บทที่ 1

บทนำ

ที่มาของปัญหาในการรู้จำเสียงพูด

Roe and Wilpon (1993) ได้แสดงความเห็นว่าภายใน 25 ปีข้างหน้านี้ การรู้จำเสียงพูดจะปรากฏให้เห็นเป็นรูปเป็นร่างมากยิ่งขึ้นในการประยุกต์ใช้งาน โดยเริ่มต้นขึ้นจากสิ่งที่เสมือนจะไร้ผลจนกระทั่งกลายเป็นสิ่งที่เป็นจริงขึ้นมาได้ในปัจจุบัน คำกล่าวข้างต้นนี้ได้แสดงให้เห็นถึงความเป็นไปได้และความสำคัญของการรู้จำเสียงพูด ที่จะเกิดขึ้นได้ในระยะเวลาอันใกล้นี้ ในปัจจุบันการรู้จำเสียงพูด (Speech Recognition) ถูกนำมาประยุกต์ใช้งานกันอย่างกว้างขวางโดยเฉพาะในระบบสื่อสารโทรคมนาคม จุดมุ่งหมายหลักของการรู้จำเสียงพูดก็คือการเพิ่มพูนความสามารถของอุปกรณ์ต่างๆ ในการรับรู้และสื่อสารโต้ตอบกับมนุษย์ได้ เพื่อเพิ่มทางเลือกในการควบคุมสั่งการอุปกรณ์เครื่องมือต่างๆ โดยเฉพาะเครื่องคอมพิวเตอร์ ซึ่งการใช้เสียงพูดควบคุมสั่งการนี้ถือได้ว่าเป็นวิธีการที่เป็นธรรมชาติมากที่สุดของมนุษย์ ศาสตร์ทางด้านการรู้จำเสียงพูดเกิดขึ้นมากว่า 40 ปีแล้ว แต่เพิ่งจะเริ่มมีการศึกษากันอย่างจริงจังในช่วงสองทศวรรษที่ผ่านมานี้เอง

อย่างไรก็ตามการวิจัยทางด้านกรรมวิธีประมวลผลสัญญาณเสียงพูดในปัจจุบันนี้ได้มีความก้าวหน้าไปมาก โดยสามารถนำเทคนิคการรู้จำเสียงพูดและการสังเคราะห์เสียงพูดไปประยุกต์ใช้ในงานหลายๆ ด้าน เช่น การให้บริการของธุรกิจธนาคาร การป้อนข้อมูล เป็นต้น อย่างไรก็ดีเทคนิคต่างๆ ที่ถูกพัฒนาขึ้นมาใช้เหล่านั้นล้วนอยู่บนพื้นฐานของหลักการออกเสียงที่ไม่ใช่เสียงภาษาไทย ถ้ามีการนำมาปรับใช้กับเสียงพูดภาษาไทยโดยตรง ด้วยเทคนิคเดียวกันย่อมมีโอกาสที่จะทำให้ผลลัพธ์ที่ได้มีประสิทธิภาพลดลง ทั้งนี้เนื่องจากลักษณะเฉพาะของแต่ละภาษาที่มีความแตกต่างกัน ดังนั้นจึงจำเป็นต้องมีการศึกษาค้นคว้า ดัดแปลง และพัฒนากรรมวิธีรู้จำเสียงพูดที่เหมาะสมกับเสียงพูดภาษาไทย เพื่อให้มีความถูกต้องสูงในการรู้จำเสียงพูดภาษาไทยอีกด้วย

