บทที่ 3

อัลกอริทึมและเทคนิคต่าง ๆ

ไดนามิคไทม์วาร์ปปิง (Dynamic Time Warping)

แม้ว่าผู้พูดบุคคลเดียวกันจะเปล่งเสียงคำ ๆ หนึ่งออกมา 2 ครั้ง แต่ในเสียงทั้ง 2 เสียงนั้นจะมีช่วงจังหวะของคำต่างกันได้ ซึ่งแสดงถึงการเปลี่ยนแปลงของช่วงเวลาอย่างไม่เชิงเส้นของคำ เพื่อให้การประมวลสัญญาณทำได้ง่ายขึ้น จึงมีการนำเสนอเทคนิคในการปรับยืดขยายรูปคลื่นสัญญาณตามแกนเวลาแบบไดนามิคที่เรียกว่า Dynamic Time Warping (DTW)

DTW เป็นเทคนิคที่อยู่บนพื้นฐานของวิธีการไดนามิคโปรแกรมมิง (Dynamic Programming, DP) [15] ซึ่งวิธีไดนามิคโปรแกรมมิงถูกนำมาประยุกต์ใช้กับ DTW ของเสียงพูดโดย Slutsker, Vintsyuk, Velichko กับ Zagoruyko ชาวรัสเซีย และ Sakoe กับ Chiba ชาวญี่ปุ่น และผลการศึกษาได้ถูกตีพิมพ์ออกไปในเวลาใกล้เคียงกัน จนทำให้วิธีการของไดนามิคโปรแกรมมิงได้รับความนิยมในการประยุกต์ใช้กับการรู้จำเสียงพูดอย่างกว้างขวาง

ตามวิธีของไดนามิคโปรแกรมมิง สมมติให้มีคู่ลำดับ (Sequence) ของเสียง 2 ชุด สำหรับที่จะนำมาเปรียบเทียบกัน ดังนี้

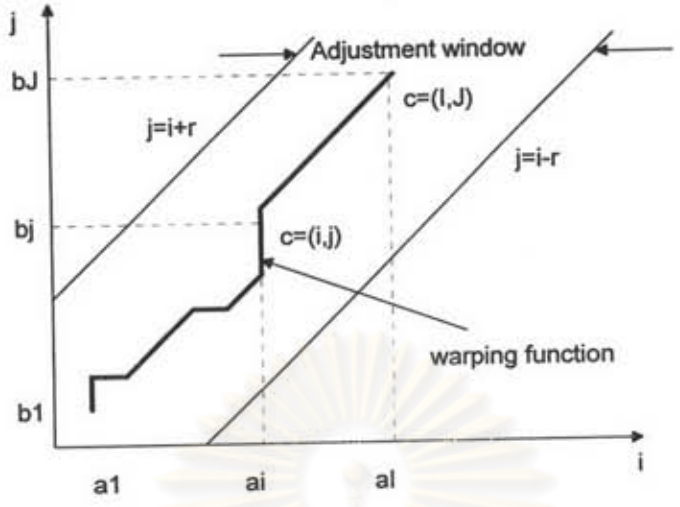
$$A = a_1, a_2, \ldots, a_I$$

และ $$B = b_1, b_2, \ldots, b_J \qquad (3.1)$$

โดยที่ A และ B แทนคู่ลำดับในแต่ละช่วงเวลา

เมื่อพิจารณา A และ B ที่วางตามตำแหน่งในระนาบ ดังรูปที่ 3.1 ซึ่งจะแสดงให้เห็นว่า คู่ลำดับ A และ B บนแกนเวลาสอดคล้องกันกับ วาร์ปปิงฟังก์ชัน (Warping Function) ซึ่งสามารถแทนได้ด้วยลำดับ c = (i, j) ดังนี้

$$F = c_1, c_2, \ldots, c_k, \ldots, c_K, \quad \text{โดยที่} \quad c_k = (i_k, j_k) \qquad (3.2)$$



รูปที่ 3.1 Dynamic Time Warping ระหว่าง A และ B

และเมื่อระยะห่างของสเปคตรัม (Spectral Distance) ระหว่าง A และ B คือ a_i และ b_j ถูกแทนโดย $d(c) = d(i,j)$ และผลรวมของระยะห่างจากจุดเริ่มต้นไปยังจุดสุดท้ายของลำดับตามฟังก์ชัน F สามารถแทนด้วย

$$D(F) = \frac{\sum_{k=1}^{K} d(c_k) w_k}{\sum_{k=1}^{K} w_k} \qquad (3.3)$$

ซึ่งที่ค่า D(F) ที่ได้ยิ่งมีค่าน้อย จะถือว่าความคล้ายกันระหว่าง A กับ B มากขึ้นเท่านั้น โดยที่ w_k เป็น ค่าการถ่วงน้ำหนักแบบค่าบวก (Positive Weighting Function) และสัมพันธ์กับฟังก์ชัน F

สมการที่ (3.3) สามารถทำให้มีค่าน้อยที่สุดได้ โดยคำนึงถึงฟังก์ชัน F ภายใต้เงื่อนไขดังต่อไปนี้

