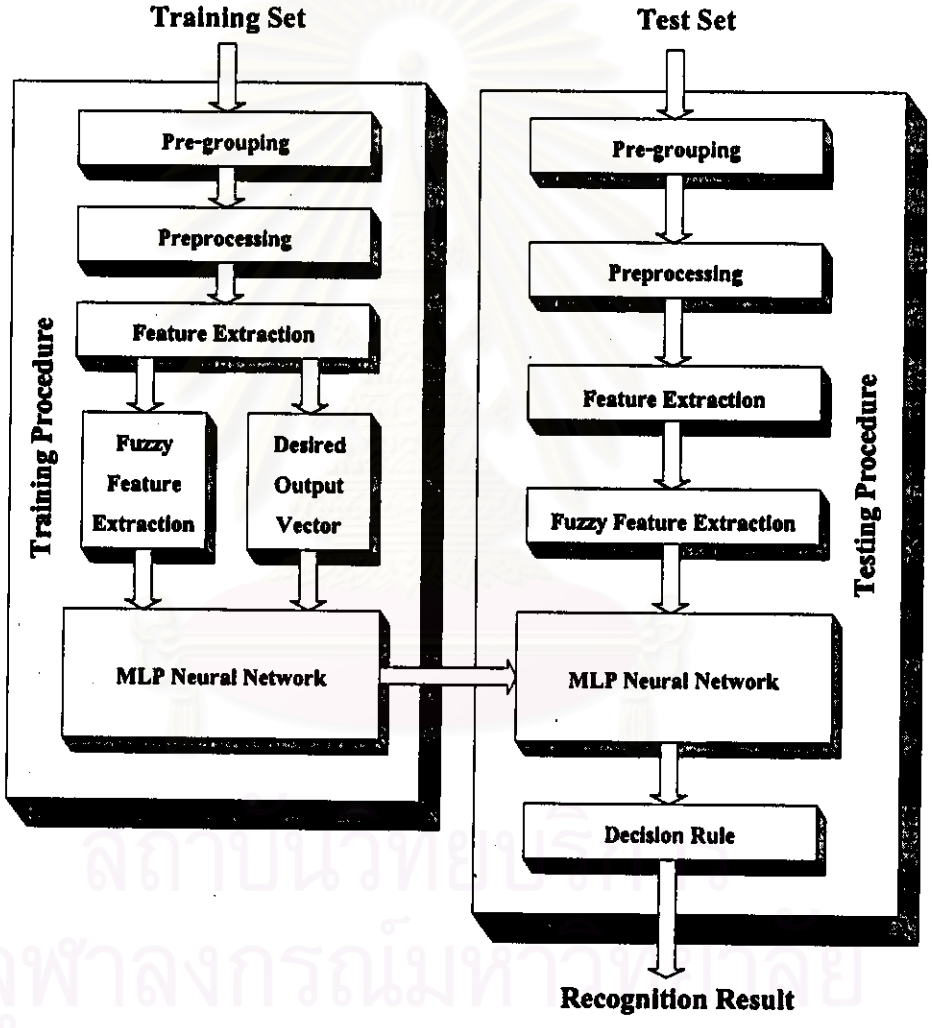
บทที่ 3
กระบวนการรู้จำเสียงที่เสนอ

งานวิจัยนี้ ได้พัฒนาโปรแกรมสำหรับการรู้จำเสียงพูดคำไทยหลายพยางค์ แบบไม่ขึ้นต่อผู้พูด (Speaker Independent) หลักการสำคัญคือ ใช้เทคนิคแบบฟัซซีร่วมกับนิวรอลเน็ตเวอร์ก ใช้คำเป็นตัวทาบเทียบ (Word Based Recognition) รูปที่ 3.1 เป็นโครงสร้างทั้งหมดของระบบที่สร้างขึ้น



รูปที่ 3.1 โครงสร้างของระบบรู้จำเสียง
โดยใช้เทคนิคแบบฟัซซีร่วมกับนิวรอลเน็ตเวอร์ก

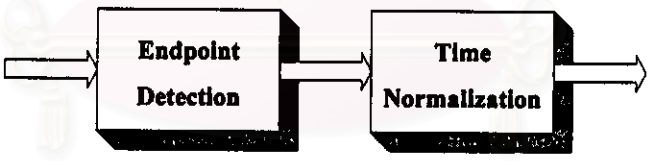
ขั้นตอนการรู้จำแบ่งออกเป็น 2 ส่วนหลักคือ ช่วงฝึกฝน (Training หรือ Learning Procedure) และช่วงรู้จำ (Recognition Procedure) หรือช่วงทดสอบการรู้จำ (Test Procedure) ดังแสดงในรูปที่ 3.1

ช่วงฝึกฝน (Training Procedure)

การฝึกฝน เป็นแบบ Supervised Learning โดยป้อนสัญญาณเสียงในชุดฝึกฝน (Training Set) พร้อมทั้งตัวเลขที่ระบุว่าเป็นเสียงของคำศัพท์ตัวใด เข้าไปในระบบ ตัวเลขที่ระบุว่าเป็นเสียงคำศัพท์ตัวใด จะถูกนำไปสร้างเป็นเวกเตอร์ข้อมูลออกที่ต้องการ (Desired Output Vector) สำหรับนิวรอลเน็ตเวอร์ก ส่วนสัญญาณเสียง ซึ่งเก็บในรูปของสัญญาณดิจิตอลด้วยเครื่องคอมพิวเตอร์อยู่แล้ว จะถูกนำไปผ่านกระบวนการประมวลผลเบื้องต้น (Preprocessing) กระบวนการสกัดค่าลักษณะเด่น (Feature Extraction) กระบวนการแปลงเป็นค่าสมาชิกภาพแบบฟัซซี (Fuzzy Membership Conversion) หรือการสกัดค่าลักษณะเด่นแบบฟัซซี (Fuzzy Feature Extraction) หลังจากนั้น จึงนำเวกเตอร์ข้อมูลเข้าที่ได้จากการสกัด และเวกเตอร์ข้อมูลออกที่ต้องการไปใช้ในการฝึกฝนนิวรอลเน็ตเวอร์กแบบ MLP รายละเอียดในแต่ละขั้นตอนมีดังนี้

