

บทที่ 3

ทฤษฎีที่เกี่ยวข้อง

3.1 การประมวลผลสัญญาณเบื้องต้น (Signal Preprocessing)

การประมวลผลสัญญาณเบื้องต้นเป็นขั้นตอนของการจัดเตรียมข้อมูลจากข้อมูลเสียงที่ได้จากการบันทึกด้วยโปรแกรมคอมพิวเตอร์ผ่านไมโครโฟนและการ์ดเสียง เนื่องจากการบันทึกสัญญาณมักจะมีสัญญาณรบกวนจากสิ่งแวดล้อมและตัวอุปกรณ์ ซึ่งมีการเปลี่ยนแปลงตามเวลา การประมวลผลสัญญาณเบื้องต้นมี 2 กรรมวิธี ดังนี้

3.1.1 กรรมวิธีเน้นล่วงหน้า (Preemphasis)

ขั้นตอนนี้เป็นการบีบอัดช่วงพิสัยพลวัต (Dynamic Range) ของสัญญาณเสียงพูด โดยการลดความแปรปรวนของสัญญาณด้วยการพิจารณาความเปลี่ยนแปลงของสัญญาณในแต่ละช่วงเวลา ดังสมการที่ 3.1

$$\tilde{s}[n] = s[n] - as[n-1] \quad (3.1)$$

จากสมการที่ 3.1 จะส่งผลให้ค่าอัตราส่วนสัญญาณต่อรบกวน (SNR) มีค่าสูงขึ้น ซึ่งในทางปฏิบัติแล้วจะนำสัญญาณมาผ่านตัวกรองดิจิทัลอันดับที่หนึ่ง (First-Order Digital Filter) ที่มีฟังก์ชันถ่ายโอนดังสมการที่ 3.2

$$H[z] = 1 - az^{-1} \quad (3.2)$$

โดยกำหนดให้สัมประสิทธิ์ของตัวกรอง a มีค่าเป็น 0.95 เนื่องจากเป็นค่าที่ให้ผลดีที่สุดสำหรับการวิเคราะห์สัญญาณเสียงพูด [14]

โดยที่	a	เป็นสัมประสิทธิ์ของตัวกรอง
	$\tilde{s}[n]$	เป็นสัญญาณเสียงพูดขาออกที่ผ่านกรรมวิธีเน้นล่วงหน้าเป็นเวลา n
	$s[n]$	เป็นค่าของสัญญาณเสียงพูดขาเข้าเป็นเวลา n

3.1.2 วงจรกรองผ่านสูง (High-pass filter)

เนื่องจากในการบันทึกเสียงมักจะมีเสียงรบกวนของแหล่งจ่ายไฟ (50 Hz) และความถี่ต่ำๆ ที่มาจากการพ่นลมขณะพูด ปนเข้ามาในสัญญาณ ซึ่งจะมีผลกระทบกับการรู้จำของระบบ เราจึงต้องกำจัดสัญญาณเหล่านี้ด้วยตัวกรองความถี่สูงผ่าน สำหรับงานวิจัยนี้ใช้ตัวกรองชนิดบัตเตอร์เวิร์ท (Butterworth Filter : BW) อันดับที่ 10 มีความถี่คัตออฟที่ 60 Hz

3.2 การตรวจหาหน่วยเสียง (Phoneme Detection)

เป็นขั้นตอนของการหาตำแหน่งที่แน่นอนของหน่วยเสียงทั้งหมดว่าอยู่ตำแหน่งใดของรูปสัญญาณในแต่ละพยางค์ จากการศึกษาลักษณะและโครงสร้างของภาษาไทยจะพบว่าการออกเสียงแต่ละคำจะประกอบไปด้วยหน่วยเสียงหลัก 4 หน่วยด้วยกัน คือ

- หน่วยเสียงพยัญชนะต้น (Initial Consonant)
- หน่วยเสียงสระ (Vowel)
- หน่วยเสียงวรรณยุกต์ (Tone)
- หน่วยเสียงพยัญชนะท้ายหรือตัวสะกด (Final Consonant)

ด้วยโครงสร้างของภาษาไทยทำให้เราทราบอย่างคร่าวๆว่าหน่วยเสียงของพยัญชนะต้นจะอยู่ส่วนหน้า หน่วยเสียงสระและวรรณยุกต์อยู่ส่วนกลาง และหน่วยเสียงตัวสะกดอยู่ส่วนท้ายของรูปสัญญาณ การตรวจหาหน่วยเสียงจะประกอบด้วยขั้นตอนย่อยดังต่อไปนี้

3.2.1 การแปลงเวฟเลต (Wavelet Transform)

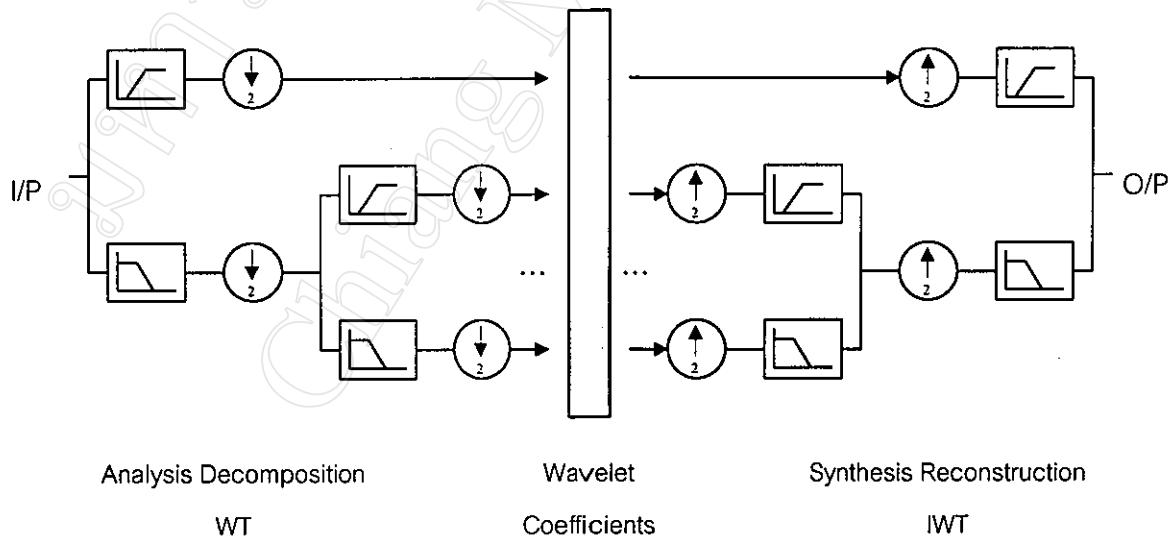
สัญญาณเสียงพูดนั้นจะมีข้อมูลที่สำคัญอยู่ทั้งในความถี่ต่ำและความถี่สูง การเปลี่ยนแปลงความถี่ ณ รอยต่อของหน่วยเสียงส่วนใหญ่จะอยู่ในย่านความถี่สูง และแตกต่างกันออกไปในแต่ละหน่วยเสียงที่ประกอบอยู่ในพยางค์ งานวิจัยนี้จึงใช้การแปลงเวฟเลตมาช่วยในการหาหน่วยเสียง ซึ่งจะทำให้ความแม่นยำของตำแหน่งที่มีการเปลี่ยนแปลงความถี่ ได้ดีกว่าการแปลงฟูเรียร์เวลาสั้น (Short Time Fourier Transform) การแปลงเวฟเลตสามารถอธิบายได้ดังสมการที่ 3.3