1. เงื่อนไขโมโนโทนี (Monotony Condition)

$$i_k \geq i_{k-1}$$

และ $$j_k \geq j_{k-1} \qquad (3.4)$$

2. เงื่อนไขความต่อเนื่อง (Continuity Condition)

$$i_k - i_{k-1} \leq 1$$

และ $j_k - j_{k-1} \le 1$ (3.5)

จากเงื่อนไขในข้อ 1 และ 2 สามารถนำมารวมกันได้เป็น

$$0 \le i_k - i_{k-1} \le 1$$

และ $0 \le j_k - j_{k-1} \le 1$ (3.6)

3. เงื่อนไขขอบเขต (Boundary Condition)

$$i_1 = j_1 = 1$$

และ $i_K = I, j_K = J$ (3.7)

4. เงื่อนไขการปรับวินโดว์ (Adjustment Window Condition)

$|i_k - j_k| \le r$, r = ค่าคงที่ (3.8)

สำหรับเงื่อนไขข้อที่ 4 มีไว้เพื่อป้องกันการยืดขยายในแกนเวลามากเกินไป และกำหนดให้ค่าการถ่วงน้ำหนักซึ่งเป็นตัวหารในสมการ (3.3) เป็นค่าคงที่ ไม่ขึ้นกับฟังก์ชัน F เช่น เมื่อกำหนดค่าถ่วงน้ำหนักให้เป็น

$$w_k = (i_k - i_{k-1}) + (j_k - j_{k-1}) \quad \text{โดยที่} \quad (i_0 = j_0 = 0) \tag{3.9}$$

ซึ่งจะได้ผลรวม

$$\sum_{k=1}^{K} w_k = J + I \tag{3.10}$$

ดังนั้นสมการ (3.3) สามารถเขียนได้ใหม่เป็น

$$D(F) = \frac{1}{I+J} \sum_{k=1}^{K} d(c_k) w_k \tag{3.11}$$

เราสามารถลดจำนวนในการหาค่าระยะห่างที่มีค่าน้อยที่สุด โดยไม่จำเป็นต้องทำการหาจากทุก ๆ ค่าที่เป็นไปได้ ของฟังก์ชัน F ดังนั้นผลรวมของลำดับ $c_1, c_2, \ldots, c_k$ เมื่อ $c_k = (i, j)$ คือ

$$g(c_k) = g(i,j) = \min_{c_1 \ldots c_{k-1}} \left[\sum_{m=1}^{k} d(c_m) w_m \right]$$

$$= \min_{c_1 \ldots c_{k-1}} \left[\sum_{m=1}^{k-1} d(c_m) w_m + d(c_k) w_k \right]$$

$$= \min_{c_{k-1}} \left[\min_{c_1 \ldots c_{k-2}} \left\{ \sum_{m=1}^{k-1} d(c_m) w_m \right\} + d(c_k) w_k \right]$$

$$= \min_{c_{k-1}} \left[g(c_{k-1}) + d(c_k) w_k \right] \qquad (3.12)$$

ซึ่งจากการจัดสมการ (3.12) ใหม่ ร่วมกับเงื่อนไขทั้งหมดข้างต้น และการหาค่าการถ่วงน้ำหนัก w สมการ (3.12) สามารถเขียนได้ใหม่ดังนี้

$$g(i,j) = \min \begin{pmatrix} g(i, j-1) + d(i,j) \\ g(i-1, j-1) + 2d(i,j) \\ g(i-1, j) + d(i,j) \end{pmatrix} \qquad (3.13)$$

ดังนั้นระยะห่างของ A และ B หลังจากการทำ DTW แล้ว จะเป็นไปตามสมการ (3.13) โดยการกำหนดเงื่อนไขเริ่มต้นให้ คือ $g(1,1) = 2d(1,1)$ และ $j = 1$ จากนั้นคำณวณสมการที่ (3.13) โดยการเปลี่ยนค่า i ที่อยู่ภายในวินโดว์ที่กำหนด และคำนวณซ้ำไปเรื่อย ๆ โดยเปลี่ยนค่า j ไปจน $j = J$ ซึ่งผลรวมของระยะห่างทั้งหมดของ A และ B จะได้เป็น $\frac{g(I,J)}{I+J}$ วิธีการดังที่กล่าวมานั้นเป็นวิธีการทำ DP Matching ซึ่งนำมาใช้กับ DTW และวาร์ปปิงฟังก์ชัน F ความคล้ายกันของ A และ B จะดูจากผลรวมระยะห่างของ A และ B ทั้งหมด

ค่าระยะห่าง (Distance) [15]

สำหรับระยะห่างระหว่าง A และ B ในแต่ละจุดหาได้จากผลต่างระหว่าง Magnitude ของ A และ B ณ ความถี่เดียวกัน

$$d(i,j) = \sum_{k=0}^{K} \left| M_{Ai}(k) - M_{Bj}(k) \right| \qquad (3.14)$$

โดยที่ M_{Ai} และ M_{Bj} คือ Magnitude ของ A เฟรมที่ i และ B เฟรมที่ j และ k คือ Sample ที่ความถี่ต่างภายในเฟรม

