บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้องกับงานวิจัยนี้ สามารถแบ่งออกได้เป็น 2 กลุ่ม คือ

2.1.1 ทฤษฎีที่เกี่ยวกับการรู้จำตัวอักษร (Pattern Recognition)

2.1.1.1 การจับคู่เปรียบเทียบกับอักขระต้นแบบ (Template matching) [1]

การจับคู่เปรียบเทียบกับอักขระต้นแบบ เป็นวิธีการรู้จำตัวอักขระที่ง่ายที่สุด ตัวอักขระนำเข้าจะถูกนำมาเปรียบเทียบรูปแบบกับตัวอักษรต้นแบบที่จัดทำขึ้นมาก่อน การที่จะรู้ว่าตัวอักขระนำเข้าเป็นอักขระตัวใด ก็โดยการดูจากผลการเปรียบเทียบอักขระนำเข้ากับอักขระต้นแบบว่าใกล้เคียงกับอักขระต้นแบบตัวใดมากที่สุด โดยค่าความผิดพลาดที่เกิดขึ้นต้องไม่เกินค่าการยอมรับได้ (Threshold) ที่กำหนด

สูตรทางคณิตศาสตร์ที่ใช้ในการคำนวณการจับคู่เปรียบเทียบกับอักขระต้นแบบมีอยู่หลายรูปแบบ เช่น สมมติให้อักขระนำเข้า

$$P = \{ p(x_1, y_1), p(x_2, y_2),..., p(x_n, y_n) \}$$

และอักขระต้นแบบ

$$R = \{ r(s_1, t_1), r(s_2, t_2),..., r(s_n, t_n) \}$$

โดยที่ $p(x, y)$ และ $r(s, t)$ คือจุดจุดหนึ่งบนระนาบ 2 มิติ สูตร Euclidean distance กำหนดว่า

$$D_E(P, R) = \sum_n \sqrt{(x_n - s_n)^2 + (y_n - t_n)^2}$$

สูตร city - block หรือ D4 กำหนดว่า

$$D_C(P, R) = \sum_n (|x_n - s_n| + |y_n - t_n|)$$

สูตร Hamming distance กำหนดว่า

$$D_H(P, R) = \sum_n e_n$$

โดยที่ e_n คือ

$$e_n = \begin{cases} 1 & \text{ถ้า } p(x_n, y_n) = r(s_n, t_n) \\ 0 & \text{ถ้า } p(x_n, y_n) = r(s_n, t_n) \end{cases}$$

2.1.1.2 การตรวจรู้จากข้อมูลทางสถิติ (Statistical Methods) [2]

การตรวจรู้อักขระวิธีนี้จะใช้ข้อมูลที่เป็นตัวเลขในการ วิเคราะห์หารหัสของอักขระตัวเลขดังกล่าวได้จากขั้นตอนการจัดเตรียมภาพก่อนการประมวลผล (Preprocessing) และใช้วิธีการวิเคราะห์และแยกภาพโดยสมการทางคณิตศาสตร์ (Deterministic Classification Technique) ในการหาหรือในบางวิธีจะใช้ทฤษฎีทางสถิติเรื่องความน่าจะเป็นเข้าช่วย เช่น วิธีของ Bayesian Estimation เป็นต้น

2.1.1.3 การตรวจรู้จากหลักโครงสร้างของอักขระ (Syntactic Methods)

การตรวจรู้อักขระวิธีนี้จะอาศัยแบบการคำนวณที่สามารถเปลี่ยนรูปแบบของอักขระให้เป็นรหัสที่สามารถใช้อ้างอิงอักขระนั้น ๆ ได้ โดยรหัสเหล่านั้นจะต้องเป็นรหัส