$$\psi_{b,a}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) \quad (3.3)$$

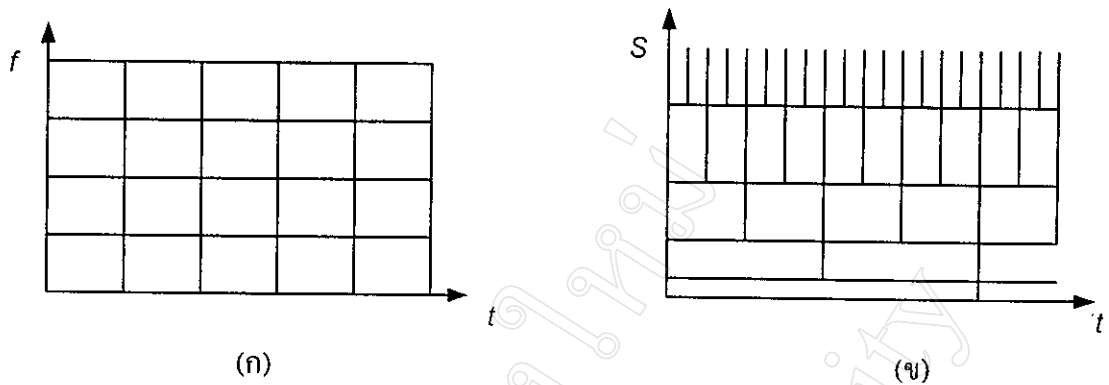
เมื่อ $\psi(t)$ เป็นฟังก์ชันเวฟเลตต้นกำเนิด ที่ถูกสเกลและถูกเลื่อนตำแหน่งโดยพารามิเตอร์ a และ b ตามลำดับ และทำการนอร์มอลไลซ์ด้วย $\frac{1}{\sqrt{a}}$

หลักการของเวฟเลตคือ การแยกสัญญาณออกเป็นส่วนประกอบเล็กๆ ที่สัมพันธ์กัน โดยที่ส่วนประกอบเล็กๆ เหล่านี้จะอยู่ในรูปของเวฟเลตที่ถูกสเกลและเลื่อนตำแหน่ง จึงกล่าวได้ว่าสัญญาณใดๆ สามารถสร้างขึ้นมาได้โดยฟังก์ชันพื้นฐาน(Basis Function) ดังนั้นการกระจายเวฟเลต (Wavelet Decomposition) ก็คือการแปลงเวฟเลต (Wavelet Transform : WT) นั่นเอง ในทำนองเดียวกันการรวมกลับเวฟเลต (Wavelet Reconstruction) จะเป็นการแปลงกลับเวฟเลต (Inverse Wavelet Transform : IWT) ซึ่งเป็นการนำเอาส่วนประกอบย่อยๆ เหล่านี้มารวมกันเพื่อประกอบเป็นสัญญาณเดิมดังรูปที่ 3.1

เนื่องจากการกระจายของเวฟเลตแต่ละระดับจะมีการลดการสุ่ม(Down Sampling)ทำให้ระนาบของเวฟเลตจะต่างไปจาก STFT ดังรูปที่ 3.2



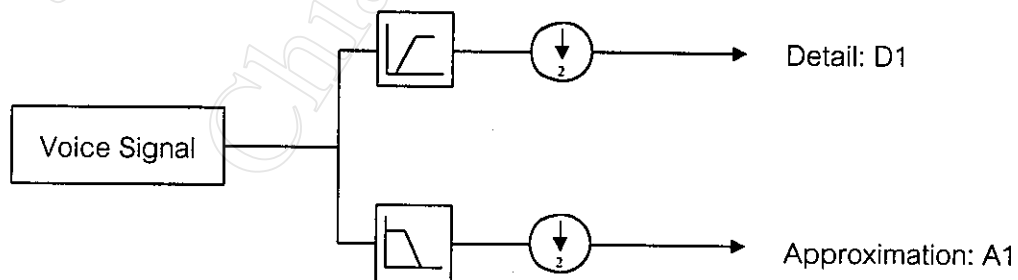
รูปที่ 3.1 การแปลงเวฟเลต (WT) และการแปลงกลับเวฟเลต (IWT)



รูปที่ 3.2 ระยะเวลา-ความถี่ ของ (ก) Short Time Fourier Transform, Δf และ Δt มีค่าคงที่ตลอดระยะเวลา (ข) Wavelet Transform, Δf กว้างขึ้นและ Δt ลดลงเมื่อ ΔS เพิ่มขึ้น

3.2.2 การหาขอบเขตของหน่วยเสียง

บริเวณรอยต่อของหน่วยเสียงพ้องจะขึ้นต้นกับสระ และสระกับพยัญชนะท้ายจะมีลักษณะเด่นที่มีการเปลี่ยนแปลงของความถี่โดยฉับพลัน การเปลี่ยนแปลงนั้นอาจเป็นการเปลี่ยนแปลงในระดับความถี่ต่ำอย่างเดียว, ในระดับความถี่สูงอย่างเดียว หรือทั้งสองระดับความถี่ สำหรับงานวิจัยนี้จะกระจายเวฟเลตออกเป็น 1 ระดับ โดยใช้ฟังก์ชันพื้นฐาน Daubechies ดังรูปที่ 3.3 จะได้องค์ประกอบย่อยๆ ของสัญญาณออกมา 2 ชุด ซึ่งจะเป็นสัญญาณในย่านความถี่ต่ำและสูง สัญญาณที่ถูกแยกออกมาด้วยวงจรกรองความถี่ต่ำผ่าน จะเรียกว่าส่วนประมาณ Approximation (A) ส่วนสัญญาณที่ถูกแยกออกมาด้วยวงจรกรองความถี่สูงผ่าน จะเรียกว่าส่วนรายละเอียด Detail (D)