การรู้จำเสียงพูดนั้นมีการวิจัยกันอย่างแพร่หลาย การศึกษาวิจัยเกี่ยวกับการรู้จำนั้นมีทั้งชนิดขึ้นกับผู้พูด (Speaker-Dependent) และชนิดไม่ขึ้นกับผู้พูด (Speaker-Independent) (Furui, 1985) โดยในกรณีการรู้จำชนิดไม่ขึ้นกับผู้พูดยังสามารถจำแนกตามเสียงพูดได้เป็น แบบคำโดด (Isolated Word) แบบคำต่อเนื่อง (Connected Word) และแบบเสียงพูดต่อเนื่อง (Continuous Speech) การรู้จำเสียงพูดนั้นมีงานวิจัยมากมายเริ่มตั้งแต่ตัวเลขภาษาอังกฤษ ตัวหนังสือภาษาอังกฤษ คำศัพท์ภาษาอังกฤษ เป็นต้น ซึ่งได้มีผู้เสนอวิธีการรู้จำหลายวิธีการด้วยกัน เช่น เทคนิค Hidden Markov Model (HMM), Dynamic Time Warping (DTW) Level Building Vocabulary Base Neural Network เป็นต้น แต่มีวิธีการหนึ่งที่รวดเร็วและมีความถูกต้องแม่นยำสูงมาก คือวิธีการเปรียบเทียบบนพื้นฐานของคำศัพท์ที่ใช้อ้างอิงโดยอาศัย

แบบจำลองฮิดเดน มาร์คอฟร่วมกับการควอนไทซ์แบบเวกเตอร์ ซึ่งก็คือการเปรียบเทียบข้อมูลเสียงพูดกับข้อมูลเสียงพูดอ้างอิงที่ได้ถูกจัดเก็บไว้ล่วงหน้า นอกจากนั้นจำนวนคำศัพท์ก็ยังเป็นตัวกำหนดความซับซ้อนของคำศัพท์ที่มีอยู่ในฐานข้อมูลคำศัพท์อีกด้วย

การค้นคว้าวิจัยในด้านการรู้จำเสียงพูดภาษาไทยที่มีในประเทศไทยนั้น ได้มีการศึกษาค้นคว้าวิจัยมาอย่างต่อเนื่องดังจะเห็นได้จากผลงานวิจัยต่างๆ เช่น การตรวจรู้เสียงพูดภาษาไทย (ทวิ ประทุมทาน, 2530) ระบบการรับรู้เสียงพูดแบบต่างบุคคล (ไพศาล ธรรมโพธิทอง, 2533) การรู้จำเสียงพูดตัวเลขเป็นภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยวิธีฮิดเดน มาร์คอฟ โมเดลและเวกเตอร์ควอนไตซ์เซชั่น (เสาวลักษณ์ อารีพงศา, 2538) การรู้จำเสียงพูดสระภาษาไทยโดดๆ ไม่ขึ้นกับผู้พูดโดยการวัดสเปกตรัมดิสแตนส์และใช้ไดนามิกไทม์วาร์ปปิ่ง (ธีระ ภัทราพรนันท์, 2538) การรู้จำเสียงพูดตัวเลขไทยโดยไม่ขึ้นกับผู้พูดโดยการใช้ไดนามิกไทม์วาร์ปปิ่ง (ระพีพัฒน์ เพ็ญศิริ, 2538) และการรู้จำคำพูดภาษาไทยโดยใช้ลักษณะบ่งความต่างของหน่วยเสียง (ณัฐกร ทับทอง, 2538) เป็นต้น โดยงานวิจัยเหล่านี้ล้วนเป็นการวิจัยเกี่ยวกับคำโดดภาษาไทยซึ่งเป็นการปูพื้นฐานไปสู่การรู้จำเสียงพูดภาษาไทยเป็นคำต่อเนื่องและเสียงพูดต่อเนื่องต่อไป ดังนั้นงานวิจัยนี้จึงทำการวิจัยเกี่ยวกับการรู้จำคำพูดภาษาไทยเป็นคำต่อเนื่อง (Thai Connected Word) ซึ่งยังไม่ได้มีการทำวิจัยมาก่อน โดยอาศัยหลักการและข้อมูลที่ได้มาจากการวิจัยเสียงพูดคำโดดภาษาไทยเป็นพื้นฐาน ซึ่งได้แนวความคิดมาจากงานวิจัยเกี่ยวกับเสียงพูดตัวเลขภาษาไทย โดยเฉพาะการนำแบบจำลองฮิดเดน มาร์คอฟมาประยุกต์ใช้การรู้จำเสียงพูดภาษาไทย (เสาวลักษณ์ อารีพงศา, 2538) เพื่อให้สามารถรู้จำเสียงพูดคำไทยหลายพยางค์ได้ด้วยการปรับเปลี่ยนโครงสร้างระบบให้สามารถรองรับความหลากหลายของข้อมูลคำศัพท์ได้