ที่ไม่ซ้ำกันสำหรับอักขระที่ต่างรูปกัน และจะมีค่าของรหัสเหมือนกันหรือใกล้เคียงกันสำหรับอักขระรูปเดียวกัน แต่ต่างแบบพิมพ์กัน เมื่อนำรหัสของอักขระทุกรูปมารวมกันก็จะได้เป็นรหัสต้นแบบที่สามารถใช้ในการเปรียบเทียบกับรหัสของอักขระที่ต้องการจะตรวจรู้ได้ วิธีการที่ใช้ในการค้นหารหัสที่จะตรวจรู้กับรหัสต้นแบบนั้น จะใช้วิธีการสร้างกฎเกณฑ์หรือหลักไวยกรณ์ (Pattern Grammar) ของรหัสของอักขระ เพื่อใช้ในการค้นหารหัสของอักขระที่ต้องการจะตรวจรู้กับรหัสต้นแบบ เพราะอักขระรูปเดียวกันแต่ต่างแบบพิมพ์กันย่อมจะมีค่ารหัสที่คล้ายกันจนสามารถใช้หลักเกณฑ์บางอย่าง ซึ่งสามารถใช้เป็นหลักไวยกรณ์ของรหัสนั้นมาอ้างอิงถึงได้

2.1.2 ทฤษฎีที่เกี่ยวกับการอ่านออกเสียงจากแฟ้มข้อความ (Text-to-speech) [3]

การแปลงข้อความให้ออกมาเป็นเสียงพูดนั้น (Text-to-speech) คือ ลักษณะของการที่เราป้อนข้อความเป็นคำๆ หรือชุดของหน่วยย่อยของคำเข้าไป แล้วเครื่องจะทำการแปลงชุดของข้อความนั้นให้ออกมาเป็นเสียงพูดตามชุดตัวอักขระนั้นได้ (Veltri, 1985) [4] หลักการหรือขั้นตอนอย่างคราวๆ ของการแปลงข้อความให้ออกมาเป็นเสียงพูด จะสามารถทำได้โดยเริ่มจากระบบจะทำการรับข้อความเข้ามา แล้วก็เปลี่ยนข้อความนั้นให้เป็นชุดของโฟนีม (phoneme) หรือสัญญลักษณ์แทนเสียงของคำแต่ละคำแทน รวมทั้งพยายามที่จะกำหนดค่าต่างๆ ที่เกี่ยวข้องกับเสียง เช่น ความยาวของคำ พยางค์ที่จะเน้น ระดับความดังและระดับของพิทช์ (pitch) ที่ใช้ในการเน้น เป็นต้น ขั้นต่อไปก็คือการแปลโฟนีมและตัวแปรต่างๆ ข้างต้นให้กลายเป็นพารามิเตอร์ในโดเมนของความถี่ รวมทั้งการจัดการเกี่ยวกับการเปลี่ยนแปลงอย่างต่อเนื่องระหว่างโฟนีมหนึ่งกับโฟนีมตัวถัดไปแล้วพารามิเตอร์เหล่านี้ก็จะถูกทำการสังเคราะห์ให้ออกมาเป็นเสียงพูดอีกที

จากหลักการข้างต้นพอจะสรุปได้ว่า การแปลงตัวอักขระให้ออกมาเป็นเสียงพูดนั้นมี 2 ขั้นตอนใหญ่ๆ คือ

2.1.2.1 ขั้นตอนที่ทำการรับข้อความเข้ามา

เป็นส่วนที่ทำการรับข้อความเข้ามาแล้วแปลงเป็นสัญญลักษณ์แทนเสียง พร้อมทั้งพารามิเตอร์ที่ควบคุมลักษณะของเสียง เช่น ระดับความถี่การเน้นเสียงที่พยางค์ เป็นต้น ในส่วนแรกนั้นเราจำเป็นจะต้องรู้ว่าข้อความที่ป้อนเข้ามานั้นอ่านออกเสียงอย่างไรจึงจะถูกต้อง แล้วจึงแทนด้วยชุดสัญญลักษณ์ตามเสียงที่ถูกต้องนั้น ซึ่งอาจจะทำได้โดยวิธีการดังนี้ (Bursky, 1985) [5]

ใช้ชุดของกฎเกณฑ์ต่างๆ ตามหลักการแปลงตัวอักษรเป็นหน่วยเสียง (Letter-to-sound rule set)