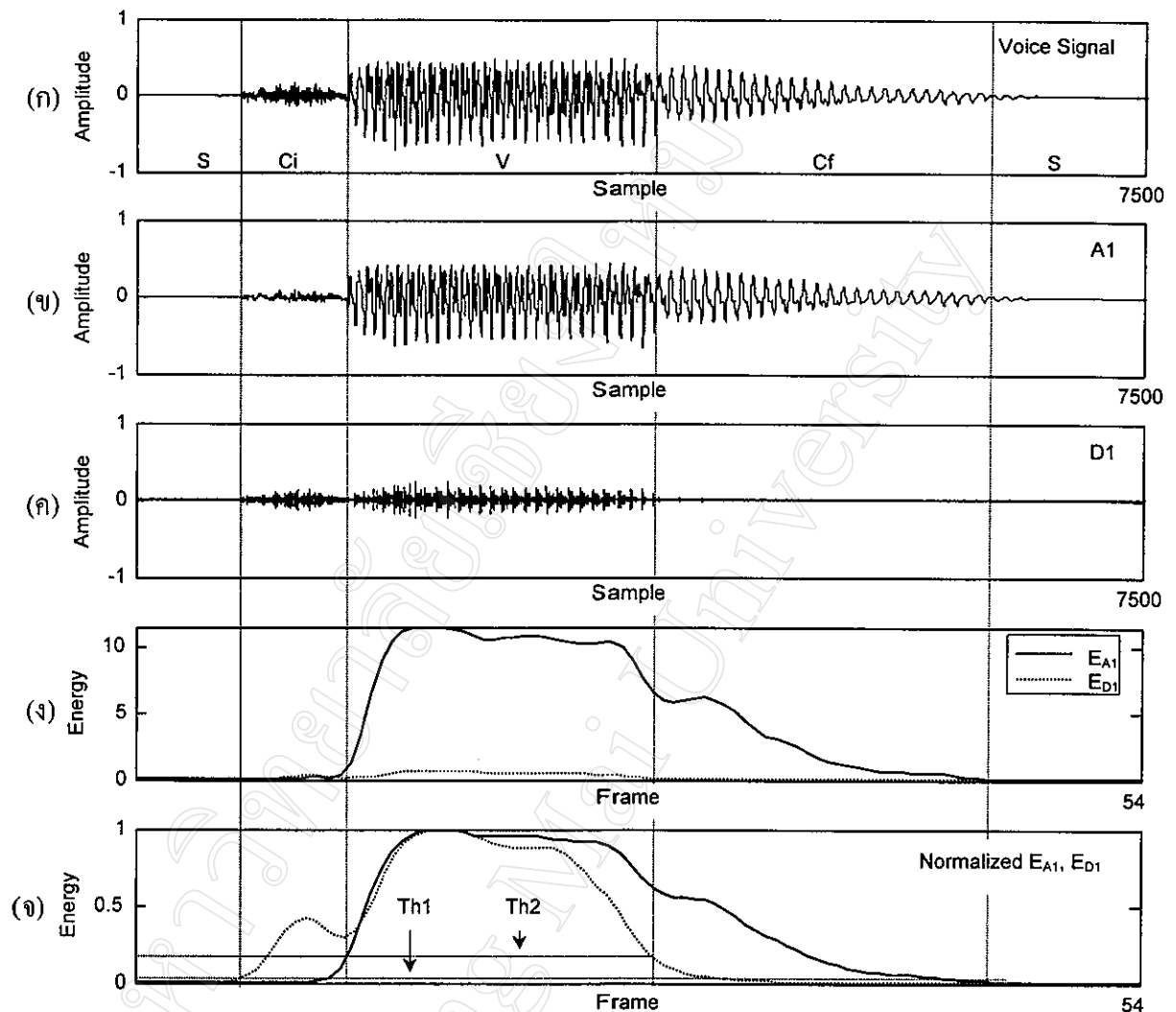


รูปที่ 3.3 การกระจายเวฟเลตของสัญญาณ 1 ระดับ

จากนั้นหาค่าพลังงานกำลังสองของทั้ง A_1 และ D_1 ดังแสดงในรูปที่ 3.4 (ง) จะเห็นว่าค่าพลังงานของ D_1 (E_{D_1}) นั้นน้อยกว่าของ A_1 (E_{A_1}) มาก เนื่องจากเสียงพูดจะประกอบไปด้วยความถี่ต่ำที่มีพลังงานสูง เนื่องจากความถี่สูงก็มีความสำคัญเท่าเทียมกับความถี่ต่ำดังนั้นจึงต้องนอร์มอลไลซ์ E_{A_1} และ E_{D_1} ให้มีขนาดเท่ากัน ดังแสดงในรูปที่ 3.4 (จ) พิจารณาที่ E_{D_1} จะเห็นความเปลี่ยนแปลงค่าพลังงานของ E_{D_1} อย่างชัดเจน

เราจะหาจุดเริ่มต้นของพยางค์ด้วยการเคลื่อนที่จากตำแหน่งที่มีค่าพลังงานสูงสุดของแต่ละ E_{A_1} และ E_{D_1} มาทางด้านหน้าของพยางค์ โดยที่จุดสุดท้ายที่มีค่าติดกับค่าขีดเริ่ม (Th_1) จะถือว่าเป็นจุดเริ่มต้นพยางค์ หาจุดเริ่มต้นของหน่วยเสียงสระ ด้วยการพิจารณาผลรวมในแต่ละ E_{A_1} และ E_{D_1} ในช่วงตั้งแต่จุดเริ่มต้นพยางค์จนถึงค่าขีดเริ่ม (Th_2) ของ E_{A_1} ถ้าพลังงานรวมของ E_{D_1} มากกว่าของ E_{A_1} แสดงว่าพยัญชนะต้นเป็นพยัญชนะที่ประกอบด้วยความถี่สูง เราจะเลือกจุดตัดระหว่าง E_{A_1} กับ Th_2 เป็นจุดเริ่มต้นหน่วยเสียงสระ แต่ถ้าพลังงานรวมของ E_{A_1} มากกว่าของ E_{D_1} แสดงว่าพยัญชนะต้นเป็นพยัญชนะที่ประกอบไปด้วยความถี่ต่ำ เราจะเลือกจุดตัดระหว่าง E_{D_1} กับ Th_2 เป็นจุดเริ่มต้นหน่วยเสียงสระ หากจุดสิ้นสุดของหน่วยเสียงสระ เริ่มจากการเคลื่อนที่จากตำแหน่งที่มีค่าพลังงานสูงสุดของ E_{D_1} ไปทางด้านท้ายของพยางค์ โดยที่จุดตัดระหว่าง E_{D_1} กับ Th_2 ถือว่าเป็นจุดสิ้นสุดของหน่วยเสียงสระ ส่วนการหาจุดสิ้นสุดของพยางค์ เริ่มจากการเคลื่อนที่จากตำแหน่งที่มีค่าพลังงานสูงสุดของ E_{A_1} ไปทางด้านท้ายของพยางค์ โดยที่จุดตัดระหว่าง E_{A_1} กับ Th_1 ถือว่าเป็นจุดสิ้นสุดของพยางค์

เมื่อทราบขอบเขตของหน่วยเสียงสระแล้ว หน่วยเสียงที่เหลือก็จะเป็นพยัญชนะต้นและพยัญชนะท้าย เราจะนำสัญญาณบริเวณส่วนกลางของหน่วยเสียงสระออกมา 30 ms ($f_s \times 30\text{ms} \cong 330$ ตัวอย่าง) เพื่อเป็นตัวแทนของหน่วยเสียงสระนั้นๆ เนื่องจากรูปสัญญาณของเสียงสระจะมีความเป็นรายคาบ คือมีรูปสัญญาณที่ซ้ำๆ กัน ดังรูปที่ 3.6 (ก)



Ci: Initial Consonant, V: Vowel, Cf: Final Consonant, S: Silence

รูปที่ 3.4 ตัวอย่างการหาขอบเขตของหน่วยเสียงของคำว่า “ทีม”

3.3 การสกัดลักษณะเด่น (Feature Extraction)

เป็นการดึงลักษณะเฉพาะของสัญญาณเสียงแต่ละเสียงที่มีไม่เหมือนกันออกมา แล้วให้ระบบจดจำลักษณะเด่นของเสียงแต่ละเสียงไว้ เมื่อสัญญาณที่เข้ามาภายหลังมีลักษณะเด่นที่เหมือนหรือใกล้เคียงกับลักษณะเด่นของเสียงใด ระบบก็จะสามารถบอกได้ว่าเป็นเสียงอะไร หรือใกล้เคียงกับเสียงใดมากที่สุด สำหรับงานวิจัยนี้ใช้ลักษณะเด่นสำหรับหน่วยเสียงต่างๆ ดังนี้

3.3.1 Linear Predictive Coding(LPC) coefficients

เป็นการหาลักษณะเด่นของเสียงสระ เนื่องจากเป็นวิธีหนึ่งที่นิยมใช้ในการรู้จำสัญญาณที่มีความเป็นรายคาบ ดังรูปที่ 3.5 ประกอบไปด้วย 4 ขั้นตอนหลัก ดังนี้

- Frame Blocking

เป็นการแบ่งข้อมูลออกเป็นเฟรม ที่มีขนาดเท่ากัน โดยที่ข้อมูลในเฟรมข้างเคียงจะมีการซ้อนทับกัน(Overlap) เพื่อลดความไม่ต่อเนื่องของสัญญาณบริเวณรอยต่อของข้างเคียง สามารถเขียนเป็นสมการได้ดังสมการที่ 3.4

$$x_l(n) = \tilde{s}(Ml + n), \quad n = 0, 1, 2, \dots, N-1, \quad l = 0, 1, 2, \dots, L-1 \quad (3.4)$$

โดยที่	N	=	ขนาดของเฟรม
	M	=	$N - \text{Overlap}$
	l	=	เฟรมที่
	L	=	จำนวนเฟรมทั้งหมด

- Windowing

เป็นขั้นตอนของการลดความไม่ต่อเนื่องกันของสัญญาณ ณ รอยต่อของเฟรมข้างเคียง ชนิดของวินโดว์ที่นิยมใช้กับสัญญาณเสียง คือ Hamming window ดังสมการที่ 3.5