3.1 การประมวลผลเบื้องต้น

มีขั้นตอนย่อยดังแสดงในรูปที่ 3.2



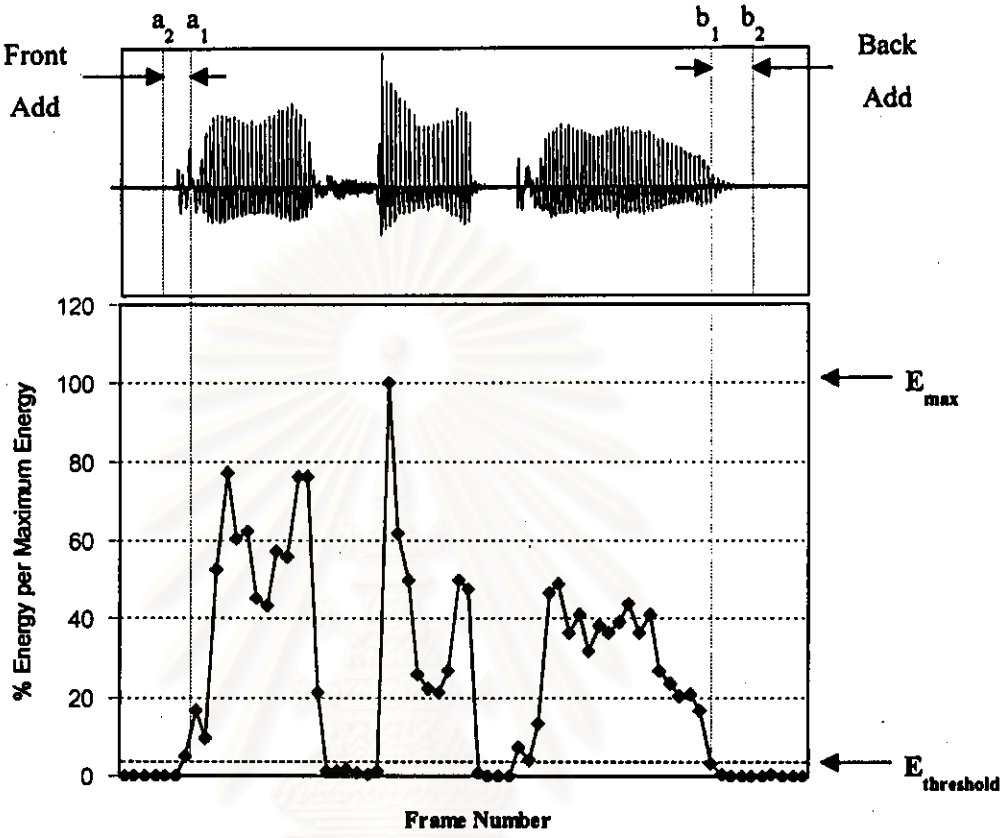
รูปที่ 3.2 โครงสร้างขั้นตอนในกระบวนการประมวลผลเบื้องต้น

3.1.1 การตัดหัวท้ายคำ (Endpoint Detection)

สำหรับในงานวิจัยนี้ ใช้ค่าพลังงานแบบปกติในการหาจุดหัวท้ายคำ โดยมีเผื่อเวลาในช่วงต้น และท้าย เพื่อให้ครอบคลุมสัญญาณเสียงที่ต่ำกว่าค่าที่กำหนด แต่มีความสำคัญในการรู้จำ

รูปที่ 3.3 เป็นตัวอย่างรูปคลื่น และพลังงานของสัญญาณเสียง ขณะทำการตัดหัวท้ายคำ ค่าระดับพลังงานที่กำหนด (Energy Threshold) $E_{threshold}$ ที่ใช้ในการตรวจสอบจุดเริ่ม และท้ายคำ จะใช้ค่าที่เป็นจำนวนเท่าของค่าพลังงานมากที่สุด (Maximum Energy) E_{max} นั่นคือ

$$E_{threshold} = kE_{max} \quad , \quad 0 \le k \le 1 \qquad (3.1)$$



รูปที่ 3.3 ตัวอย่างรูปคลื่น และพลังงานของสัญญาณเสียง

เมื่อค่าพลังงานของส่วนย่อยมีค่าสูงกว่า $E_{threshold}$ ติดกันมากกว่าจำนวนส่วนย่อยที่กำหนด จะถือว่าเป็นจุดเริ่มต้นของคำ a_1 และเช่นเดียวกันจะได้จุดท้ายคำ b_1 หลังจากนั้นจะทำการบวกเผื่อเวลาในช่วงต้น และท้ายเพื่อให้ครอบคลุมสัญญาณเสียงที่ต้องการ จะได้สัญญาณหลังจากผ่านการตัดหัวท้ายคำ เริ่มต้นตั้งแต่จุด a_2 และสิ้นสุดที่ b_2 สำหรับค่าเวลาที่บวกเผื่อในตอนต้นและท้ายจะเก็บข้อมูลโดยพิจารณาจากรูปคลื่นเสียงคำทั้งหมดที่จะรู้จำ

3.1.2 การนอร์มอลไลซ์ทางเวลา (Time Normalization)

งานวิจัยนี้ เลือกใช้วิธีการนอร์มอลไลซ์ทางเวลาโดยการประมาณค่าในช่วงเชิงเส้น (Linear Interpolation) เนื่องจากใช้เวลาในการคำนวณน้อย ไม่ต้องใช้หน่วยความจำมากนัก และ