นั่นคือการรวบรวมกฎเกณฑ์ข้อบังคับต่างๆ ที่บัญญัติไว้ในหลักภาษาของการอ่าน เพื่อจะได้จัดเตรียมชุดสัญญลักษณ์ทางเสียงได้ถูกต้อง ดังนั้นจะมีกฎเกณฑ์เหล่านี้อยู่มากมาย แต่อย่างไรก็ตามจะต้องมีคำที่แปลกออกไปไม่ได้อ่านออกเสียงตามกฎเกณฑ์เหล่านั้น เป็นคำที่อยู่นอกเหนือกฎเกณฑ์ ทำให้การใช้วิธีการนี้เพียงอย่างเดียวไม่สามารถครอบคลุมการอ่านคำต่างๆ ได้อย่างถูกต้องทั้งหมด

การใช้ระบบพจนานุกรม (Dictionary)

เป็นการแก้ปัญหาจากการให้กฎเกณฑ์ต่างๆ ได้เป็นบางส่วน นั่นคือ ในกรณีที่เป็นคำที่อยู่นอกเหนือกฎเกณฑ์ หรืออ่านออกเสียงผิดไปจากหลักการตามปกติ ก็จะจัดคำเหล่านั้นเป็นกลุ่มๆ พร้อมทั้งกำหนดคำอ่านที่ถูกต้องของแต่ละคำเอาไว้รวบรวมแยกไว้เป็นพจนานุกรมในแฟ้มข้อมูล ก็จะสามารถค้นหาสัญญลักษณ์แทนเสียงที่ถูกต้องของแต่ละคำได้ แต่อย่างไรก็ตามก็ยังมีข้อยกเว้นอยู่บ้าง เช่น ในกรณีของคำพ้องรูป ที่สามารถอ่านออกเสียงได้หลายแบบโดยขึ้นกับความหมายของคำที่อยู่รอบๆ คำนั้น ซึ่งจะไปเกี่ยวข้องเข้าสู่ปัญหาในแง่ของปัญญาประดิษฐ์ ทำให้ยุ่งยากขึ้นไปอีกหลายระดับ

2.1.2.2 ขั้นตอนที่ทำหน้าที่สังเคราะห์เสียงตามชุดสัญญลักษณ์แทนเสียงและพารามิเตอร์

เป็นส่วนที่ใช้พารามิเตอร์สังเคราะห์เสียงขึ้นมาตามชุดสัญญลักษณ์แทนเสียงให้ออกมาเป็นเสียงพูดที่เราได้ยินกัน สำหรับในส่วนที่สองนี้เป็นส่วนของการสังเคราะห์เสียงพูดโดยตรง จึงเป็นส่วนที่สำคัญมากสำหรับระบบการแปลงข้อความเป็นเสียงพูดนี้ เพราะเป็นส่วนที่จะทำให้เครื่องสามารถกำเนิดเสียงพูดออกมาได้ เทคนิคในการทำให้เครื่องสามารถสร้างเสียงพูดออกมาได้นั้นมีอยู่หลายวิธี แต่วิธีที่มีความยืดหยุ่นมากที่สุดวิธีหนึ่งก็คือ การใช้ลักษณะของการเก็บการแทนค่าของคำพูดต้นฉบับเอาไว้ (store a textual representation of the utterance) ซึ่งมีวิธีการที่แตกต่างกันออกไป 2 วิธีใหญ่ๆ คือ (Witten, 1982) [6]

การสร้างเสียงโดยการเก็บบันทึกเสียงพูด (speech storage)

เป็นวิธีของการเก็บบันทึกเสียงจริงๆ ของมนุษย์เอาไว้อย่างตรงไปตรงมา โดยอาจจะแบ่งเก็บเป็นข้อมูลของหน่วยย่อยๆ ของคำหรือพยางค์ ซึ่งเมื่อมีการสร้างเสียงพูดขึ้นมาก็จะมีการอ้างถึงข้อมูลของเสียงหน่วยย่อยๆ นี้นำมาเรียงต่อๆ กันเกิดเป็นคำหรือประโยคขึ้นมา

