บทที่ 2

แนวความคิดและทฤษฎี

ในบทนี้จะได้กล่าวถึงประเภทของวิธีการบีบขนาดข้อมูลโดยทั่วไป ประเภทของวิธีการบีบข้อมูลเสียง หลักการในการเข้ารหัสแบบทำนายเชิงเส้น หลักการบีบข้อมูลเสียงพูดด้วยวิธีการทำนายเชิงเส้นโดยการกระตุ้นด้วยรหัส (Code-Excited Linear Prediction - CELP) และหลักการบีบข้อมูลเสียงพูดด้วยวิธีการทำนายช่วงยาวโดยการกระตุ้นด้วยพัลส์เศษค้าง (Residual Pulse Excitation - Long Term Prediction - RPE-LTP) ในรายละเอียดจะเน้นที่การเข้ารหัสแบบทำนายเชิงเส้นซึ่งเป็นหลักการพื้นฐานที่ใช้ในวิธีการบีบข้อมูลเสียงแบบอื่น ๆ โดยเริ่มตั้งแต่การสร้างแบบจำลองทางคณิตศาสตร์ของอวัยวะที่ใช้ในการเปล่งเสียงของมนุษย์ พารามิเตอร์ต่าง ๆ ของแบบจำลองรวมทั้งวิธีการหาค่าพารามิเตอร์เหล่านี้จากสัญญาณเสียง

วิธีการทั่วไปในการบีบขนาดข้อมูล

Pohlmann (1991) ได้กล่าวถึงวิธีการต่าง ๆ ที่ใช้ในการบีบขนาดข้อมูลซึ่งสามารถแบ่งได้เป็น 2 ประเภทใหญ่ ๆ ด้วยกัน วิธีแรกเป็นวิธีการบีบขนาดข้อมูลโดยไม่สูญเสียรายละเอียดข้อมูล (lossless compression) ส่วนอีกวิธีคือวิธีการบีบขนาดข้อมูลโดยมีการสูญเสียรายละเอียดข้อมูล (lossy compression)

1. การบีบขนาดข้อมูลโดยไม่สูญเสียรายละเอียดข้อมูล

การบีบขนาดข้อมูลประเภทนี้เมื่อบีบข้อมูลและคลายคืนมาจะได้ข้อมูลที่เหมือนเดิมทุกประการ เหมาะสำหรับข้อมูลที่ต้องการความถูกต้องห้ามผิดพลาดแม้แต่น้อย เช่น โปรแกรมคอมพิวเตอร์ แฟ้มข้อมูลของฐานข้อมูล ฯลฯ วิธีบีบขนาดข้อมูลแบบนี้ได้แก่ วิธีฮัฟแมน (Huffman coding) วิธีซิฟ-เลมเพล (Ziv-Lempel coding) หรือโปรแกรมบีบข้อมูล pkzip ที่มีใช้กันบนคอมพิวเตอร์พีซีก็จัดอยู่ในประเภทนี้ แม้ว่าวิธีนี้จะให้ความถูกต้อง

เที่ยงตรงแต่ก็มีข้อเสียคืออัตราส่วนการบีบข้อมูลไม่สูงนัก วิธีประเภทนี้มักไม่เหมาะต่อการบีบขนาดข้อมูลเสียง รายละเอียดของขั้นตอนวิธีเหล่านี้จะไม่กล่าวถึงในที่นี้

2. <u>การบีบขนาดข้อมูลโดยมีการสูญเสียรายละเอียดข้อมูล</u>

การบีบขนาดข้อมูลประเภทนี้จะทำให้เกิดการผิดเพี้ยนของข้อมูลไปบ้างเมื่อทำการคลายข้อมูลคืน เหมาะสำหรับข้อมูลที่ไม่จำเป็นต้องมีความถูกต้องของข้อมูลเหมือนต้นฉบับทุกประการ ข้อมูลที่เหมาะกับวิธีประเภทนี้ได้แก่ข้อมูลจำพวกภาพและเสียงเพราะถึงแม้จะเกิดการผิดเพี้ยนไปบ้าง ภาพหรือเสียงนั้นก็ยังสามารถนำไปใช้ได้ ตัวอย่างวิธีบีบขนาดข้อมูลประเภทนี้ได้แก่ การทำคอมแพนดิงด้วยวิธีเอลลอว์ (A law) หรือมิวลอว์ (μ law) และการเข้ารหัสแบบทำนายเชิงเส้น เป็นต้น โดยทั่วไปวิธีประเภทนี้จะให้อัตราการบีบขนาดข้อมูลดีกว่าแบบแรก

<u>วิธีการในการบีบข้อมูลเสียง</u>

มีการศึกษาวิธีการต่าง ๆ ในการเข้ารหัสข้อมูลเสียง บางวิธีที่มีประสิทธิภาพสูงก็กลายเป็นมาตรฐานที่ใช้กันอย่างแพร่หลาย อย่างเช่นในด้านเครือข่ายสื่อสารซึ่งไม่จำเป็นต้องใช้คุณภาพของเสียงในระดับความคมชัดสูง (high fidelity) มาตรฐานที่มีใช้กันได้แก่ การทำคอมแพนดิงตามมาตรฐาน CCITT G.711 โดยใช้มิวลอว์ซึ่งมีใช้ในประเทศสหรัฐอเมริกาและญี่ปุ่น หรือใช้เอลลอว์ในประเทศแถบยุโรปรวมทั้งประเทศไทยด้วย การทำคอมแพนดิงด้วยเอลลอว์และมิวลอว์เป็นมาตรฐานที่ยังใช้กันอยู่ในระบบโทรศัพท์ในทุกวันนี้ อัตราข้อมูลที่ได้จากการทำคอมแพนดิงดังกล่าวคือ 64 Kbps (Frerking, 1994) ต่อมามีการพัฒนาวิธีการที่สามารถบีบขนาดข้อมูลลงต่ำกว่า 64 Kbps ได้แก่วิธีเอดีพีซีเอ็ม (ADPCM - Adaptive Differential Pulse Code Modulation) ที่อัตราข้อมูลต่าง ๆ เช่น 32 Kbps และ 16 Kbps วิธีแอลดี-ซีอีแอลพี (LD-CELP - Low Delay CELP) ที่อัตราข้อมูล 16 kbps (Kondos, 1994) สำหรับในกลุ่มที่ให้คุณภาพของเสียงในระดับสูงก็ได้แก่เอซี3 (AC-3 - Audio-Compression-3) พัฒนาโดยห้องปฏิบัติการดอลบี (Dolby) สำหรับสร้างเสียงแบบรอบทิศทางในระบบเสียงระดับสูง (Nath, 1996) เป็นต้น

การบีบขนาดข้อมูลเสียงพูดเพื่อใช้ในงานด้านการสื่อสารที่ทำได้ในอัตราบีบอัดข้อมูลที่สูงและยังมีคุณภาพของเสียงพูดที่จัดว่าดีเกิดขึ้นหลังจากที่มีผู้คิดวิธีการเข้ารหัสแบบทำนายเชิงเส้นหรือแอลพีซีขึ้น การประยุกต์ใช้ที่เห็นได้ชัดได้แก่การบีบข้อมูลเสียงพูดในระบบโทรศัพท์เคลื่อนที่ซึ่งใช้วิธีการเข้ารหัสสัญญาณเสียงพูดที่มีพื้นฐานมาจากแอลพีซีเพื่อให้ได้อัตรา

