บทที่ 3

การบีบย่อเสียงพูดโดยวิธี CELP

การบีบย่อเสียงพูด

การบีบย่อสัญญาณเสียงพูดนั้นอาจแบ่งได้เป็น 2 วิธีคือ การเข้ารหัสรูปคลื่น (waveform coding) ซึ่งเป็นการเข้ารหัสสัญญาณเสียงพูด หรือ spectrum ของสัญญาณเสียงพูดโดยตรง และ การเข้ารหัสเสียง (voice coding) ซึ่งเป็นการเข้ารหัสสัญญาณเสียงพูดเป็นชุดของ parameter ซึ่งแสดงถึงเสียงพูดในแต่ละ frame วิธีการเข้ารหัสเสียงนี้เป็นที่นิยมในการบีบย่อสัญญาณเสียงพูดในปัจจุบัน เนื่องจากสามารถให้อัตราการบีบย่อที่สูงกว่าวิธีการเข้ารหัสรูปคลื่นที่คุณภาพเสียงใกล้เคียงกัน

วิธีการบีบย่อสัญญาณเสียงพูดโดยการเข้ารหัสเสียงโดยทั่วไปแล้วจะมีพื้นฐานอยู่ที่การใช้ LPC (Linear Predictive Coding) และ VQ (Vector Quantization) โดยมีสมมุติฐานว่าสัญญาณเสียงพูดนั้นมีลักษณะเป็น Quasi-Stationary หรือเป็นสัญญาณที่มีลักษณะคงตัวในช่วงเวลาสั้นๆ 20-30 ms โดยในช่วงเวลานี้สัญญาณจะมีองค์ประกอบในเชิงความถี่ที่เหมือนหรือคล้ายคลึงกัน(Bristow 1984, Deller et al., 1993) ดังนั้นจะทำการแบ่งสัญญาณเสียงพูดเป็น frame ย่อยๆขนาด 20-30 ms ทำการวิเคราะห์หา LPC parameters ของสัญญาณเสียงพูด ซึ่งอาจกล่าวได้ว่า LPC parameters นี้ เป็นตัวแสดงถึงผลตอบเชิงความถี่ของเส้นเสียง(Vocal Tract) ในระบบการสร้างเสียงของมนุษย์(Bristow, 1984) โดยเขียนสมการ transfer function ของ LPC filter ได้ดังนี้ (Bristow, 1984)

$$A(z) = 1 + \sum_{k=1}^{p} a(k) z^{-k} \qquad (3\text{-}1)$$

โดยที่ $\{a(k), 1 \le k \le p\}$ คือ linear predictor coefficients

สัญญาณเสียงพูดที่จะนำมาวิเคราะห์นั้น จะต้องถูกทำ windowing เพื่อแบ่งเป็น frame ดังนี้

$$x(n) = \begin{Bmatrix} w(n)s(n), & 0 \le n < N \\ 0, & otherwise \end{Bmatrix} \quad (3\text{-}2)$$

โดยที่ N คือขนาดของ frame

วิธีในการวิเคราะห์หาค่า LPC parameters ที่นิยมใช้กันได้รับการพัฒนาโดย Levinson และ Durbin ซึ่งเรียกว่า Levinson-Durbin recursive method มีขั้นตอนดังนี้ (Bristow, 1984)

$$E_0 = R(0)$$

$$For\ m = 1, 2, \ldots, p$$

$$K_m = -\left[R(m) + \sum_{k=1}^{m-1} a_{m-1}(k) R(m-k) \right] \Big/ E_{m-1} \quad (3\text{-}3)$$

$$a_m(m) = K_m$$

$$a_m(k) = a_{m-1}(k) + K_m a_{m-1}(m-k), \qquad 1 \le k \le m-1 \quad (3\text{-}4)$$

$$E_m = (1 - K_m^2) E_{m-1} \quad (3\text{-}5)$$

โดยที่ autocorrelation $R(i) = \sum_{n=1}^{i-9} x(n)x(n-i), \qquad 0 \le i \le p \quad (3\text{-}6)$

สัมประสิทธิ์ $\{a_m(k),\ 1 \le k \le m\}$ ก็คือ LPC parameters ที่ order m ซึ่งผลลัพธ์สุดท้ายคือ $\{a_p(k),\ 1 \le k \le p\}$ โดยที่ p คือ order ของ LPC parameters ที่ต้องการ