หรือจะใช้ผลต่างยกกำลังสองเป็นค่าระยะห่าง

$$d(i,j)=\sum_{k=0}^{K}\left(M_{Ai}(k)-M_{Bj}(k)\right)^2 \qquad (3.15)$$

สำหรับค่าระยะห่างที่ได้ทำการทดลองใช้ในวิทยานิพนธ์นี้ คือ

$$d(i,j)=\sum_{k=0}^{K}|M_{Ai}(k)-M_{Bj}(k)|\cdot w_k \qquad (3.16)$$

โดยที่ค่าการถ่วงน้ำหนัก w_k จะเท่ากับ Magnitude ที่มีค่ามากกว่าระหว่าง $A(k)$ กับ $B(k)$

$$w_k=\max\begin{pmatrix}M_{Ai}(k)\\M_{Bj}(k)\end{pmatrix} \qquad (3.17)$$

ถ้า M_{Ai} และ M_{Bj} = 0 ; $w_k=1$

ซึ่งจะเป็นการให้ตำแหน่งที่มี Magnitude สูง ๆ มีความสำคัญมากกว่า ซึ่งจะเป็นการดึงเอาลักษณะเฉพาะตัวของเสียงนั้น ๆ ออกมา

การแบ่งกลุ่มแบบอ้างอิง

ในกรณีที่มีแบบอ้างอิงหรือคำศัพท์อ้างอิงอยู่เป็นจำนวนมาก จะพบปัญหาของการทดสอบแบบทดสอบกับแบบอ้างอิงเป็นจำนวนมาก ซึ่งทำให้เสียเวลาในการทดสอบกับแบบอ้างอิงทุกตัว ดังนั้นการแบ่งแบบอ้างอิงออกเป็นกลุ่มๆ โดยแบบอ้างอิงภายในกลุ่มเดียวกันมีความคล้ายคลึงกัน ซึ่งจะเป็นการคัดเลือกแบบอ้างอิงที่เหมาะสมไปทำการทดสอบ จึงทำให้สามารถลดเวลาในการทดสอบลงได้ [12]

สำหรับในวิทยานิพนธ์นี้ ได้ทำการจัดกลุ่มแบบอ้างอิงโดยอาศัยสเปคตรัมที่เกิดในช่วงความถี่ 0 - 4 kHz ออกเป็น 4 ช่วง คือ 0 - 1 kHz, 1 - 2 kHz, 2 - 3 kHz และ 3 - 4 kHz แทนด้วย F_1 F_2 F_3 และ F_4 ซึ่งเสียงที่มีสเปคตรัมเกิดขึ้นในช่วงความถี่ F_1 F_2 F_3 และ F_4 เหมือนกัน จะอยู่กลุ่มเดียวกัน โดยแบ่งแต่ละช่วงความถี่ F_1 F_2 F_3 และ F_4 ออกเป็น 8 แถบความถี่ย่อย ๆ และถ้ามีพลังงานเกินค่าที่กำหนดไว้ในแต่ละแถบความถี่ย่อยแถบใด ๆ ใน 8 แถบของ F_1 F_2 F_3 หรือ F_4 จะถือว่ามีสเปคตรัมเสียงเกิดขึ้นในช่วง F_1 F_2 F_3 หรือ F_4 นั้น ๆ ซึ่งค่าที่ได้จะเป็นตัว

กำหนดกลุ่มของเสียง และจากการที่ได้ทำการแบ่งกลุ่มโดยใช้แถบความถี่ช่วงละ 1 kHz ทำให้มีกลุ่มของแบบอ้างอิงได้ทั้งหมด ดังนี้

กลุ่มที่	F_1	F_2	F_3	F_4
0	0	0	0	0
1	0	0	0	1
2	0	0	1	0
3	0	0	1	1
4	0	1	0	0
5	0	1	0	1
6	0	1	1	0
7	0	1	1	1
8	1	0	0	0
9	1	0	0	1
10	1	0	1	0
11	1	0	1	1
12	1	1	0	0
13	1	1	0	1
14	1	1	1	0
15	1	1	1	1

โดยให้

1	แทนการมีสเปคตรัมเสียงเกิดในช่วงความถี่นั้น
0	แทนการไม่มีสเปคตรัมเสียงเกิดในช่วงความถี่นั้น
F_1	แทนช่วงแถบความถี่ 0 kHz ถึง 1 kHz
F_2	แทนช่วงแถบความถี่ 1 kHz ถึง 2 kHz
F_3	แทนช่วงแถบความถี่ 2 kHz ถึง 3 kHz
F_4	แทนช่วงแถบความถี่ 3 kHz ถึง 4 kHz

แต่ตามความเป็นจริงแล้ว จะพบว่าในช่วงความถี่ 0 ถึง 1 kHz นั้นจะมีสเปคตรัมของเสียงเกิดขึ้นอยู่เสมอ ซึ่งกรณีของกลุ่ม 0 ถึง 7 สามารถตัดออกไปได้ จึงเหลือแต่เพียง 8 กลุ่มเท่านั้น คือ