ข้อมูลต่ำเพราะในระบบโทรศัพท์เคลื่อนที่นั้นมีช่วงความถี่ที่ใช้ได้จำกัด การที่จะให้บริการได้มากช่องสัญญาณขนาดไหนก็ขึ้นอยู่กับว่าสามารถบีบข้อมูลเสียงได้มากน้อยเพียงใดด้วย Frerking (1994) ได้กล่าวถึงตัวอย่างวิธีการบีบขนาดข้อมูลเสียงพูดที่มีพื้นฐานจากแอลพีซีที่ใช้กันอยู่ได้แก่วิธี วีเอสอีแอลพี (VSELP - Vector Sum Excited LPC) ที่อัตราข้อมูล 8.0 Kbps มีใช้ในระบบดิจิตอลเซลลูลาร์ของอเมริกา วิธีซีอีแอลพีที่อัตราข้อมูล 4.8 Kbps (มาตรฐานเฟดเดอรัล 1016) และวิธีอาร์พีอี-แอลทีพี ในระบบโทรศัพท์เคลื่อนที่จีเอสเอ็มเป็นต้น Frerking ยังได้กล่าวถึงวิธีการบีบขนาดข้อมูลเสียงพูดซึ่งอาศัยหลักการแอลพีซีว่าได้มีผู้ทำการวิจัยจนได้อัตราข้อมูลต่ำในระดับ 400-800 bps หรือแม้กระทั่งต่ำกว่านั้น แต่ปัญหาที่เกิดขึ้นกับวิธีที่ให้อัตราข้อมูลต่ำพวกนี้คือคุณภาพของเสียงพูดไม่เป็นที่ยอมรับ

Kondos (1995) ได้จำแนกวิธีการเข้ารหัสข้อมูลเสียงซึ่งมีการทำวิจัยอย่างกว้างขวางในปัจจุบันไว้เป็นสามประเภทด้วยกัน ได้แก่ กลุ่มตัวเข้ารหัสรูปคลื่น (waveform coder) กลุ่มตัวเข้ารหัสเสียง (vocoder = voice + coder) และกลุ่มตัวเข้ารหัสลูกผสม (hybrid coder)

1. ประเภทตัวเข้ารหัสรูปคลื่น

วิธีประเภทนี้เมื่อทำการบีบและคลายข้อมูลเสียงคืนมันจะพยายามรักษารูปคลื่นของสัญญาณเสียงไว้ให้เหมือนเดิมมากที่สุด โดยที่ไม่สนใจว่ารูปคลื่นของสัญญาณเป็นอย่างไร นั่นคือไม่สนใจลักษณะเฉพาะของสัญญาณเสียงพูด วิธีการบีบข้อมูลประเภทนี้ได้แก่ การทำคอมแพนดิง เอดีพีซีเอ็ม และ ซีวีเอสดี (CVSD - Continuously Variable Slope Delta Modulation) เป็นต้น โดยทั่วไปสำหรับวิธีประเภทนี้เมื่อมีสัญญาณรับเข้า (input) หนึ่งตัวอย่างสัญญาณก็สามารถสร้างสัญญาณส่งออก (output) ได้หนึ่งสัญญาณทันที การวัดคุณภาพของสัญญาณเสียงที่ได้จากวิธีประเภทนี้สามารถวัดได้ในลักษณะตัวอย่างสัญญาณส่งออกเทียบกับตัวอย่างสัญญาณรับเข้าได้เลย วิธีที่ใช้วัดได้แก่ การวัดอัตราส่วนสัญญาณต่อสัญญาณรบกวน (signal to noise ratio - SNR) เป็นต้น ลักษณะเด่นของวิธีประเภทนี้คือให้คุณภาพสัญญาณสูงแต่ข้อเสียก็คืออัตราข้อมูลก็สูงด้วย

2. ประเภทตัวเข้ารหัสเสียง

วิธีประเภทนี้เป็นวิธีที่แทบจะตรงข้ามกับวิธีประเภทแรก นั่นคือตัวเข้ารหัสเสียงจะพยายามดึงลักษณะเฉพาะของเสียงออกมาเป็นพารามิเตอร์ในขณะที่มันทำการวิเคราะห์สัญญาณเสียง พารามิเตอร์เหล่านี้จะถูกนำไปใช้ในการสังเคราะห์สัญญาณเสียงขึ้นมาใหม่ การวัดคุณภาพของเสียงที่สังเคราะห์ออกมาด้วยวิธี SNR ใช้ไม่ได้กับวิธีประเภทนี้ วิธีทั่วไปที่ใช้ในการวัดคือวิธีคะแนนความเห็นเฉลี่ย (Mean Opinion Score - MOS) ตัวอย่างวิธีการบีบข้อมูลเสียงในประเภทนี้ได้แก่แอลพีซี ตัวเข้ารหัสเสียงเป็นวิธีหลักในการบีบข้อมูลเสียงที่ความถี่

ต่ำกว่า 4.8 Kbps ลงไป ลักษณะเด่นของวิธีประเภทนี้คือให้อัตราข้อมูลต่ำ แต่คุณภาพของสัญญาณเสียงที่ได้ก็ต่ำกว่าวิธีประเภทตัวเข้ารหัสรูปคลื่นและตัวเข้ารหัสลูกผสม

3. ประเภทตัวเข้ารหัสลูกผสม

วิธีประเภทนี้เป็นวิธีที่นำข้อดีของสองประเภทแรกมารวมกัน ทำให้ได้อัตราข้อมูลต่ำกว่าวิธีประเภทตัวเข้ารหัสรูปคลื่น แต่ขณะเดียวกันก็ให้คุณภาพของสัญญาณสูงกว่าวิธีประเภทตัวเข้ารหัสเสียง วิธีประเภทนี้ได้แก่ ซีอีแอลพี และ อาร์พีอี-แอลทีพี ซึ่งจะได้กล่าวถึงต่อไป สำหรับการวัดคุณภาพของสัญญาณเสียงที่ได้ก็ใช้วิธีคะแนนความเห็นเฉลี่ยเช่นกัน

การบีบข้อมูลเสียงพูดโดยวิธีการเข้ารหัสแบบทำนายเชิงเส้น

การเข้ารหัสแบบทำนายเชิงเส้นหรือแอลพีซี เป็นวิธีหนึ่งที่สามารถบีบขนาดข้อมูลเสียงพูดลงได้มาก ข้อดีที่ได้คือทำให้ระบบมีราคาถูกและใช้ทรัพยากรอย่างคุ้มค่า ตัวอย่างการประยุกต์การเข้ารหัสแอลพีซีที่มีใช้กันได้แก่ การมัลติเพล็กซ์สัญญาณเสียงพูดที่เข้ารหัสด้วยวิธีแอลพีซีที่อัตราข้อมูล 2400 bps จำนวน 4 ช่องสัญญาณโดยใช้เทคนิค QAM (Quadrature Amplitude Modulation) ผ่านสายโทรศัพท์แบบอนาล็อกที่อัตราบอด (baud rate) 9600 bps สำหรับข้อเสียของการเข้ารหัสแบบนี้ที่อัตราข้อมูลต่ำกว่า 2400 bps ก็คือเสียงพูดที่ได้หลังการถอดรหัสขาดความเป็นธรรมชาติและฟังคล้ายเสียงพูดของหุ่นยนต์ (Frerking, 1994)

การเข้ารหัสด้วยวิธีแอลพีซีมักจะมีประโยชน์กับระบบที่มีแบนด์วิดธ์จำกัด และไม่สามารถที่จะใช้การเข้ารหัสแบบพีซีเอ็มได้ ตัวอย่างเช่นการส่งสัญญาณเสียงในช่วงความถี่วิทยุย่าน HF ซึ่งแบนด์วิดธ์ของสัญญาณมักจะถูกจำกัดไว้เพียง 3 KHz

การเข้ารหัสด้วยวิธีแอลพีซีในการใช้งานจริงมักจะใช้กันที่อัตราข้อมูล 2400 bps ถึงแม้ว่าจะมีการพัฒนาจนสามารถทำได้ที่ความถี่ต่ำกว่านี้ กระทั่งลงไปจนถึง 400 bps แล้วก็ตาม นอกจากนี้ยังสามารถใส่ข้อมูลเพื่อตรวจสอบแก้ไขความผิดพลาดของข้อมูลแบบง่าย ๆ ลงไปในสัญญาณที่อัตราความถี่ 2400 bps ทำให้สัญญาณสามารถทนต่อความผิดพลาดของข้อมูลที่เข้ารหัสแล้วในระดับ 1-2 เปอร์เซ็นต์ได้ (Frerking, 1994)

1. การใช้เสียงพูดในการสื่อสารของมนุษย์

Deller, Proakis และ Hansen (1993) ได้กล่าวถึงรายละเอียดการทำงานของอวัยวะที่ใช้ในการเปล่งเสียงซึ่งพอสรุปได้โดยย่อดังนี้ การพูดเกิดขึ้นเมื่อผู้พูดเกิดความคิดขึ้น