เมื่อผ่านสัญญาณเสียงพูด s(n) เข้าไปใน Linear Predictive filter เราจะได้ สัญญาณเศษเหลือ หรือ residual signal r(n) ออกมาดังสมการต่อไปนี้

$$r(n) = e(n) + \sum_{k=1}^{p} a(k)e(n-k) \tag{3-7}$$

สัญญาณเศษเหลือนี้จะมีขนาดเล็กกว่าสัญญาณเสียงพูดที่นำมาวิเคราะห์มาก และจะถูกเข้ารหัสโดยใช้ Vector Quantition (VQ)

สำหรับการสังเคราะห์สัญญาณเสียงพูดกลับมานั้น สามารถทำได้โดยผ่านสัญญาณเศษเหลือเข้าไปใน inverse filter ของ linear predictive filter ซึ่งเขียน transfer function ได้ดังนี้ (Bristow, 1984)

$$A^{-1}(z) = \frac{1}{1 + \sum_{k=1}^{p} a(k)z^{-k}} \tag{3-8}$$

โดยที่ $\{a(k),\ 1 \le k \le p\}$ คือ linear predictive coefficients

ซึ่งจะได้สัญญาณ r(n) ดังนี้คือ

$$e(n) = -\sum_{k=1}^{p} a(k)e(n-k) + r(n) \tag{3-9}$$

โดยการใช้ LPC และ VQ นี้ มีผู้พัฒนาวิธีการบีบย่อสัญญาณเสียงขึ้นมาหลายวิธีซึ่งใช้พื้นฐานของวิธีดังกล่าว ซึ่งได้แก่ RELP(Residual Excited Linear Prediction), Multipulse LPC Vocoder, CELP(Code-Excited Linear Prediction), VSELP(Vector Sum Excited Linear Prediction) เป็นต้น ซึ่งวิธีที่ให้ผลดีและนิยมนำมาใช้กันในปัจจุบันคือ CELP โดยมีผู้นำมาดัดแปลงเพื่อใช้ในการบีบสัญญาณเสียงได้แก่ วิธี ACELP (Adaptive Code-Excited Linear Prediction) โดย Galand ในปีค.ศ.1992 (Galand et al., 1992) และที่ออกเป็นมาตรฐานคือ CCITT recommendation G.728 เพื่อการบีบย่อสัญญาณเสียงพูดที่ 16,000 bps ในปีค.ศ.1992 (ITU, 1992) โดยใช้วิธี LD-CELP(Low Delay Code-Excited Linear Prediction) ซึ่งวิธีการบีบย่อสัญญาณเสียงพูดที่นำเสนอในวิทยานิพนธ์นี้ก็ดัดแปลงมาจากวิธี CELP นี้เช่นเดียวกัน

หลักการบีบย่อเสียงพูดโดยวิธี CELP

การบีบย่อเสียงพูดโดยวิธี Code-Excited Linear Prediction (CELP) เป็นวิธีการแบบวิเคราะห์โดยการสังเคราะห์ (Analysis-by-Synthesis) โดยมีพื้นฐานอยู่ที่ filter สองแบบที่ต่อกันอยู่แบบ cascade ซึ่งก็คือ short-term predictive filter หรือ linear predictive filter และ long-term predictive filter

linear predictive filter จะเป็นตัวสร้าง spectral envelope ของสัญญาณเสียงพูด หรืออาจเรียกได้ว่าจำลองแบบของระบบการออกเสียงของมนุษย์ ในขณะที่ long-term predictive filter จะเป็นตัวทำนายการซ้ำกันเป็นคาบของสัญญาณเสียงพูด

parameters ของ filter ทั้งสองชนิดนี้จะถูก update เป็นระยะๆ โดยที่ parameters เหล่านี้จะถูกคำนวณมาเพื่อลดค่า prediction error energy ให้ต่ำที่สุด ซึ่งโดยทั่วไปแล้ว linear predictive filter มักใช้ order เป็น 10-12 และ update ทุกๆ 20-30 msec