กลุ่มที่	F_1	F_2	F_3	F_4
8	1	0	0	0
9	1	0	0	1

10	1	0	1	0
11	1	0	1	1
12	1	1	0	0
13	1	1	0	1
14	1	1	1	0
15	1	1	1	1

สเปคตรัมเสียงเกิดขึ้นในแถบความถี่ F_1, F_2, F_3 และ F_4 ของกลุ่มที่ 15 นั้นเกิดจากค่าพลังงานของแถบความถี่ย่อยใดแถบหนึ่งใน 8 แถบหรือมากกว่าหนึ่งแถบใน F_1, F_2, F_3 และ F_4 มีค่าเกินค่า threshold ที่กำหนด จึงถือว่ามีสเปคตรัมเสียงเกิดขึ้นทุกช่วงความถี่ F_1, F_2, F_3 และ F_4 ดังนั้นค่า threshold จึงเป็นตัวกำหนดว่าจะมีสเปคตรัมเสียงเกิดขึ้นในช่วงแถบความถี่นั้น ๆ หรือไม่

ตารางที่ 3.1 สระที่ถูกจัดอยู่ในกลุ่มต่าง ๆ เมื่อ threshold = 0.5 เท่าของค่าเฉลี่ยพลังงานทั้งหมด

กลุ่ม	อะ	อา	อิ	อี	อึ	อือ	อุ	อู	เอะ	เอ	แอะ	แอ	เอาะ	ออ	โอะ	โอ	เออะ	เออ	เอีย	อัว	เอือ	อำ	อัย	เอา
1000			1	1			4	4	1															
1001																								
1010			2	2						3														
1011			10	9			3	3	1	3					3	2								
1100		4		1	1	5	16	17			1		3	5	12	14	3	3	3	15	4	10		15
1101	1														3	2				1				2
1110	11	7	5	4	13	8	1		8	6	7	10	10	10	2		8	12	5	2	7	2	10	1
1111	12	13	6	7	10	11			14	12	16	14	11	9	4	6	13	9	16	6	13	12	14	6

ตารางที่ 3.2 สระที่ถูกจัดอยู่ในกลุ่มต่าง ๆ เมื่อ threshold = 1.0 เท่าของค่าเฉลี่ยพลังงานทั้งหมด

กลุ่ม	อะ	อา	อิ	อี	อึ	อือ	อุ	อู	เอะ	เอ	แอะ	แอ	เอาะ	ออ	โอะ	โอ	เออะ	เออ	เอีย	อัว	เอือ	อำ	อัย	เอา
1000			6	9	8	10	16	15	3	1		3			10	10		3	15		7		3	
1001								2								2			1					
1010			3	1	3	3		1	8	8	3	3							1		1			
1011			13	14			7		4	3	3	3			5				7				3	
1100	13	17	1		7	4	5	6	3	5	7	3	18	19	9	12	18	15		21	10	17	13	19
1101		1				1							1							1	1		2	4
1110	7	6	1		1	1			1	1	7	10					3	1				5		
1111	4				5	5			5	6	4	2	5	5			3	5		2	5	2	3	1

ตารางที่ 3.3 สระที่ถูกจัดอยู่ในกลุ่มต่าง ๆ เมื่อ threshold = 1.5 เท่าของค่าเฉลี่ยพลังงานทั้งหมด

กลุ่ม	อะ	อา	อิ	อี	อึ	อือ	อุ	อู	เอะ	เอ	แอะ	แอ	เอาะ	ออ	โอะ	โอ	เออะ	เออ	เอีย	อัว	เอือ	อำ	อัย	เอา
1000		1	17	18	17	19	21	24	11	12	12	16			22	23	4	8	19	22	20	10	24	3
1001					1	2	1								1				2					
1010			2		1	2	1		7	6	2	2							1					
1011			5	6	1														2					
1100	24	23			3	1	1				6	3	24	24	1	1	17	14		2	1	14		21
1101					1												1							
1110									4	6	4	3									3			
1111									2								2	1						

ตารางที่ 3.4 สระที่ถูกจัดอยู่ในกลุ่มต่าง ๆ เมื่อ threshold = 0.6 ถึง 1.4 เท่าของค่าเฉลี่ยพลังงานทั้งหมด

กลุ่ม	อะ	อา	อิ	อี	อึ	อือ	อุ	อู	เอะ	เอ	แอะ	แอ	เอาะ	ออ	โอะ	โอ	เออะ	เออ	เอีย	อัว	เอือ	อำ	อัย	เอา
1000			62	72	62	87	114	142	29	31	9	18			85	91	10	24	94	39	62	13	68	4
1001				12		1	4	3							8	4			4					
1010			28	11	15	21	1	1	45	55	25	33			7	3			15	1	4		8	
1011			104	103	2		22	11	40	37	11	6			31	17			44				14	
1100	129	149	7	3	62	33	75	59	21	18	57	50	156	170	83	94	126	111	16	150	83	151	61	176
1101	2	2			4	11							4	2	1	3	3		1	4	3	1	2	14
1110	49	43	9	11	34	25			33	31	58	71	19	9		1	35	38	5	6	28	23	21	
1111	36	22	6	4	37	38			48	44	56	38	37	35	1	3	42	43	37	16	36	28	42	22

