บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในการดำเนินชีวิตประจำวันของมนุษย์นั้นมีการติดต่อสื่อสารกับคอมพิวเตอร์อยู่ตลอด อาจจะโดยการติดต่อด้วยมือผ่านอุปกรณ์นำเข้า (Input Devices) ที่ปัจจุบันได้มีการพัฒนาขึ้นมาหลายชนิด เช่น คีย์บอร์ด หรือ เมาส์ หรือ ปุ่มที่มีอยู่หลากหลายบนแป้นโทรศัพท์ แต่การติดต่อที่เป็นธรรมชาติมากที่สุดคือการใช้เสียงพูด การติดต่อกับคอมพิวเตอร์ด้วยเสียงพูดนั้นช่วยอำนวยความสะดวกในการใช้งานคอมพิวเตอร์ให้กับผู้ใช้ต่าง ๆ ได้ เช่น ผู้สูงอายุที่ไม่มีความชำนาญในการใช้อุปกรณ์นำเข้าที่มีอยู่มากมายหลายชนิด และ ผู้ที่มีความทุพพลภาพทางร่างกาย ได้แก่ ผู้พิการทางมือและผู้พิการทางตา

การใช้เสียงพูดเพื่อติดต่อกับคอมพิวเตอร์มีส่วนเกี่ยวข้องกับระบบรู้จำเสียงพูด (Speech Recognition) และการสังเคราะห์เสียงพูด (Speech Synthesis) การรู้จำเสียงพูด เป็นกระบวนการในการเปลี่ยนสัญญาณเสียงที่อยู่ในรูปของคลื่นให้อยู่ในรูปของข้อความที่คอมพิวเตอร์สามารถประมวลผลได้ โดยการรู้จำเสียงอาจจะมีขอบเขตจากการรู้จำเพียงแค่คำสั่งง่าย ๆ ที่สามารถเข้าใจได้ไปสู่การรู้จำข้อมูลทั้งหมดในสัญญาณเสียงพูด เช่น รู้จำคำทุกคำ การหาความหมาย และรู้จำเสียงพูดที่มีอารมณ์ ส่วนการสังเคราะห์เสียง เป็นกระบวนการที่ทำงานในตรงกันข้าม คือเปลี่ยนข้อความให้อยู่ในรูปของสัญญาณเสียงที่มนุษย์เข้าใจได้ ในช่วงเวลาที่ผ่านมา การรู้จำเสียงถูกนำไปประยุกต์กับงานในชีวิตประจำวันอย่างมากมาย อาทิเช่น ระบบช่วยแปล (Machine Translation) ระบบถอดข้อความของข่าว (News Transcription) เป็นต้น แต่อย่างไรก็ตาม ในการรู้จำเสียงนั้น เสียงทั้งหมดที่เปล่งออกมาจะถูกรู้จำออกมาทั้งหมด ซึ่งรวมถึงเสียงที่ไม่ได้ตั้งใจเปล่งออกมา หรือ เสียงที่เป็นคำทั่ว ๆ ไป ที่ไม่มีนัยสำคัญ

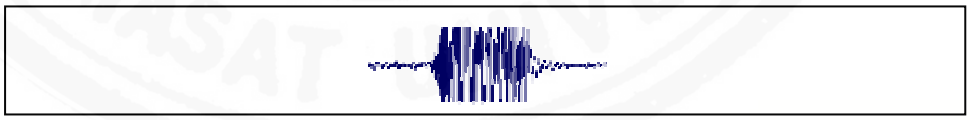
การค้นหาคำหลัก (Keyword Spotting) เป็นแขนงการวิจัยที่ได้พัฒนามาจากการรู้จำเสียง โดยสนใจที่จะรู้จำเฉพาะคำที่เป็นคำหลักเท่านั้น ซึ่งเหมาะกับงานที่ต้องการความรวดเร็วและงานสนใจในเรื่องการเข้าใจความหมาย การค้นหาคำหลัก (Keyword Spotting) จะมีความแตกต่างจากการรู้จำเสียงตรงที่ ในการสร้างโมเดลหรือแบบจำลองทางเสียงเพื่อการรู้จำนั้น อาจจะมีแบบจำลองของเสียงอยู่ 2 ส่วนด้วยกัน นั่นคือ แบบจำลองของคำหลักและแบบจำลองของคำอื่น เช่น คำทั่ว ๆ ไป คำอุทาน เสียงเงียบ เสียงกระแอม เป็นต้น