และต้องการที่จะถ่ายทอดความคิดนั้นไปยังผู้รับฟัง สมองจะเปลี่ยนโครงสร้างความคิดไปเป็นโครงสร้างทางภาษาโดยคำนึงถึงหลักไวยากรณ์ทางภาษาตามที่ได้เรียนรู้มาและถูกส่งเป็นสัญญาณไปกระตุ้นกล้ามเนื้ออวัยวะที่ใช้ในการเปล่งเสียงก่อให้เกิดเป็นเสียงพูดขึ้นมา เสียงที่เปล่งออกมาเป็นคลื่นความดันอากาศส่งออกไปยังผู้ฟังและตัวผู้พูดเองก็ได้ยินด้วย อวัยวะที่ใช้ในการรับฟังของผู้พูดจะสั่นไปตามคลื่นความดันอากาศที่ตัวเขาเองเปล่งออกมา ก่อให้เกิดเป็นสัญญาณพัลส์ (pulse) ป้อนกลับไปยังสมอง สัญญาณที่ป้อนกลับมานี้ทำให้ผู้พูดใช้เป็นข้อมูลในการควบคุมอวัยวะที่ใช้เปล่งเสียงทำให้เขาสามารถส่งเสียงพูดได้อย่างเหมาะสมต่อเนื่องกันไป สำหรับสมองของผู้ฟังนั้นก็จะมีการตอบสนองต่อสัญญาณพัลส์ที่เกิดจากเสียงของผู้พูดโดยมีการตีความหมายของสัญญาณตามไวยากรณ์ทางภาษาที่ได้จากกระบวนการเรียนรู้ในอดีตและดึงเอาความคิดของผู้พูดออกมา จากการศึกษาของนักวิทยาศาสตร์และผู้ที่วิจัยทางด้านการเปล่งเสียงทำให้ทราบจุดที่น่าสนใจในระบบการพูดและรับฟังของมนุษย์หลาย ๆ ประการ อย่างเช่น ในกรณีที่รับฟังผู้พูดหลายคน ผู้ฟังก็สามารถเลือกว่าจะฟังใครได้โดยอาศัยการแยกแยะความแตกต่างของเฟส (phase) ของเสียงที่ได้ยินโดยหูซ้ายและหูขวา ดังนั้นจึงพบว่าสำหรับผู้ที่หูข้างใดข้างหนึ่งพิการนั้นจะขาดความสามารถนี้ จุดอ่อนจุดหนึ่งของระบบการรับฟังของมนุษย์ได้แก่ไม่สามารถแยกแยะความแตกต่างของความถี่สองความถี่เมื่อมันมีค่าใกล้เคียงกันโดยจะได้ยินความถี่เป็นความถี่เดียวซึ่งเท่ากับความแตกต่างของสองความถี่นั้น เช่นถ้ามีแหล่งกำเนิดเสียงสองแหล่งมีความถี่ 1000 Hz และ 1040 Hz เสียงที่ได้ยินจะกลายเป็นเสียงความถี่ 40 Hz เพียงความถี่เดียว

2. <u>หลักการเข้ารหัสแบบทำนายเชิงเส้น</u>

หลักการเข้ารหัสแบบทำนายเชิงเส้นมาจากการศึกษาการทำงานของอวัยวะที่ใช้ในการเปล่งเสียงของมนุษย์ แล้วพยายามทำแบบจำลองเชิงคณิตศาสตร์ขึ้นมาเพื่อเลียนแบบการพูดของคน Frerking (1994) ได้กล่าวถึงรายละเอียดของหลักการแอลพีซีซึ่งจะได้กล่าวถึงรายละเอียดดังต่อไปนี้ การบีบข้อมูลเสียงพูดด้วยวิธีนี้เริ่มต้นโดยนำข้อมูลเสียงพูดมาแบ่งเป็นช่วง ๆ ช่วงละประมาณ 20 ms แล้วทำการหาพารามิเตอร์ของแบบจำลองในช่วงเวลาดังกล่าว แล้วแทนข้อมูลเสียงพูดในช่วงนั้นด้วยพารามิเตอร์เหล่านี้ ซึ่งแน่นอนว่ามีขนาดน้อยกว่าข้อมูลเสียงพูดรับเข้า ส่วนการคลายข้อมูลก็ทำโดยนำพารามิเตอร์มาสร้างแบบจำลองแล้วก็ใส่การกระตุ้นที่เหมาะสมเข้าไป ผลที่ได้ออกมาก็คือเสียงพูดเสมือนเสียงต้นกำเนิดนั่นเอง

ในระหว่างที่เปล่งเสียงออกมา เส้นเสียงจะสั่นที่ความถี่ค่าหนึ่งซึ่งเรียกว่าพิทช์ (pitch) ผลที่ได้ทำให้เกิดเป็นพัลซ์กระตุ้นซึ่งเต็มไปด้วยฮาร์โมนิค (harmonic) ซึ่งเป็นสัญญาณที่มีความถี่เป็นทวีคูณของความถี่พิทช์ สัญญาณกระตุ้นที่เกิดขึ้นจะถูกมอดูเลต (modulate) ด้วยโพรงต่าง ๆ ได้แก่ ช่องในลำคอ ปาก และโพรงจมูก ความถี่ต่าง ๆ จะขยายเพิ่มมากขึ้นหรือลดลงทำให้เกิดเป็นเสียงขึ้นมา ความถี่กำทอน (resonance frequency) ของ

เสียงในระบบนี้เรียกว่าฟอร์แมนต์ (formant) พิทช์เฉลี่ยสำหรับเสียงผู้ชายประมาณ 130 Hz ในขณะที่เสียงของผู้หญิงจะสูงกว่าประมาณหนึ่งเท่า

นอกจากเสียงที่เกิดจากการกระตุ้นของเส้นเสียงซึ่งเรียกเสียงที่เกิดขึ้นนี้ว่าเสียงโฆษะ (voiced) แล้วก็ยังมีเสียงอโฆษะ (unvoiced) เช่นอักษร 'f' ในคำว่า 'five' และ 's' ในคำว่า 'six' และ 'ส' ในคำว่า 'เสือ' เป็นต้น ซึ่งในกรณีนี้จะไม่มีการสั่นของเส้นเสียง แต่จะมีการกระตุ้นของอวัยวะกำเนิดเสียงอื่น ๆ ทำให้เกิดการไหลวนของอากาศและเกิดเป็นสัญญาณไวท์นอยส์ (white noise) ซึ่งประกอบด้วยทุกความถี่ สัญญาณไวท์นอยส์ดังกล่าวก็จะถูกมอดูเลตโดยโพรงต่าง ๆ ที่กล่าวไว้ข้างต้นด้วยเช่นกัน

แนวความคิดพื้นฐานของการเข้ารหัสด้วยวิธีแอลพีซีก็คือการจำลองการทำงานของระบบเปล่งเสียงของมนุษย์ การที่ระบบสามารถบีบขนาดของข้อมูลได้เนื่องจากไม่ต้องส่งสัญญาณเสียงไปทางปลายทางแต่จะส่งพารามิเตอร์ต่าง ๆ ของระบบที่จำเป็นต้องใช้ในการสร้างเสียงต้นฉบับกลับคืนมาไปแทน พารามิเตอร์ต่าง ๆ ได้แก่

- พิทช์
- การตัดสินใจว่าเป็นสัญญาณที่กระตุ้นโดยเส้นเสียงหรือไวท์นอยส์
- สัมประสิทธิ์ของตัวกรองสังเคราะห์ซึ่งจะได้กล่าวถึงในส่วนถัดไป
- อัตราการขยายสัญญาณ

3. <u>แบบจำลองของระบบการเปล่งเสียงของมนุษย์</u>

ในทางคอมพิวเตอร์เมื่อนำสัญญาณเสียงมาใช้ เสียงนั้นก็จะเป็นลำดับของสัญญาณที่ได้จากการแปลงสัญญาณจากอนาล็อกมาเป็นดิจิตอลและเก็บเรียงต่อเนื่องกันไปในหน่วยความจำ ในการอ้างอิงเวลาของสัญญาณจำเป็นจะต้องรู้อัตราการสุ่มสัญญาณว่ามีค่าเป็นเท่าไร สามารถใช้ตัวเลขจำนวนเต็มที่แทนระยะห่างระหว่างตัวอย่างสัญญาณเป็นตัวบอกเวลาได้ ตัวอย่างเช่นการบอกคาบของพิทช์ (ระยะเวลาระหว่างพัลส์ที่ติดกันของการกระตุ้นของเสียงแบบโฆษะ) สามารถบอกเป็นตัวเลขเช่น 132 ถ้าอัตราการสุ่มสัญญาณคือ 22000 ตัวอย่าง/วินาที จะได้คาบของพิทช์เป็น:

