARCOR) ซึ่งย่อมาจาก สหสัมพันธ์บางส่วน (Partial-Correlation) คือ การทำนายสัญญาณด้วยตัวทำนายเชิงเส้น (Linear Predictor) ที่มีออร์เดอร์เท่ากับ m ทั้งสองทาง คือทำนายไปข้างหน้าและทำนายย้อนหลัง จากรูป 2.14 แสดงชุดของสัญญาณสุ่ม x(kT) จำนวน m+2 ค่า การทำนายไปข้างหน้าหมายถึงการทำนายค่า x(n) จากค่า x(n-1),(n-2),...,x(n-m) ค่าผิดพลาดของการทำนายที่ออร์เดอร์ m แทนด้วย $x_m^+(n)$ เท่ากับ

$$x_m^+(n) = x(n) - \{ -\sum_{i=1}^{m} a_{mi} x(n-i) \}$$

$$= \sum_{i=0}^{m} a_{mi} x(n-i) \qquad a_{m0} = 1 \qquad (2.38)$$



รูป 2.14 ชุดของสัญญาณสุ่มและการทำนายทั้งสองทาง

การทำนายย้อนหลัง คือ การทำนายค่า x(n-m-1) จากค่า x(n-m),x(n-m+1),...,x(n-1) ค่าผิดพลาดที่ออร์เดอร์ m แทนด้วย $x_m^-(n)$ เท่ากับ

$$x_m^-(n) = x(n-m-1) - \{ -\sum_{i=1}^{m} b_{mi}x(n-1)\}$$

$$= \sum_{i=1}^{m+1} b_{mi}x(n-i) \quad ; \quad b_{m,m+1} = 1 \qquad (2.39)$$

จากนั้นใช้เงื่อนไขค่าผิดพลาดยกกำลังสองรวมน้อยที่สุดกับการทำนายทั้งสองทิศทางพร้อมกัน ให้ α_m และ β_m คือ ค่าผิดพลาดยกกำลังสองรวมของการทำนายไปข้างหน้าและการทำนายไปข้างหลังตามลำดับ

$$\alpha_m = \sum_{n=n_0}^{n_1} \{x_m^+(n)\}^2 \qquad (2.40a)$$

$$\beta_m = \sum_{n=n_0}^{n_1} \{x_m^-(n)\}^2 \qquad (2.40b)$$

โดย m เป็นตัวแปรมีค่าเท่ากับ 1,2,...,M เมื่อ M เป็นจำนวนออร์เดอร์ทั้งหมดของตัวทำนาย ใช้เงื่อนไขค่าผิดพลาดยกกำลังสองรวมน้อยที่สุดดังนั้น

$$\frac{\partial \alpha_m}{\partial a_{mi}} = 0 \quad , \quad \frac{\partial \beta_m}{\partial b_{mi}} = 0$$

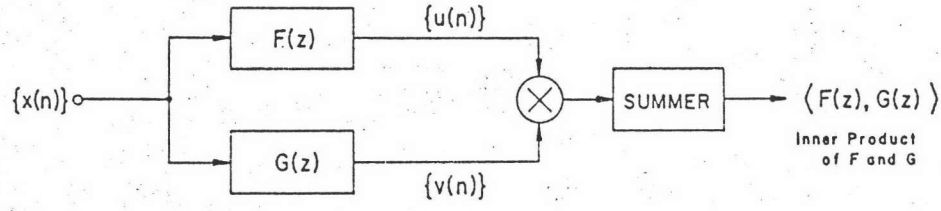
$$m = 1,2,\ldots,m \; ; \; i = 1,2,\ldots,m$$

เงื่อนไขนี้ยังไม่สามารถนำไปใช้ตั้งสมการเพื่อหาค่าของสัมประสิทธิ์ตัวทำนายได้โดยตรง จำเป็นต้องมีเงื่อนไขเพิ่มเติม กรรมวิธีดั้งเดิมในการหาคำตอบซึ่งพัฒนาโดย F.Itakura จะมีคำตอบอยู่ในรูปของ สัมประสิทธิ์พาร์คอร์ (PARCOR Coefficient) แทนด้วยสัญญลักษณ์ k_m มีรายละเอียดอยู่ใน [16] ส่วนวิธีของ J.D Markel และ A.H.Gray ก็เป็นการหาคำตอบอยู่ในรูปของ k_m เช่นเดียวกัน แต่ด้วยวิธีที่แตกต่างกัน Markel และ Gray อาศัยหลักการอินเนอร์โพรดักท์และคุณสมบัติออโธโกนัลของตัวทำนาย นำไปใช้ในการหาค่า k_m

2.3.5 หลักการอินเนอร์โพรดักท์และคุณสมบัติออโธโกนัลของตัวทำนาย

อาศัยนิยามและทฤษฎีเกี่ยวกับอินเนอร์โพรดักท์ในวิชาพีชคณิตเชิงเส้นประยุกต์มาใช้กับสัญญาณเสียง สมมุติให้ {x(n)} คือ ชุดของสัญญาณสุ่มที่เป็นอินพุทร่วมของฟิลเตอร์ F(z) และ G(z) เอาท์พุทของฟิลเตอร์ F(z) คือ {u(n)} และ เอาท์พุทของฟิลเตอร์ G(z) คือ {v(n)} ดังในรูป 2.15 อินเนอร์โพรดักท์ของฟิลเตอร์ F(z) กับ G(z) ในช่วง $n = n_0$ ถึง $n = n_1$ แทนด้วยสัญญลักษณ์ <F(z),G(z)> จะเท่ากับผลรวมของผลคูณระหว่าง u(n) กับ v(n)

$$\langle F(z),G(z)\rangle = \sum_{n=n_0}^{n_1} u(n)v(n) \tag{2.41}$$



รูป 2.15 อินเนอร์โพรดักท์ของฟิลเตอร์ F(z) กับ G(z)

ถ้าให้ F(z) และ G(z) เป็นรีลฟิลเตอร์ (Real Filter) อยู่รูป

$$F(z) = \sum_{i=0}^{\infty} f_i z^{-i} \quad ; \quad G(z) = \sum_{i=0}^{\infty} g_i z^{-i}$$

อินเนอร์โพรดักท์ของ F(z) กับ G(z) จะเท่ากับ

$$\langle F(z),G(z)\rangle = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} f_i \{\sum_{n=n_0}^{n_1} x(n-i)x(n-j)\} g_i \tag{2.42}$$

ถ้า f และ g เท่ากับ 1 ตลอดจะได้

$$\langle z^{-i}, z^{-j} \rangle = \sum_{n=n_0}^{n_1} x(n-i)x(n-j) \qquad (2.43)$$

ด้วยวิธีโควาเรียนซ์ กำหนด $n_0 = M$ และ $n_1 = n-1$ ได้

$$\langle z^{-i}, z^{-j} \rangle = C(i,j) = C(j,i) \qquad (2.44)$$

ด้วยวิธี อัตสหสัมพันธ์ กำหนด $n_0 = -\infty$ และ $n = \infty$ โดยให้ $x(n)$ มีค่าเท่ากับ 0 เมื่อ $N<0$ และ $n>N-1$ จะได้

$$\langle z^{-i}, z^{-j} \rangle = r(j-i)$$

ให้ค่าผิดพลาดจากการทำนายทั้งสองทาง คือ $x_m^+(n)$ และ $x_m^-(n)$ เป็นเอาท์พุทของฟิลเตอร์ $A_m(z)$ และ $B_m(z)$ ซึ่งคล้องจองกับสมการ (2.38) และ (2.39)

$$A_m(z) = \sum_{i=0}^{m} a_{mi} z^{-i} \quad ; a_{m0} = 1 \qquad (2.45ก)$$

$$B_m(z) = \sum_{i=0}^{m+1} b_{mi} z^{-i} \quad ; b_{m,m+1} = 1 \qquad (2.45ข)$$

ด้วยนิยามของอินเนอร์โพรดักท์ ค่าผิดพลาดยกกำลังสองรวมสามารถแสดงได้ในรูปของอินเนอร์โพรดักท์คือ

$$\alpha_m = \langle A_m(z), A_m(z) \rangle = \| A_m(z) \|^2 \qquad (2.46ก)$$

$$\beta_m = \langle B_m(z), B_m(z) \rangle = \| B_m(z) \|^2 \qquad (2.46ข)$$

สมมุติเงื่อนไขค่าผิดพลาดยกกำลังสองรวมน้อยที่สุดกับการทำนายไปข้างหน้า ดังนั้น α_m ในสมการ (2.46 ก) จะมีค่าน้อยที่สุด เนื่องจาก $A_m(z)$ เป็นโพลีโนเมียลของ z^{-i}; i=0,1,2,..,m ดังนั้นถ้านำค่าใดๆ ซึ่งเป็นโพลีโนเมียล ของ z^{-j}; j=1,2,..,m บวกเข้ากับ $A_m(z)$ นอร์มสแควร์ (Norm Square) ที่ได้จะต้องมีค่ามากกว่า หรือเท่ากับ α_m

$$\|A_m(z)+Cz^{-j}\|^2 \geq \|A_m(z)\|^2 \quad ; j = 1,2,\ldots,m \qquad (2.47)$$

โดย C คือ ค่าใดๆ แยกอสมการ (2.47) และย้ายข้างจะได้

$$2C\langle A_m(z),z^{-j}\rangle + C^2\langle z^{-j},z^{-j}\rangle \geq 0$$

เลือก $C = -\langle A_m(z),z^{-j}\rangle / \langle z^{-j},z^{-j}\rangle$ โดยให้ $\langle z^{-j}, z^{-j}\rangle > 0$