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \quad (3.5)$$

เมื่อคูณกับสัญญาณในแต่ละเฟรมแล้ว จะได้

$$\tilde{x}_l(n) = x_l(n)w(n), \quad 0 \leq n \leq N-1 \quad (3.6)$$

- Autocorrelation Analysis

เป็นการนำสัญญาณในแต่ละวินโดว์มาทำอัตสหสัมพันธ์ ดังสมการที่ 3.7

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m), \quad m = 0, 1, 2, \dots, p \quad (3.7)$$

เมื่อ $r_i(m)$ คือ ค่าที่ได้จากอัตราสัมพันธ์ ซึ่งจะมีขนาดเท่ากับ $p+1$ ค่า โดยที่ p เป็นอันดับของการวิเคราะห์ LPC

- LPC Analysis

เป็นการแปลงค่าที่ได้มาจากขั้นตอนอัตราสัมพันธ์มาเป็นสัมประสิทธิ์ LPC โดยใช้วิธีของDurbin ดังสมการที่ 3.8-3.12 (เพื่อความสะดวกเราจะละตัวห้อย l)

$$E^{(0)} = r(0) \quad (3.8)$$

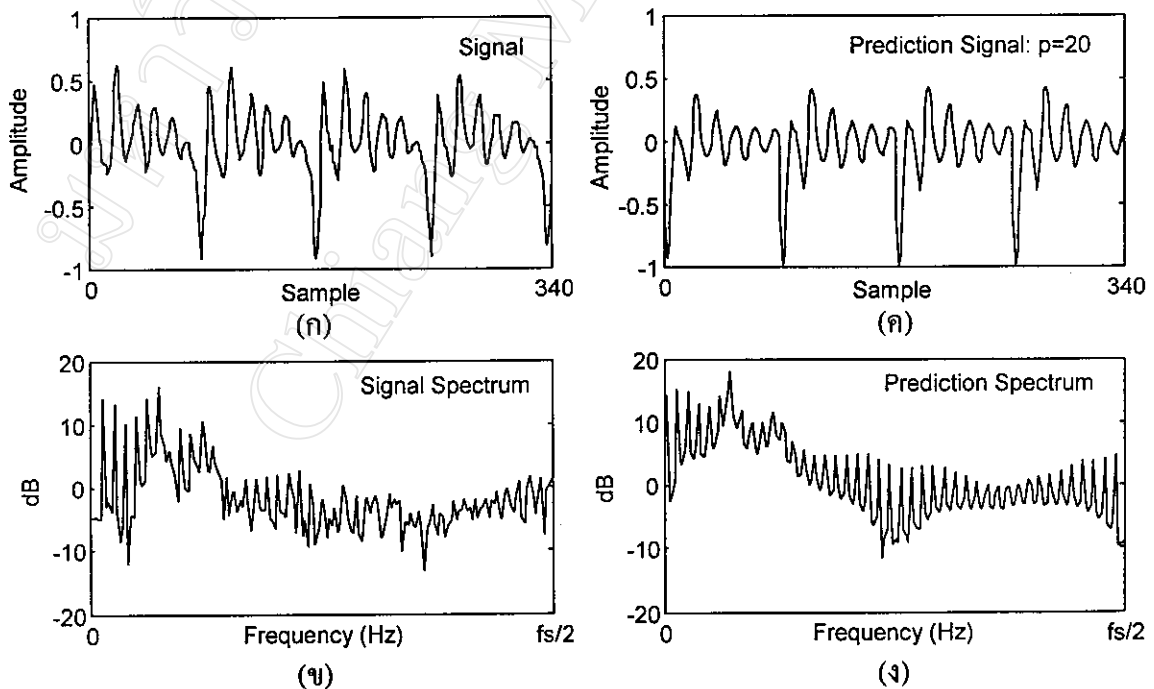
$$k_i = \left\{ r(i) - \sum a_j^{(i-1)} r(i-j) \right\} / E^{(i-1)}, \quad 1 \leq i \leq p \quad (3.9)$$

$$a_i^{(i)} = k_i \quad (3.10)$$

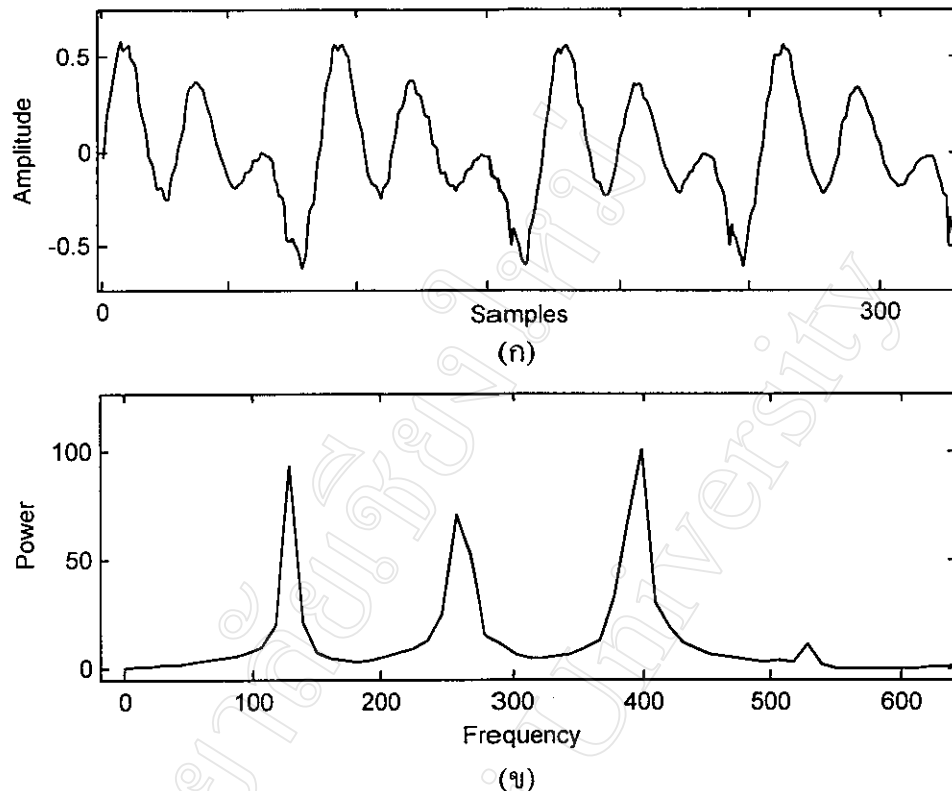
$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \quad (3.11)$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \quad (3.12)$$

โดยที่ a_m = สัมประสิทธิ์ LPC , $1 \leq m \leq p$



รูปที่ 3.5 ตัวอย่างของ (ก) สัญญาณเสียงสระ ซึ่งมี (ข) สเปกตรัม จากนั้น (ค) ทำนายสัญญาณจากสัมประสิทธิ์ Linear Predictive Coding ซึ่งมี (ง) สเปกตรัมของสัญญาณที่ได้จากการทำนาย



รูปที่ 3.6 (ก) สัญญาณเสียงสระในโดเมนเวลา (ข) สัญญาณเสียงรูป ก) ที่ผ่านการแปลงฟูเรียร์

3.3.2 องค์ประกอบความถี่ (Frequency Component)

เป็นลักษณะเด่นของหน่วยเสียงสระ พยัญชนะต้น และพยัญชนะท้ายที่มีเสียงยาว ลักษณะเด่นนี้เกิดขึ้นจากลักษณะการสั่นของสายเสียง ซึ่งจะมีจำนวนองค์ประกอบความถี่แตกต่างกันไป องค์ประกอบความถี่หาได้ดังนี้

- การแปลงฟูเรียร์ (Fourier Transform)

การแปลงฟูเรียร์เป็นการมองสัญญาณอยู่ในแกนความถี่กับกำลังงานของความถี่ ทำให้เราสามารถเห็นช่วงของความถี่ที่มีกำลังงานมากอย่างชัดเจน ดังรูปที่ 3.6 (ข)