คาบ (วินาที) = คาบ (ตัวเลข) / อัตราการสุ่มสัญญาณ (ตัวอย่าง/วินาที)
= 132/22000
= .006 วินาที

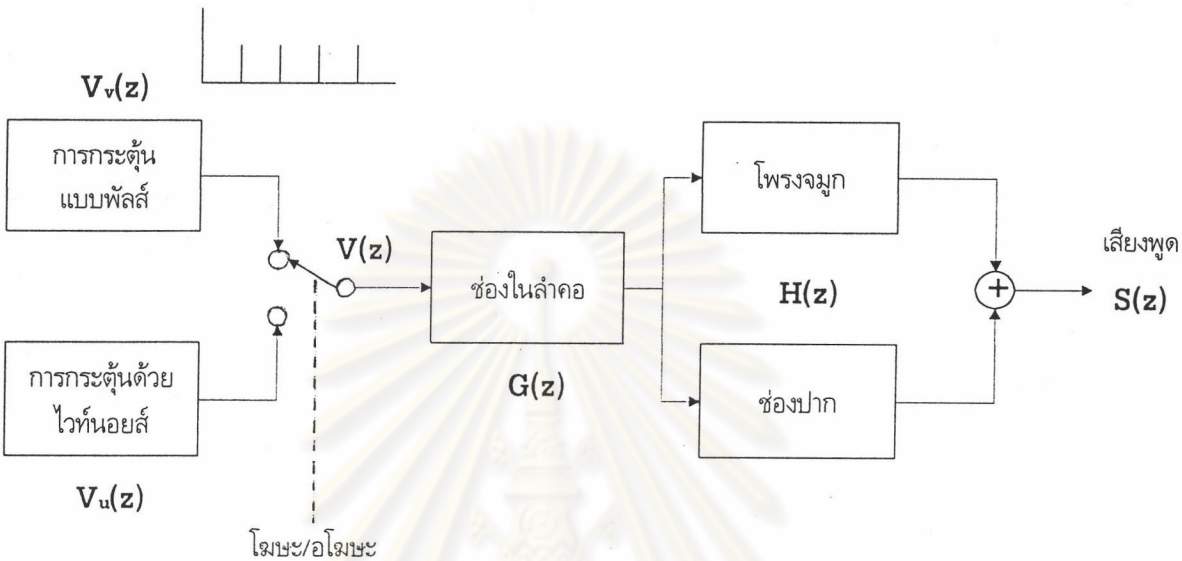
หรือถ้าต้องการบอกเป็นความถี่

ความถี่ (Hz) = 1/คาบ (วินาที)

= อัตราการสุ่มสัญญาณ (ตัวอย่าง/วินาที) / คาบ (ตัวเลข)

= 22000/132

= 166.67 Hz



รูปที่ 2.1 แสดงแบบจำลองในการเปล่งเสียงของมนุษย์

Frerking (1994) ได้แสดงให้เห็นถึงกลไกในระหว่างที่คนเปล่งเสียงดังที่แสดงในรูปที่ 2.1 จุดกำเนิดของเสียงเริ่มต้นขึ้นจากการกระตุ้นในสองลักษณะได้แก่ การกระตุ้นแบบโฆษะและการกระตุ้นแบบอโฆษะ การกระตุ้นแบบโฆษะเกิดจากการสั่นของเส้นเสียงทำให้เกิดเป็นสัญญาณพัลส์ในทุก ๆ คาบเวลาหนึ่ง สัญญาณกระตุ้นแบบโฆษะสามารถประมาณได้โดยอนุกรมของพัลส์ซึ่งมีสมการต่อไปนี้

$$V_v(z) = \delta \sum_{n=0}^{\infty} \left(z^{-p}\right)^n \qquad (1)$$

โดยที่ V_v คือสัญญาณกระตุ้นแบบโฆษะ

p คือคาบของพิทช์

δ คือฟังก์ชันซึ่งกำหนดโดย

$\delta(n)$ = 1 ถ้า n = 0

= 0 ถ้า n ไม่เท่ากับ 0

และ n คือหมายเลขแสดงลำดับของตัวอย่างเสียง

สำหรับการกระตุ้นในแบบอโฆษะ สัญญาณการกระตุ้น V_u สามารถแทนได้ด้วยสัญญาณรบกวนไวท์นอยส์ ซึ่งถือว่าเป็นสัญญาณที่มีองค์ประกอบทางความถี่ทุกความถี่ การ

สร้างสัญญาณนี้ในคอมพิวเตอร์ทำได้โดยกำเนิดตัวเลขสุ่มไปเก็บไว้อย่างเป็นลำดับในแอเรย์ (array) ของสัญญาณกระตุ้น

สำหรับภาษาส่วนใหญ่รวมทั้งภาษาไทยและภาษาอังกฤษในระหว่างที่คนพูดจะเกิดการกระตุ้นสองแบบนี้สลับกันเป็นช่วง ๆ ในช่วงที่มีการเปลี่ยนจากการกระตุ้นแบบโฆษะเป็นอโฆษะหรือกลับกัน มักจะเป็นช่วงที่วิเคราะห์หาพารามิเตอร์ของระบบได้ยาก และมักทำให้เกิดการวิเคราะห์ผิดพลาดได้ง่าย

สัญญาณการกระตุ้นจะผ่านช่องภายในลำคอไปยังปากและโพรงจมูกซึ่งในส่วนนี้สามารถแทนได้ด้วยฟังก์ชันโอนย้าย (transfer function) G(z) ที่มีสมการเป็น

$$G(z) = \frac{1}{\left(1 - e^{-CT} z^{-1}\right)^2} \qquad (2)$$

นอกจากนี้สัญญาณยังถูกมอดูเลตด้วยโพรงจมูกและช่องปาก ทำให้เกิดเป็นความถี่ฮาร์โมนิคต่าง ๆ ของความถี่พิทช์ บางความถี่ก็ถูกขยายบางความถี่ก็ถูกลดทอนขนาดสัญญาณจนเกิดเป็นเสียงขึ้นมา ในส่วนของโพรงจมูกและช่องปากสามารถแสดงเป็นแบบจำลองได้โดยฟังก์ชันโอนย้ายดังนี้

$$H(z) = \frac{1}{\prod_{j=1}^{k}\left[1 - 2e^{-C_j T}\cos\left(B_j T\right) z^{-1} + e^{-2C_j T} z^{-2}\right]} \qquad (3)$$

โดยที่ ฟอร์แมนต์อันดับที่ j คือ

$$f_j = \frac{B_j}{2\pi}$$

และแบนด์วิดธ์ของฟอร์แมนต์ดังกล่าวคือ

$$B_j = \frac{C_j}{2\pi}$$

ดังนั้นสัญญาณเสียงพูดที่ผ่านแบบจำลองทั้งสองจึงสามารถเขียนได้เป็น

$$S(z) = G(z)H(z)GV(z) \qquad (4)$$

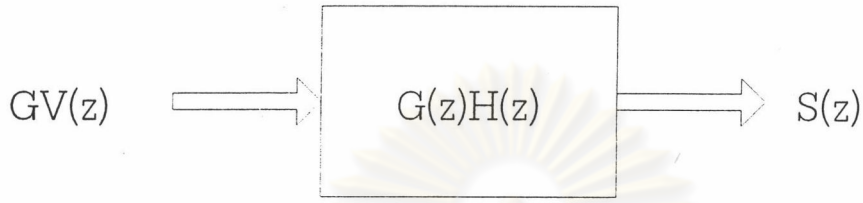
โดยที่

S(z) คือสัญญาณเสียงผลลัพธ์

G(z) คือแบบจำลองของช่องในลำคอ

H(z) คือแบบจำลองของช่องปากและโพรงจมูก

GV(z) คือสัญญาณกระตุ้น



รูปที่ 2.2 แสดงแบบจำลองของการเปล่งเสียงในรูปของฟังก์ชันโอนย้าย

ในส่วนของ G(z) และ H(z) สามารถยุบรวมเป็นฟังก์ชันโอนย้ายเดียวได้ ฟังก์ชันโอนย้ายใหม่นี้เรียกว่าตัวกรองสังเคราะห์ (synthesis filter) 1/A(z) และสำหรับ A(z) ก็เรียกว่าตัวกรองวิเคราะห์ (analysis filter)