จะได้ $-\{\langle A_m(z),z^{-j}\rangle\}^2 \geq 0$

เนื่องจาก $A_m(z)$ เป็นรีลฟิลเตอร์ (Real Filter) ดังนั้นอสมการเป็นจริงต่อเมื่อ

$$\langle A_m(z),z^{-j}\rangle = 0 \quad ; j =1,2,\ldots,m \qquad (2.48)$$

ด้วยวิธีการเดียวกันสามารถแสดงให้เห็นว่า

$$\langle B_m(z),z^{-j}\rangle = 0 \quad ; j = 1,2,\ldots,m \qquad (2.49)$$

สรุปก็คือถ้าใช้เงื่อนไขค่าผิดพลาดยกกำลังสองรวมน้อยที่สุดกับตัวทำนาย ผลที่ตามมาคือ $A_m(z)$ และ $B_m(z)$ จะออโธโกนัล กับ z^{-j}; j=1,2,..,m

2.3.6 การหาสัมประสิทธิ์พาร์คอร์

เนื่องจาก $A_{m-1}(z)$ เป็นโพลีโนเมียลของ z^{-j} เมื่อ J=0,1,..,m-1 และ $B_{m-1}(z)$ เป็นโพลิโนเมียลของ z^{-j} เมื่อ j=1,2,..,m ดังนั้นผลรวมเชิงเส้นของ $A_{m-1}(z)$ กับ $B_{m-1}(z)$ จะเป็นโพลีโนเมียลของ z^{-j} เมื่อ j=0,1,2,..,m เพราะฉนั้น $A_m(z)$ จึงสามารถหาได้จากผลรวมเชิงเส้นของ $A_{m-1}(z)$ กับ $B_{m-1}(z)$

$$A_m(z) = A_{m-1}(z) + k_m B_{m-1}(z) \qquad (2.50)$$

โดยที่ k_m เป็นค่าที่ทำให้ $A_m(z)$ อยู่ในเงื่อนไขค่าผิดพลาดยกกำลังสองรวมน้อยที่สุดหรือ

$$\langle A_m(z), z^{-j} \rangle = 0 \qquad ; j = 1,2,\ldots,m$$

ดังนั้นจะได้

$$\langle A_m(z), z^{-m} \rangle = \langle A_{m-1}(z), z^{-m} \rangle + k_m \langle B_{m-1}(z), z^{-m} \rangle = 0 \qquad (2.51)$$

จากการกำหนดฟิลเตอร์ $A_m(z)$ และ $B_m(z)$ สมการที่ (2.45 ก) และ (2.45 ข)

$$A_m(z) = \sum_{i=0}^{m} a_{mi} z^{-i} \qquad ; a_{m0} = 1$$

$$B_m(z) = \sum_{i=1}^{m+1} b_{mi} z^{-i} \qquad ; b_{m,m+1} = 1$$

อินเนอร์โพรดักท์ระหว่าง $A_{m-1}(z)$ กับ $B_{m-1}(z)$ สามารถเขียนโดยแตก $B_{m-1}(z)$ เป็นเทอมย่อยๆ ได้

$$\langle A_{m-1}(z), B_{m-1}(z) \rangle = \langle A_{m-1}(z), b_{m-1,1} z^{-1} \rangle + \ldots + \langle A_{m-1}(z), z^{-m} \rangle$$

จากหลักการอินเนอร์โพรดักท์เราสามารถพิสูจน์ได้ว่า $A_{m-1}(z)$ และ $B_{m-1}(z)$ ออโธโกนัลกับ z^{-j} ; j=1,2,..,m-1 ดังนั้น เทอมทางขวามือจะเหลือเพียง

$$\langle A_{m-1}(z), B_{m-1}(z) \rangle = \langle A_{m-1}(z), z^{-m} \rangle \qquad (2.52)$$

ในทำนองเดียวกันแตกโพลีโนเมียลของ $B_{m-1}(z)$ และใช้คุณสมบัติออโธโกนัล สามารถแสดงให้เห็นได้ว่า

$$\langle B_{m-1}(z), B_{m-1}(z) \rangle = \langle B_{m-1}(z), z^{-m} \rangle = \| B_{m-1}(z) \|^2 = \beta_{m-1} \qquad (2.53)$$

ด้วยวิธีการเดียวกันสามารถแสดงให้เห็นว่า นอร์มสแควร์ของ $A_{m-1}(z)$ จะเท่ากับ

$$\|A_{m-1}(z)\|^2 = \alpha_{m-1} = \langle 1, A_{m-1}(z) \rangle \qquad (2.54)$$

จากสมการ (2.51) k_m มีค่าเท่ากับ

$$k_m = - \frac{\langle A_{m-1}(z), z^{-m} \rangle}{\langle B_{m-1}(z), z^{-m} \rangle} \qquad (2.55)$$

แทนค่าสมการ (2.55) จาก สมการ (2.52) และ (2.53) ได้

$$k_m = - \frac{1}{\langle B_{m-1}(z), B_{m-1}(z) \rangle} \langle A_{m-1}(z), B_{m-1}(z) \rangle$$

$$= - \frac{1}{\beta_{m-1}} \sum_{n=n_0}^{n_1} x^+_{m-1}(n) x^-_{m-1}(n) \qquad (2.56)$$

k_m ที่ได้คือ สัมประสิทธิ์พาร์คอร์ ซึ่งใกล้เคียงกับที่ Saito และ Itakura ตั้งขึ้นมาซึ่งอยู่ในรูป

$$k_m = \frac{\sum_{n=-\infty}^{\infty} x^+_{m-1}(n) x^-_{m-1}(n)}{\{ \sum_{n=-\infty}^{\infty} \{x^+_{m-1}(n)\}^2 \sum_{n=-\infty}^{\infty} \{x^-_{m-1}(n)\}^2 \}^{\frac{1}{2}}}$$

k_m มีความหมาย เป็นค่านอร์มัลไลซ์ของ สหสัมพันธ์ข้ามระหว่างค่าผิดพลาดในการทำนายไปข้างหน้าและย้อนหลัง หรือเรียกว่าเป็นค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน (Partial-Correlation) ซึ่งเป็นที่มาของคำว่า พาร์คอร์ (PARCOR) นอกจากนี้ในแง่ของวิชา Acoustics พารามิเตอร์ k_m ยังคล้องจองกับสัมประสิทธิ์การสะท้อนเสียงในท่อนำเสียงหรือ Reflection - Coefficient ดังนั้น บางครั้งจึงเรียก k_m ว่า Reflection Coefficient จากสมการ (2.56) จะเห็นว่า k_m มีค่าอยู่ระหว่าง -1 ถึง +1 ซึ่งเป็นข้อดีในการนำไปใช้งาน การคำนวณสมการทำได้ด้วย Fixed Point Arithematic และมีความแน่นอนทางด้านเสถียรภาพในการคำนวณ

2.3.7 อินเวอสฟิลเตอร์จากสัมประสิทธิ์พาร์คอร์

ถ้าอาศัยวิธี อัตสหสัมพันธ์ ในการแก้สมการเงื่อนไขค่าผิดพลาดยกกำลังสองรวมน้อยที่สุดกับการทำนายไปข้างหน้าและการทำนายย้อนหลัง ความสัมพันธ์ที่ได้ตามมาก็คือ [6]

$$\min \alpha_m = \min \beta_m \qquad (2.57)$$

ผลตามมาก็คือ สัมประสิทธิ์ตัวทำนายของทั้งสองทางจะสัมพันธ์กัน

$$b_{mk} = a_{m,m+1-k} \quad ; k = 1,2,\ldots,m+1 \qquad (2.58)$$

สรุปก็คือ โพลีโนเมียล $A_m(z)$ และ $B_m(z)$ จะมีสัมประสิทธิ์ที่มีค่าเท่ากันแต่อยู่ในลำดับที่ตรงกันข้ามกัน จากนิยามของ $B_m(z)$

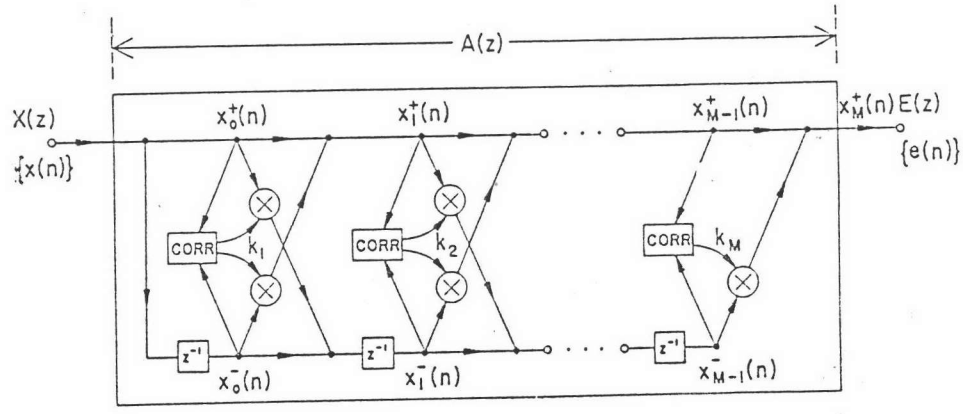
$$B_m(z) = \sum_{i=1}^{m+1} b_{mi} z^{-i}$$