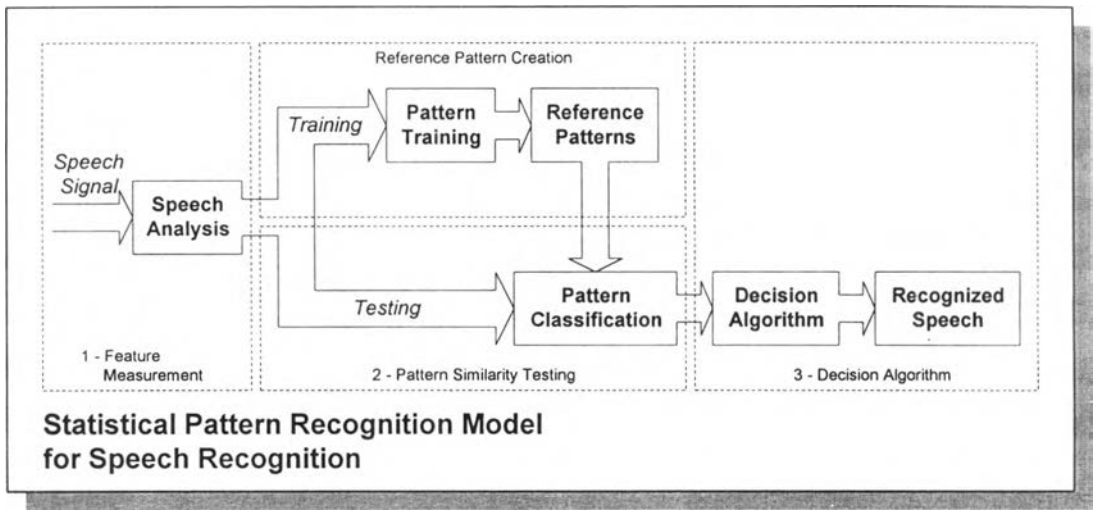
วัตถุประสงค์

1. เพื่อศึกษาถึงวิธีการรู้จำคำพูดภาษาไทยโดยมีลักษณะต่างกันที่จำนวนพยางค์
2. เพื่อศึกษาเทคนิคแบบจำลองฮิดเดน มาร์คอฟ ในการรู้จำเสียงพูดภาษาไทย
3. เพื่อพัฒนากรรมวิธีที่เหมาะสมในการรู้จำเสียงพูดภาษาไทย
4. เพื่อให้เครื่องคอมพิวเตอร์สามารถรู้จำคำพูดภาษาไทยหลายพยางค์ได้
5. เพื่อเป็นแนวทางสำหรับการศึกษาวิธีการรู้จำวิธีการอื่นๆ

แนวความคิด

ระบบการรู้จำเสียงพูดจะอาศัยการฝึกฝน (Training) เพื่อจดจำรูปแบบอ้างอิง (Reference Patterns) ไว้ใช้ในการเปรียบเทียบกับเสียงพูดที่ยังไม่ทราบรูปแบบ และใช้รูปแบบอ้างอิงเหล่านั้นในการตัดสินใจเลือกรูปแบบที่ใกล้เคียงกับเสียงพูดที่ถูกนำมาเปรียบเทียบมากที่สุด ขั้นตอนในการรู้จำเสียงพูดโดยทั่วไปแบ่งได้เป็น 3 ขั้นตอน (Rabiner and Levinson, 1981) ได้แก่

1. การวิเคราะห์และวัดค่าลักษณะสำคัญ (Feature Measurement)
2. การจำแนกรูปแบบ (Pattern Classification) หรือ
   การทดสอบความคล้ายคลึงกันของรูปแบบ (Pattern Similarity Testing)
3. ขั้นตอนวิธีการตัดสินใจ (Decision Algorithm)



รูปที่ 1.1 แบบจำลองรูปแบบการรู้จำทางสถิติที่ใช้ในการรู้จำเสียงพูด (Rabiner and Levinson, 1981)