สามารถให้ผลอัตราการรู้จำได้ดี โดยแบ่งชุดคำที่จะรู้จำออกเป็น 3 ชุดคือ ชุดของเสียงคำหนึ่งพยางค์ สองพยางค์ และสามพยางค์ แต่ละชุดจะทำการนอร์มอลไลซ์ทางเวลาด้วยจำนวนจุดข้อมูลที่ต้องการต่างกันไป เนื่องจากเสียงคำที่มีจำนวนพยางค์มากกว่า ก็ควรจะมีระยะเวลา (Duration Time) นานกว่า นั่นคือควรจะมีจำนวนจุดข้อมูลมากกว่าเสียงคำที่มีจำนวนพยางค์น้อยกว่า นอกจากนี้ การแบ่งคำออกเป็นชุดย่อยๆ และใช้นิวรอลเน็ตเวอร์กในการรู้จำแต่ละชุดย่อยแยกกันไป จะช่วยเพิ่มอัตราการรู้จำได้มาก

เนื่องจากจำนวนข้อมูลหลังจากผ่านการนอร์มอลไลซ์ทางเวลาแล้ว มีผลต่ออัตราการรู้จำค่อนข้างมาก ดังนั้นจึงต้องมีการเก็บผลอัตราการรู้จำ เมื่อใช้จำนวนข้อมูลที่ต้องการหลังการนอร์มอลไลซ์ทางเวลาต่างๆ กันไป เพื่อเลือกจำนวนจุดข้อมูลที่เหมาะสมสำหรับแต่ละชุดคำศัพท์

3.2 การสกัดค่าลักษณะเด่น (Feature Extraction)

ขั้นตอนการสกัดลักษณะเด่นในงานวิจัยนี้ จะใช้วิธีการประมาณพันธะเชิงเส้น (Linear Predictive Coding) สัญญาณที่ป้อนเข้าเป็นสัญญาณที่ผ่านการนอร์มอลไลซ์แล้ว โดยมีขั้นตอนในการสกัดแสดงในรูปที่ 3.4

รูปที่ 3.4 โครงสร้างขั้นตอนกระบวนการสกัดค่าลักษณะเด่น

3.2.1 การเน้นล่วงหน้า (Preemphasis)

เป็นการอัดพิสัยพลวัตของสัญญาณ (Signal Dynamic Range) มีผลให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (S/N Ratio) มีค่าสูงขึ้น โดยใช้วงจรกรองแบบดิจิตอลอันดับหนึ่ง ดังนี้

$$H(z) = 1 - az^{-1} \qquad ; a = 0.95 \qquad (3.2)$$

ซึ่งจะได้สัญญาณขาออก $\tilde{s}(n)$ คือ

$$\tilde{s}(n) = s(n) - as(n-1) \qquad (3.3)$$

3.2.2 การแบ่งสัญญาณเสียงเป็นส่วนย่อย (Block into Frame)

เนื่องจากสัญญาณเสียงมีคุณสมบัติเปลี่ยนแปลงทางสถิติตลอดเวลา (Non-stationary) จึงไม่สามารถจำลองการกระจายของสัญญาณเสียงทางสถิติได้เลย วิธีแก้ไขคือ แบ่งสัญญาณเสียงออกเป็นส่วนย่อย (Frame) ขนาดความยาวประมาณ 15 ถึง 50 มิลลิวินาที ในช่วงความยาวขนาดนี้ ถือได้ว่าสัญญาณเสียงมีคุณสมบัติเปลี่ยนแปลงทางสถิติน้อยมาก หรือแทบไม่มีเลย (Stationary) ดังนั้นสามารถคำนวณค่าใดๆ ทางสถิติในแต่ละส่วนย่อยนี้ได้

การแบ่งสัญญาณเสียงออกเป็นส่วนย่อย ส่วนย่อยแต่ละส่วนจะมีการเหลื่อมกับสัญญาณเสียงในส่วนย่อยที่อยู่ติดกัน เพื่อให้ค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้นของแต่ละส่วนมีความต่อเนื่อง โดยจะทำการเหลื่อมด้วยความยาวในช่วง 0 ถึง 0.5 ของความยาวส่วนย่อย ถ้าแบ่งสัญญาณเสียงของคำหนึ่งๆ ออกเป็น L ส่วนย่อย แต่ละส่วนมี N จุดข้อมูล และมีการเหลื่อมกัน P จุดข้อมูล จะได้ว่าทุกส่วนย่อยจะมีจุดเริ่มต้น เลื่อนไปจากจุดเริ่มต้นของส่วนย่อยก่อนหน้าอยู่ M จุดข้อมูล โดยที่ $M = N - P$ ดังนั้น สัญญาณเสียงสำหรับส่วนย่อยที่ l ใดๆ $X_l(n)$ จะเขียนได้ดังสมการ

$$X_l(n) = \tilde{s}(Ml + n) \qquad \begin{matrix} ; n = 0, 1, ..., N-1 \\ l = 0, 1, ..., L-1 \end{matrix} \tag{3.4}$$

3.2.3 การลดขอบด้วยฟังก์ชันหน้าต่าง (Smoothing Window)

เป็นการนำสัญญาณเสียงแต่ละส่วนย่อยมาผ่านฟังก์ชันหน้าต่างที่กำหนด เพื่อลดทอนค่าแอมปลิจูดที่ขอบทั้งสองด้านของส่วนย่อย เป็นการป้องกันการเปลี่ยนแปลงอย่างรวดเร็วที่บริเวณขอบ และเป็นการเน้นให้มีการประมวลผลในช่วงกลางของส่วนย่อยเป็นหลัก สัญญาณที่ผ่านการลดขอบด้วยฟังก์ชันหน้าต่างจะเขียนได้ดังสมการ

$$\tilde{X}_l(n) = X_l(n).w(n) \tag{3.5}$$

โดยที่ $w(n)$ เป็นฟังก์ชันหน้าต่าง ในที่นี้จะใช้ฟังก์ชันหน้าต่างแบบแฮมมิง (Hamming Window) (Furui, 1989) ซึ่งมีสมการดังนี้