และจากสมการ (2.58) สามารถแสดงให้เห็นได้ว่า

$$B_m(z) = z^{-(m+1)} A_m\left(\frac{1}{z}\right) \qquad (2.59)$$

เมื่อนำไปแทนค่าในสมการ (2.50) ร่วมกับความสัมพันธ์ตาม (2.58) จะได้

$$B_m(z) = z^{-1}\{k_m A_{m-1}(z) - B_{m-1}(z)\} \qquad (2.60)$$



รูป 2.16 โครงสร้างของอินเวิร์สฟิลเตอร์ที่ใช้วิธีพาร์คอร์

จากนิยามค่าผิดพลาด

$$x_m^+(z) = A_m(z)x(z)$$

และ $$x_m^-(z) = B_m(z)x(z)$$

จากสมการ (2.50) จะได้

$$x_m^+(z) = x_{m-1}^+(z) + k_m x_{m-1}^-(z) \qquad (2.61)$$

และจากสมการ (2.60) จะได้

$$x_m^-(z) = z^{-1}\{k_m x_{m-1}^+(z) + x_{m-1}^-(z)\} \qquad (2.62)$$

สมการ (2.61) และ (2.62) แสดงในรูปของแซมเบิลโดเมน คือ

$$x_m^+(n) = x_{m-1}^+(n) + k_m x_{m-1}^-(n) \qquad (2.63\text{ก})$$

$$x_m^-(n) = k_m x_{m-1}^+(n-1) + x_{m-1}^-(n-1) \qquad (2.63\text{ข})$$

โดยที่ $x_0^+(n) = x(n) \quad ; \quad x_0^-(n) = x(n-1)$ (2.64)

รูป 2.16 แสดงโครงสร้างของอินเวิร์สฟิลเตอร์ที่ใช้วิธีพาร์คอร์ ซึ่งมีลักษณะโครงสร้างแบบแลตทิซ (Lattice Structure) กล่องสี่เหลี่ยมที่เขียนไว้ว่า CORR เป็นส่วนที่ทำหน้าที่คำนวณหาสัมประสิทธิ์พาร์คอร์ หรือค่า k_m ด้วยสมการ (2.56) ความสัมพันธ์ของสัญญาณในแต่ละสเตจเป็นไปตามสมการ (2.63 ก) และ (2.63 ข)

2.3.8 กรรมวิธีคำนวณค่าสัมประสิทธิ์พาร์คอร์

ตามวิธีของ Markel และ Gray การคำนวณสมการทำได้ทั้งวิธีโควาเรียนซ์หรือวิธี อัตสหสัมพันธ์ ส่วนวิธีของ Itakura ใช้ได้เฉพาะกับวิธี อัตสหสัมพันธ์ ในวิทยานิพนธ์นี้อาศัยวิธีพิสูจน์สมการของ Markel และ Gray ร่วมกับเงื่อนไขของ Itakura การคำนวณจึงใช้วิธี อัตสหสัมพันธ์ จากความสัมพันธ์ของอินเวิร์สฟิลเตอร์ $A_m(z)$ ดังในสมการ(2.50) จะเห็นว่า $A_m(z)$ สามารถหาได้จาก $A_{m-1}(z)$ และ $B_{m-1}(z)$ จากนิยาม $A_0(z)=1$ และ $B_0(z)=z^{-1}$ ดังนั้นจึงสามารถคำนวณ $A_m(z)$ ที่ m=1 ได้คือ

$$A_1(z) = A_0(z) + k_0 B_0(z)$$

$$= 1 + k_0 B_0(z)$$

หา $A_m(z)$ ที่ m=2

$$A_2(z) = A_1(z) + k_1 B_1(z)$$

$$= 1 + k_0 B_0(z) + k_1 B_1(z)$$

ดังนั้น $A_m(z)$ สามารถหาได้จากการคำนวณต่อๆ กันไปแบบรีเคอร์ซีฟ (Recersive) เทอมทั่วไปคือ

$$A_m(z) = 1 + \sum_{i=1}^{m} k_i B_i(z) \qquad (2.65)$$

นอร์มสแควร์ของสมการ (2.65) เมื่อย้าย 1 มาข้างซ้ายจะได้

$$\| A_m(z) - 1 \|^2 = \sum_{i=1}^{m} k_i^2 B_{i-1}$$

แตกเทอมทางซ้ายมือจะได้เป็น

$$\| A_m(z) - 1 \|^2 = \| A_m(z) \|^2 - 2 \langle A_m(z), 1 \rangle + \| 1 \|^2$$

เช่นเดียวกับสมการ (2.54) สามารถแสดงให้เห็นได้ว่า $\| A_m(z) \|^2 = \langle A_m(z), 1 \rangle = \alpha_m$ ดังนั้น

$$\alpha_m = \| 1 \|^2 - \sum_{i=1}^{m} k_i^2 \beta_{i-1} \qquad (2.66)$$

แทนค่า m ด้วย m+1 จะได้ α_{m+1} ลบออกด้วย α_m จะได้

$$\alpha_{m+1} = \alpha_m - k_{m+1}^2 \beta_m \qquad (2.67)$$

ใช้เงื่อนไข $\min \alpha_m = \min \beta_m$

จะได้ $$\alpha_{m+1} = (1 - k_{m+1}^2) \alpha_m \qquad (2.68)$$

ในการหาค่า k สมการ (2.56) สมการเขียนใหม่ได้เป็น

$$k_{m+1} = -\frac{1}{\beta_m}\langle A_m(z), B_m(z)\rangle$$

แทน β_m ด้วย α_m และจาก (2.52) แทน $\langle A_m(z), B_m(z)\rangle$ ด้วย $\langle A_m(z), z^{-(m+1)}\rangle$ ได้

$$k_{m+1} = -\frac{1}{\alpha_m}\langle A_m(z), z^{-(m+1)}\rangle$$

$$= -\frac{1}{\alpha}\sum_{i=0}^{m} r(m+1-i)a_{mi} \qquad (2.69)$$

จาก (2.50) สามารถเขียนใหม่ได้

$$A_{m+1}(z) = A_m(z) + k_{m+1}B_m(z) \qquad (2.70)$$

แทน k_{m+1} จากสมการ (2.69) ลงไปในสมการ (2.70) แล้วเทียบสัมประสิทธิ์ทั้งสองข้างจะได้

$$a_{m+1,0} = 1 \qquad (2.71ก)$$

$$a_{m+1,0} = a_{mi} + k_{m+1}b_{mi} \quad ; i = 1,2,\ldots,m \qquad (2.71ข)$$

$$a_{m+1,m+1} = k_{m+1} \qquad (2.71ค)$$

โดย b_{mi} ในสมการ (2.71 ข) หาได้จากความสัมพันธ์ในสมการ (2.58)
ซึ่งนำมาเขียนใหม่

$$b_{mi} = a_{m,m+1-i} \quad ; i = 1,2,\ldots,m+1 \qquad (2.72)$$

สรุปสมการที่ใช้ในการคำนวณค่า k_m ประกอบด้วยสมการ (2.68),(2.69),(2.71) และ สมการ (2.72) ซึ่งสมการเหล่านี้อยู่ในรูปการคำนวณแบบรีเคอร์ซีฟทั้งสิ้น ในการนำสมการต่างๆ มาคำนวณด้วยเครื่องคอมพิวเตอร์ สามารถสรุปเป็นขั้นตอนต่างๆ ได้ดังนี้

ขั้นที่ 1 คำนวณค่าอัตสหสัมพันธ์

กำหนดให้ในหนึ่งเฟรมของการวิเคราะห์มีจำนวนสัญญาณสุ่มเท่ากับ N แซมเปิล และฟิลเตอร์มีจำนวนออร์เดอร์ เท่ากับ M ตามวิธี อัตสหสัมพันธ์ ค่า อัตสหสัมพันธ์ ที่ใช้ในการคำนวณขั้นต่อๆ ไปจะมีจำนวนเท่ากับ M+1 ค่า คือ

$$r(k) = \sum_{n=0}^{N-1-k} x(n)x(n+k) \; ; \; k = 0,1,2,\ldots,M \qquad (2.73)$$

โดย $x(n)$; $n=0,1,2,..,N-1$ คือสัญญาณเสียงที่สุ่มมาในหนึ่งเฟรม

$$k_1 = -\frac{1}{\alpha_0} r(1) a_{00}$$

ขั้นที่ 2 คำนวณค่าเริ่มต้น

จากสมการ (2.69) แทน $m=0$ จะได้

$$\alpha_0 = \langle A_0(z), A_0(z) \rangle = r(0)$$

ซึ่งจากเงื่อนไขเริ่มต้นตามสมการ (2.45 ก) $a_{00}=1$ และ $\alpha_0 = \langle A_0(z), A_0(z) \rangle = r(0)$ ดังนั้น

$$k_1 = -r(1)/r(0) \qquad (2.74)$$

จากสมการ (2.71 ก) และสมการ (2.71 ค) แทน $m=0$ ได้

$$a_{10} = 1 \; ; \; a_{11} = k_1 \qquad (2.75)$$

จากสมการ (2.68) แทน $m=0$ จะได้

$$\alpha_1 = (1-k_1^2)\alpha_0 = (1-k_1^2)r(0) \qquad (2.76)$$

ขั้นที่ 3 ทำการคำนวณแบบรีเคอร์ซีฟ

ขั้นตอนนี้จะนำค่าเริ่มต้นจากขั้นตอนที่แล้วมาคำนวณต่อๆ กันแบบรีเคอร์ซีฟโดยอาศัยสมการ (2.68),(2.69),(2.71) และสมการ (2.72) ในการหาค่าสัมประสิทธิ์พาร์คอร์จนครบจำนวนออร์เดอร์ คือ k_i ; i=1,2,..,M และพร้อมกันก็จะได้ค่าสัมประสิทธิ์ตัวทำนาย a_i ; i=0,1,2,..,M จะต้องมีการคำนวณต่อๆ กันไปทั้งหมด M-1 รอบ คือแทนค่า m=1,2...,M-1 เมื่อคำนวณเสร็จขั้นที่ m = M-1 จะได้คำตอบครบตามที่ต้องการรวมทั้งค่า α_m หรือค่าผิดพลาดยกกำลังสองรวมในเฟรม ซึ่งจะใช้เป็นพารามิเตอร์ควบคุมพลังงานของ Excitation ในภาคสังเคราะห์เสียง