จาก block diagram ของตัวเข้ารหัสแบบ CELP ที่แสดงอยู่ในรูป 3-1 นั้น จะเห็นได้ว่าสัญญาณเสียงพูดที่ถูกสังเคราะห์ขึ้นมาจะถูกนำไปเปรียบเทียบกับสัญญาณเสียงพูดขาเข้า ผลต่างของสัญญาณนี้จะเรียกว่าเป็น Residual error signal ซึ่ง residual error signal นี้จะถูกนำไปผ่าน Perceptual weighting filter ซึ่งจะให้ความสำคัญของสัญญาณมากที่ความถี่ซึ่งตรงกับความถี่ Harmonic ของเสียงพูดใน frame นั้น โดยสมบัติของ perceptual weighting filter นั้นจะถูกกำหนดโดย linear predictive filter A(z) ซึ่งอาจแสดง transfer function ของ perceptual weighting filter ได้ดังนี้ (Deller et. al., 1993)

$$(z) = \frac{A(z)}{A(z\ c)} \quad \text{โดย } 0 < c \leq \qquad (3\text{-}10)$$

ซึ่งในทางปฏิบัติได้มีการพิสูจน์แล้วว่า ค่า c ระหว่าง 0.7 และ 0.9 ให้ผลดี (Deller et. al., 1993)

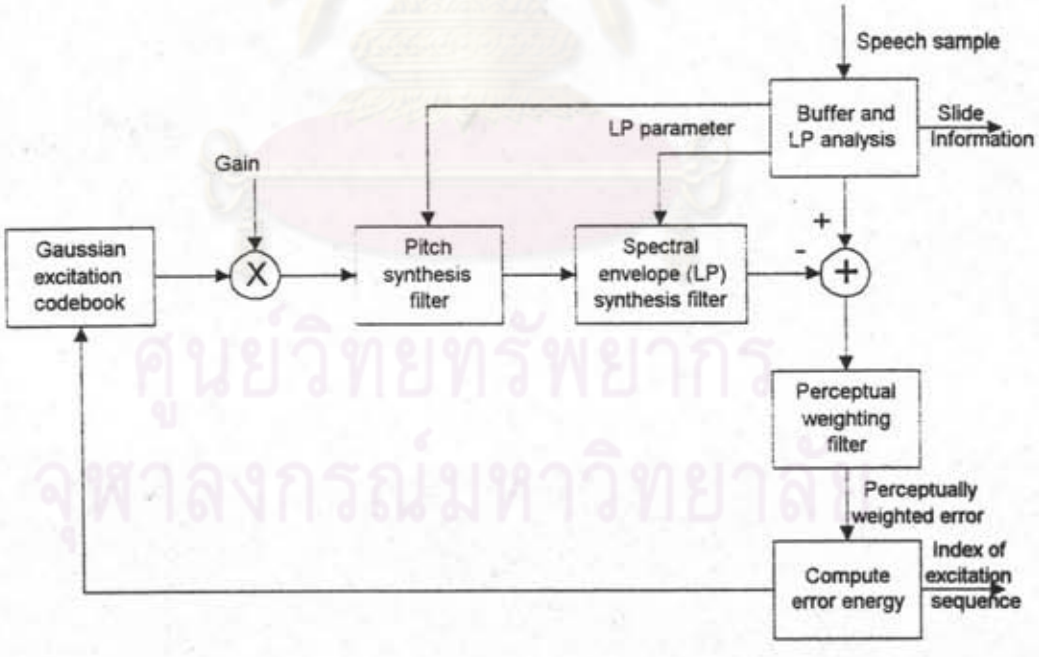
residual error signal ที่ผ่าน perceptual weighting filter มาแล้ว จะถูกแบ่งเป็น block ย่อยๆ แล้วนำไปเปรียบเทียบกับ VQ codebook ซึ่งเราจะได้สัญญาณกระตุ้น (excitation signal) ที่ทำให้ได้พลังงานของ error น้อยที่สุด หลังจากนั้นค่า gain ที่จะทำให้มีพลังงานของ error น้อยที่สุด ก็จะถูกคำนวณสำหรับแต่ละ codeword

สำหรับ long-term predictive filter จะเป็นตัวทำนายการซ้ำกันเป็นคาบของสัญญาณ ซึ่งสามารถเขียน long-term predictive filter ได้ดังนี้ (Hussain et al., 1991)