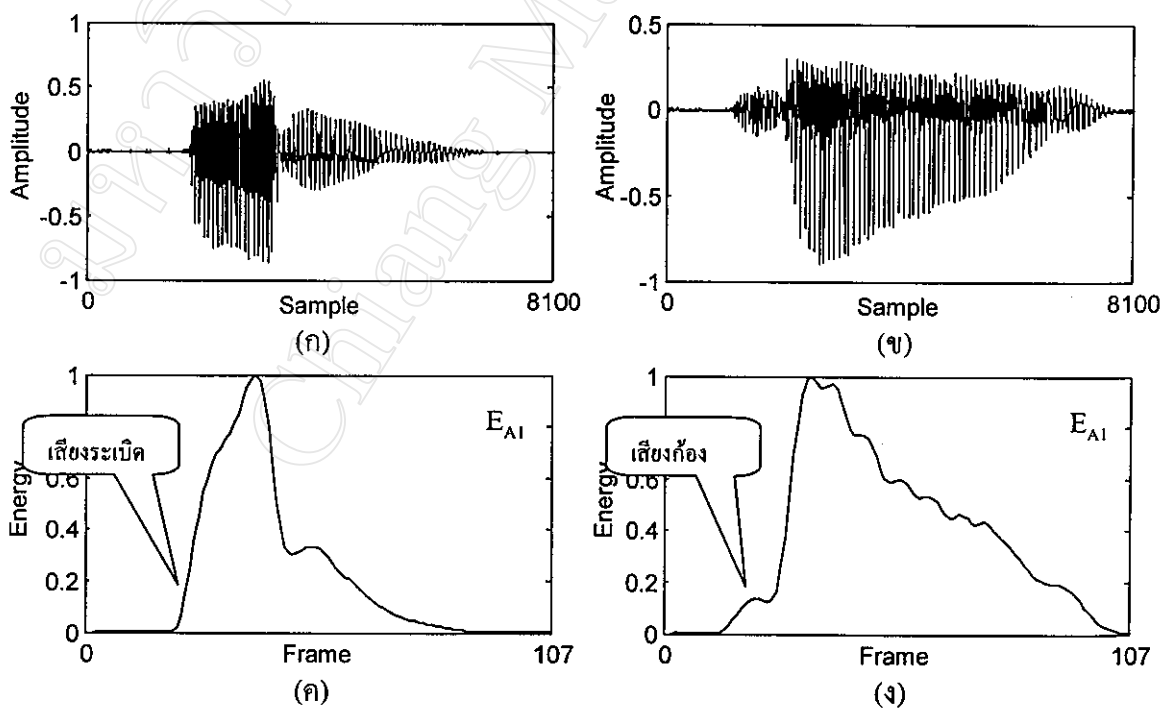
- วิธีนับองค์ประกอบความถี่

จากรูปที่ 3.6 (ข) จะเห็นจำนวนองค์ประกอบความถี่อย่างชัดเจนว่ามี 3 องค์ประกอบ สำหรับการนับนั้นเราจะหาค่ากำลังสูงสุด(Global Maximum) แล้วตั้งค่าขีดเริ่มต้นที่ 15 เปอร์เซ็นต์ของค่ากำลังสูงสุด สร้างเฟรมที่มีความกว้าง 75 Hz (ต่ำกว่าความถี่มูลฐานของมนุษย์) แล้วค่อยๆ เพิ่มขนาดความกว้างของเฟรมครั้งละ 25 Hz แล้วหาค่าสูงสุดในเฟรม จนได้ค่าสูงสุด(Local Maximum) ที่ไม่ต่ำกว่า 80 เปอร์เซ็นต์ของค่าสูงสุด เนื่องจากส่วนใหญ่กำลังของความถี่มูลฐานจะมีค่าสูงที่สุดหรือไม่ก็ไม่ต่ำกว่า 80 เปอร์เซ็นต์ของค่าสูงสุด เมื่อได้ตำแหน่งของความถี่มูลฐานมาแล้ว จึงสร้าง

เฟรมที่มีความกว้าง 1.5 เท่าของความถี่มูลฐาน แล้วเลื่อนเฟรมไปเรื่อยๆ จนถึงจุดสิ้นสุด โดยมีการทับซ้อน 75 เปอร์เซ็นต์ ในการเลื่อนแต่ละครั้งจะหาค่าสูงสุดของเฟรมทุกครั้ง ค่าสูงสุดที่สูงกว่าค่าขีดเริ่มจะถือว่าเป็นจุดยอด ท้ายที่สุดเราจะได้ตำแหน่งที่มีเป็นจุดยอดทั้งหมด อาจมีบางตำแหน่งที่ซ้ำซ้อนกันแต่เราจะนับเป็นจุดยอดเดียว จุดยอดที่นับได้ก็จะเป็นจำนวนองค์ประกอบความถี่นั่นเอง

3.3.3 เสียงระเบิด (Plosive)

เป็นลักษณะเด่นของหน่วยเสียงพยัญชนะต้นและพยัญชนะท้าย ส่วนในพยัญชนะท้ายจะเรียกว่าเสียงกักลม (Stop) เสียงระเบิดนั้นเกิดมาจากการเปลี่ยนแปลงพลังงานอย่างฉับพลัน เกิดขึ้นในกรณีที่มีการกักลมไว้โดยริมฝีปาก ลิ้น โคนลิ้น และสายเสียงเช่นในเสียง ป, ท, ข และ อ ตามลำดับ การตรวจสอบเสียงระเบิดนั้นทำได้โดยการวัดการเปลี่ยนแปลงพลังงานในช่วงที่สนใจ ซึ่งก็คือความชันของค่าพลังงานนั่นเอง ถ้ามีความชันมากก็แสดงว่ามีการเปลี่ยนแปลงของพลังงานมาก พลังงานที่เราสนใจนั้นจะอยู่ที่พลังงานของความถี่ต่ำ (E_{A1}) ที่ได้มาจากการกระจายเวฟเลขในกระบวนการข้างต้น ดังรูปที่ 3.7 (ค) ช่วงของการระเบิดจะอยู่ในขณะที่เริ่มเกิดเสียงสระ จนถึงจุดที่สระเริ่มมีพลังงานคงที่ ค่าความชันสูงสุดของพลังงานจะเป็นค่าของการระเบิด



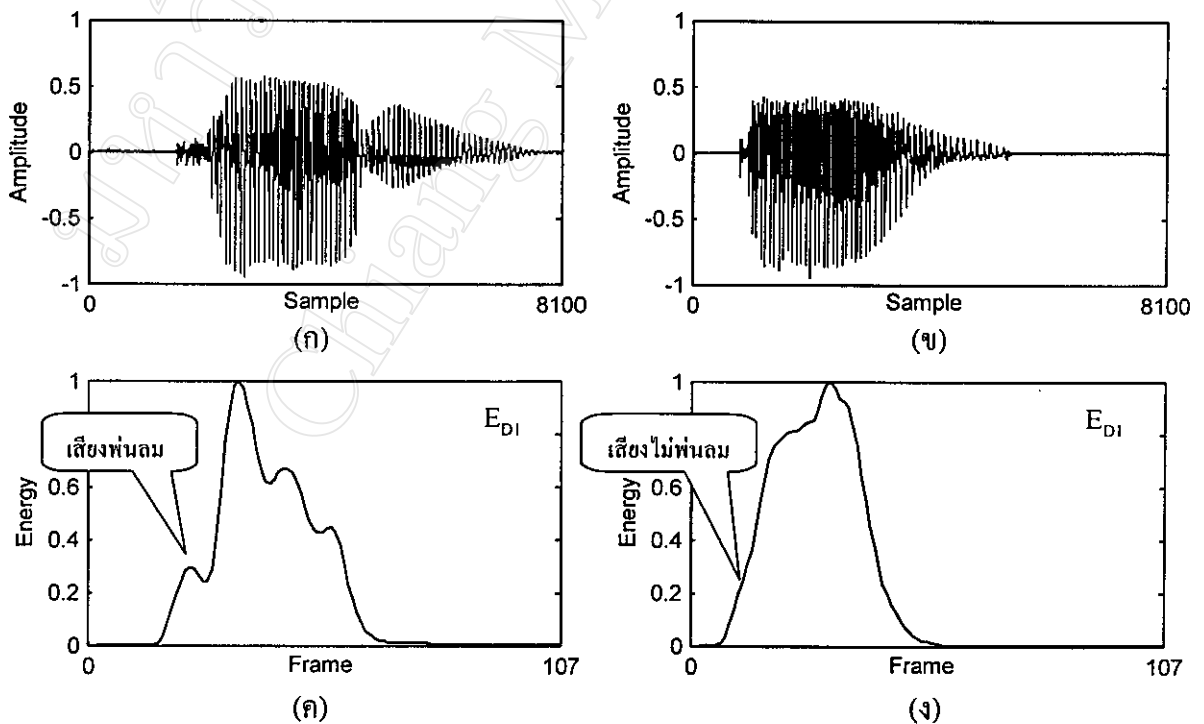
รูปที่ 3.7 ตัวอย่างเสียงระเบิด จากคำว่า “เป็น” และ เสียงก้อง จากคำว่า “ดาว”