ข้อดี ก็คือ สามารถจะสร้างเสียงพูดที่มีคุณภาพสูงมากๆ หรือใกล้เคียงกับสำเนียงของมนุษย์ได้เป็นอย่างดี

ข้อเสีย ก็คือ ใช้หน่วยเก็บข้อมูลมากตามคุณภาพของเสียงที่ได้และคำศัพท์ที่มีอยู่ในพจนานุกรมเสียง

การสร้างเสียงโดยการสังเคราะห์เสียงพูด (speech Synthesis)

เป็นการสร้างเสียงพูดขึ้นมาโดยที่เครื่องจะเป็นตัวสร้างเสียงพูดขึ้นมาโดยตัวมันเอง โดยไม่จำเป็นต้องมีการบันทึกเสียงของมนุษย์สำหรับเสียงพูดที่ต้องการจะให้เครื่องพูดออกมาเอาไว้เลย

ข้อดี ใช้หน่วยความจำน้อยกว่าวิธีแรกมาก ทำให้สามารถมุ่งไปสู่การสร้างระบบสังเคราะห์เสียงที่ไม่จำกัดจำนวนคำศัพท์ และที่สำคัญที่สุดก็คือสามารถทำได้บนเครื่องคอมพิวเตอร์ขนาดเล็กและสิ้นเปลืองค่าใช้จ่ายน้อย

ข้อเสีย โปรแกรมในการสังเคราะห์เสียงพูดมีความยุ่งยากและเสียงที่ได้ออกมามีคุณภาพอยู่ในระดับปานกลาง

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 งานวิจัยที่เกี่ยวกับการรู้จำตัวอักษร (Pattern Recognition) [7]

ในปี 1984 พิพัฒน์ หิรัณย์วณิชชากรและคนอื่นๆ ได้เสนอเทคนิคการวิเคราะห์เส้นแสดงขอบของอักษร และการหาค่าความคล้ายระหว่างส่วนโค้ง โดยใช้ลักษณะสำคัญได้แก่ จำนวนหัว จำนวนส่วนโค้ง ลักษณะการแยกกันของอักษร และความยาวของแต่ละส่วนโค้งย่อย เพื่อใช้ในการรู้จำตัวพิมพ์อักษรไทยในระบบออฟ-ไลน์กับตัวอักษรรูปแบบเดียวที่มีขนาด 50 x 50 จุด ซึ่งผลของการรู้จำตัวอักษรมีความถูกต้อง 99.40 % (พิพัฒน์ หิรัณย์วณิชชากร, 1984)

ในปี 1986 ชมทิพ พรพนมชัย ได้เสนอเทคนิคการเปลี่ยนเส้นแสดงโครงร่างอักษรให้อยู่ในรูปของรหัส และการเปรียบเทียบความเหมือนของรหัสที่ได้กับรหัสต้นแบบ โดยใช้ลักษณะสำคัญได้แก่ การกระจายของจุดตามแนวแถว และแนวสดมภ์ เพื่อใช้ในการรู้จำตัวพิมพ์อักษรไทยในระบบออฟ-ไลน์ กับตัวอักษรรูปแบบเดียวที่มีขนาด 20 x 20 จุด ซึ่งผลของการรู้จำตัวอักษรมีความถูกต้อง 70.00 % (ชมทิพ พรพนมชัย, 1986)

ในปี 1987 ชม กิ้มปาน และคนอื่นๆ ได้เสนอเทคนิคการกระจายแบบคาร์ฮูเนนโลบ และการสร้างฟังก์ชันการตัดสินใจแบบเชิงเส้นบนระนาบของไอเกนเวคเตอร์ โดยใช้ลักษณะสำคัญได้แก่ การกระจายของจุดที่อยู่ภายในเมตริกซ์ของอักษร เพื่อใช้ในการรู้จำตัวอักษรไทยในระบบออฟ-ไลน์กับตัวอักษรรูปแบบเดียวที่มีขนาด 128 x 64 จุด ซึ่งผลของการรู้จำตัวพิมพ์อักษรมีความถูกต้อง 98.00 % (ชม กิ้มปาน, 1987)