สำหรับปัญหาของระบบค้นหาคำหลักในปัจจุบัน คือ ระบบการค้นหาคำหลัก จะต้องมีการฝึกฝนคำหลักใหม่เพิ่มเติมเพื่อให้ระบบรู้จำคำหลักใหม่ที่เกิดใหม่ ซึ่งเป็นการสิ้นเปลื้องเวลาในการฝึกฝน

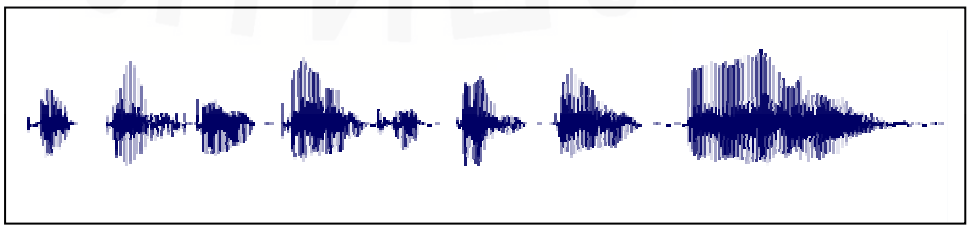
ในการวัดประสิทธิภาพของการค้นหาคำหลัก (Keyword Spotting) ในเสียงพูดนั้น จะมีเกี่ยวข้องกับปัจจัยต่าง ๆ ได้แก่

1. ลักษณะของสัญญาณเสียงพูด ว่าเป็นเสียงพูดแบบคำโดดหรือเป็นเสียงพูดแบบต่อเนื่อง การรู้จำเสียงพูดแบบคำโดด หรือเสียงพูดแบบต่อเนื่องนั้น มีความยากง่ายที่แตกต่างกัน โดยการออกเสียงพูดแบบคำโดด คือการออกเสียงคำหลักที่ต้องการเป็นคำ แล้วนำไปให้คอมพิวเตอร์ทำการรู้จำ ส่วนเสียงพูดแบบต่อเนื่องคือเสียงพูดปกติ จึงทำให้เสียงที่ได้นั้นมีทั้งคำหลัก (Keyword) และคำพูดที่ไม่ใช่คำหลัก (Non-Keyword) ดังภาพที่ 1.1 และภาพที่ 1.2 จะเห็นได้ว่าลักษณะของเสียงพูดแบบคำโดด จะมีขอบเขตของคำที่ชัดเจน ส่วนลักษณะของเสียงพูดแบบต่อเนื่องจะมีเสียงของคำหลักและคำที่ไม่ใช่คำหลักประสมกันอยู่ ด้วยเหตุนี้จึงทำให้ระบบค้นหาคำหลัก (Keyword Spotting) ในเสียงพูดแบบต่อเนื่อง จะต้องสามารถค้นหาคำหลักในเสียงพูดแบบต่อเนื่องให้ได้ ซึ่งในเสียงพูดที่นำมาค้นหาคำหลักนั้นอาจจะมีคำหลักมากกว่า 1 คำ หรือไม่มีคำหลักอยู่เลย

ภาพที่ 1.1
ลักษณะเสียงพูดแบบคำโดด

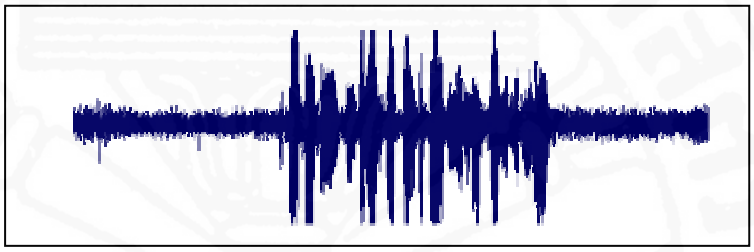


ภาพที่ 1.2
ลักษณะเสียงพูดแบบต่อเนื่อง



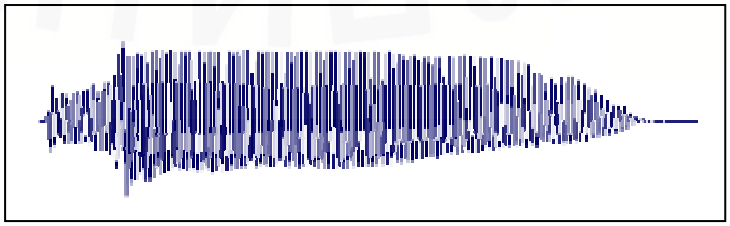
2. สัญญาณเสียงพูดที่จะนำมาค้นหาคำหลัก ถ้าคุณภาพเสียงที่ได้ไม่ดี เช่น มีเสียงรบกวนจากเครื่องปรับอากาศ เสียงรบกวนจากรถยนต์ เข้ามาปนอยู่ในสัญญาณเสียง ดังภาพที่ 1.3 ก็จะทำให้ผลลัพธ์ในการค้นหานั้นได้ผลที่ไม่ดี