ขั้นตอนทั้งสามนี้ถูกนำมาเขียนเป็นโปรแกรมย่อยชื่อ AUTO เพื่อใช้ในการวิเคราะห์เสียงพูดหาพารามิเตอร์ของฟิลเตอร์ในแต่ละเฟรม จำนวนแซมเปิลสัญญาณที่ใช้วิเคราะห์ในหนึ่งเฟรมกำหนดที่ตัวแปร N จำนวนออร์เดอร์ของฟิลเตอร์กำหนดที่ตัวแปร M รูป2.17.1แสดงชื่อตัวแปรในโปรแกรมเทียบกับสัญญลักษณ์ในสมการ รูป2.17.2แสดงโปรแกรมย่อย AUTO และคำอธิบายเทียบกับขั้นตอนที่กล่าวมา

TEXT SYMBOL	SUBROUTINE AUTO SYMBOL
N	N
x(n) ; n=0,1,...,N-1	x(n)
M	M
a_{mi} ; $0 \le i \le m \le M$	A(i)
α_m ; m=0,1,...,M	ALPHA
k_m ; m=1,2,...,M	RC(m)
r(i) ; i=0,1,...,M	R(i)
β_m ; i=0,1,...,M-1	ALPHA
b_{mi} ; $1 \le i \le m+1 \le M$	B(i)

รูป2.17.1ชื่อตัวแปรในโปรแกรม AUTO เทียบกับสัญลักษณ์ในสมการ

```
1  REM   ** SUBROUTINE AUTO
2  FOR K = 0 TO M
3 R(K) = 0
4 NK = N - 1 - K
5  FOR NP = 0 TO NK
6 R(K) = R(K) + X(NP) * N(NP + K)
7  NEXT NP
8  NEXT K
9 ALPHA = R(0)
10 RC(1) =  - R(1) / R(0)
11 A(0) = 1
12 A(1) = RC(1)
13 ALPHA = ALPHA - RC(1) * RC(1) * ALPHA
14  FOR MI = 1 TO M - 1
15  FOR J = 1 TO MI + 1
16 JB = MI + 1 - J
17 B(J) = A(JB)
18  NEXT J
19 S = 0
20  FOR IP = 0 TO MI
21 S = S + R(MI + 1 - IP) * A(IP)
22  NEXT IP
23 MA = MI + 1
24 RC(MA) =  - S / ALPHA
25  FOR IP = 1 TO MI
26 A(IP) = A(IP) + RC(MA) * B(IP - 1)
27  NEXT IP
28 A(MA) = RC(MA)
29 ALPHA = ALPHA - RC(MA) * RC(MA) * ALPHA
30  IF ALPHA <  = 0 THEN  PRINT "INSUFFICIENT ACCURACY"
31  NEXT MI
32  RETURN
```

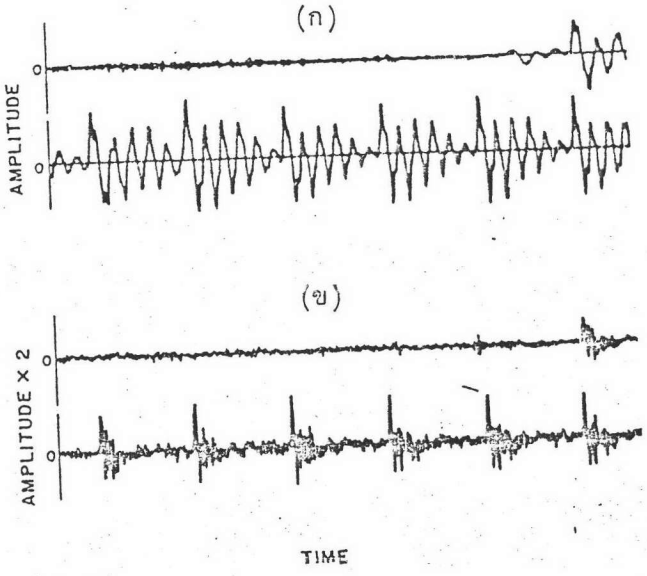
รูป 2.17.2 โปรแกรมย่อย AUTO

คำอธิบายโปรแกรม AUTO

บรรทัดที่	หน้าที่
2 - 8	คำนวณค่า อัตสหสัมพันธ์ ด้วยสมการ (2.73)
9 - 10	แทนค่าเริ่มต้น α_0 และ k_1 สมการ (2.74)
11 - 12	แทนค่า a_{00}, a_{10}, a_{11} สมการ (2.75)
13	คำนวณ α_1 จากสมการ (2.76)
14 - 31	ขั้นตอนที่ 3 คำนวณทั้งหมด M-1 รอบ
15 - 18	หาค่า b_{mi} จากสมการ (2.72)
19 - 24	คำนวณ k_1 จากสมการ (2.69)
25 - 27	หาค่า a_{mi} จากสมการ (2.71 ข)
28	หาค่า $a_{m+1,m+1}$ จากสมการ (2.71 ค)
29	หาค่า α_m จากสมการ (2.68)
30	ถ้าค่า α มีค่าน้อยกว่า 0 แสดงว่าการคำนวณผิดพลาด

2.4 การหาคาบของสัญญาณเสียงพูด

การหาคาบของเสียง (Pitch Period) สามารถทำได้หลายวิธี [18] ในที่นี้จะกล่าวเฉพาะที่เกี่ยวกับกรรมวิธี Simplified Inverse Filter Tracking หรือ SIFT ซึ่งใช้หลักการของวิธีแอลพีซี กรรมวิธี SIFT นี้พัฒนาโดย J.D Markel [14]

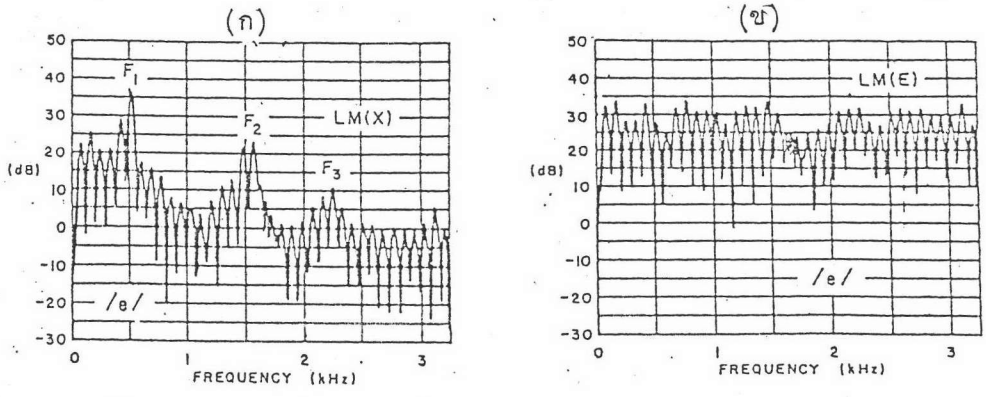


รูป 2.18 สัญญาณเสียงพูดและสัญญาณค่าผิดพลาดของคำว่า "Shade"

2.4.1 สัญญาณค่าผิดพลาดของเสียงพูด

ดังที่ได้กล่าวไว้ว่าเสียงพูดมี 2 ลักษณะคือ เสียงก้อง (Voiced Sound) และเสียงไม่ก้อง (Unvoiced Sound) รูป 2.18 ก แสดงสัญญาณของเสียงพูดคำว่า "Shade" ช่วงต้นของคำคือ /Se/ เป็นเสียงไม่ก้อง สังเกตได้ที่ส่วนบนของรูป 2.18 ก สัญญาณจะมีรูปร่างไม่แน่นอน ซึ่งเป็นลักษณะของสัญญาณรบกวน ส่วนของเสียงสระ /a/ ซึ่งเป็นเสียงก้องจะมีลักษณะเป็นคาบ สัญญาณค่าผิดพลาด (Error Signal) ของเสียงพูดคือเอาท์พุทของอินเวิร์สฟิลเตอร์ ในรูป 2.18 ข. แสดงสัญญาณค่าผิดพลาดของเสียงคำเดียวกัน จะเห็นว่าในส่วนเสียงไม่ก้อง สัญญาณค่าผิดพลาดจะมีลักษณะ

เป็นเสียงรบกวนคล้ายกับสัญญาณเสียง ที่เป็นเช่นนี้เพราะสัญญาณรบกวนมีความไม่แน่นอน ตัวทำนายไม่สามารถทำนายได้ทำให้เกิดค่าผิดพลาดที่ไม่มีความแน่นอนเช่นกัน ในกรณีของเสียงก้อง ความสามารถในการทำนายได้ของเสียงทำให้มีค่าผิดพลาดสูงเฉพาะในช่วงต้นของคาบ ด้วยธรรมชาติของเสียงพูดอันนี้ทำให้แบบจำลองการผลิตเสียงพูดใช้สัญญาณ Pulse Train ที่มีความห่างของพัลส์เท่ากับคาบของเสียง หรือ Random Pulse (ซึ่งทำหน้าที่เป็น White Noise) เป็นสัญญาณ Excitation ป้อนสู่ฟิลเตอร์แทนสัญญาณค่าผิดพลาด ข้อมูลที่เกี่ยวกับสัญญาณ Excitation จึงเหลือเพียงตัวกำหนดให้สัญญาณเป็น Random Pulse หรือ Pulse Train และคาบของ Pulse Train