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{3.6}$$

3.2.4 การวิเคราะห์ค่าอัตสหสัมพันธ์ (Autocorrelation Analysis)

ค่าอัตสหสัมพันธ์ลำดับที่ m ในส่วนย่อยที่ l ใดๆ หาได้จาก

$$R_l(m) = \sum_{n=0}^{N-1-|m|} \tilde{X}_l(n)\tilde{X}_l(n+m) \qquad ; m = 0, 1, ..., p \qquad (3.7)$$

เมื่อ p เป็นอันดับ (Order) ของการวิเคราะห์ โดยทั่วไป ค่าอันดับที่ใช้ในการวิเคราะห์จะมีค่าอยู่ในช่วง 8 ถึง 12 (Rabiner and Levinson, 1981) การเลือกค่าอันดับขึ้นอยู่กับความถูกต้องแม่นยำของการแทนเอนเวโลปเชิงสเปกตรัม เวลาในการคำนวณ และเนื้อที่หน่วยความจำ

3.2.5 การวิเคราะห์ค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น (LPC Analysis)

จะนำค่าอัตสหสัมพันธ์ที่ได้จากขั้นตอนที่ 3.2.4 มาคำนวณหาค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น (LPC) ด้วยวิธีของ Levinson-Durbin ดังที่ได้กล่าวมาแล้วในหัวข้อที่ 2.5.4 ผลลัพธ์สุดท้ายจะได้ $a_m(1), a_m(2), ..., a_m(p)$ เป็นค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้นจำนวน p ค่าในแต่ละส่วนย่อย ดังนั้นสัญญาณเสียงแต่ละคำจะถูกแทนด้วยชุดของสัมประสิทธิ์การประมาณพันธะเชิงเส้น p อันดับ จำนวน L ชุดต่อกัน ชุดข้อมูลนี้จะใช้เป็นเวกเตอร์ข้อมูลเข้าของนิวรอลเน็ตเวอร์กต่อไป

3.3 การสกัดลักษณะเด่นแบบฟัซซี (Fuzzy Feature Extraction)

เวกเตอร์ข้อมูลเข้าที่ได้จากขั้นตอนการสกัดค่าลักษณะเด่น ซึ่งเขียนได้ในรูป

$$\underline{i} = [i_1 \quad i_2 \quad i_3 \quad ..... \quad i_n]^t \qquad (3.8)$$

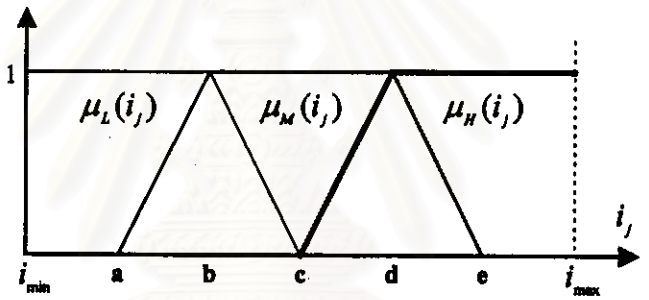
โดยที่ i_j ใดๆ ($j = 1, 2, ..., n$) เป็นค่าของสัมประสิทธิ์การประมาณพันธะเชิงเส้น n เป็นจำนวนสัมประสิทธิ์ทั้งหมดที่แทนสัญญาณเสียงหนึ่งคำ ซึ่งมีจำนวนเท่ากับ $p \times L$ ค่า จะถูกเปลี่ยนเป็นเวกเตอร์ใหม่ที่ประกอบด้วยค่าสมาชิกภาพแบบฟัซซี ของค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น บนคุณสมบัติทางภาษา 3 ระดับ { น้อย (L), ปานกลาง (M), มาก (H)} โดยอาศัยฟังก์ชันสมาชิกภาพแบบฟัซซี บนคุณสมบัติทางภาษาทั้งสาม ดังนั้นจะได้เวกเตอร์ข้อมูลเข้าแบบใหม่ มีมิติเป็น 3 เท่าของมิติของเวกเตอร์ข้อมูลเข้าแบบเดิม ดังนี้

$$\underline{\tilde{I}} = [\ \mu_L(i_1)\ \ \mu_M(i_1)\ \ \mu_H(i_1)\ \ \mu_L(i_2)\ \ \mu_M(i_2)\ \ \mu_H(i_2)\ \ \ldots\ldots\ \ \mu_L(i_n)\ \ \mu_M(i_n)\ \ \mu_H(i_n)\ ]^t \quad (3.9)$$

ฟังก์ชันสมาชิกภาพแบบฟัซซีที่ใช้ มี 3 ชนิดได้แก่

3.3.1 ชนิดสี่เหลี่ยมคางหมู (Trapezoidal)