ในปี 1990 วัลนพ ตันฤดี ได้เสนอเทคนิคของการรับรู้ลายมือเขียนอักษรไทย โดยมีขั้นตอนต่างๆ ดังนี้ ขั้นตอนที่ 1 คือการแบ่งกลุ่มของรูปแบบอย่างคร่าวๆ โดยการกำหนดขอบเขตของระดับการเขียนไว้ล่วงหน้า ขั้นตอนที่ 2 คือการหาลักษณะสำคัญของรูปแบบโดยใช้รหัสแบบลูกโซ่ของ ฟรีแมน ขั้นตอนที่ 3 คือการใช้ไดนามิคโปรแกรมมิ่งมาหาความแตกต่างกันที่น้อยที่สุดระหว่างรูปแบบ หลังจากนั้นจะได้ผลลัพธ์ที่ใกล้เคียงที่สุด ซึ่งผลของการรับรู้ลายมือเขียนมีความถูกต้อง 98.5 % (วัลนพ ตันฤดี, 1990)

ในปี 1991 Jou I. -C. ได้เสนอเทคนิคของนิวรอลเน็ตเวอร์ค (Neural Network) การรู้จำตัวอักษรนี้มี 2 ขั้นตอน ดังนี้ ขั้นตอนที่ 1 คือ การหาลักษณะสำคัญของตัวอักษร โดยใช้ลักษณะสำคัญได้แก่ เวคเตอร์ของเส้นแสดงตัวอักษร ส่วนขั้นตอนที่ 2 คือ การจับคู่เปรียบเทียบกับตัวอักษรต้นแบบ โดยการจับคู่เปรียบเทียบนี้เป็นการเปรียบเทียบเวคเตอร์ของเส้นแสดงตัวอักษร กับตัวอักษรที่ใช้เป็นตัวอักษรต้นแบบ เพื่อใช้ในการรู้จำตัวอักษรจีนในระบบออน-ไลน์ ซึ่งผลของการรู้จำตัวอักษรมีความถูกต้อง 91.00 % และมีความเร็วในการรู้จำอักษรจีน 1 ตัวอักษรในเวลา 0.25 วินาที (Jou, I. -C., 1991)

ในปี 1992 มนลดา บุญสุวรรณ ได้เสนอเทคนิคการวิเคราะห์เส้นแสดงขอบของอักษรและวิธีการเปรียบเทียบไดนามิคโปรแกรมมิ่ง โดยใช้ลักษณะสำคัญได้แก่ จำนวนหัว จำนวนส่วนโค้ง อัตราส่วนความกว้างต่อความสูง และความยาวระหว่างจุดเปลี่ยนทิศทางแต่ละจุด เพื่อใช้ในการรู้จำตัวพิมพ์อักษรไทยในระบบออฟ-ไลน์ กับตัวอักษรรูปแบบเดียวที่มีขนาด 40 x 40 จุด ซึ่งผลของการรู้จำตัวอักษรมีความถูกต้อง 94.70 % (มนลดา บุญสุวรรณ, 1992)

2.2.2 งานวิจัยที่เกี่ยวกับการอ่านออกเสียงจากแฟ้มข้อความ (Text-to-speech)[3]