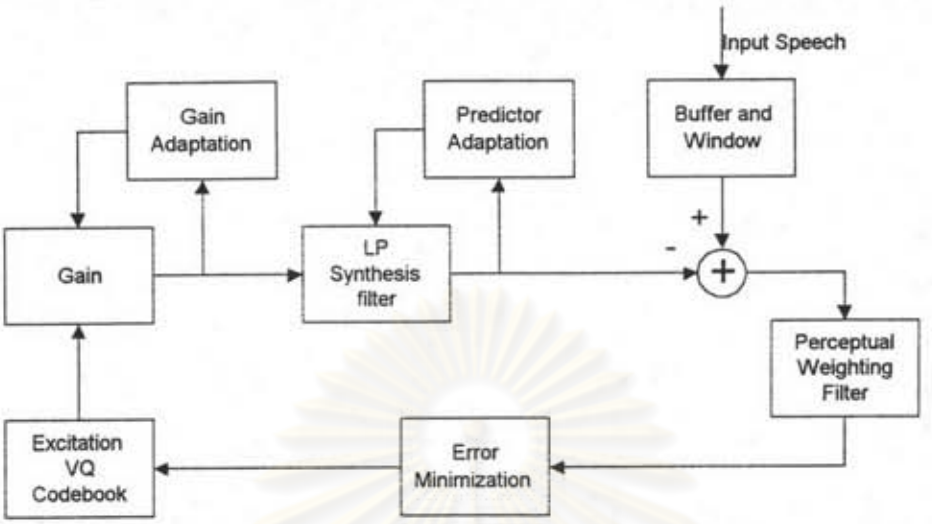
$$L(z) = g \cdot z^{-p} \tag{3-11}$$

โดยที่ parameter g คือ pitch gain และ parameter p คือ pitch period

โดยวิธี CELP นี้จะใช้ long-term predictive filter ในการทำนายสัญญาณเสียงพูดใน frame ปัจจุบัน โดยอาศัย long-term parameters และ สัญญาณเสียงพูดจาก frame ก่อนๆ ผลต่างของสัญญาณเสียงพูดที่ได้จากการทำนาย และสัญญาณเสียงพูดจริงจะผ่าน linear predictive filter ผลลัพธ์ที่ได้ก็คือ สัญญาณเศษเหลือ (residual signal) ซึ่งจะถูกนำไปเข้ารหัสโดย VQ แล้วจึงทำกระบวนการย้อนกลับเพื่อให้ได้สัญญาณเสียงพูดที่จะนำไปใช้ในการทำนายสัญญาณเสียงพูดใน frame ต่อๆไป



รูปที่ 3-1 แสดง block diagram ของตัวเข้ารหัสแบบ CELP (Deller et. al., 1993)



รูปที่ 3-2 แสดงตัวเข้ารหัสแบบ LD-CELP (Deller et. al., 1992)

วิธี CELP ได้มีผู้ดัดแปลงนำไปใช้ในการบีบย่อเสียงพูดอีกหลายวิธี โดย Galand ได้เสนอวิธีการใช้ Adaptive Code-Excited Linear Prediction (ACELP) (Galand et. al., 1992) ซึ่งใช้ codebook ที่แบ่งเป็น 2 ส่วน โดยมีส่วนหนึ่งคงที่ และอีกส่วนหนึ่งปรับตามสัญญาณใน block ก่อนๆ สามารถให้ bit-rate ต่ำถึง 7,200 bps นอกจากนี้ยังได้มีผู้พัฒนาเป็น LD-CELP (Low-Delay Code-Excited Linear Prediction) ซึ่งเป็นวิธีที่ใช้สัญญาณจาก block ก่อนๆ มาวิเคราะห์เพื่อทำนาย parameters ของ filter ต่างๆ ซึ่งทำให้ไม่จำเป็นต้องทำ buffer สัญญาณเอาไว้นานมาก ทำให้ได้ระบบที่มี delay เพียง 2 msec ซึ่งตรงตามมาตรฐานของระบบสื่อสารในปัจจุบัน ซึ่งกำหนดให้มี delay อย่างมากที่สุดไม่เกิน 5 msec ซึ่งจะทำให้ไม่ต้องใช้วงจรตัดเสียงก้อง (echo canceller) ซึ่งได้ออกเป็นมาตรฐานของระบบสื่อสารโดย ITU เป็นมาตรฐาน CCITT recommendation G.728 (ITU, 1992) ซึ่งให้ bit-rate ที่ 16,000 bps ในขณะที่ Galand ก็ได้ประยุกต์ใช้ ACELP เข้ากับ LD-CELP ได้ bit-rate ที่ 13,800 bps (Galand et. al., 1992) สำหรับ block diagram ของตัวเข้ารหัสโดยใช้ LD-CELP แสดงอยู่ในรูป 3-2