$$\frac{1}{A(z)} = \frac{S(z)}{GV(z)} \qquad (5)$$

โดยที่ 1/A(z) คือตัวกรองสังเคราะห์

ดังนั้นเมื่อจัดสมการเสียใหม่ก็จะได้

$$S(z) = \frac{1}{A(z)} GV(z) \qquad (6)$$

ตัวกรองสังเคราะห์นี้เป็นแบบจำลองที่ใช้แสดงการเปลี่ยนแปลงรูปร่างของอวัยวะต่าง ๆ ที่ใช้ในการเปล่งเสียงในระหว่างที่มีการพูด ตั้งแต่ช่องในลำคอ โพรงจมูก ช่องปากรวมไปถึงลักษณะของริมฝีปากขณะที่พูดด้วย มีการศึกษาและได้พบว่าฟังก์ชันโอนย้ายตัวกรองสังเคราะห์นี้สามารถประมาณได้ด้วยตัวกรองที่มีแต่โพล (pole) อย่างเดียวได้ ตัวกรองสังเคราะห์ดังกล่าวเขียนได้เป็น

$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{P} a_i z^{-i}} \qquad (7)$$

โดยที่

P ในที่นี้เป็นอันดับ (order) ของตัวกรองซึ่งก็คือจำนวนของโพลนั่นเอง

a_i คือสัมประสิทธิ์การทำนาย (prediction coefficient) ที่ i

เมื่อแทนตัวกรองสังเคราะห์ในสมการที่ (7) ลงในสมการที่ (6) แล้วจัดรูปเสียใหม่ก็จะได้สมการที่แสดงความสัมพันธ์ของสัญญาณเสียงกับสัญญาณกระตุ้นและสัมประสิทธิ์การทำนายดังนี้

$$S(z) = \sum_{i=1}^{P} a_i z^{-i} S(z) + GV(z) \qquad (8)$$

สมการข้างบนนี้สามารถเขียนให้อยู่ในโดเมนเวลา (time domain) ได้เป็น

$$S(n) = \sum_{i=1}^{P} a_i S(n-i) + GV(n) \qquad (9)$$

จากสมการนี้จะเห็นว่าสามารถคำนวณค่าตัวอย่างสัญญาณเสียงปัจจุบันได้จากตัวอย่างสัญญาณเสียงในอดีตที่ผ่าน ๆ มา s(n-i) ถ้ารู้ค่าของสัมประสิทธิ์การทำนาย a_i อัตราขยาย G และสัญญาณกระตุ้น V(n)

ถึงจุดนี้สามารถสรุปว่าพารามิเตอร์สำหรับการเข้ารหัสสัญญาณเสียงพูดด้วยวิธีแอลพีซี นั้นมีอะไรบ้าง พารามิเตอร์เหล่านี้เป็นสิ่งที่จะต้องวิเคราะห์หาจากสัญญาณเสียงพูดรับเข้า เพื่อใช้ในการเข้ารหัสสัญญาณ พารามิเตอร์เหล่านี้ได้แก่

— สัมประสิทธิ์การทำนาย สัญญาณเสียงพูดรับเข้าจะถูกแบ่งออกเป็นช่วง ๆ ละประมาณ 20 ms แต่ละช่วงจะถูกนำมาวิเคราะห์หาสัมประสิทธิ์การทำนาย จำนวนของสัมประสิทธิ์ที่นิยมใช้มักจะเป็น 10 หรือ 12 ซึ่งตัวเลขนี้ก็คืออันดับของตัวกรอง (filter order) นั่นเอง ในส่วนถัดไปจะได้กล่าวถึงวิธีการในการหาค่าสัมประสิทธิ์นี้

— อัตราการขยายของสัญญาณกระตุ้น

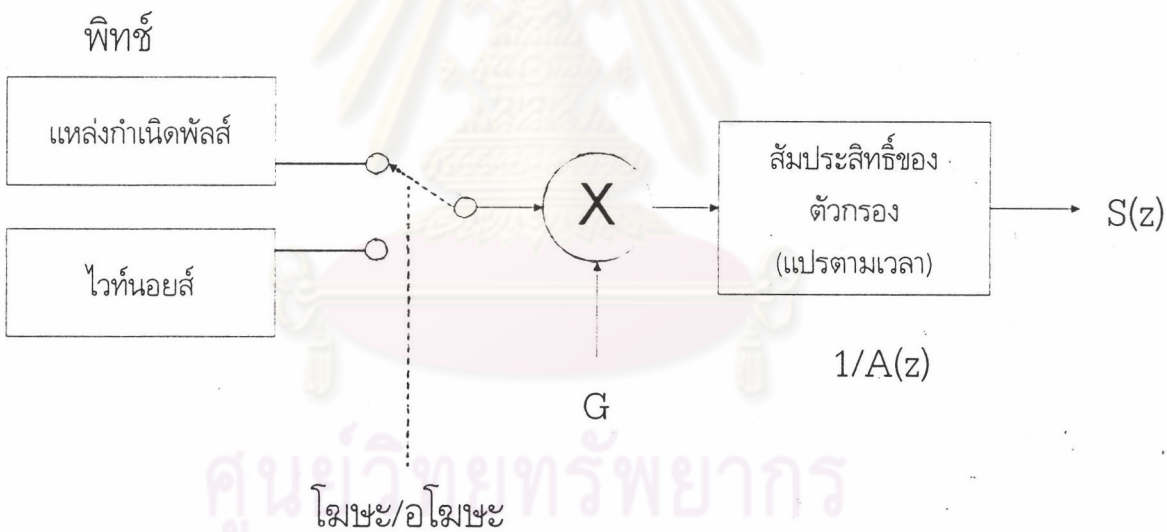
— พิทช์

— การตัดสินใจว่าสัญญาณเสียงพูดรับเข้าในช่วงนั้นเป็นแบบโฆษะหรืออโฆษะ

Frerking (1994) ได้กล่าวถึงตัวอย่างมาตรฐานการเข้ารหัสสัญญาณเสียงพูดโดยวิธีแอลพีซีที่มีใช้กันอยู่มาตรฐานหนึ่งได้แก่มาตรฐาน แอลพีซี10 มาตรฐานนี้กำหนดช่วงของการวิเคราะห์ไว้เป็น 22.5 ms เพื่อที่จะทำการหาพารามิเตอร์ข้างต้นมาหนึ่งชุด (สัมประสิทธิ์การทำนายมีจำนวน 10 ตัว) และสามารถนำพารามิเตอร์เหล่านี้มาเข้ารหัสด้วยจำนวนบิตเพียง 54 บิต ดังนั้นจึงได้อัตราข้อมูลส่งออก 54 บิต/22.5 ms นั่นคือ 2400 bps โดยมาตรฐานนี้ถือว่าสัญญาณเสียงพูดรับเข้ามีอัตราความถี่ของการสุ่มสัญญาณ 8000 ตัวอย่าง/วินาที ถ้าสมมติให้

อุปกรณ์แปลงสัญญาณมีขนาด 14 บิต อัตราส่วนของการบีบข้อมูลก็จะเป็น 8000x14/2400 = 46.67 เท่าซึ่งนับว่าเป็นการบีบขนาดข้อมูลลงไปได้มากทีเดียว