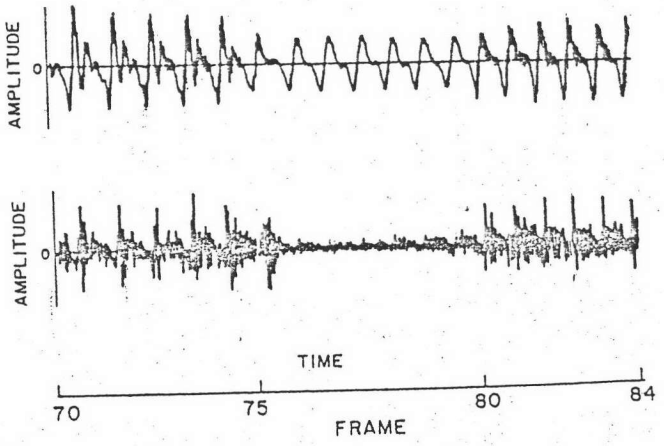


รูป 2.19 (ก) สเปกตรัมของสัญญาณเสียง และ (ข) สัญญาณค่าผิดพลาดของเสียงสระอี

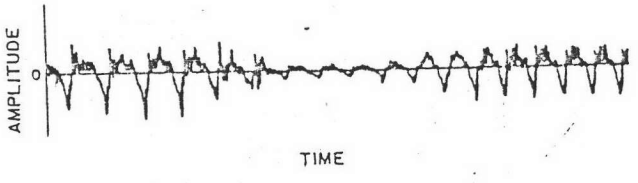
2.4.2 หลักการหาคาบของสัญญาณ

จุดสำคัญของการหาคาบ คือ การตัดส่วนประกอบทางความถี่อื่นๆ ออกไปให้เหลือแต่ความถี่หลักมูล (Fundamental Frequency) ส่วนประกอบทางความถี่ที่มีอิทธิพลมากต่อรูปร่างของสัญญาณคือ ความถี่ฟอร์แมนท์ที่หนึ่ง (First Formant Frequency) ดังในรูป 2.18 ก ซึ่งการจะหาคาบจากรูปคลื่นทำได้ยาก เมื่อนำสัญญาณเสียงมาผ่านอินเวิร์สฟิลเตอร์จะได้สัญญาณค่าผิดพลาด ซึ่งส่วนประกอบของความถี่ของฟอร์แมนท์ต่างๆ ถูกตัดออกไปมาก พิจารณาส่วนประกอบความถี่รูป 2.19 ข. ซึ่งเป็นสเปกตรัมของสัญญาณค่าผิดพลาด จะเห็นว่าส่วนประกอบทางความถี่ค่อนข้างเรียบไม่มียอดของความถี่ฟอร์แมนท์ ดังนั้นเราจึงสามารถหาคาบได้สะดวก อย่างไรก็ตามมีข้อจำกัด

เกิดขึ้นเมื่อเสียงที่ทำการวิเคราะห์เป็นเสียงนาสิก รูป 2.20 แสดงสัญญาณเสียง /n/ ของคำว่า "Linear" จะเห็นว่า ยอดของสัญญาณค่าผิดพลาดซึ่งแสดงความเป็นคาบหายไป ที่เป็นเช่นนี้เนื่องจากเสียง /n/ มี Zero ที่มีความถี่ระหว่าง 900-1400Hz ทำให้ความสามารถในการทำนายของรูปคลื่นลดลงไป การแก้ไขคือเพิ่มภาคกรองความถี่ผ่านต่ำ (Lowpass Filter) เพื่อลดอิทธิพลของส่วนประกอบความถี่สูงก่อนนำสัญญาณผ่านอินเวิร์สฟิลเตอร์ ในทางปฏิบัติกระทำได้โดยนำสัญญาณมาหาความแตกต่างระหว่างแซมเปิล (Difference) หรือ คูณสัญญาณด้วยฟังก์ชัน $1-Z^{-1}$ นำผลที่ได้ไปคำนวณหาพารามิเตอร์ของฟิลเตอร์ เพื่อป้อนเข้าสู่อินเวิร์สฟิลเตอร์ขณะที่อินพุทของอินเวิร์สฟิลเตอร์รับสัญญาณเสียงโดยตรง ซึ่งการกระทำเช่นนี้ได้ผลเท่ากับการนำ Integrator มาใช้งานแทนภาคกรองความถี่ผ่านต่ำ ผลที่ได้คือสัญญาณค่าผิดพลาดในช่วงเสียง /n/ เห็นเป็นคาบมากขึ้นดังในรูป 2.21



รูป 2.20 สัญญาณเสียงและค่าผิดพลาดของเสียง /n/ ในคำว่า "Linear"

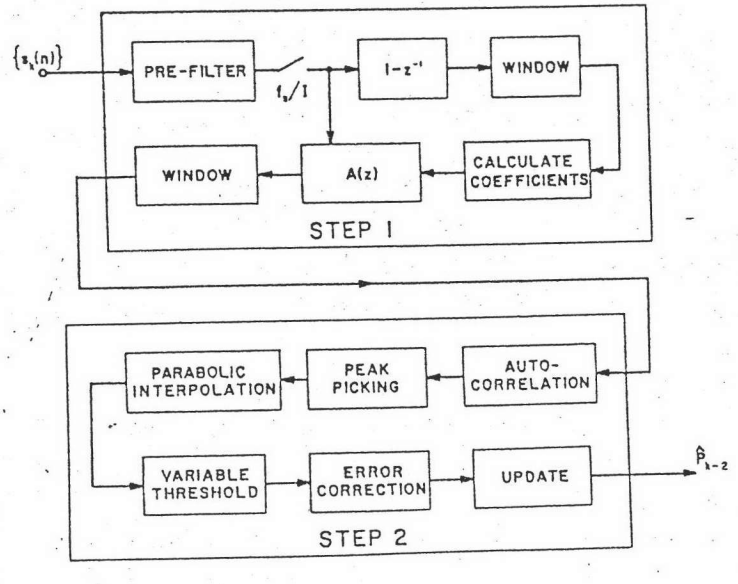


รูป 2.21 สัญญาณค่าผิดพลาดของเสียง /n/ ที่ผ่านการคูณด้วยฟังก์ชัน $1 - Z^{-1}$

เดิมทีการหาคาบของสัญญาณอาศัยฟังก์ชัน Short-time Autocorrelation ดังที่กล่าวไว้ในหัวข้อ 2.2.2 แต่วิธีนี้ยังไม่สามารถตัดอิทธิพลของความถี่เป็นฟอร์แมนท์ได้ จึงได้มีการประยุกต์เอาฟังก์ชัน อัตสหสัมพันธ์ ใช้กับสัญญาณค่าผิดพลาด ทำให้การหาคาบของเสียงถูกต้องยิ่งขึ้น หลักการนี้ J.D. Markel พัฒนามาเป็นกรรมวิธี SIFT [19]

2.4.3 กรรมวิธี SIFT

กรรมวิธี Simplified Inverse Filter Tracking หรือ SIFT เป็นกรรมวิธีการหาคาบของสัญญาณให้ความถูกต้องในช่วง 50-250Hz หรือค่าของคาบอยู่ระหว่าง 4-20 ms ความถี่ของสัญญาณสุ่ม (Sampling Frequency) เท่ากับ 10 kHz ในการวิเคราะห์แต่ละเฟรมใช้สัญญาณ 400 แซมเปิล กระบวนการทำงานแบ่งเป็นสองขั้นคือ STEP1 และ STEP2 รูป 2.22 แสดงขั้นตอนของกรรมวิธี ซึ่งมีรายละเอียดดังนี้



รูป 2.22 ขั้นตอนของกรรมวิธี SIFT

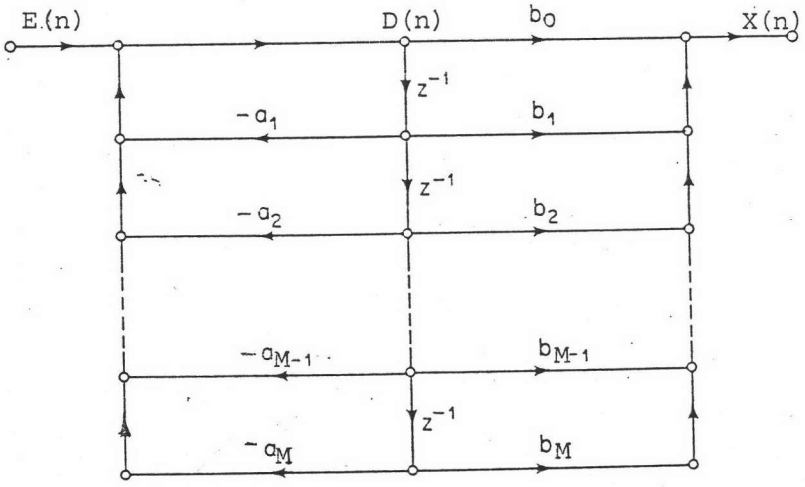
ขั้นที่ 1 (STEP1) สัญญาณเสียงพูด {S(n)} ซึ่งสุ่มมาด้วยความถี่ 10 kHz นำมาผ่าน Pre-filter ซึ่งเป็นภาคกรองความถี่ผ่านต่ำความถี่คัทออฟที่ 1 kHz ทำหน้าที่ป้องกันการเกิด Aliasing ในการทำ Down Sampling ในขั้นต่อไป การทำ Down Samplingหมายถึงการสุ่มสัญญาณจากสัญญาณที่สุ่มมาแล้วอีกทีเพื่อลดความถี่ในการสุ่ม ในที่นี้ลดลง 5 เท่า (fs/I = 2 kHz) สัญญาณเสียงจากเดิมมีจำนวน 400 แซมเปิลจะเหลือ 80 แซมเปิล เหตุผลในการกระทำ Down Sampling เพื่อตัดความถี่สูงไม่ให้มารบกวนรวมทั้งยังลดการคำนวนลง สัญญาณที่ได้จากการกระทำ Down Sampling นำไปหาค่าแตกต่างระหว่างแซมเปิล (Difference) แล้วผ่าน Hamming Window ก่อนเข้ากระบวนการคำนวนเพื่อหาสัมประสิทธิ์ตัวทำนาย a_i สัมประสิทธิ์ตัวทำนายที่ได้จะใช้เป็นพารามิเตอร์ของอินเวิร์สฟิลเตอร์ A(z) อินเวิร์สฟิลเตอร์ที่ใช้มีจำนวน Order เท่ากับ 4 ซึ่งเพียงพอในการคลุมย่านความถี่เสียง 1 kHz เพราะเสียงในย่าน 1 kHz ประกอบด้วยความถี่หลักมูลกับความถี่ฟอร์แมนท์ที่หนึ่ง อินพุทของอินเวิร์สฟิลเตอร์คือ สัญญาณจากการกระทำ Down Sampling ส่วนเอาท์พุทของอินเวิร์สฟิลเตอร์หรือสัญญาณค่าผิดพลาดจะนำไปผ่าน Hamming Window อีกทีก่อนส่งให้ขั้นต่อไป