3.3.4 เสียงก้อง และไม่ก้อง (Voiced and Voiceless)

เป็นลักษณะเด่นของหน่วยเสียงพยัญชนะต้น เสียงก้องคือเสียงที่มีความเป็นรายคาบ เช่นเสียง ค, ร, ม, บ เป็นต้น ส่วนเสียงไม่ก้องคือเสียงที่ไม่มีความเป็นรายคาบ เช่นเสียง พ, ฟ เป็นต้น หรือไม่มีเสียง เช่นเสียง ป, จ เป็นต้น สามารถหาได้จากพลังงานของสัญญาณที่ผ่านการกระจายเวฟเลทในขั้นตอนข้างต้น เนื่องจากเสียงก้องในภาษาไทยนั้นจะอยู่ในช่วงความถี่ต่ำ ดังนั้นสามารถวิเคราะห์ค่าพลังงานจาก E_{A1} ได้ ดังรูปที่ 3.7 (ง) ส่วนเสียงไม่ก้องจะอยู่ในช่วงความถี่สูง หรือไม่มีก็อยู่ในช่วงความถี่เป็นศูนย์ สามารถวิเคราะห์ค่าพลังงานได้จาก E_{D1} ค่าที่วิเคราะห์ได้นี้จะเป็นค่าของพลังงานในช่วงที่สนใจ

3.3.5 เสียงพ่นลม และไม่พ่นลม (Aspirated and Unaspirated)

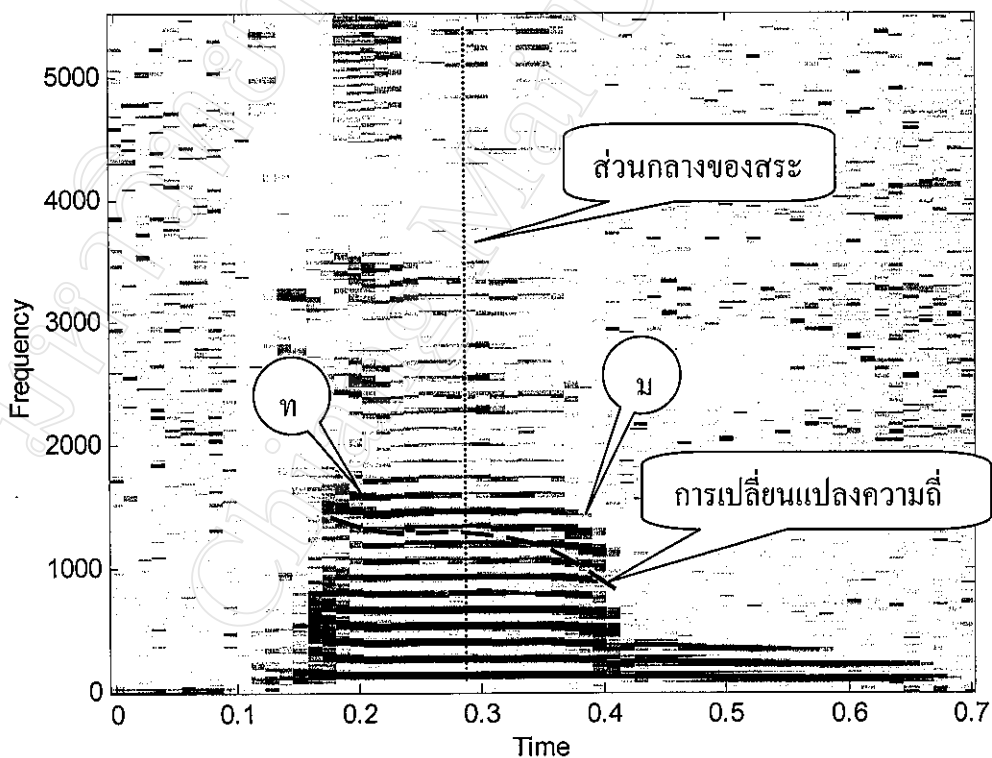
เป็นลักษณะเด่นของหน่วยเสียงพยัญชนะต้น เกิดจากขณะออกเสียงนั้นจะมีการพ่นลมออกมา เช่น พ, ข, ฉ เป็นต้น การพ่นลมออกมานั้นทำให้เกิดสัญญาณความถี่สูง และมีพลังงานมากกว่าสัญญาณรบกวน ดังนั้นสามารถวิเคราะห์ได้จากค่าพลังงานของสัญญาณ E_{D1} ในช่วงที่สนใจ ดังรูปที่ 3.8 (ค) และ (ง)



รูปที่ 3.8 ตัวอย่างเสียงพ่นลม จากคำว่า “เทอม” และเสียงไม่พ่นลม จากคำว่า “ก้าว”

3.3.6 เสียงที่เกิดจากริมฝีปากกับริมฝีปากและเสียงที่เกิดจากลิ้นกับปุ่มเหงือก (Labial and Alveolar)

เป็นลักษณะเด่นของหน่วยเสียงพยัญชนะต้นและพยัญชนะท้าย เกิดขึ้นก่อนออกเสียง ริมฝีปากจะปิดสนิท ในการออกเสียง ป, พ, บ, ม, และห่อริมฝีปากในเสียง ว ในกรณีจบการออกเสียงพยัญชนะท้ายริมฝีปากจะปิดสนิท ในการออกเสียง แม่ กบ, กม และห่อริมฝีปากในแม่ เกว และเกิดขึ้นก่อนการออกเสียง ลิ้นจะสัมผัสปุ่มเหงือกในการออกเสียง ต, ท, ด, น เป็นต้นและลิ้นเกือบสัมผัสเพดานปากในเสียง ย ในกรณีจบการออกเสียงใน แม่ กด, กน และลิ้นเกือบสัมผัสเพดานปากในแม่ เกย วิเคราะห์ลักษณะเด่นนี้จากความถี่ในย่าน 1000-2000 Hz ที่เปลี่ยนแปลงในช่วงต้นของหน่วยเสียงสระในการวิเคราะห์พยัญชนะต้น และช่วงท้ายของหน่วยเสียงสระในการวิเคราะห์พยัญชนะท้าย ลักษณะการเปลี่ยนแปลงความถี่ที่แสดงในรูปที่ 3.9



รูปที่ 3.9 การเปลี่ยนแปลงความถี่ในหน่วยเสียงสระของคำว่า “thom”

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

(ก)