ในปี 1990 อาทร นันทิยกุล ได้เสนอเทคนิคการสังเคราะห์เสียงพูดจากข้อความภาษาไทย เป็นการสร้างเสียงพูดขึ้นมาจากข้อความภาษาไทยที่ถูกป้อนเป็นอินพุทเข้าสู่ระบบ ซึ่งได้ประยุกต์ลงบนไมโครคอมพิวเตอร์และมีภาคการแปลงสัญญาณระหว่างสัญญาณอนาล็อกกับสัญญาณดิจิตอลรวมอยู่ด้วยโดยใช้หลักของการวิเคราะห์หน่วยย่อยของเสียงพูดคือ พยางค์ของเสียงต้นแบบมาทำการคัดตัวอย่างด้วยความถี่ 10 kHz และนำมาทำการวิเคราะห์ด้วยเทคนิคการทำนายแบบเชิงเส้น (LPC : Linear Prediction Coding) แบบออร์เดอร์ 10 ทีละเฟรม (เฟรมละ 200 จุดหรือ 20 มิลลิวินาที) ได้เอาท์พุทออกมาเป็นชุดพารามิเตอร์ประกอบด้วย 1) ค่าความผิดพลาดเฉลี่ย 2) ค่าคาบของพิทช์ และ 3) ค่าสัมประสิทธิ์ของการทำนาย (10 ค่า ต่อหนึ่งเฟรม) เก็บเอาไว้ในหน่วยเก็บความจำสำรองในรูปแบบของพจนานุกรมข้อมูล ซึ่งสามารถจะทำการแก้ไขพารามิเตอร์เพื่อปรับปรุงให้ได้เสียงสังเคราะห์ที่ดีขึ้น จากนั้นก็จะทำการสร้างเสียงพูดสังเคราะห์ขึ้นมา โดยนำเอาชุดพารามิเตอร์ของหน่วยย่อยของเสียงที่ได้มาจากการค้นหาในพจนานุกรมข้อมูลตามข้อความที่ป้อนเข้ามา มาทำการสังเคราะห์ผ่านตัวกรองแลททิส (Lattice filter) และภาคการแปลงสัญญาณดิจิตอลเป็นสัญญาณอนาล็อก ออกลำโพงกลับคืนมาเป็นเสียงพูด โดยคุณภาพของเสียงที่สังเคราะห์ออกมาอยู่ในระดับปานกลาง ซึ่งวัดผลได้จากการรับฟังของกลุ่มตัวอย่าง และเสียงสังเคราะห์นี้จะมีคุณสมบัติของเสียงวรรณยุกต์ ซึ่งเป็นลักษณะสำคัญของภาษาไทยด้วย

ในปี 2535 วิสุทธิ์ สุวรรณสุขโรจน์ ได้จัดทำโครงงานคอมพิวเตอร์ช่วยคนตาบอด (Computer Aid Blind) ขึ้น โดยการทำงานของระบบจะเป็นการรอรับอินพุทจากแป้นพิมพ์ เมื่อผู้ใช้กดแป้นตัวใดโปรแกรมก็จะอ่านออกเสียงของแป้นนั้นออกมาทางลำโพง โปรแกรมนี้สามารถอ่านออกเสียงได้ทั้งแป้นภาษาไทยและภาษาอังกฤษ รวมทั้งแป้นควบคุมอื่นๆ เช่น ปุ่มลูกศร, ปุ่ม Back Space, ปุ่มฟังก์ชัน F1, F2,... เป็นต้น นอกจากนั้นในระบบยังมีโปรแกรมอ่านออกเสียงจากแฟ้มข้อความ (Text File) อีกด้วย โดยจะอ่านออกมาทีละตัวอักษร

เทคนิคที่ใช้ในโครงงานนี้ คือ จะทำการบันทึกเสียงพูดของแป้นพิมพ์แต่ละตัวเป็นแฟ้มเสียงไว้ก่อน เมื่อสั่งให้ระบบเริ่มทำงาน โปรแกรมจะทำการอ่านข้อมูลเสียงทั้งหมดเข้ามาไว้ในหน่วยความจำหลัก แล้วฝังตัวเองลงไปในหน่วยความจำ (Terminal but Stay Resident) เพื่อดักรอการกดแป้นพิมพ์จากผู้ใช้ เมื่อผู้ใช้กดแป้นพิมพ์ตัวใดก็จะวิ่งไปค้นหาข้อมูลเสียงของแป้นนั้น แล้วทำการอ่านออกเสียง คุณภาพของเสียงจัดอยู่ในระดับดี เพราะเสียงที่ได้มาจากการบันทึกเสียงมนุษย์จริง [8]