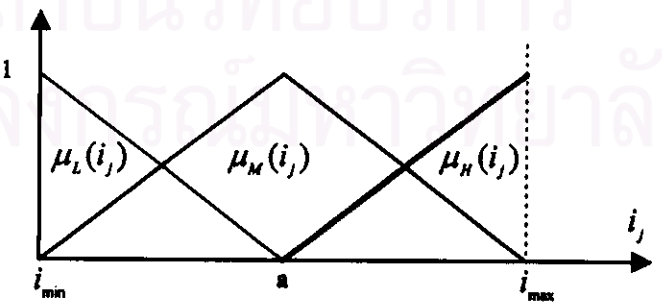
มีกราฟแสดงฟังก์ชันสมาชิกภาพแบบฟัซซี บนคุณสมบัติทางภาษา 3 ระดับ ดังรูปที่ 3.5 โดยที่ i_{max} และ i_{min} เป็นขอบเขตของปริภูมิ (Space) ของค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้นทั้งหมด หาได้โดยบวกเผื่อจากค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้นสูงสุด และต่ำสุดที่สกัดได้จากสัญญาณเสียงคำทุกคำในชุดฝึกฝน ค่า a, b, c, d และ e หาได้จากการแบ่งช่วงจาก i_{min} ถึง i_{max} ออกเป็น 6 ส่วนเท่าๆ กัน



รูปที่ 3.5 กราฟแสดงฟังก์ชันสมาชิกภาพแบบฟัซซีชนิดสี่เหลี่ยมคางหมู บนคุณสมบัติทางภาษา 3 ระดับ

3.3.2 ชนิดสามเหลี่ยม (Triangular)

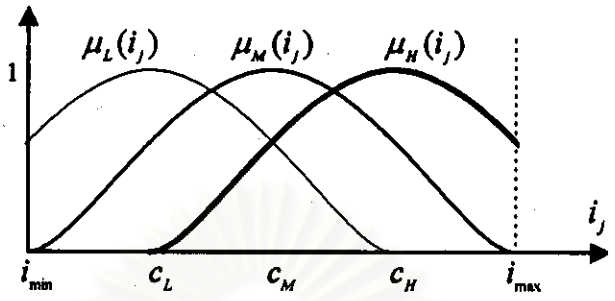
มีกราฟแสดงได้ดังรูปที่ 3.6 โดยที่จุด a คือจุดกึ่งกลางระหว่างค่า i_{min} ถึง i_{max}



รูปที่ 3.6 กราฟแสดงฟังก์ชันสมาชิกภาพแบบฟัซซีชนิดสามเหลี่ยม บนคุณสมบัติทางภาษา 3 ระดับ

3.3.3 ชนิด Pi

มีกราฟแสดงได้ ดังรูปที่ 3.7



รูปที่ 3.7 กราฟแสดงฟังก์ชันสมาชิกภาพแบบฟัซซีชนิด Pi บนคุณสมบัติทางภาษา 3 ระดับ

การคำนวณหาค่าสมาชิกภาพแบบฟัซซี จะอาศัยสมการที่ 2.18 โดยที่ ค่า c และ λ สำหรับแต่ละฟังก์ชันบนคุณสมบัติทางภาษาทั้งสามระดับ สามารถหาได้ดังนี้

$$\begin{aligned}
\lambda_M &= \frac{1}{2}(i_{max} - i_{min}) \\
c_M &= i_{min} + \lambda_M \\
\lambda_L &= \frac{1}{fdenom}(c_M - i_{min}) \\
c_L &= c_M - \frac{1}{2}\lambda_L \\
\lambda_H &= \frac{1}{fdenom}(i_{max} - c_M) \\
c_H &= c_M + \frac{1}{2}\lambda_H
\end{aligned} \qquad (3.10)$$

fdenom เป็นค่าที่ปรับสัดส่วนในการทับซ้อนกัน (Overlapping) ของกราฟ

3.4 เวกเตอร์ข้อมูลออกที่ต้องการ (Desired Output Vector)

เวกเตอร์ข้อมูลที่ต้องการ ซึ่งสอดคล้องกับเวกเตอร์ข้อมูลเข้า สำหรับฝึกฝนนิวรอลเน็ตเวอร์กแบบ MLP จะมีมิติเท่ากับจำนวนคำศัพท์ที่จะรู้จำ จากการศึกษางานวิจัยที่ใช้นิวรอลเน็ตเวอร์กแบบ MLP จะมีการใช้ข้อมูลออกที่ต้องการในการฝึกฝนได้หลายรูปแบบ ได้แก่

1) เป็นเลขฐานสอง (Binary)

เมื่อเวกเตอร์ข้อมูลเข้า $\underline{i}$ สกัดมาจากสัญญาณเสียงคำศัพท์ตัวที่ j ใดๆ ในชุดฝึกฝน เมื่อ $j = 1, 2, ..., c$ โดยที่ c เท่ากับจำนวนคำศัพท์ที่จะรู้จำ จะกำหนดให้ค่าข้อมูลออก ณ ตำแหน่งที่ j มีค่าเท่ากับ 1 นอกนั้นจะให้มีค่าเท่ากับ 0

2) เป็นค่าสมาชิกภาพของแต่ละคำศัพท์ (Class Membership)

เวกเตอร์ข้อมูลออกที่ต้องการจะอยู่ในรูป

$$\underline{t} = [\, \mu_1(i) \quad \mu_2(i) \quad ..... \quad \mu_c(i)]^t \qquad (3.11)$$

โดยที่ $\mu_k(i)$ เป็นค่าสมาชิกภาพของเวกเตอร์ข้อมูลเข้า $\underline{i}$ ในคำศัพท์ตัวที่ k สามารถคำนวณได้ดังสมการที่ 2.21 และสมการที่ 2.22 สำหรับเวกเตอร์ข้อมูลเข้า $\underline{i}$ ซึ่งสกัดมาจากสัญญาณเสียงคำศัพท์ตัวที่ j $\mu_k(i)$ จะมีค่ามากที่สุดเมื่อ $k = j$ และจะมีค่าน้อยเมื่อ $k \neq j$

3) เป็นค่าสมาชิกภาพของแต่ละคำศัพท์ ในกรณีกำกวมที่สุด (Fuzziest Case)

ในกรณีที่ข้อมูลออกที่ต้องการแบบ $\mu_k(i)$ มีความกำกวม จะทำการปรับค่าใหม่เป็นค่า $\mu_{INT(k)}(i)$ ดังที่แสดงในสมการที่ 2.23