กำหนดค่า threshold นั้นมีความสำคัญต่อการกำหนดกลุ่ม ถ้ากำหนดค่า threshold ไว้สูงเกินไป จะมีสเปคตรัมเกินค่า threshold เฉพาะในช่วง 0 - 1 kHz หรือ F_1 เป็นส่วนมาก ทำให้แบบอ้างอิงส่วนใหญ่จะตกไปอยู่ในกลุ่ม 1000 มาก ดังตารางที่ 3.3 เพราะเสียงส่วนใหญ่จะมีพลังงานในช่วง 0 ถึง 1 kHz สูงกว่าช่วงความถี่อื่นๆ ในทางกลับกันถ้ากำหนด threshold ไว้ต่ำเกินไปจะทำให้เสียงส่วนใหญ่มีพลังงานเกิน threshold ที่กำหนดไว้ทุกแถบความถี่ แบบอ้างอิงจะตกไปอยู่ในกลุ่ม 1111 เป็นส่วนใหญ่ ดังตารางที่ 3.1 ดังนั้นการจัดกลุ่มแบบอ้างอิงควรจะให้ครอบคลุมความน่าจะเป็นที่จะเกิดสเปคตรัมเสียงในช่วงความถี่ของกลุ่มต่าง ๆ ได้ทั้งหมด ดังในตาราง 3.4 ที่ใช้ค่า threshold ตั้งแต่ 0.6 เท่าของค่าเฉลี่ยพลังงานทั้งหมด จนถึงค่า 1.4 เท่าของค่าเฉลี่ยพลังงานทั้งหมด เพื่อให้ครอบคลุมกรณีต่าง ๆ มากที่สุด มิฉะนั้นจะทำให้เปอร์เซนต์ความถูกต้องในการรู้จำลดลง

ซึ่งกลุ่มอ้างอิง 8 กลุ่มประกอบไปด้วยแบบอ้างอิงของเสียงต่างๆ ดังนี้

1. กลุ่ม 1000 มี 20 แบบอ้างอิง ประกอบด้วย

อิ, อี, อึ, อือ, อุ, อู, เอ, เอ, แอะ, แอ, โอะ, โอ เออะ, เออ, เอีย, อัว, เอือ, อำ, อัย, เอา

2. กลุ่ม 1001 มี 9 เสียง แบบอ้างอิง ประกอบด้วย
อิ, อี, อึ, อือ, อุ, อู โอะ, โอ, เอีย

3. กลุ่ม 1010 มี 16 แบบอ้างอิง ประกอบด้วย
อิ, อี, อึ, อือ, อุ, อู, เอะ, เอ, แอะ, แอ, โอะ, โอ, เอีย, อัว, เอือ, อัย

4. กลุ่ม 1011 มี 14 แบบอ้างอิง ประกอบด้วย
อิ, อี, อึ, อือ, อุ, อู, เอะ, เอ, แอะ, แอ, โอะ, โอ, เอีย, อัย

5. กลุ่ม 1100 มี 24 แบบอ้างอิง ประกอบด้วย
อะ, อา, อิ, อี, อึ, อือ, อุ, อู, เอะ, เอ, แอะ, แอ, เอาะ, ออ, โอะ, โอ, เออะ, เออ, เอีย, อัว, เอือ, อำ, อัย, เอา

6. กลุ่ม 1101 มี 16 แบบอ้างอิง ประกอบด้วยย
อะ, อา, อึ, อือ, เอาะ, ออ, โอะ, โอ, เออะ, เออ, เอีย, อัว, เอือ, อำ, อัย, เอา

7. กลุ่มที่ 1110 มี 21 แบบอ้างอิง ประกอบด้วย
อะ, อา, อิ, อี, อึ, อือ, เอะ, เอ, แอะ, แอ, เอาะ, ออ, โอะ, โอ, เออะ, เออ, เอีย, อัว, เอือ, อำ, อัย

8. กลุ่ม 1111 มี 22 แบบอ้างอิง ประกอบด้วย
อะ, อา, อิ, อี, อึ, อือ, เอะ, เอ, แอะ, แอ, เอาะ, ออ, โอะ, โอ, เออะ, เออ, เอีย, อัว, เอือ, อำ, อัย, เอา