การเลือกขนาดของช่วงเวลาสำหรับการวิเคราะห์หรือขนาดของเฟรมมีผลต่ออัตราส่วนการบีบอัดข้อมูลและคุณภาพของเสียงพูดที่ได้ ถ้าเลือกขนาดเฟรมใหญ่ก็จะได้อัตราการบีบขนาดข้อมูลที่สูง แต่ถ้าเฟรมมีขนาดใหญ่เกินไปก็จะมีผลเสียต่อคุณภาพของเสียงพูดเนื่องจากพารามิเตอร์ของระบบอาจเปลี่ยนไปแล้ว เช่นอัตราขยายสัญญาณกระตุ้นเปลี่ยนไปหรืออวัยวะที่ใช้เปล่งเสียงบางส่วนได้เปลี่ยนตำแหน่งไปจากเดิมทำให้สัมประสิทธิ์ของตัวกรองที่วิเคราะห์หาได้เบี่ยงเบนไปจากความเป็นจริงมาก ผลก็คือพารามิเตอร์ที่วิเคราะห์ออกมาได้จะขาดความแม่นยำไป ในทางกลับกันถ้าพยายามเลือกช่วงเวลาการวิเคราะห์ที่สั้นไปอัตราการบีบขนาดข้อมูลก็จะเลวลงและนอกจากนี้ก็ต้องอาศัยการคำนวณที่ต้องการความเร็วมากขึ้นด้วยซึ่งจุดนี้จะมีผลต่อการทำงานในลักษณะทันทีเพราะต้องอาศัยตัวประมวลผลสัญญาณที่มีความสามารถสูง



รูปที่ 2.3 แสดงจุดที่มีการนำพารามิเตอร์ต่าง ๆ ไปใช้ในการสังเคราะห์เสียงพูดด้วยวิธีแอลพีซี

4. <u>การหาสัมประสิทธิ์การทำนายของตัวกรองสังเคราะห์</u>

Kondos (1995) ได้กล่าวถึงวิธีการหาสัมประสิทธิ์การทำนายของตัวกรองสังเคราะห์ไว้หลายวิธี ได้แก่วิธีออโตโครีเลชัน (auto-correlation) วิธีโคแวเรียนซ์ (corariance) วิธีผลึก (lattice) ทุกวิธีให้ผลที่ใกล้เคียงกันโดยมีข้อดีข้อเสียแตกต่างกันบ้างเล็กน้อย เช่นวิธีโคแวเรียนต์มีความได้เปรียบในแง่ความถูกต้องของการวิเคราะห์สูงกว่าเพราะว่าไม่ต้องใช้ฟังก์ชันวินโดว์ (window function) ในการทำให้สัญญาณมีค่าเป็นศูนย์ที่ขอบของกรอบการวิเคราะห์ แต่ขณะเดียวกันก็มีการคำนวณในช่วงการวิเคราะห์มากกว่าและต้องการความสามารถ

ของตัวประมวลผลสัญญาณที่สูงกว่าด้วย ในเชิงพาณิชย์มักนิยมใช้หลักการของออโตโครีเลชันมากกว่าเพราะว่าสามารถทำการวิเคราะห์แบบทันทีได้ง่ายกว่า วิธีเหล่านี้มีหลักการพื้นฐานเหมือนกัน นั่นคือพยายามหาค่าของสัมประสิทธิ์การทำนายที่ทำให้ค่าผิดพลาดจากการทำนายนี้มีค่าต่ำสุด ความผิดพลาดจากการทำนายแสดงได้โดย

$$e(n) = S(n) - \hat{S}(n) \qquad (10)$$

โดยที่ $\hat{S}(n)$ หมายถึงค่าของสัญญาณที่ได้จากการทำนาย

การหาค่าของสัมประสิทธิ์การทำนาย a_i ทำได้โดยการประมาณค่าของ a_i ให้มีค่าเป็น α_i ดังนั้นค่าของ $\hat{S}(n)$ ที่ได้จากการทำนายจะมีค่าเป็น

$$\hat{S}(n) = \sum_{i=1}^{P} \alpha_i S(n-i) \qquad (11)$$

เมื่อเขียนในรูปแซดโดเมน (Z domain) จะได้

$$\hat{S}(z) = \sum_{i=1}^{P} \alpha_i S(z) z^{-i} \qquad (12)$$

ความผิดพลาดซึ่งเกิดจากการทำนายนี้ก็คือ

$$\begin{aligned} e(n) &= S(n) - \hat{S}(n) \\ &= S(n) - \sum_{i=1}^{P} \alpha_i S(n-i) \end{aligned}$$

เมื่อเขียนในรูปแซดโดเมน

$$\begin{aligned} e(z) &= S(z) - \sum_{i=1}^{P} \alpha_i S(z) z^{-i} \\ &= S(z)\left(1 - \sum_{i=1}^{P} \alpha_i z^{-i}\right) \end{aligned} \qquad (13)$$

ถ้าสามารถหาค่าของ α_i ได้อย่างแม่นยำนั่นคือได้ $\alpha_i = a_i$ ก็จะได้ค่าความผิดพลาดเป็น

$$\begin{aligned} e(z) &= S(z)\left(1 - \sum_{i=1}^{P} a_i z^{-i}\right) \\ &= s(z) A(z) \\ &= GV(z) \qquad \text{...จากสมการ (6)} \end{aligned}$$

สมการนี้หมายความว่าความผิดพลาดที่ได้จากการประมาณโดยใช้ค่าสัมประสิทธิ์การทำนายที่ถูกต้องก็คือสัญญาณกระตุ้น (หรือเรียกว่าสัญญาณเศษค้าง (residual signal) ในวิธีการบีบข้อมูลเสียงบางวิธี) ซึ่งก็สมเหตุสมผลเนื่องจากส่วนที่ไม่สามารถทำนายได้ก็คือสัญญาณการกระตุ้นนั่นเอง

การหาค่าของ α_i ทำได้โดยหาค่าความผิดพลาดจากการทำนายนำมายกกำลังสองแล้วทำการหาอนุพันธ์เทียบกับ α_i เพื่อหาค่า α_iที่ทำให้ได้ความผิดพลาดต่ำที่สุด ค่าความผิดพลาดสามารถเขียนได้เป็น

$$
\begin{aligned}
E &= \sum^{N} e^2(n) \\
&= \sum^{N} \left[S(n) - \hat{S}(n) \right]^2 \\
&= \sum^{N} \left[S(n) - \sum_{i=1}^{P} \alpha_i S(n-i) \right]^2
\end{aligned}
$$

สิ่งที่จะต้องทำก็คือหาค่า α_i ที่ทำให้ $\frac{\partial(E)}{\partial(\alpha_i)}$ มีค่าเป็นศูนย์

ผลจากการทำอนุพันธ์ดังกล่าวจะทำให้ได้สมการจำนวน P สมการ โดยที่ P เป็นอันดับของตัวกรอง เพื่อที่จะหาค่า α_i จำเป็นต้องแก้สมการ P สมการเหล่านี้ เลวินสันและเดอบินได้ศึกษาวิธีการในการหาสัมประสิทธิ์เหล่านี้ และได้เสนอขั้นตอนวิธีการหาดังนี้ (Rabiner และ Shafer, 1978 อ้างถึงใน Frerking, 1994)

$$k_i' = \frac{R_s(i) - \sum_{j=1}^{i-1} \alpha_i R_s(i-j)}{E} \qquad (14)$$

$$\alpha_i = k_i' \qquad (15)$$

$$\alpha_j' = \alpha_i - k_i' \alpha_{i-j} \qquad j=1, 2, 3, ..., i-1 \qquad (16)$$

$$E' = \left(1 - (k_i')^2\right) E \qquad (17)$$

โดยที่

k_i= สัมประสิทธิ์การสะท้อน

α_i=สัมประสิทธิ์การทำนาย

$R_s(i)$=ออโตโครีเลชัน (autocorrelation) ตำแหน่งที่ i ของสัญญาณเสียงพูดรับเข้า

E=ค่าเฉลี่ยความผิดพลาดกำลังสองของสัญญาณที่ได้จากการทำนายเทียบกับค่าจริง

เครื่องหมายไพรม์ (′) หมายถึงค่าของตัวแปรนั้นในรอบใหม่

สมการข้างต้นทั้งสี่สมการนี้จะถูกคำนวณเป็นจำนวน P รอบโดยที่ P เป็นอันดับของตัวกรอง เริ่มต้นตั้งแต่ i เท่ากับ 1 ไปจนถึง P ในรอบแรกกำหนดให้ E มีค่าเป็น $R_s(0)$ ในแต่ละรอบเมื่อคำนวณหาค่าสัมประสิทธิ์การสะท้อน k_i' ได้แล้วมันจะถูกกำหนดค่าให้กับสัมประสิทธิ์การทำนายที่ i และสัมประสิทธิ์การทำนายตัวก่อน ๆ จะถูกคำนวณใหม่ตามสมการที่ (16) ตั้งแต่ตัวที่ 1 ไปจนถึงตัวที่ i-1 ค่าผิดพลาดจะถูกคำนวณใหม่ตามสมการที่ (17) เพื่อใช้ในรอบถัดไป