แบบจำลองการรู้จำเสียงพูดคำต่อเนื่องบนพื้นฐานของคำศัพท์นั้น ดังแสดงในรูปที่ 1.1 เริ่มต้นจากการสร้างและเก็บลักษณะสำคัญ (Feature) จากเสียงพูดที่ยังไม่ทราบรูปแบบ แล้วนำไปเปรียบเทียบกับรูปแบบของคำศัพท์แต่ละคำที่ได้เก็บไว้แล้ว เพื่อหารูปแบบที่ใกล้เคียงกันมากที่สุดกับเสียงพูดที่เข้ามา ถ้ารูปแบบทั้งสองใกล้เคียงกันมากพอ ระบบจะตัดสินให้เป็นคำดังกล่าวทันที แต่ถ้ารูปแบบทั้งสองไม่ใกล้เคียงกัน ระบบจะไม่ตัดสินว่าเป็นคำใด แต่จะให้ผู้พูดพูดซ้ำอีกครั้งหนึ่ง วิธีการหาลักษณะสำคัญของสัญญาณเสียงเพื่อใช้ในการเปรียบเทียบนั้นมีหลายวิธีการด้วยกัน (Rabiner and Wilpon, 1979; Rabiner and Levinson, 1981; Rabiner, 1994) เช่น Digital Filter Bank, Fourier Transform, Cepstral Coefficient, Linear Prediction Coefficient, LP-Derived Filter Bank, LP-Derived Cepstral Coefficient เป็นต้น โดยการวิเคราะห์จะสามารถแบ่งได้เป็น 2 ประเภท (Rabiner, 1994) ได้แก่ การวิเคราะห์ในเชิงเวลา (Time Domain Analysis) และการวิเคราะห์ในเชิงความถี่ (Frequency Domain Analysis) ซึ่งในงานวิจัยนี้จะใช้วิธีการวิเคราะห์ในเชิงเวลาเป็นหลัก

ขั้นตอนในการเปรียบเทียบรูปแบบระหว่างรูปแบบที่ยังไม่ทราบ กับรูปแบบของคำศัพท์ที่ได้จัดเก็บไว้แล้วนั้น จัดอยู่ในขั้นตอนการจำแนกประเภทรูปแบบ (Pattern Classification) (Rabiner and Levinson, 1981; Levinson and Roe, 1990; Roe and Wilpon, 1993) ดังแสดงในรูปที่ 1.1 จากการวิจัยเกี่ยวกับ

เสียงพูดที่ผ่านมานั้น วิธีการจำแนกประเภทรูปแบบในการรู้จำเสียงพูดสามารถแบ่งได้เป็น 4 วิธี ได้แก่ การเข้าคู่ต้นแบบ (Template Matching) ระบบตามกฎเกณฑ์ (Rule-Based System) ระบบแบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Model, HMM) และเครือข่ายนิวรอล (Neural Network)

ข้อแตกต่างของวิธีการจำแนกรูปแบบทั้ง 4 ประเภทก็คือ วิธีการเข้าคู่รูปโดยส่วนใหญ่จะใช้ร่วมกับวิธีการ Dynamic Time Warping ซึ่งใช้งานได้ดีกับการรู้จำเสียงพูดเป็นคำโดด ระบบตามกฎเกณฑ์นั้นจะอาศัยเงื่อนไขในการตัดสิน ซึ่งเมื่อระบบมีขนาดใหญ่และซับซ้อนมากขึ้นจะทำให้การตัดสินใจผิดพลาดได้ง่ายตั้งแต่ขั้นตอนแรก ระบบแบบจำลองฮิดเดน มาร์คอฟนั้นเป็นวิธีการที่เป็นที่ถูกนำมาประยุกต์ใช้งานมากที่สุดในขณะนี้ ซึ่งโดยส่วนใหญ่นำมาใช้กับการรู้จำคำพูดต่อเนื่องและเสียงพูดต่อเนื่องเป็นหลัก (Rabiner and Wilpon, 1979; Rabiner and Levinson, 1981; Furui, 1985; Levinson and Roe, 1990; Roe and Wilpon, 1993; Rabiner, 1994) เนื่องจากสามารถเก็บข้อมูลและแบบจำลองทางสถิติของเสียงพูดไว้ได้มากที่สุด นอกจากนั้นวิธีการนี้ยังเป็นขั้นตอนวิธีการทั่วไปของวิธีการ Dynamic Programming อีกด้วย ส่วนระบบเครือข่ายนิวรอลนั้นเพิ่งจะเริ่มมีการนำมาประยุกต์ใช้งานเมื่อไม่นานมานี้ โดยใช้กับการรู้จำเสียงพูดต่อเนื่อง ซึ่งจะช่วยให้ระบบสามารถเรียนรู้และปรับปรุงตัวเองได้เมื่อมีรูปแบบใหม่ๆ เข้ามา