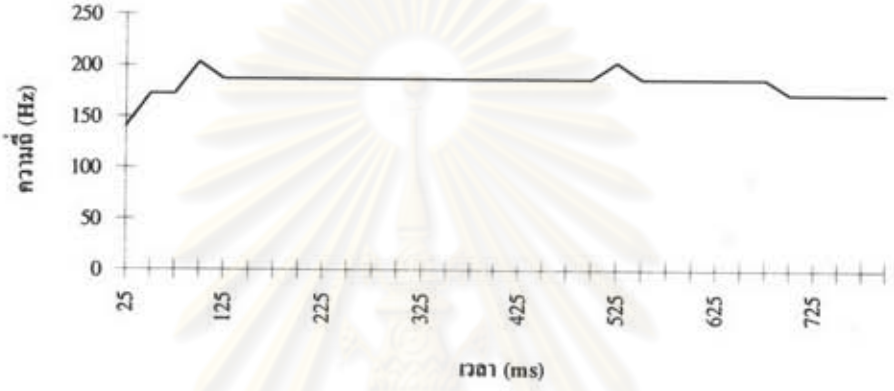
สำหรับการคัดเลือกนั้นจะทำการเลือกคู่เสียง ทั้งสระเสียงสั้น-ยาวไปด้วยกันเป็นคู่ ในกรณีที่การคัดเลือกแบบอ้างอิงเลือกสระเสียงสั้นหรือเสียงยาวตัวใดตัวหนึ่ง จากการคัดเลือกจะเห็นว่าเสียงแบบอ้างอิงของสระบางตัว เช่น อะ อา ที่มีสเปคตรัมในช่วง 1-2 kHz จะไม่ปรากฏในกลุ่ม 10xx เลย ถือได้ว่า การแบ่งกลุ่มแบบอ้างอิงสามารถแยกกลุ่มของเสียงสระบางตัวได้อย่างชัดเจน สำหรับเสียงสระบางตัวที่ไม่สามารถแบ่งกลุ่มได้อย่างชัดเจน อาจจะมีกระจายอยู่ในกลุ่มต่าง ๆ ทั้งหมดหรือเกือบทั้งหมด เพื่อให้การแบ่งกลุ่มแบบอ้างอิงสามารถครอบคลุมการเลือกกลุ่มของแบบทดสอบได้ทุกกรณี

การจำแนกเสียงวรรณยุกต์

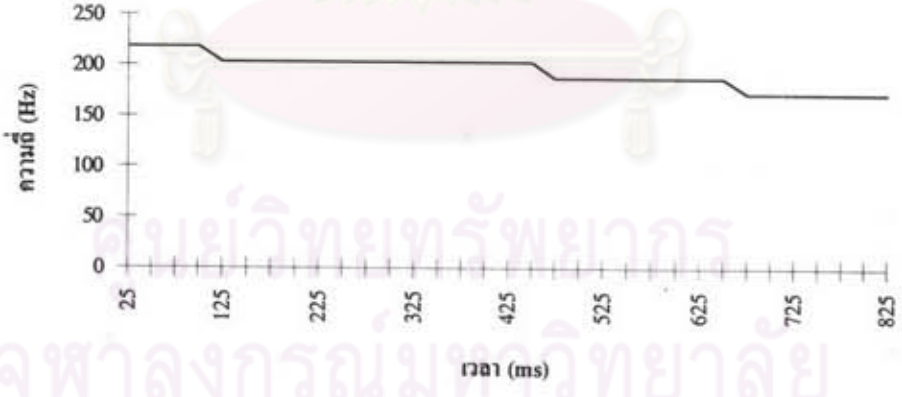
เสียงวรรณยุกต์ คือลักษณะของเสียงที่มีระดับเสียง สูง, ต่ำ คงระดับตลอดทั้งพยางค์ หรือเสียงที่มีการเปลี่ยนแปลงระดับเสียงตามช่วงเวลาต่าง ๆ แตกต่างกันออกไป [11] ดังรูปที่ 2.24 ซึ่งผลต่างของความถี่มูลฐานสูงสุดกับความถี่มูลฐานต่ำสุดที่เกิดขึ้น เรียกว่า Pitch Range ความถี่มูลฐานของเสียงวรรณยุกต์ในเพศชายจะมีความถี่ที่ต่ำกว่าในเพศหญิงแต่

มีลักษณะการเปลี่ยนระดับเสียงเหมือนกัน [10] ดังนั้นการแยกแยะเสียงวรรณยุกต์สามารถใช้การเปลี่ยนระดับเสียงสัมพัทธ์ (Relative Pitch) ได้

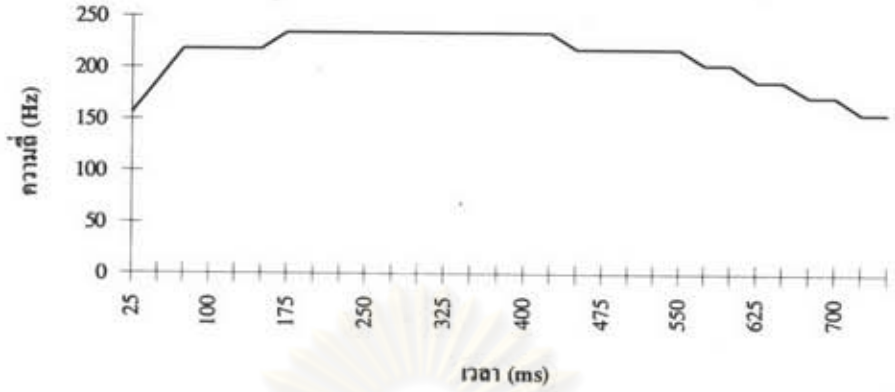
การหาการเปลี่ยนระดับเสียงในวรรณยุกต์ต่าง ๆ ทำได้โดยตรวจจับส่วนยอดของสเปคตรัมเสียงในช่วง 50 - 500 Hz ซึ่งจะเริ่มตรวจจากเฟรมที่มีพลังงานสูงสุดไปยังด้านหน้าและด้านหลัง โดยจะตรวจหาค่าที่สูงสุดและอยู่ในขอบเขตที่กำหนด แล้วนำมาเมื่อนำมาพลอตจะได้ดัง ตัวอย่างในรูปที่ 3.4, 3.5, 3.6, 3.7 และ 3.8 ซึ่งมีลักษณะใกล้เคียงกับรูปที่ 2.23