ขั้นที่ 2 (STEP1) เริ่มด้วยการคำนวน อัตสหสัมพันธ์ ของสัญญาณจากขั้นที่ 1 ฟังก์ชันอัตสหสัมพันธ์ r(i) ของสัญญาณจริงใดๆ จะมีค่าสูงสุดที่ i=0 และจะลดลงเมื่อ i เพิ่มขึ้น กรณีของสัญญาณที่มีคาบและซ้ำกันอยู่ภายในเฟรม ค่าอัตสหสัมพันธ์นอกจากมีค่าสูงสุดที่ i=0 แล้ว ฟังค์ชันอัตสหสัมพันธ์ ยังมียอด (Peak) อยู่ที่บริเวณ i = nT n คือ 1,2,3 ... และ T คือคาบของสัญญาณ ดังนั้นการหาจุดยอดแรกของ r(i) จะได้คาบของสัญญาณโดยมีค่าเท่ากับ i จากรูป 2.22 Peak Picking คือส่วนของการหาจุดยอดแรกภายในช่วง 2.5 ms ถึง 15.5 ms ระยะของจุดยอดจะผ่าน Parabolic Interpolator เพื่อเพิ่ม Resolution จากนั้นนำยอดมาตัดสินด้วยความสูงของยอดว่าขนาดความสูงนี้จะถือเป็นยอดหรือไม่ เนื่องจากเมื่อมีความถี่สูงขึ้น ค่าอัตสหสัมพันธ์ที่จุดยอดจะลดลง ดังนั้น Threshold ในการตัดสินยอดจึงเปลี่ยนไปตามความถี่ด้วย ขั้นสุดท้ายคือ การแก้ไขข้อผิดพลาดโดยอาศัยค่าของคาบจากเฟรมที่ผ่านมาและเฟรมที่ถัดไป ถ้าคาบที่ได้มีการเปลี่ยนแปลงระหว่างเฟรมมากกว่าที่กำหนด จะถือว่าคาบที่ได้ไม่ถูกต้องและจะบังคับให้มีค่าเท่ากับ 0 อีกกรณีหนึ่งคือค่าของเฟรมระหว่างกลางมีค่าเป็น 0 ขณะที่ค่าของเฟรมก่อนหน้าและเฟรมต่อมาที่มีค่าใกล้เคียงกัน จะถือว่าการคำนวณในเฟรมกลางนี้ผิดพลาดและค่าของคาบจะมีค่าเท่ากับค่าเฉลี่ยของเฟรมก่อนหน้าและเฟรมต่อมา ในการคำนวนหาคาบในเฟรมหนึ่งจะอาศัยข้อมูลทั้งหมด 3 เฟรมติดต่อกัน ดังนั้นค่าของคาบที่ได้จะช้าไป 2 เฟรม ถ้าผลลัพธ์ได้ค่าของคาบเท่ากับ 0 หมายความว่าเฟรมนั้นสัญญาณไม่มีคาบ หรือสรุปว่าเสียงไม่ก้อง จากกรรมวิธีทั้งสองขั้นที่กล่าวมานำไปเขียนเป็นโปรแกรมย่อยด้วยภาษาเบสิคชื่อ โปรแกรมย่อย Step 1 และโปรแกรมย่อย Step 2 ทั้งสองโปรแกรมย่อยทำงานร่วมกันจะสามารถคำนวณหาคาบของเสียงในแต่ละเฟรมได้

การได้คำตอบ จะต้องทำงานอย่างน้อย 3 เฟรมขึ้นไป โปรแกรมย่อย Step 1 มีการเรียกโปรแกรมย่อย Auto และโปรแกรมย่อย Direct โปรแกรมย่อย Auto ทำหน้าที่คำนวณพารามิเตอร์ของฟิลเตอร์เพื่อป้อนให้อินเวิร์สฟิลเตอร์ รายละเอียดของโปรแกรมย่อย Auto ได้กล่าวไว้แล้วในหัวข้อ 2.3.8 ส่วนโปรแกรมย่อย Direct คือฟิลเตอร์ตามโครงสร้างที่เรียกว่า Direct Form ดังในรูปที่ 2.23 ซึ่งมีสมการอยู่ในรูป

$$D(n) = E(n) - \sum_{i=1}^{M} a_i D(n-i) \qquad (2.77)$$

$$x(n) = \sum_{i=0}^{M} P_i D(n-i) \qquad (2.78)$$



รูป 2.23 โครงสร้างของฟิลเตอร์แบบ Direct Form

การเรียกโปรแกรมย่อย Direct ครั้งแรกในโปรแกรมย่อย Step1 โปรแกรมย่อย Direct จะทำหน้าที่เป็นภาคกรองความถี่ผ่านต่ำ ออร์เดอร์เท่ากับ 3 ความถี่คัทออฟที่ 1 kHz ตามพารามิเตอร์ที่กำหนดไว้ตอนต้นของโปรแกรมย่อย Step1 ในการเรียกโปรแกรมย่อย Direct ครั้งที่สองโปรแกรมย่อยจะทำงานเป็นอินเวอสฟิลเตอร์ทำหน้าที่หาสัญญาณค่าผิดพลาด (Error Signal) ซึ่งฟิลเตอร์มีจำนวนออร์เดอร์เท่ากับ 4 รูป 2.24 แสดงโปรแกรมย่อย STEP 1 รูป 2.25 แสดงโปรแกรมย่อย DIRECT และรูป 2.26 แสดงโปรแกรมย่อย STEP 2 ทั้งโปรแกรมย่อย Step 1 และโปรแกรมย่อย Step 2 ถูกนำไปใช้เป็นส่วนหนึ่งของโปรแกรม SIFTX ซึ่งใช้ในการวิเคราะห์คาบของเสียงพูดเป็นคำๆ

```
1  REM  ** SUBROUTINE STEP1
2 A(0) = 1:A(1) =  - 1.340366:A(2) = 1.0119:A(3) =  - .614109
3 PD(0) = 0.035708:PD(1) = .006996:PD(2) = .006996:PD(3) = .035708
4 MD = 3:M1 = 4: FOR KK = 0 TO M1:D(KK) = 0: NEXT KK
5  FOR I = 0 TO WS - 1
6  FOR J = 1 TO DS
7 LN = LN + 2
8 XI = ( PEEK (LN + 1) * 256 +  PEEK (LN) - 2048) / 2048
9  GOSUB 5000: REM  GOSUB DIRECT
10  NEXT J
11 PU(I) = 80
12 C =  COS ((I - 1) * 7.28318 / (WS - 1))
13 X(I) = (80 - UP) * (.54 - .46 * C)
14 UP = 80
15  NEXT I
16 NZ = N:MZ = M:N = WS:M = M1
17  GOSUB 4000: REM  GOSUB AUTO
18 N = NZ:M = MZ
19 PD(0) = 1: FOR KK = 1 TO M1:PD(KK) = 0: NEXT
20  FOR KK = 0 TO M1:D(KK) = 0: NEXT
21 MD = M1
22  FOR I = 0 TO WS - 1
23 XI = PU(I)
24  GOSUB 5000: REM  GOSUB DIRECT
25  IF I <  = M1 THEN 28
26 C =  COS ((I - M1 - 1) * 7.28318 / (WS - M1 - 1))
27 PU(I - M1) = 80 * (.54 - .64 * C)
28  NEXT
29  RETURN
```