จากประเภทของการจำแนกรูปแบบข้างต้น การรู้จำคำพูดภาษาไทยบนพื้นฐานของคำศัพท์ในงานวิจัยนี้จะใช้วิธีการหาสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น (Linear Prediction Coefficient) ด้วยวิธีการวิเคราะห์อัตสหสัมพันธ์ (Autocorrelation Analysis) เพื่อจัดเก็บลักษณะสำคัญของสัญญาณเสียงพูด (Rabiner and Wilpon, 1979; Rabiner and Levinson, 1981; Furui, 1985) และใช้วิธีการควอนไทซ์แบบเวกเตอร์เพื่อลดจำนวนข้อมูลโดยการสร้างชุดรหัสเพื่อแทนกลุ่มข้อมูลลักษณะสำคัญ ส่วนวิธีฮิดเดนมาร์คอฟ โมเดลจะใช้ในการจำแนกรูปแบบเพื่อคัดเลือกคำศัพท์ที่มีความใกล้เคียงมากที่สุด ในขั้นตอนการทดสอบความคล้ายคลึงกันของรูปแบบดังแสดงในรูปที่ 1.1 โดยใช้การวัดค่าความเพี้ยนแบบค่าความผิดพลาดกำลังสองเฉลี่ย (Mean-Square Error, MSE) ส่วนในขั้นตอนวิธีการตัดสินใจนั้นจะใช้วิธีการ Viterbi Algorithm (Rabiner and Wilpon, 1979; Rabiner and Levinson, 1981; Roe and Wilpon, 1993) ในการตัดสินใจเลือกคำศัพท์ที่ถูกต้องเหมาะสมและตรงกับเสียงพูดที่เข้ามา

ปัญหาของการรู้จำเสียงพูดภาษาไทยที่ผ่านมา

1. ขั้นตอนเทคนิคการหาจุดสิ้นสุดเสียงพูด (Endpoint Detection) ขั้นตอนการหาจุดสิ้นสุดเสียงพูดจัดเป็นขั้นตอนที่สำคัญที่สุดขั้นตอนหนึ่งในการรู้จำเสียงพูด การหาจุดสิ้นสุดเสียงพูดที่เหมาะสมและมีประสิทธิภาพจะช่วยให้อัตราการรู้จำเพิ่มสูงขึ้น

2. ขนาดของชุดรหัสที่เหมาะสม จำนวนชุดรหัสที่เหมาะสมกับจำนวนคำศัพท์จะช่วยในการจดจำรูปแบบของเสียงพูดแต่ละเสียงได้อย่างละเอียดครบถ้วนซึ่งช่วยให้อัตราการรู้จำเพิ่มสูงขึ้น ดังนั้นจึงต้องมีการปรับแต่งจำนวนชุดรหัสของคำศัพท์แต่ละคำให้เหมาะสม

3. จำนวนสถานะของแบบจำลองฮิดเดน มาร์คอฟที่เหมาะสม จำนวนสถานะที่เหมาะสมกับเสียงพูดแต่ละชุดมีความสำคัญอย่างยิ่งและมีความเกี่ยวข้องโดยตรงกับความถูกต้องในการรู้จำเสียงพูด ดังนั้นการเลือกจำนวนสถานะที่เหมาะสมจะช่วยให้อัตราการรู้จำเพิ่มสูงขึ้น จึงต้องมีการปรับแต่งจำนวนสถานะของระบบให้เหมาะสมกับชุดคำศัพท์ที่ระบบจะต้องเรียนรู้ต่อไป