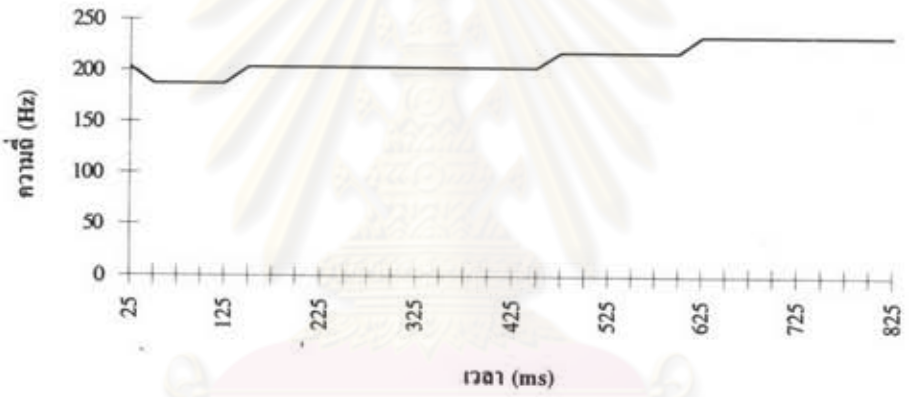
รูปที่ 3.4 ความถี่มูลฐานของเสียง "อา"



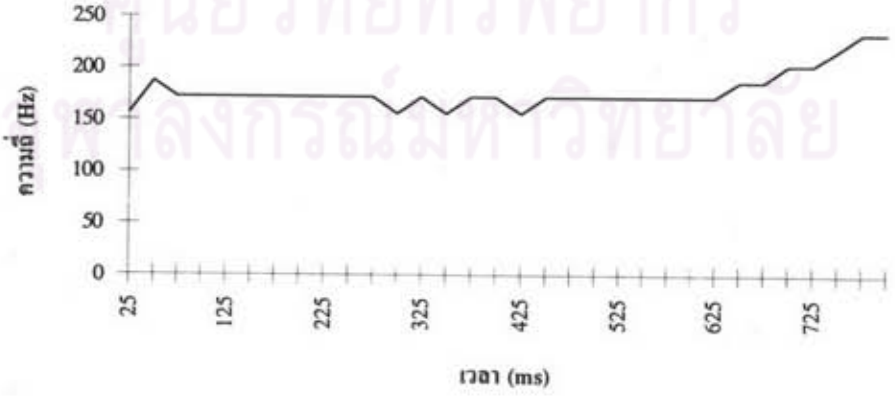
รูปที่ 3.5 ความถี่มูลฐานของเสียง "อ่า"



รูปที่ 3.6 ความถี่มูลฐานของเสียง "อ้า"



รูปที่ 3.7 ความถี่มูลฐานของเสียง "อ๊า"



รูปที่ 3.8 ความถี่มูลฐานของเสียง "อำ"

จากรูปที่ 3.4 และ 3.5 พบว่าวรรณยุกต์สามัญและวรรณยุกต์เอก มีลักษณะใกล้เคียงกันมาก คือระดับเสียงตกลงทั้งคู่ แต่วรรณยุกต์สามัญจะมีระดับเสียงคงที่มากกว่าครึ่งหนึ่งของระยะทั้งหมดโดยเริ่มเกิดในช่วงแรกของช่วงเวลา และจะตกลงเล็กน้อยในตอนท้ายของช่วงเวลา ส่วนวรรณยุกต์เอกจะมีการตกลงของเสียงในช่วงต้น ซึ่งสอดคล้องตามรูปที่ 2.24 วรรณยุกต์โทจะมีการเปลี่ยนระดับสูงขึ้นเล็กน้อยก่อนที่จะตกลง วรรณยุกต์ตรีจะมีระดับเสียงที่สูงขึ้น และวรรณยุกต์จัตวาจะมีระดับเสียงตกลงในช่วงแรกและเปลี่ยนระดับสูงขึ้นในช่วงต่อมา