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

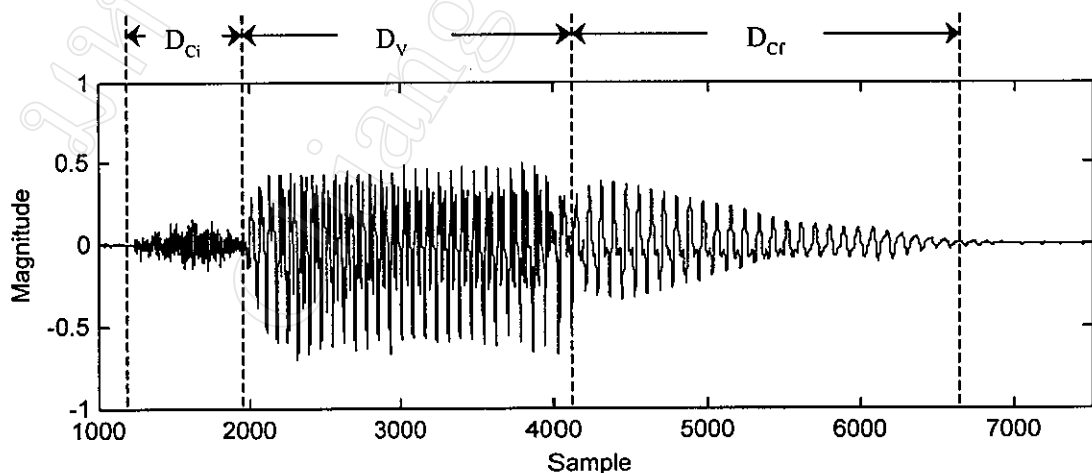
(ข)

รูปที่ 3.10 ตัวกรองภาพสเปกโตรแกรม (ก) บนซ้าย-ล่างขวา และ (ข) ล่างซ้าย-บนขวา

เนื่องจากลักษณะของความถี่ที่มีพลังงานสูงอยู่ในแนวลาดเอียง ดังนั้นในการตรวจสอบทิศทางของการลาดเอียงนั้นเราจะใช้ตัวกรองในแนวทะแยง กรองภาพสเปกโตรแกรม ในช่วงที่เราสนใจ ผลรวมที่มากกว่าของพลังงานที่ผ่านตัวกรองจะบอกถึงแนวของการลาดเอียงของความถี่ ตัวกรองมี 2 แบบ คือ ตัวกรองในแนวทะแยงจาก บนซ้าย-ล่างขวา กับ ล่างซ้าย-บนขวา ตัวกรองในที่นี่จะมีลักษณะเป็นเมตริกซ์ทะแยงที่มีขนาด 5×5 ดังรูปที่ 3.10

3.3.7 ระยะเวลาการเกิดเสียง (Time Duration)

เป็นลักษณะเด่นของทั้งหน่วยเสียงพยัญชนะและสระ วิเคราะห์ได้จากการวัดระยะเวลาของหน่วยเสียงนั้นๆ ว่าเป็นเสียงสั้น หรือเสียงยาว มีหน่วยเป็นวินาที ดังรูปที่ 3.11



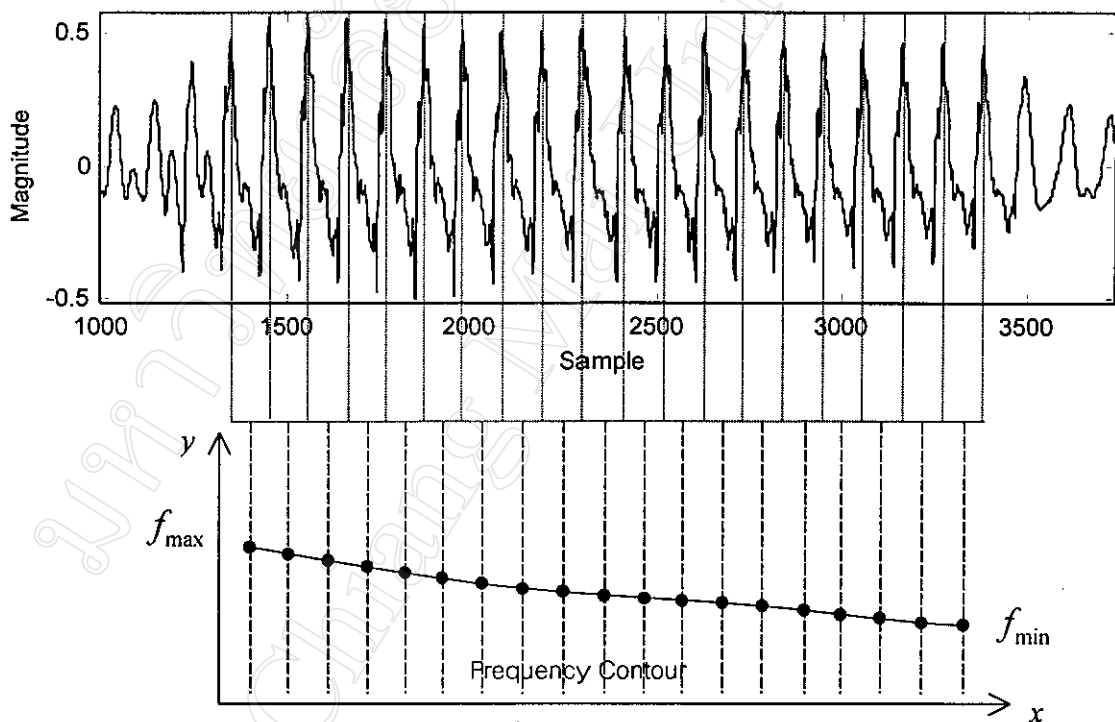
D_{Ci} : Initial Consonant Duration, D_V : Vowel Duration, D_{Cr} : Final Consonant Duration

รูปที่ 3.11 ระยะเวลาของการเกิดเสียงของหน่วยเสียงต่างๆ

3.3.8 การประมาณส่วนโค้ง (Curve Fitting)

เป็นลักษณะเด่นของเสียงวรรณยุกต์ เป็นขั้นตอนของการวิเคราะห์หาเสียงวรรณยุกต์จากหน่วยเสียงวรรณยุกต์ที่ปรากฏอยู่ในหน่วยเสียงสระ [9] เสียงวรรณยุกต์จะมีลักษณะของการเปลี่ยนแปลงความถี่ในรูปแบบต่างๆ ไปตามเวลา เสียงวรรณยุกต์ทั้ง 5 เสียง [1] มีรูปแบบดังรูปที่ 2.1

เราจะใช้ระดับเสียง (Pitch) ในการจำแนกวรรณยุกต์ทั้ง 5 ออกจากกัน คุณลักษณะเด่นของเสียงวรรณยุกต์ คือ การเปลี่ยนแปลงของความถี่ในย่านความถี่มูลฐานไปตามเวลา ซึ่งก็คือพิสัยการเปลี่ยนระดับเสียง (Pitch Range) นั่นเอง ดังรูปที่ 3.12 หลังจากทราบพิสัยการเปลี่ยนระดับเสียงแล้วเราจะใช้ การประมาณส่วนโค้ง ขนาด 20 จุด เพื่อป้อนให้โครงข่ายประสาทเทียม