รูป 2.24 โปรแกรมย่อย Step 1

คำอธิบายโปรแกรมย่อย Step 1

บรรทัดที่	หน้าที่
1-2	กำหนดค่าพารามิเตอร์สำหรับภาคกรองความถี่ผ่านต่ำ ความถี่คัทออฟ 1 kHz
4	กำหนดค่าเริ่มต้นของตัวแปรในโปรแกรมย่อย Direct
5-15	นำสัญญาณเสียงมาผ่านภาคกรองความถี่ผ่านต่ำ (บรรทัดที่ 9) คูณด้วย Hamming Window และหาค่าแตกต่างระหว่างแซมเปิล (บรรทัดที่ 12-13) ในขณะเดียวกันก็ทำการ Down Sampling ไปด้วย (บรรทัดที่ 6-10) และจะวนทำจนสัญญาณมีจำนวนครบเฟรม...
16-18	คำนวณพารามิเตอร์ของอินเวิร์สฟิลเตอร์
19-20	กำหนดค่าเริ่มต้นของตัวแปรในโปรแกรมย่อย Direct
21-28	คำนวณอินเวิร์สฟิลเตอร์ (บรรทัดที่ 24) พร้อมกับนำผลที่ได้คูณกับ Hamming Window (บรรทัดที่ 26-27)

```
1  REM  ** SUBROUTINE DIRECT
2 SO = 0
3 D(0) = XI
4  FOR JJ = 1 TO MD
5 JI = MD - JJ + 1
6 SO = SO + D(JI) * FD(JI)
7 D(0) = D(0) - A(JI) * D(JI)
8 D(JI) = D(JI - 1)
9  NEXT
10 SO = SO + D(0) * FD(0)
11  RETURN
```

รูป 2.25 โปรแกรมย่อย Direct

คำอธิบายโปรแกรมย่อย Direct

บรรทัดที่	หน้าที่
2	กำหนดค่าเริ่มต้น
3	กำหนดสัญญาณเข้า
6	คำนวณตามสมการ (2.78)
7	คำนวณตามสมการ (2.77)
8	คำนวณ Delay ของ D(n)
10	คำนวณตามสมการ (2.78) ซึ่งค่าที่ได้จะเป็นเอาท์พุทของฟิลเตอร์

```
1  REM  ** SUBROUTINE STEP2
2  FOR I = 1 TO 33
3 J = I - 1
4 KK = 76 - J
5 SM = 0
6  FOR L = 1 TO KK
7 II = L + J
8 SM = SM + PU(L) * PU(II)
9  NEXT L
10 AB(I) = SM
11  NEXT I
12 P1 = PI(1)
13 P2 = PI(2)
14 P3 = PI(3)
15 AM = AB(6)
16 C = 6
17  FOR I = 6 TO 32
18  IF AB(I) >  = AM THEN AM = AB(I):C = I
19  NEXT I
20  IF AM = 0 THEN 35
21  IF AB(C) < AB(C - 1) THEN 35
22 AA = (AB(C + 1) + AB(C - 1) - 2 * AB(C)) / 2
23 BB = (AB(C + 1) - AB(C - 1)) / 4
24 AP = AB(C) - BB * BB / AA
25 AC = C - BB / AA
26 V = AP / AB(1)
27  IF C >  = 19 THEN 30
28 D =  - 1 * (C - 6) / 13 + 2
29  GOTO 31
30 D =  - .1 * (C - 19) / 13 + 1
31 V = V / D
32  IF V >  = .35 THEN 37
33  IF P1 = 0 THEN 35
34  IF V >  = .30 THEN 37
35 P0 = 0
36  GOTO 38
37 P0 = AC
38  IF  ABS (P1 - P3) <  = .375 * P3 THEN P2 = (P1 + P3) / 2
39  IF P3 <  > 0 THEN 43
40  IF P2 = 0 THEN 43
41  IF  ABS (P0 - P1) > .2 * P1 THEN 43
42 P2 = 2 * P1 - P0
43  IF P1 <  > 0 THEN 45
44  IF  ABS (P2 - P3) > .375 * P3 THEN P2 = 0
45 PI(3) = P2:PI(2) = P1:PI(1) = P0
46 PT(FI) = (PI(3) - 1) * DS
47  RETURN
```

รูป 2.26 โปรแกรมย่อย Step 2

คำอธิบายโปรแกรมย่อย Step 2

บรรทัดที่	ทำหน้าที่
2-11	คำนวณค่า อัตสหสัมพันธ์ ของสัญญาณจาก Step 1
15-19	หาตำแหน่งของจุดยอดของค่าอัตสหสัมพันธ์
20	ตรวจว่าจุดยอดเท่ากับ 0
21	ถ้า Slope ก่อนถึงจุดยอดเป็นลบ แสดงว่าไม่มีจุดยอด
22-26	Parabolic Interpolation
27-31	ปรับค่าของจุดยอดตามความถี่
32-37	ตัดสินความเป็นเสียงก้องหรือไม่ก้อง
38-44	แก้ข้อผิดพลาดโดยเทียบกับข้อมูลของเฟรมก่อนหน้า P3 และเฟรมต่อมา P1 และประมาณคำตอบของเฟรมปัจจุบัน คือ P2
45	เตรียมค่าสำหรับเฟรมต่อไป
46	คำตอบค่าของคาบเป็นจำนวนแซมเปิล ถ้าต้องการแปลงเป็นมิลลิวินาที คูณคำตอบด้วย 0.1

2.5 การคำนวนเพื่อสังเคราะห์เสียง

ข้อมูลเสียงที่ได้จากการวิเคราะห์ แบ่งได้เป็น 3 ส่วนคือ

1) พารามิเตอร์ของฟิลเตอร์หรือสัมประสิทธิ์พาร์คอร์ มีจำนวนข้อมูลในเฟรมหนึ่งๆ เท่ากับจำนวนออร์เดอร์ของฟิลเตอร์
2) ค่าผิดพลาดยกกำลังสองรวม ใช้เป็นข้อมูลบอกถึงปริมาณพลังงานของสัญญาณ Excitation ที่จะป้อนเข้าสู่ฟิลเตอร์ขณะทำการสังเคราะห์เสียง
3) ค่าคาบของเสียง ถ้าคาบมีค่าเท่ากับ 0 หมายความว่าเสียงไม่มีคาบ หรือเป็นเสียงไม่ก้อง สัญญาณ Excitation ใช้ในการสังเคราะห์เสียงจะกำหนดให้เป็น Random Pulse กรณีที่คาบของเสียงมีค่าอยู่ในช่วงที่ถูกต้องตามกรรมวิธี SIFT คืออยู่ระหว่าง 4-20 ms หมายความว่าเสียงเป็นเสียงก้อง สัญญาณ Excitation จะกำหนดให้เป็น Pulse Train ที่มีระยะระหว่างพัลส์เท่ากับคาบของเสียง

2.5.1 ขนาดของสัญญาณ Excitation

ตามสมการการสังเคราะห์เสียง เสียงที่ผ่านออกมาจากฟิลเตอร์สังเคราะห์หรือส่วนกลับของอินเวิร์สฟิลเตอร์ จะต้องมีพลังงานรวมในหนึ่งเฟรมเท่ากับพลังงานรวมของเสียงจริงในหนึ่งเฟรมด้วย นี่คือเงื่อนไขที่เรียกว่า Energy Matching เขียนเป็นสมการได้

$$\sum_{n=n_0}^{n} \hat{s}^2(n) = \sum_{n=n_0}^{n} s^2(n) \qquad (2.79)$$

$\hat{s}(n)$ คือสัญญาณที่สังเคราะห์ขึ้น และ $s(n)$ คือสัญญาณสุ่มของเสียงจริง อย่างไรก็ตามเงื่อนไขนี้ไม่เหมาะสมในแง่ปฏิบัติเพราะมีความยุ่งยากในการคำนวนหาค่าอัตราขยาย (Gain) Markel และ Gray ได้เสนอวิธีที่สะดวกในการคำนวนขึ้นซึ่งใช้กับวิธี อัตสหสัมพันธ์ โดยอาศัย ค่าผิดพลาดยกกำลังสองรวมหรือกำหนดให้ $\delta = \alpha^{\frac{1}{2}}$ สมมติในหนึ่งเฟรมมีสัญญาณอยู่ N แซมเปิล และคำนวนค่าผิดพลาดยกกำลังสองรวมได้ α

ในกรณีของเสียงไม่ก้องใช้ Random Number $\{g(n)\}$ ซึ่งมีความแปรปรวนเท่ากับ σ_g^2 เป็นสัญญาณ Excitation ความสัมพันธ์ของพลังงานในหนึ่งเฟรม คือ

$$Ne^2(n) = g^2(n)\sigma^2/\sigma_g^2$$

หรือ

$$e(n) = g(n)\sigma/(\sigma_g N^{\frac{1}{2}}) \qquad (2.80)$$

$e(n)$ คือ สัญญาณ Excitation ถ้าค่า $g(n)$ กระจายอย่างสม่ำเสมอในช่วง $-b$ ถึง b จะได้

$$\sigma_g = b/\sqrt{3}$$

ถ้าให้ $b=1$ จะได้ $e(n) = g(n)\sigma\sqrt{3/N}$ (2.81)

ในกรณีที่เสียงเป็นเสียงก้องสัญญาณ Excitation จะเป็น Pulse Train ห่างกันเท่ากับคาบของเสียงซึ่งแปลงเป็นจำนวนแซมเปิลเท่ากับ I ความสัมพันธ์ของพลังงานคือ

$$e^2(n)\delta_{n,\ell_I}{}^{N/I} = \sigma^2 \qquad n = 0,1,2,3,\ldots$$

หรือ

$$e(n) = \begin{cases} \sigma(I/N)^{\frac{1}{2}} & n = 0,I,2I,\ldots \\ 0 & n \neq 0,I,2I,\ldots \end{cases} \qquad (2.82)$$

ในกรณีนี้จะเห็นว่าค่าเฉลี่ยของสัญญาณ Excitation ไม่เท่ากับ 0 ซึ่งอาจทำให้เกิดการเพิ่มพูนของ D.C bias ทำให้เกิดเสียงรบกวนความถี่ต่ำ การแก้ไขปัญหาคือ จัดให้ค่าเฉลี่ยของสัญญาณ Excitation เป็น 0 คือ

$$e(n) = \begin{cases} \sigma(I/N)^{\frac{1}{2}} & n = 0,I,2I,\ldots \qquad (2.83\ \text{ก}) \\ -\sigma(I/N)^{\frac{1}{2}}/(I-1) & n \neq 0,I,2I,\ldots \qquad (2.83\ \text{ข}) \end{cases}$$