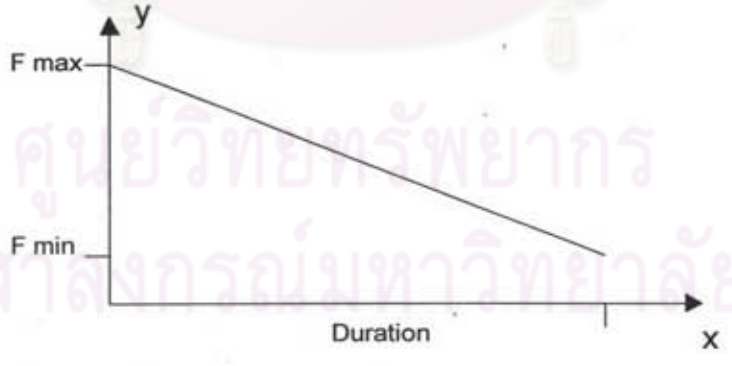
สำหรับวิธีการจำแนกเสียงวรรณยุกต์ เริ่มจากนำค่า Magnitude ในช่วง 50 - 500 Hz ที่ได้ มาหาค่าพารามิเตอร์ต่างๆ เช่น ผลต่างการเปลี่ยนระดับเสียง (Pitch Range), ช่วงความยาวเสียงทั้งหมด (Duration) เพื่อนำมาสร้างสมการแทนวรรณยุกต์อ้างอิงเพื่อทำ Curve Fitting กับรหัสวรรณยุกต์แบบทดสอบ โดยการหาค่าระยะห่างยกกำลังสองเฉลี่ย (Mean Squar Error) ระหว่างแบบทดสอบกับแบบอ้างอิง และเลือกค่าที่น้อยที่สุดไปพิจารณาอีกครั้ง สมการแทนแบบอ้างอิงเหล่านี้มีจะถูกดัดแปลงให้มีลักษณะใกล้เคียงกับการเปลี่ยนแปลงของเสียงวรรณยุกต์ตามรูปที่ 2.24 โดยใช้พารามิเตอร์ที่ได้จากแบบทดสอบ ซึ่งประกอบด้วย

1. สมการเส้นตรง

$$y = f_{\max} - ax \qquad (3.19)$$

กำหนดให้ใช้แทนแบบอ้างอิงวรรณยุกต์สามัญ และวรรณยุกต์เอก โดย

a = พิสัยการเปลี่ยนระดับ / ช่วงความยาวเสียงทั้งหมด
y = แกนความถี่
x = แกนเชิงเวลา
$f_{\max}$ = ความถี่สูงสุดที่เกิดในการเปลี่ยนระดับ



รูปที่ 3.9 รูปกราฟของสมการเส้นตรง $y = f_{\max} - ax$

2. สมการเส้นตรง

$$y = f_{\min} + ax \qquad (3.20)$$

กำหนดให้ใช้แทนแบบอ้างอิงวรรณยุกต์ตรี โดย

a = พิสัยการเปลี่ยนระดับ / ช่วงความยาวเสียงทั้งหมด

y = แกนความถี่

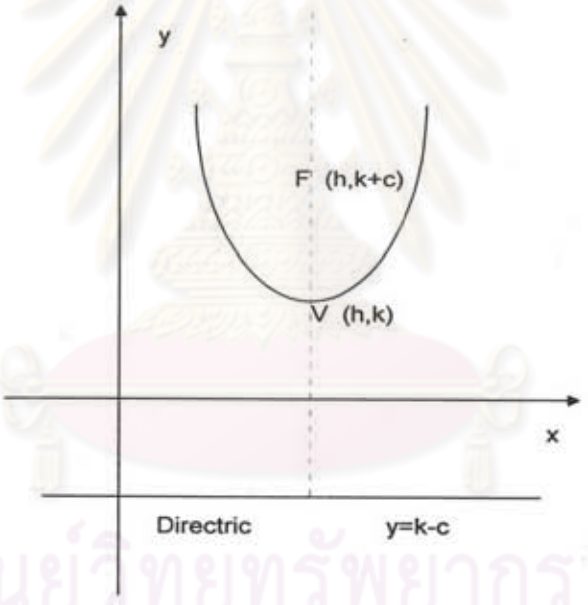
x = แกนเชิงเวลา

$f_{\min}$ = ความถี่ต่ำสุดที่เกิดในการเปลี่ยนระดับ

3. สมการพาราโบลา

$$(x-h)^2 = 4C(y-k) \qquad (3.21)$$

เมื่อ $(y-k) \geq 0$ จะได้ $C \geq 0$



รูปที่ 3.10 รูปกราฟของสมการพาราโบลา $(x-h)^2 = 4C(y-k)$

จัดรูปสมการใหม่ได้เป็น

$$y = f_{\min} - \left(\frac{1}{duration + L.cw}\right)(x-cw)^2 \qquad (3.22)$$

โดย $f_{\min} = k$, $cw = h$ และ $duration + L.cw = -4c$

เมื่อ cw เป็นจุดยอดของเสียงวรรณยุกต์แบบทดสอบ และ L เป็นค่าคงที่ขึ้นอยู่กับเสียงสระ

4. สมการพาราโบลา

$$y = f_{max} + \left(\frac{1}{duration + L.cw}\right)(x - cw)^2 \qquad (3.23)$$

โดย $f_{max} = k$, $cw = h$ และ $duration + L.cw = 4c$

เมื่อ cw เป็นจุดยอดของเสียงวรรณยุกต์แบบทดสอบ และ L เป็นค่าคงที่ขึ้นอยู่กับเสียงสระ

ผลลัพธ์ที่ได้จากการทดสอบจะถูกนำมาพิจารณาอีกขั้นตอนหนึ่ง โดยการตรวจ Pitch Range, ช่วงความยาวของระดับเสียงคงที่ เพื่อยืนยันผลลัพธ์ที่ได้