Frerking (1994) ได้กล่าวถึงการวิเคราะห์และการสังเคราะห์สัญญาณเสียงพูดในทางปฏิบัติว่าโดยมากจะใช้สัมประสิทธิ์การสะท้อนแทนที่จะเป็นสัมประสิทธิ์การทำนายเนื่องจากสามารถตรวจสอบเสถียรภาพของตัวกรองได้จากสัมประสิทธิ์การสะท้อน และใช้ตัวกรองผลึก (lattice filter) ในการวิเคราะห์และสังเคราะห์ดังกล่าว รายละเอียดการทำงานของตัวกรองผลึกจะได้กล่าวถึงในบทถัดไป

5. <u>การวิเคราะห์หาพิทช์ของสัญญาณเสียงพูด</u>

จุดที่ยากจุดหนึ่งในเข้ารหัสสัญญาณเสียงพูดด้วยวิธีแอลพีซีก็คือการหาพิทช์ของสัญญาณเสียงพูดอย่างถูกต้อง มีหลาย ๆ วิธีที่ได้ถูกนำมาลองใช้และส่วนใหญ่ก็สามารถทำงานได้เป็นอย่างดีกับเสียงพูดที่มีพิทช์ที่ชัดเจน ปัญหาส่วนใหญ่จะเกิดกับช่วงที่กำลังเปลี่ยนจากเสียงแบบโฆษะเป็นเสียงแบบอโฆษะ จะกล่าวถึงบางวิธีที่มีใช้อยู่ดังต่อไปนี้

5.1 <u>วิธีเอเอ็มดีเอฟ</u>

วิธีเอเอ็มดีเอฟ (Average Magnitude Difference Function - AMDF) เป็นวิธีที่ใช้ในตัวเข้ารหัสเสียง (vocoder) ของรัฐบาลสหรัฐและนาโต้ วิธีนี้ถูกพัฒนาขึ้นมาในระหว่างที่การคำนวณแบบคูณยังต้องใช้เวลาในการประมวลผลมาก วิธีนี้จึงหลีกเลี่ยงการคำนวณหาออโตโครีเลชันซึ่งต้องใช้การคูณมากครั้ง หลักการพื้นฐานของวิธีการนี้จะเปรียบเทียบสัญญาณเสียงกับตัวมันเองที่ถูกหน่วงเวลาออกไป และพยายามหาค่าเวลาหน่วงที่ทำให้ที่ทำให้ได้ค่าเฉลี่ยความแตกต่างต่ำที่สุด โดยยึดถือว่าถ้าการหน่วงนั้นเป็นจำนวนเท่าของคาบของพิทช์ก็จะทำให้ความแตกต่างมีค่าน้อย ฟังก์ชันเอเอ็มดีเอฟได้ถูกเสนอโดย Tremain (Tremain, 1982 อ้างถึงใน Frerking, 1994) ดังนี้

$$AMDF(\tau) = \sum_{n=1}^{N} |s(i) - s(i+\tau)| \qquad (18)$$

ก่อนที่จะนำสัญญาณมาวิเคราะห์หาเอเอ็มดีเอฟ ก็จะต้องมีการปรับสัญญาณให้มีความเหมาะสมเสียก่อนได้แก่การนำสัญญาณมากรองส่วนที่เป็นความถี่สูงเกิน 800 Hz ทิ้งไป เพื่อกำจัดข้อมูลส่วนเกินทิ้งไปเป็นต้น

5.2 วิธีออโตโครีเลชัน

วิธีการนี้จะต้องนำสัญญาณมาหาออโตโครีเลชันเสียก่อน แล้วจึงวิเคราะห์หาพิทช์ ออโตโครีเลชันของสัญญาณมีสมการเป็นดังนี้

$$R_s(\tau) = \sum_{i=0}^{N-1} s(i)s(i+\tau) \qquad (19)$$

การทำออโตโครีเลชันก็คือการหาผลรวมของการคูณกันระหว่างสัญญาณเสียงรับเข้ากับตัวมันเองที่ถูกเลื่อนออกไป ค่าของออโตโครีเลชันจะมีค่ามากถ้าหากสัญญาณที่มาคูณกันมีรูปร่างหน้าตาคล้ายกัน นั่นคือถ้าหากคาบของการเลื่อนของสัญญาณนั้นเป็นทวีคูณของคาบของพิทช์ก็จะได้ผลรวมค่าสูง การหาค่าพิทช์ก็จะทำการวิเคราะห์ออโตโครีเลชันในช่วงประมาณ 3-15 ms ภายในกรอบการวิเคราะห์ เพื่อหาจุดที่ให้ค่าออโตโครีเลชันสูงสุด เวลาที่จุดนั้นนับจากจุดเริ่มต้นของกรอบวิเคราะห์ก็คือคาบของพิทช์นั่นเอง

6. การหาค่าอัตราขยายของสัญญาณกระตุ้น

เนื่องจากตลอดเวลาที่กำลังพูดอยู่อัตราขยายของสัญญาณกระตุ้นจะมีการเปลี่ยนแปลงอยู่ตลอดเวลา จำเป็นต้องหาค่าอัตราขยายนี้ ซึ่งทำได้โดยใช้สมการต่อไปนี้

$$G^2 = R_s(0) - \sum_{i=1}^{P} \alpha_i R_s(i) \qquad (20)$$

สมการนี้ได้มาจากการแทนค่าสมการ (9) ซึ่งก็คือ

$$S(n) = \sum_{i=1}^{P} a_i S(n-i) + GV(n)$$

ลงในสูตรการหาค่าออโตโครีเลชันของสัญญาณ (19)

$$R_s(\tau) = \sum_{n=0}^{N-1} S(n)S(n+\tau)$$

ดังนั้นออโตโครีเลชันของสัญญาณจึงเขียนใหม่ได้เป็น

$$R_s(\tau) = \sum_{n=0}^{N-1}\left[\sum_{i=1}^{P} a_i S(n-i) + GV(n)\right]S(n+\tau) \qquad (21)$$

สำหรับกรณีที่เป็นการกระตุ้นแบบโฆษะค่าของสัญญาณกระตุ้นจะมีค่าเป็น G ที่ n=0 ส่วนที่ n ค่าอื่นๆสัญญาณกระตุ้นจะมีค่าเป็น 0 กรณีที่ τ มีค่าเป็น 0 สมการข้างบนจะเขียนได้ในรูปที่ง่ายขึ้นเป็น

$$R_s(0) = \sum_{n=0}^{N-1}\left[\sum_{i=1}^{P} a_i S(n-i)S(n) + GV(n)S(n)\right]$$

$$R_s(0) = \sum_{i=1}^{P}\sum_{n=0}^{N-1} a_i S(n)S(n-i) + \sum_{n=0}^{N-1} GV(n)S(n) \qquad (22)$$

$$R_s(0) = \sum_{i=1}^{P} a_i R_s(i) + GS(0) \qquad (23)$$

แต่จากสมการ (9) รู้ว่า S(0) มีค่าเป็น G ดังนั้น

$$R_s(0) = \sum_{i=1}^{P} a_i R_s(i) + G^2 \qquad (24)$$

สมการข้างบนก็คือที่มาของหลักเกณฑ์การกำหนดค่าอัตราขยายสัญญาณกระตุ้นที่เสนอไว้ในสมการ (20) นั่นเอง สมการ (20) นี้สามารถใช้กับเสียงแบบโฆษะโดยถือว่าในช่วงการวิเคราะห์มีพัลส์กระตุ้นที่ n=0 เท่านั้นที่ n ค่าอื่น ๆ มีค่าสัญญาณกระตุ้นเป็น 0