ภาพที่ 1.3
ลักษณะสัญญาณเสียงพูดที่มีเสียงรบกวน



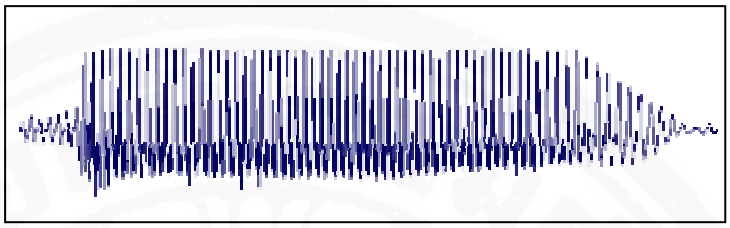
3. ผู้พูดก็เป็นอีกปัจจัยหนึ่ง เนื่องจากเสียงที่ได้จะมีคุณสมบัติเฉพาะตามผู้พูดแตกต่างกันตามวัยและเพศ เป็นต้น ดังรูปที่ 1.4 และ รูปที่ 1.5 แสดงให้เห็นลักษณะของเสียงผู้ชายและเสียงผู้หญิงที่แตกต่างกัน ถึงแม้ว่าการออกเสียงจะเป็นคำคำเดียวและบันทึกเสียงในสภาวะแวดล้อมเดียวกัน ดังนั้นการเรียนรู้เสียงจากผู้พูดเพียงคนเดียว หรือการเรียนรู้จากเสียงของผู้พูดหลายคน จึงทำให้มีความยุ่งยากซับซ้อนในการค้นหาคำหลักต่างกัน โดยการเรียนรู้เสียงแบบไม่ขึ้นอยู่กับผู้พูดจำเป็นต้องใช้ตัวอย่างเสียงมาเรียนรู้มากกว่าการเรียนรู้เสียงจากผู้พูดเพียงคนเดียว เพื่อให้ระบบสามารถสกัดค่าลักษณะสำคัญ (Feature Extraction) ของเสียงมาเรียนรู้

ภาพที่ 1.4
ลักษณะสัญญาณเสียงของผู้พูดเพศหญิง เมื่อออกเสียง “บอ”



ภาพที่ 1.5

ลักษณะสัญญาณเสียงของผู้พูดเพศชาย เมื่อออกเสียง ”บอ”



4. โดเมน (Domain) หรือ ขอบเขตของเรื่อง เนื่องจากคำหลักในโดเมนหนึ่งอาจจะไม่ใช่คำหลักในอีกโดเมนก็ได้ เช่น คำว่า “พืชสวน” เป็นคำหลักในโดเมน (Domain) ของ “งานมหกรรมพืชสวนโลก” แต่ไม่ใช่คำหลักในโดเมน (Domain) ของ “มหาวิทยาลัยธรรมศาสตร์”

ดังนั้นงานวิจัยในวิทยานิพนธ์ฉบับนี้จะทำการวิจัยเรื่องการค้นหาคำหลักในสัญญาณเสียง โดยปราศจากเสียงรบกวนต่าง ๆ และสกัดคำหลักจากเสียงพูดแบบต่อเนื่องโดยไม่ขึ้นกับผู้พูด นอกจากนี้ขอเสนอความสามารถในการดึงคำหลักจากเสียงพูดที่พูดในขอบเขตเรื่องต่าง ๆ กันได้ แต่อย่างไรก็ตามจะใช้ตัวอย่างของเสียงที่ได้จากสภาพแวดล้อมที่ปราศจากเสียงรบกวน

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อเสนอวิธีการสร้างโมเดลคำหลักแบบอัตโนมัติ ซึ่งจะเป็นการลดขั้นตอนในการฝึกอบรมเสียงคำหลัก เพื่อให้ระบบรู้จำเพิ่มเติมใหม่ทุกครั้ง ในกรณีที่ระบบไม่รู้จักคำหลักที่ต้องการค้นหา
2. เพื่อทดสอบและวัดประสิทธิภาพในการค้นหาคำหลักจากโมเดลคำหลักที่สร้างขึ้น

1.3 ขอบเขตของงานวิจัย

ในการค้นหาคำหลักจากโมเดลคำหลักแบบอัตโนมัติ แบ่งออกเป็น

1. ขั้นตอนในการจัดเตรียมแฟ้มเสียงที่นำมาใช้ในวิทยานิพนธ์นี้ แบ่งออกเป็นบันทึกแฟ้มเสียงโดยผู้วิจัยเอง และแฟ้มเสียงที่นำมาจากคลังข้อมูลเสียง