4. จำนวนเสียงพูดต้นแบบ (Training Set) จากการค้นคว้าวิจัยที่ผ่านมานั้นจำนวนเสียงพูดต้นแบบที่ใช้ในขั้นตอนการฝึกฝนมีจำนวนน้อยเกินไป จึงไม่ครอบคลุมการแปรเปลี่ยนทั้งหมดของเสียงพูดทำให้อัตราการรู้จำที่ได้จึงมีค่าต่ำ ดังนั้นจึงจำเป็นต้องมีการเพิ่มจำนวนเสียงพูดให้มากขึ้นซึ่งจะช่วยให้อัตราการรู้จำเพิ่มสูงขึ้นด้วย

เป้าหมายและขอบเขตของงานวิจัย

1. สามารถรู้จำคำพูดต่อเนื่องภาษาไทยได้ตรงตามชุดคำศัพท์ที่กำหนดไว้
2. อัตราความแม่นยำในการรู้จำมากกว่า 85% ขึ้นไป

ขั้นตอนและวิธีการดำเนินการ

1. ศึกษาคุณลักษณะของเสียงพูด และแบบจำลองการเปล่งเสียงพูด
2. ค้นคว้าและเก็บรวบรวมข้อมูลรายละเอียดที่เกี่ยวข้องกับวิธีการดังต่อไปนี้
   1) การประมวลผลสัญญาณเบื้องต้น (Signal Preprocessing)
   2) การวิเคราะห์และวัดค่าลักษณะสำคัญ (Feature Measurement)
   3) การทดสอบความคล้ายคลึงกันของรูปแบบ (Pattern Similarity Testing)
   4) ขั้นตอนวิธีการตัดสินใจ (Decision Algorithm)
3. เก็บข้อมูลเสียงพูดคำศัพท์ภาษาไทยของกลุ่มตัวอย่าง ได้แก่

   1) กลุ่มตัวอย่างเสียงพูดเพื่อฝึกฝน (Training Group) สำหรับใช้ในการฝึกฝนระบบเพื่อสร้างรูปแบบอ้างอิงของฐานข้อมูลคำศัพท์

   2) กลุ่มตัวอย่างเสียงพูดเพื่อทดสอบ (Testing Group) สำหรับใช้ในการทดสอบรูปแบบอ้างอิงที่สร้างขึ้นมาจากกลุ่มตัวอย่างเสียงพูดเพื่อฝึกฝน

4. วิเคราะห์และพัฒนาโปรแกรมในแต่ละส่วน
5. ทำการฝึกฝนพร้อมทั้งจำแนกกลุ่มคำเพื่อสร้างรูปแบบคำพูดอ้างอิงจากเสียงพูดของกลุ่มตัวอย่าง
6. ทำการทดสอบอัตราความแม่นยำในการรู้จำเสียงพูด โดยใช้เสียงพูดของทั้งสามกลุ่ม
7. ปรับปรุงแก้ไขโปรแกรมหรือเพิ่มจำนวนการฝึกฝนเพื่อเพิ่มอัตราความถูกต้องในการรู้จำ
8. สรุปรวบรวมผลการวิจัยทั้งหมด พร้อมทั้งจัดทำเอกสารเกี่ยวกับวิทยานิพนธ์

ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1. ทราบถึงวิธีการรู้จำคำพูดภาษาไทยและลักษณะทางภาษาศาสตร์ของคำพูดภาษาไทย
2. ทราบถึงกรรมวิธีที่เหมาะสมในการรู้จำเสียงพูดภาษาไทย
3. เป็นการเพิ่มเติมความสามารถให้แก่เครื่องคอมพิวเตอร์ในการรู้จำคำพูดภาษาไทย
4. เพื่อเป็นประโยชน์แก่ผู้พิการทางสายตาและผู้ที่ไม่สามารถช่วยตนเองได้ ให้สามารถใช้งานเครื่องมือและอุปกรณ์ต่างๆ ด้วยการสั่งงานเป็นเสียงพูดได้