สำหรับเสียงแบบอโฆษะก็มีหลักเกณฑ์ในการคำนวณค่าอัตราขยายสัญญาณกระตุ้นที่คล้าย ๆ กัน โดยถือว่าสัญญาณกระตุ้นเป็นกระบวนการไวท์นอยส์ที่มีฟังก์ชันออโตโครีเลชันของสัญญาณ $R_u(0) = 1$ และ $R_u(\tau) = 0$ สำหรับ $\tau \neq 0$ (Frerking, 1994) ซึ่งฟังก์ชันออโตโครีเลชันนี้ก็หาได้ด้วยวิธีการเดียวกันกับที่ใช้ในเสียงแบบโฆษะ ดังนั้นสมการ (20) จึงสามารถนำมาใช้กับเสียงแบบอโฆษะได้ด้วย

<u>การบีบข้อมูลเสียงพูดโดยวิธีการทำนายเชิงเส้นโดยการกระตุ้นด้วยรหัส (ซีอีแอลพี)</u>

ซีอีแอลพีจัดเป็นการเข้ารหัสสัญญาณเสียงในกลุ่มการวิเคราะห์โดยการสังเคราะห์ หมายถึงมีการสังเคราะห์สัญญาณเสียงขึ้นมาจากสัญญาณกระตุ้นซึ่งอาจคัดเลือกมาจากสัญญาณสุ่มจำนวนหลาย ๆ สัญญาณ สัญญาณเสียงที่ได้จากการสังเคราะห์ถูกนำมาเปรียบเทียบเพื่อคัด

เลือกสัญญาณกระตุ้นที่ทำให้ได้สัญญาณเสียงที่ใกล้เคียงสัญญาณเสียงต้นฉบับมากที่สุด พารามิเตอร์ที่ได้ก็คือดัชนีที่ชี้ไปยังสัญญาณกระตุ้นในโคดบุค (codebook) นั่นเอง ด้วยวิธีการนี้ทำให้สามารถเข้ารหัสสัญญาณได้อย่างมีประสิทธิภาพ โดยทั่วไปการบีบข้อมูลหรือสัญญาณเสียงพูดด้วยวิธีนี้จึงให้อัตราข้อมูลรหัสที่บีบแล้วค่อนข้างต่ำแต่ยังคงคุณภาพของเสียงพูดไว้ได้ในระดับที่ดี

ขั้นตอนการทำงานแบ่งออกได้คร่าว ๆ เป็นสามส่วนด้วยกันได้แก่ การวิเคราะห์แอลพีซี การวิเคราะห์ตัวทำนายช่วงยาวหรือการวิเคราะห์แอลทีพี (long-term predictor analysis - LTP) และส่วนที่ค้นหาโคดบุคเพื่อค้นหาสัญญาณกระตุ้นที่เหมาะสม สัญญาณเสียงพูดที่กำลังวิเคราะห์จะถูกแบ่งออกเป็นเฟรม ๆ ละประมาณ 20-30 ms แต่ละเฟรมจะถูกนำไปวิเคราะห์หาสัมประสิทธิ์แอลพีซีอย่างที่ได้กล่าวมาแล้ว สัมประสิทธิ์ที่ได้จะถูกนำไปใช้ในตัวทำนายช่วงสั้นหรือตัวทำนายเอสทีพี (short-term predictor - STP) เมื่อวิเคราะห์แอลพีซีจนได้สัมประสิทธิ์แอลพีซีมาแล้วขั้นต่อไปเป็นการวิเคราะห์แอลทีพี ในขั้นนี้การวิเคราะห์จะทำในช่วงเฟรมสั้น ๆ ภายในเฟรมของการวิเคราะห์แอลพีซีเช่นอาจเป็น 5-10 ms การวิเคราะห์แอลทีพีจะทำกับสัญญาณเศษค้าง (residual signal) ผลของการวิเคราะห์แอลทีพีจะได้พารามิเตอร์ช่วงยาว (long-term parameter) ได้แก่ดีเลย์ D และสัมประสิทธิ์ช่วงยาว β_i โดยที่ i เป็นเป็นตัวแสดงเลขหมายของแทพ (tap) ของตัวกรองช่วงยาว Kondos (1994) ได้กล่าวถึงวิธีการวิเคราะห์แอลทีพีว่าสามารถทำได้ด้วยวิธีลูปปิดหรือลูปเปิดก็ได้ เมื่อได้พารามิเตอร์ของตัวกรองทั้งช่วงสั้นและช่วงยาวมาแล้วก็สามารถกำหนดสัญญาณกระตุ้นได้ โดยทั่วไปการปรับค่าของสัญญาณกระตุ้นจะทำทุก ๆ เฟรมย่อยเช่นกัน การเลือกสัญญาณกระตุ้น ในวิธีซีอีแอลพีมาตรฐานจะเลือกจากโคดบุคของอนุกรมของค่าสุ่ม ขั้นตอนในการเลือกสัญญาณกระตุ้นที่ดีที่สุดประกอบด้วยขั้นตอนพื้นฐานสามขั้นตอนด้วยกันคือ การกรองอนุกรมสัญญาณสุ่ม การคำนวณค่าผิดพลาดที่เกิดจากแต่ละอนุกรม และขั้นตอนสุดท้ายคือเลือกอนุกรมที่ให้ค่าผิดพลาดต่ำสุดและอัตราขยายที่สอดคล้องกัน

<u>การบีบข้อมูลเสียงพูดโดยวิธีการทำนายช่วงยาวโดยการกระตุ้นด้วยพัลส์เศษค้าง (อาร์พีอี-แอลทีพี)</u>

การบีบข้อมูลเสียงพูดด้วยวิธีซีอีแอลพีที่ได้กล่าวไปแล้วนั้นแม้ว่าจะให้คุณภาพเสียงพูดที่ดีโดยที่มีอัตราข้อมูลที่บีบแล้วค่อนข้างต่ำแต่ก็ต้องใช้ความสามารถในการคำนวณของ

ตัวประมวลผลที่สูงมากโดยเฉพาะในช่วงการค้นหาอนุกรมของค่าสุ่มในโคดบุคเพื่อสร้างสัญญาณกระตุ้นที่ดีและเหมาะสมที่สุด สำหรับการบีบข้อมูลเสียงพูดด้วยวิธีอาร์พีอี-แอลทีพีมีหลักการทำงานบางส่วนที่คล้ายคลึงกับวิธีซีอีแอลพีแต่ในส่วนการวิเคราะห์และสร้างสัญญาณกระตุ้นกลับง่ายกว่าและต้องการความสามารถในการคำนวณน้อยกว่า

ในวิธีนี้เฟรมสัญญาณเศษค้างที่ได้หลังผ่านตัวกรองแอลพีซีย้อนกลับ (LPC inverse filter) จะถูกแบ่งเป็นช่วงสั้น ๆ จำนวนสี่เฟรมย่อย แต่ละเฟรมย่อยก็จะถูกนำไปวิเคราะห์เพื่อหาพารามิเตอร์ช่วงยาวอันได้แก่ดีเลย์และสัมประสิทธิ์ช่วงยาว พารามิเตอร์ดังกล่าวถูกนำไปใช้ในการวิเคราะห์หาสัญญาณกระตุ้นซึ่งสัญญาณที่ได้จะถูกดาวน์แซมปลิง (down sampling) ลงและเข้ารหัสซึ่งต่างจากวิธีซีอีแอลพีที่สัญญาณกระตุ้นได้มาจากการค้นหาในโคดบุค เนื่องจากในการวิจัยนี้ได้เลือกวิธีอาร์พีอี-แอลทีพีสำหรับการบีบข้อมูลเสียงพูดแบบทันทีดังนั้นรายละเอียดการทำงานของการบีบข้อมูลเสียงด้วยวิธีอาร์พีอี-แอลทีพีจะได้กล่าวถึงโดยละเอียดในส่วนถัดไป คุณภาพของสัญญาณเสียงพูดที่ได้หลังจากจากบีบและคลายด้วยวิธีนี้จัดว่าดีแต่อย่างไรก็ตามวิธีนี้มีอัตราข้อมูลที่บีบแล้วสูงกว่าเมื่อเทียบกับวิธีซีอีแอลพี