มีงานวิจัยหลายงานที่เสนอให้ใช้ข้อมูลออกที่ต้องการเป็นค่าสมาชิกภาพของแต่ละคำศัพท์ $\mu_k(i)$ ดังที่แสดงในหัวข้อที่ 2.4.2 ร่วมกับข้อมูลเข้าเป็นค่าสมาชิกภาพแบบฟัซซี แต่จากการทดลองเบื้องต้น ผู้ทดลองได้เสนอสมมติฐานประการหนึ่งคือ

" สำหรับงานวิจัยนี้ การใช้ข้อมูลออกที่ต้องการเป็นเลขฐานสอง ร่วมกับข้อมูลเข้าเป็นค่าสมาชิกภาพแบบฟัซซีในการฝึกฝน น่าจะให้อัตราการรู้จำสูงกว่าการใช้ข้อมูลออกที่ต้องการเป็นค่าสมาชิกภาพของแต่ละคำศัพท์ "

สำหรับสาเหตุที่เสนอสมมติฐานดังกล่าว จะวิเคราะห์ให้เห็นในหัวข้อที่ 4.3

3.5 นิวรอลเน็ตเวอร์กแบบ MLP

เวกเตอร์ข้อมูลเข้า และเวกเตอร์ข้อมูลออกที่ต้องการ ที่ได้จากขั้นตอนที่ 3.3 และ 3.4 จะนำมาฝึกฝนนิวรอลเน็ตเวอร์กแบบ MLP ซึ่งมีรายละเอียดดังต่อไปนี้

3.5.1 โครงสร้างของนิวรอลเน็ตเวอร์กแบบ MLP ที่ใช้

นิวรอลเน็ตเวอร์กแบบ MLP ที่ใช้ในงานวิจัยนี้ ประกอบด้วย ระดับชั้นข้อมูล (Input Layer) ซึ่งมีจำนวนโหนดเท่ากับมิติของเวกเตอร์ข้อมูลเข้า ระดับชั้นซ่อนตัว (Hidden Layer)

1 ระดับ เนื่องจากเพียงพอสำหรับการประมาณฟังก์ชันต่อเนื่องใดๆ ได้ (Schalkoff, 1992) โดยมีจำนวนโหนดที่เหมาะสม หาได้จากการทดลอง และระดับชั้นข้อมูลออก (Output Layer) ซึ่งมีจำนวนโหนดเท่ากับมิติของเวกเตอร์ข้อมูลออก หรือมีจำนวนเท่ากับจำนวนคำศัพท์ที่จะรู้จำด้วยนิวรอลเน็ตเวอร์กชุดนี้นั่นเอง

3.5.2 ตัวแปรสำคัญในการฝึกฝน

การฝึกฝนจะอาศัยวิธีแพร่กระจายย้อนกลับ (Back-propagation) ดังที่ได้กล่าวมาในหัวข้อ 2.2.2 โดยมีตัวแปรสำคัญดังนี้

1) ค่าอัตราการเรียนรู้ (Learning Rate) ε เป็นตัวกำหนดสัดส่วนในการปรับค่าน้ำหนักการเชื่อมต่อ โดยทั่วไปจะสุ่มค่าระหว่าง 0 ถึง 1

2) อัตราโมเมนตัม (Momentum Rate) α เป็นตัวกำหนดสัดส่วนในการนำค่าน้ำหนักการเชื่อมต่อในรอบก่อนมาใช้ในการปรับในรอบปัจจุบัน ช่วยป้องกันการแกว่ง (Oscillation) ในการปรับมากเกินไป ในงานวิจัยนี้ ใช้ค่าอัตราโมเมนตัมเท่ากับ 0.9 (Pornsukchantra, 1996)

3) ระดับความผิดพลาดที่ยอมรับได้ (Error Threshold) E_t สำหรับช่วงฝึกฝน แต่ละรอบ ในการปรับค่าน้ำหนักการเชื่อมต่อจะมีการคำนวณค่าความผิดพลาดที่ระดับชั้นข้อมูลออก E ดังแสดงในสมการที่ 2.5 การฝึกฝนด้วยเสียงคำแต่ละคำในชุดฝึกฝน จะจบลงก็ต่อเมื่อ $E < E_t$ ค่า E_t ที่เหมาะสมขึ้นอยู่กับค่าข้อมูลออกที่ต้องการที่ใช้ในการฝึกฝน สำหรับในงานวิจัยนี้จะเลือกใช้ $E_t = 0.01$ สำหรับข้อมูลออกที่ต้องการเป็นเลขฐานสอง และจะใช้ $E_t = 0.000001$ สำหรับข้อมูลออกที่ต้องการเป็นค่าสมาชิกภาพของแต่ละคำศัพท์ ซึ่งเป็นค่าที่ทดลองแล้วเหมาะสม

4) จำนวนคำที่ฝึกฝนสำเร็จที่ยอมรับได้ ในทางปฏิบัติ เราจะฝึกฝนนิวรอลเน็ตเวอร์ก ด้วยชุดฝึกฝน ซึ่งประกอบด้วยเสียงคำหลายๆ เสียง พร้อมกันทีเดียวทั้งหมด ดังนั้น นอกจากการพิจารณาค่าความผิดพลาดของเสียงคำแต่ละคำแล้ว จำนวนของเสียงคำที่ฝึกฝนสำเร็จต้องมีค่าเกินกว่าจำนวนคำที่กำหนดด้วยจึงจะถือว่าการฝึกฝนเสร็จสมบูรณ์ ในงานวิจัยนี้กำหนดให้จำนวนคำที่ฝึกฝนสำเร็จต้องมีจำนวนมากกว่า 95 เปอร์เซ็นต์ของจำนวนคำทั้งหมดในชุดฝึกฝน จึงจะหยุดการฝึกฝน