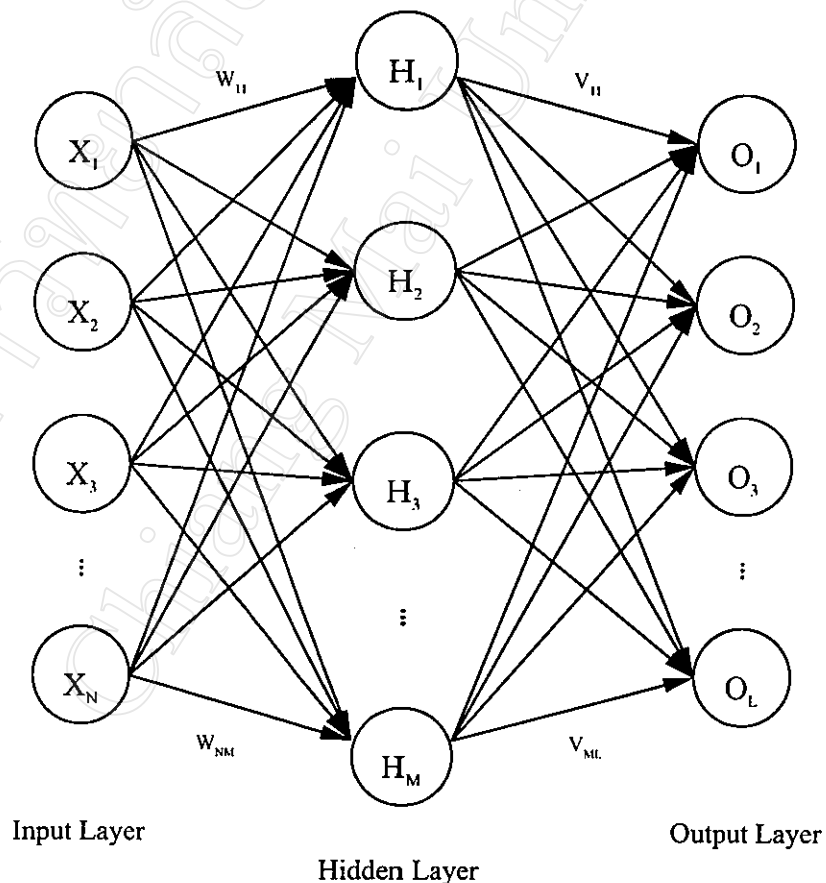
รูปที่ 3.12 รูปแบบเสียงวรรณยุกต์สามัญ ของคำว่า “นาม” ที่ได้จากการวิเคราะห์พิสัยการเปลี่ยนระดับเสียง

3.4 การหาความคล้ายคลึงกันของรูปแบบ (Pattern Similarity Determination)

เป็นขั้นตอนของการเปรียบเทียบหาความคล้ายคลึงกันระหว่างลักษณะเด่นของหน่วยเสียงที่ได้มาจากขั้นตอนข้างต้นกับลักษณะเด่นของหน่วยเสียงที่ระบบได้จดจำไว้ก่อนแล้ว สำหรับงานวิจัยนี้จะใช้โครงข่ายประสาทเทียมในการจดจำลักษณะเด่นทั้งหมด

โครงข่ายประสาทเทียมแบบแพร่ไปข้างหน้า (Feed Forward Neural Network)

สำหรับงานวิจัยนี้จะใช้โครงข่ายประสาทเทียมในการรู้จำหน่วยเสียงทั้งหมด โดยเลือกใช้โครงข่ายประสาทเทียมแบบแพร่ไปข้างหน้า ซึ่งเหมาะกับงานด้านการรู้จำเสียงพูด ลักษณะของโครงข่ายดังแสดงในรูปที่ 3.13



รูปที่ 3.13 โครงข่ายประสาทเทียมแบบแพร่ไปข้างหน้า

จากรูปที่ 3.11 $X = \{X_1, X_2, \dots, X_N\}$ เป็นเซตของชั้นอินพุต (Input Layer) , $H = \{H_1, H_2, \dots, H_M\}$ เป็นเซตของชั้นซ่อน (Hidden Layer) , $O = \{O_1, O_2, \dots, O_L\}$ เป็นเซตของชั้นเอาต์พุต (Output Layer) , W เป็นเมตริกซ์ของค่าน้ำหนักที่เชื่อมต่อระหว่างชั้นอินพุตกับชั้นซ่อน , V เป็นเมตริกซ์ของค่าน้ำหนักที่เชื่อมต่อระหว่างชั้นซ่อนกับชั้นเอาต์พุต จากรูปเป็นการเชื่อมโยงกัน 3 ชั้น ประกอบไปด้วยชั้นอินพุต ซึ่งมีเซลล์ประสาทอยู่ N โหนด ถัดมาเป็นชั้นซ่อน ซึ่งมีเซลล์ประสาทอยู่ M โหนด และสุดท้ายคือ ชั้นเอาต์พุต ซึ่งมีเซลล์ประสาทอยู่ L โหนด การเชื่อมโยงจะมีทิศทางไปข้างหน้าเสมอ ไม่มีการเชื่อมโยงกันในชั้นเดียวกัน และการเชื่อมโยงจะต่อถึงกันกับทุก ๆ โหนดที่อยู่ในชั้นถัดไป เราสามารถเพิ่มประสิทธิภาพด้วยการเพิ่มชั้นซ่อนเข้าไปอีกได้ แต่ก็ไม่ได้หมายความว่ายังมีชั้นซ่อนมากแล้วจะทำให้ประสิทธิภาพจะสูงตาม ซึ่งจำนวนชั้นซ่อนนั้นจะต้องเหมาะสมกับงานด้วย

ขั้นตอนการทำงานของโครงข่ายแบบแพร่ไปข้างหน้า เริ่มต้นจากการนำค่าอินพุตมาคำนวณหาผลรวมของแต่ละโหนด แล้วส่งผลลัพธ์ที่ได้ต่อไปยังชั้นถัดไป ทิศทางการไหลของข้อมูลจะมีลักษณะเคลื่อนไปข้างหน้าจนปรากฏผลลัพธ์ที่ชั้นเอาต์พุต วิธีการคำนวณในลักษณะนี้เรียกว่า การป้อนไปข้างหน้า (Feed Forward) ขั้นตอนต่อไปเป็นการคำนวณแบบแพร่กลับ (Back Propagation) ซึ่งจะนำผลลัพธ์ที่ได้จากชั้นเอาต์พุตมาเปรียบเทียบกับเอาต์พุตเป้าหมาย (Target Output) ค่าความผิดพลาดที่ได้จะถูกส่งกลับมายังชั้นเอาต์พุต แต่ละโหนดจะได้รับค่าความผิดพลาดไม่เท่ากัน ขึ้นกับโหนดนั้นมีผลทำให้เกิดความผิดพลาดมากน้อยแค่ไหน และจะทำการนี้กับทุกโหนดในชั้นถัดมาจนทุกๆ โหนดได้รับส่วนแบ่งค่าความผิดพลาดนั้น ขั้นสุดท้ายจะนำส่วนแบ่งค่าความผิดพลาดนั้นมาปรับค่าน้ำหนักของแต่ละโหนด ซึ่งค่าน้ำหนักจะเปลี่ยนไปมากหรือน้อยขึ้นอยู่กับปริมาณของค่าความผิดพลาดที่ได้รับ กระบวนการจะกลับไปทำซ้ำในขั้นตอนแรก และจะวนทำซ้ำจนกระทั่งค่าความผิดพลาดที่ได้มีค่าน้อยกว่าค่าความผิดพลาดค่าสุดท้ายที่กำหนดไว้

3.5 วิธีการตัดสินใจ (Decision Algorithm)

เป็นขั้นตอนที่เกี่ยวข้องกับกฎเกณฑ์และเงื่อนไขต่างๆ ที่ใช้ในการตัดสินใจเลือกลักษณะเด่นที่มีความคล้ายคลึงกันมากที่สุดระหว่างหน่วยเสียงที่ป้อนให้กับระบบกับหน่วยเสียงอ้างอิงที่ได้จัดเก็บเอาไว้ล่วงหน้าแล้ว

หลังจากได้ทำการเลือกหน่วยเสียงอ้างอิงที่มีความคล้ายคลึงกันทุกหน่วยเสียง ก็จะนำอักขระที่แทนหน่วยเสียงอ้างอิงนั้นๆ มาจัดเรียงและแสดงออกมาทางจอภาพเป็นคำอ่านภาษาไทยที่ตรงกับเสียงพูดที่ป้อนให้กับระบบ