Markel และ Gray [14] เสนอให้ใช้ค่า σ เป็นข้อมูลเสียงที่ส่งจากภาควิเคราะห์ไปสู่ภาคสังเคราะห์ และการคำนวนเกี่ยวกับขนาดของสัญญาณ Excitation จะอยู่ในส่วนของการสังเคราะห์

2.5.2 ฟิลเตอร์สังเคราะห์ตามวิธี PARCOR

จากสมการ (2.63 ก) และ (2.63 ข) และรูป 2.16 แสดงโครงสร้างและความสัมพันธ์ของสัญญาณในแต่ละสเตจของอินเวิร์สฟิลเตอร์ ซึ่งอินพุทคือสัญญาณสุ่มของเสียง $\{x(n)\}$ และเอาท์พุท คือ ค่าผิดพลาด $\{e(n)\}$ ถ้าเรามองสัญญาณย้อนหลังจากสเตจ M ไปหาสเตจ 1 โดยพิจารณาเฉพาะสัญญาณการทำนายไปข้างหน้า หรือ $x_m^+(n)$ จากสมการ (2.63 ก) จะได้

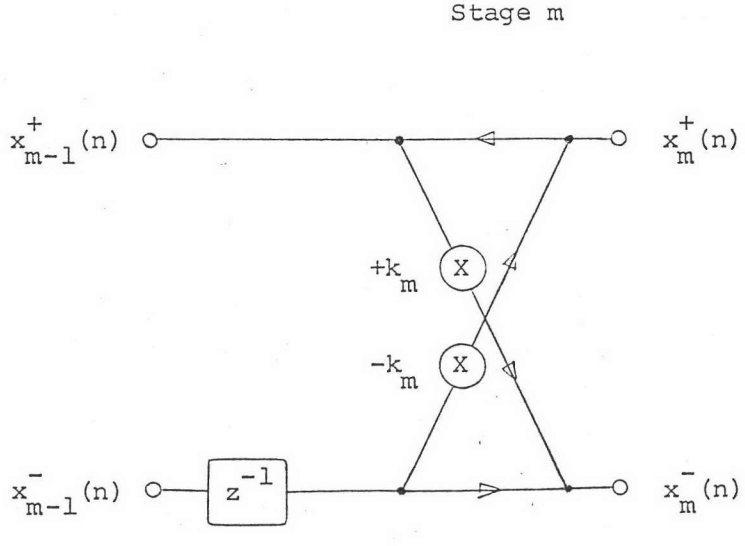
$$x_{m-1}^+(n) = x_m^+(n) - k_m x_{m-1}^-(n) \tag{2.84}$$

โดยที่ $x_M^+(n) = e(n)$ และ $x(n) = x_0^+(n)$ (2.85)

ในส่วนของการทำนายย้อนหลังหรือ $x_{m-1}^-(n)$ ยังคงเป็นอย่างเดิมคือ

$$x_m^-(n) = k_m x_{m-1}^+(n-1) + x_{m-1}^-(n-1) \tag{2.86}$$

โดยที่ $x_0^-(n) = x_0^+(n-1)$ (2.87)



รูป 2.27 โครงสร้างของฟิลเตอร์สังเคราะห์

จากสมการ (2.84) และ(2.86) นำไปสร้างเป็น Flow Diagram ของฟิลเตอร์สังเคราะห์ ได้ดังรูป 2.27 เมื่อเปรียบเทียบระหว่างฟิลเตอร์วิเคราะห์กับฟิลเตอร์สังเคราะห์ จะมีโครงสร้างเหมือนกันทุกประการ ที่แตกต่างกันคือทิศทางการเดินทางของสัญญาณ ในอินเวิร์สฟิลเตอร์หรือฟิลเตอร์วิเคราะห์ อินพุทคือสัญญาณเสียงพูดจริง x(n) เข้าที่สเตจที่ 1 ทิศทางของสัญญาณ ทำนายไปข้างหน้า หรือ $x_m^+(n)$ จะเดินทางจากซ้ายไปขวา เอาท์พุทของฟิลเตอร์ คือค่าผิดพลาด e(n) ซึ่งเท่ากับสัญญาณการทำนายไปข้างหน้าที่สเตจ M สัมประสิทธิ์พาร์คอร์ k_m ใช้ค่าบวกคูณในการทำนายทั้งสองทิศทาง ในฟิลเตอร์สังเคราะห์ อินพุท คือสัญญาณ Excitation เข้าที่สเตจ M สัญญาณจะวิ่งจากขวาไปซ้าย ในส่วนของการทำนายไปข้างหน้า ($x_m^+(n)$) ตามสมการ (2.83) โดยสัมประสิทธิ์ k_m ที่คูณกับสัญญาณทำนายย้อนหลังใช้ค่าลบ ในส่วนของการทำนายย้อนหลัง ($x_m^-(n)$) การเดินทางของสัญญาณจากซ้ายไปขวาเหมือนอินเวอสฟิลเตอร์ สัมประสิทธิ์ k_m คูณด้วยค่าบวก เอาท์พุทของฟิลเตอร์สังเคราะห์หรือเสียงที่สังเคราะห์ขึ้น จะเท่ากับสัญญาณทำนายไปข้างหน้าที่สเตจ 0 หรือ $x_0^+(n)$ ซึ่งป้อนย้อนกลับผ่าน Delay เป็น $x_0^-(n) = x_0^+(n-1)$ สมการ(2.84) ถึงสมการ (2.86) ถูกนำมาเขียนเป็นโปรแกรมย่อย TWOMUL ทำหน้าที่เป็นฟิลเตอร์สังเคราะห์ ซึ่งคำนวณทีละแซมเปิล รูป 2.28 แสดงโปรแกรมย่อย TWOMUL และคำอธิบาย

โปรแกรมย่อย SYNT คือ โปรแกรมทำหน้าที่สังเคราะห์เสียงพูดในหนึ่งเฟรม ซึ่งมีแซมเปิลทั้งหมดเท่ากับ N อาศัยสมการต่างๆ ที่คำนวณขนาดของสัญญาณ Excitation ร่วมกับโปรแกรมย่อย TWOMUL รูป 2.29 แสดงโปรแกรมย่อย SYNT และคำอธิบายเทียบกับสมการต่างๆ โปรแกรมย่อย SYNT ถูกนำไปใช้ในโปรแกรม SYNTX ซึ่งใช้ในการสังเคราะห์เสียงเป็นคำๆ สำหรับการทดสอบ

```
1  REM   ** SUBROUTINE TWOMUL
2  FOR I = 1 TO M
3 II = M - I
4 JJ = II + 1
5 DRV = DRV - RC(JJ) * RBUF(II)
6 RBUF(JJ) = RBUF(II) + RC(JJ) * DRV
7  NEXT I
8 RBUF(0) = DRV
9 YOUT = RBUF(0)
10  RETURN
```

รูป 2.28 โปรแกรมย่อย TWOMUL

บรรทัดที่	หน้าที่
2-7	คำนวณซ้ำจนครบ M สเตจ
5	คำนวณสัญญาณการทำนายไปข้างหน้าตามสมการ (2.84)
6	คำนวณสัญญาณการทำนายย้อนหลังตามสมการ (2.85)
8	เทียบค่าตามสมการ (2.87) เพื่อการคำนวณในแซมเบิลต่อไป
9	เทียบค่าตามสมการ (2.85) YOUT เท่ากับเอาท์พุทที่ได้

```
1  REM    ** SUBROUTINE SYNT
2 V% = 1
3 SG = SQR(AL):NC = SG * SQR(3/N)
4  IF PT(FI) < 30 THEN V% = 0:GOTO 7
5 IP = PT(FI)
6 AM = SG * SQR(IP/N):AU = - AM / IP - 1
7  FOR J = 1 TO N
8  IF V% = 0 THEN DRV = NC * 2 * ( RND (1) - 0.5): GOTO 12
9 DRV = AU
10  IF IL > IX THEN IX = IP:IL = 1:DRV = AM
11 IL = IL + 1
12  GOSUB 3000: REM    GOSUB TWOMUL
13 X(J) = YOUT
14  IF  ABS (X(J)) >  ABS (XM) THEN XM = X(J)
15  NEXT
16  RETURN
```

รูป 2.29 โปรแกรมย่อย SYNT

คำอธิบายโปรแกรม SYNT

บรรทัดที่	หน้าที่
3	คำนวณ SIGMA = $(ALPHA)^{\frac{1}{2}}$ และ Gain ของ Noise ตามสมการ (2.81)
4	ตรวจสอบว่า เสียงเป็นเสียงก้องหรือเสียงไม่ก้อง
5-6	คำนวณตามสมการ (2.83 ก) และ (2.83 ข)
8	ถ้าเป็นเสียงก้อง สัญญาณ Excitation จะเป็นไปตามสมการ (2.83) ถ้าเป็นเสียงไม่ก้อง สัญญาณ Excitation จะเป็นไปตามสมการ (2.81)
10	กรณีเสียงก้อง นับจำนวนแซมเปิลเพื่อเลือกใช้สมการ (2.83 ก) หรือ สมการ (2.83 ข)
12	คำนวณฟิลเตอร์สังเคราะห์ด้วยโปรแกรมย่อย TWOMUL
14	ค่าสูงสุดของสัญญาณที่สังเคราะห์ขึ้นเพื่อนอร์มัลไลซ์สัญญาณเสียงทั้งคำ