ในการฝึกฝน การกำหนดค่าน้ำหนักการเชื่อมต่อเริ่มต้น (Initial Weight) จะกำหนดด้วยการสุ่ม (Random) เพื่อป้องกันการสมมาตร (Symmetry) ของการปรับค่าน้ำหนักการเชื่อมต่อ ซึ่งส่งผลให้ไม่มีทางฝึกฝนสำเร็จได้เลย

3.5.3 นิวรอลเน็ตเวอร์กย่อย

เนื่องจากการใช้นิวรอลเน็ตเวอร์กแบบ MLP ในการรู้จำ จำนวนคำศัพท์ที่จะรู้จำมีผลอย่างมากต่ออัตราการรู้จำ จึงต้องมีการแยกชุดคำศัพท์ออกเป็นกลุ่มย่อยๆ แต่ละกลุ่มจะใช้นิวรอลเน็ตเวอร์กแบบ MLP ในการรู้จำกลุ่มละชุดแยกกันไป ผลการฝึกฝนจะได้ค่าน้ำหนักการเชื่อมต่อสำหรับนิวรอลเน็ตเวอร์กย่อยแต่ละชุด เก็บไว้สำหรับขั้นตอนการทดสอบการรู้จำต่อไป

ช่วงทดสอบการรู้จำ (Test Phase)

ช่วงทดสอบการรู้จำมีขั้นตอนย่อยแสดงในรูปที่ 3.1 (ในกรอบเส้นประ) โดยที่ช่วงการประมวลผลเบื้องต้น (Preprocessing) การสกัดค่าลักษณะเด่น (Feature Extraction) การสกัดค่าลักษณะเด่นแบบฟัซซี (Fuzzy Feature Extraction) และนิวรอลเน็ตเวอร์กแบบ MLP มีรายละเอียดเหมือนกับช่วงฝึกฝน ดังนั้นจะกล่าวถึงรายละเอียดในส่วนที่แตกต่างจากช่วงฝึกฝนคือ การแบ่งกลุ่มเบื้องต้น (Pre-grouping) และกฎเกณฑ์การตัดสินใจ (Decision Rule)

3.6 การแบ่งกลุ่มเบื้องต้น (Pre-grouping)

เป็นขั้นตอนในการแบ่งชุดคำศัพท์ที่จะรู้จำทั้งหมดเป็นกลุ่มย่อยๆ เพื่อลดจำนวนคำศัพท์ที่นิวรอลเน็ตเวอร์กชุดหนึ่งๆ จะต้องรู้จำ ช่วยเพิ่มอัตราการรู้จำ และลดระยะเวลาในการฝึกฝน และรู้จำได้มาก วิธีการแบ่งกลุ่มคำศัพท์จะอาศัยลักษณะบ่งความต่าง (Distinctive Feature) ที่เด่นชัดของคำศัพท์ที่จะรู้จำ ในงานวิจัยนี้ได้เสนอวิธีการแบ่งกลุ่ม 2 วิธีดังนี้

3.6.1 การตรวจสอบจำนวนพยางค์ (Syllable Detection)

สำหรับงานวิจัยนี้ ต้องการจะรู้จำเสียงคำไทยหลายพยางค์ (Polysyllabic Word) โดยมีคำศัพท์ที่จะรู้จำขนาดตั้งแต่ 1 ถึง 3 พยางค์ ดังที่ได้ทราบแล้วว่า นิวรอลเน็ตเวอร์กแบบ MLP ชุดหนึ่งๆ จะต้องมีจำนวนโหนดในระดับชั้นข้อมูลเข้าคงที่เสมอ ส่งผลให้จำเป็นต้องนำสัญญาณเสียงคำที่จะรู้จำด้วยนิวรอลเน็ตเวอร์กชุดเดียวกัน มาผ่านการนอร์มอลไลซ์ทางเวลาให้มีความยาวทางเวลาเท่ากัน จึงมีความเหมาะสมที่จะแยกคำศัพท์ทั้งหมดออกเป็น 3 กลุ่มคือ กลุ่มคำที่มี 1, 2 และ 3 พยางค์ กลุ่มที่มีจำนวนพยางค์มากก็จะนำไปผ่านการนอร์มอลไลซ์ทางเวลาให้มีความยาวมากกว่ากลุ่มที่มีจำนวนพยางค์น้อย และสามารถใช้นิวรอลเน็ตเวอร์กที่มีจำนวนโหนดในระดับชั้นข้อมูลเข้าต่างกัน ในการรู้จำคำศัพท์แต่ละกลุ่มแยกกันได้

การตรวจสอบจำนวนพยางค์ ๆ จะเป็นผลพลอยได้ของการตัดแบ่งพยางค์ซึ่งมีวิธีการหลักๆ ดังที่ได้กล่าวมาแล้วในหัวข้อ 2.5.1

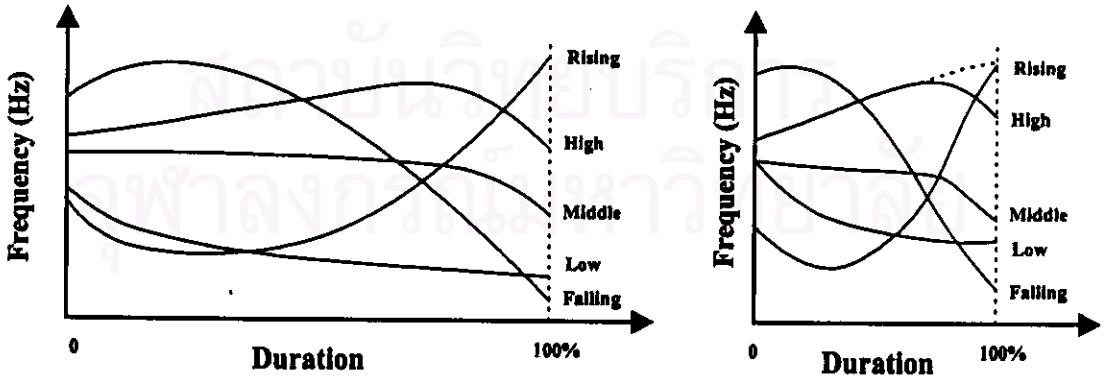
3.6.2 การตรวจสอบเสียงวรรณยุกต์ (Tone Detection)