1.1 ทำการบันทึกแฟ้มเสียง บันทึกการออกเสียงพยัญชนะและสระ โดยทำการคัดเลือกบุคคลที่มีช่วงอายุระหว่าง 25-40 ปี ทั้งเพศชาย และหญิง ที่พูดอย่างชัดถ้อยชัดคำ สภาพร่างกายของผู้พูดต้องอยู่ในสภาวะปกติ ไม่เป็นโรคที่มีผลต่อเสียงพูดเช่น หวัด และทำการบันทึกเสียงในภาวะแสดล้อมที่ปราศจากเสียงรบกวน เพื่อนำเสียงที่บันทึกมาสร้างโมเดลเสียง (Acoustic Model)

1.2 ทำการบันทึกเสียงการลงทะเบียนวิชาของนักศึกษาระดับปริญญาตรี ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยธรรมศาสตร์ ทั้งเพศชายและหญิงรวม 40 คน

1.3 นำแฟ้มเสียงจากคลังข้อมูลเสียงจากศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ได้แก่ คลังเสียง LOTUS speech corpus และ คลังเสียง NECTEC Hotel Reservation Data

1.4 คุณสมบัติของแฟ้มเสียง แฟ้มเสียงที่ใช้ในการวิจัยนี้ ใช้แฟ้มเสียงที่ทำการบันทึกในรูปแบบของ PCMwav โดยแต่ละแฟ้มเสียงจะมีค่าความถี่ที่ 16 กิโลเฮิร์ต 16 บิต โมโน

1.5 จัดกลุ่มของแฟ้มเสียงตามโดเมนที่กำหนด ดังนี้ “ระบบลงทะเบียนนักศึกษามหาวิทยาลัยธรรมศาสตร์” “LOTUS speech corpus” “ระบบจองห้องพักในโรงแรม”

2. ขั้นตอนการสร้างโมเดลเสียง (Acoustic Model) จะนำแฟ้มเสียงในแต่ละโดเมนปริมาณ 2 ใน 3 มาสร้างโมเดลเสียง
3. ขั้นตอนการหาผลการทดลอง นำแฟ้มเสียงที่เหลือ (จำนวน 1 ใน 3) ในแต่ละโดเมน มาหาผลการทดลองจากโมเดลคำหลักที่สร้างขึ้น ผลลัพธ์ที่ได้จะได้รายการคำหลัก (Keyword List) ตามฐานความรู้คำหลักที่กำหนดไว้
4. ทำการพัฒนาโปรแกรมเพื่อทดลองสร้างโมเดลคำหลักแบบอัตโนมัติ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้วิธีการสร้างโมเดลคำหลักอัตโนมัติ ซึ่งสามารถนำไปประยุกต์ใช้สำหรับการค้นหาแฟ้มเสียงว่าแฟ้มเสียง แฟ้มใดบ้างที่มีคำหลักตามที่กำหนดปรากฏอยู่

2. สามารถนำงานวิจัยไปประยุกต์ใช้งานกับระบบต่าง ๆ เช่น ระบบค้นหา (Search Engine) การสั่งงานหุ่นยนต์ (Order Robot) และการสรุปคำหลัก (Keyword Summarization) เป็นต้น
3. สามารถนำผลลัพธ์ที่ได้ไปพัฒนาเป็นงานวิจัยต่อไป เช่น นำรายการคำหลัก(Keyword List) ที่ได้ ไปสอนหุ่นยนต์ให้รู้จำ เพื่อให้สามารถทำการโต้ตอบกับมนุษย์ เป็นต้น

1.5 รายละเอียดของวิทยานิพนธ์

เนื้อหาของวิทยานิพนธ์ฉบับนี้ ถูกแบ่งออกได้เป็น 5 บท โดยบทที่ 1 เป็นบทนำโดยมีรายละเอียดดังที่เสนอไปแล้วนั้น บทที่ 2 เป็นส่วนของทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง โดยอธิบายทฤษฎีที่จะนำมาใช้ และสรุปข้อดีข้อเสียของงานวิจัยที่เกี่ยวข้อง บทที่ 3 จะอธิบายถึงวิธีการดำเนินงานวิจัย ว่ามีองค์ประกอบใดบ้าง และมีเครื่องมือใดที่จำเป็นต้องใช้ ส่วนบทที่ 4 คือการทดลองและผลการทดลอง และสุดท้ายคือบทที่ 5 จะเป็นส่วนสรุปผลการศึกษาและนำเสนอข้อเสนอแนะ