สำหรับกลุ่มคำศัพท์ที่แยกด้วยการตรวจสอบจำนวนพยางค์แล้ว ยังสามารถนำมาแยกกลุ่มด้วยวิธีตรวจสอบเสียงวรรณยุกต์ได้อีก จะยิ่งทำให้กลุ่มคำศัพท์เล็กลง ช่วยเพิ่มอัตราการรู้จำให้สูงขึ้นไปอีก หลักการตรวจสอบเสียงวรรณยุกต์ของเสียงพยางค์ในภาษาไทย จะอาศัยลักษณะการเปลี่ยนแปลงของค่าความถี่มูลฐาน (Fundamental Frequency) F_0 Abramson (ณัฐกร ทับทอง 2538) ได้ศึกษา และสรุปไว้ว่า เสียงวรรณยุกต์ในภาษาไทยมี 5 เสียง แบ่งได้เป็น 2 กลุ่มได้แก่

1) เสียงคงระดับ (Static Tones) ซึ่งมีการเปลี่ยนแปลงค่าความถี่มูลฐานน้อย ได้แก่
   - เสียงสามัญ หรือเสียงกลาง (Middle Tone) มีค่าความถี่มูลฐานเฉลี่ยปานกลาง
   - เสียงเอก หรือเสียงต่ำ (Low Tone) มีค่าความถี่มูลฐานเฉลี่ยต่ำ
   - เสียงตรี หรือเสียงสูง (High Tone) มีค่าความถี่มูลฐานเฉลี่ยสูง
2) เสียงไม่คงระดับ (Dynamic Tones) มีการเปลี่ยนแปลงค่าความถี่มูลฐานมาก ได้แก่
   - เสียงโท หรือเสียงตก (Falling Tone) ค่าความถี่มูลฐานจะโค้งลงอย่างมาก
   - เสียงจัตวา หรือเสียงขึ้น (Rising Tone) ค่าความถี่มูลฐานจะโค้งขึ้นอย่างมาก

เขียนเป็นกราฟของค่าความถี่มูลฐาน เทียบกับเวลาในการเปล่งเสียงหนึ่งพยางค์ โดยแบ่งเป็นพยางค์ที่ประกอบด้วยสระเสียงยาว และเสียงสั้น ดังแสดงในรูปที่ 3.8 ดังนั้นเราสามารถตรวจสอบเสียงวรรณยุกต์ในภาษาไทยได้ โดยพิจารณา 2 ส่วนคือ ทิศทางของค่าความถี่มูลฐาน (F_0 Direction) ซึ่งพิจารณาได้จากค่าความถี่มูลฐานที่จุดเริ่มต้น และที่จุดท้ายพยางค์ และระดับของค่าความถี่มูลฐาน (F_0 Height) ซึ่งพิจารณาได้จากค่าความถี่มูลฐานเฉลี่ยตลอดทั้งพยางค์

การตรวจสอบเสียงวรรณยุกต์ของคำภาษาไทย ได้เคยมีงานวิจัยมาแล้วคือ งานวิจัยของณัฐกร ทับทอง (2530)



(ก) สระเสียงยาว (ข) สระเสียงสั้น

รูปที่ 3.8 กราฟแสดงค่าความถี่มูลฐานเทียบกับเวลา

3.7 กฎเกณฑ์การตัดสินใจ (Decision Rule)

เมื่อทำการแพร่ค่าในเวกเตอร์ข้อมูลเข้า เข้าไปในนิวรอลเน็ตเวอร์กแบบ MLP โดยใช้ค่าน้ำหนักการเชื่อมต่อที่ได้จากช่วงฝึกฝน จะได้เวกเตอร์ข้อมูลออกที่มีค่าข้อมูลออกใกล้เคียงกับค่าในเวกเตอร์ข้อมูลออกที่ต้องการในช่วงฝึกฝน ดังนั้นการตัดสินใจว่าเป็นคำศัพท์ตัวใดจะพิจารณาจากตำแหน่งของข้อมูลออกที่มีค่าข้อมูลออกมากที่สุด (Maximum Likelihood) ดังสมการที่ 2.11

ในขั้นตอนนี้ จะทำการคำนวณค่าอัตราการรู้จำ (Recognition Rate) โดยสำหรับกลุ่มคำศัพท์ที่จะรู้จำใดๆ จะหาได้ดังสมการนี้

$$\text{Recognition Rate}\,(\%) = \frac{No.\ of\ Correct\ Word}{No.\ of\ Total\ Word} \times 100 \qquad (3.12)$$

แต่ถ้ามีการทดลองหาอัตราการรู้จำเฉลี่ยจากอัตราการรู้จำของนิวรอลเน็ตเวอร์กย่อยที่ใช้กับกลุ่มคำศัพท์ย่อยแต่ละกลุ่ม จะหาได้จากสมการนี้

$$\text{Average Reconition Rate}\ (\%) = \frac{\sum_{n}(\text{Re}conition\ Rate\ of\ Group\ n)(No.of\ Vocab\ in\ Group\ n)}{\sum_{n}(No.of\ Vocab\ in\ Group\ n)} \qquad (3.13)$$