บทคัดย่อ

งานวิจัยนี้นำเสนอภาพรวมของระบบการสืบค้นข้อมูลด้วยเสียงสำหรับข่าวภาษาไทย โดยนำฐานข้อมูลเสียงข่าวภาษาไทยของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) มาใช้ในการสร้างระบบที่ใช้เป็นการรู้จำเสียงพูดต่อเนื่องที่ใช้คำศัพท์จำนวนมากสำหรับข่าว (Broadcast News - Large vocabulary continuous speech recognition: BN-LVCSR) สำหรับข้อมูลจากไฟล์เสียงเป็นคำสำคัญ (Keyword) จากนั้นนำคำสำคัญที่ได้นี้ไปค้นหาข้อมูลข่าวที่มีการถอดความแล้ว ผลลัพธ์ที่ได้จะเป็นไฟล์เสียงที่มีคำสำคัญนั้น และทำการปรับปรุงชุดข้อมูลฝึกฝนจนได้แบบจำลองที่เหมาะสมกับชุดข้อมูลการทดสอบ คือใช้ LOTUS Corpus ข้อมูลเสียงที่อยู่ในรูปแบบการพูดโดยอ่านจากหนังสือพิมพ์สำหรับฝึกฝนแบบจำลองหน่วยเสียง และคลังข้อมูลข่าวภาษาไทย (Thai Broadcast News) ใช้ข้อมูลที่ถอดความจากรายการข่าวภาษาไทยสำหรับฝึกฝนแบบจำลองภาษา แบบจำลองที่เหมาะสมนั้นได้เปอร์เซ็นต์ความถูกต้องเท่ากับ 78.32% และเปอร์เซ็นต์ความแม่นยำเท่ากับ 74.34% อยู่ในเกณฑ์ที่น่าพึงพอใจ และใช้วิธีการ 10-best และคำสำคัญย่อย (Sub keyword) มาช่วยในการค้นหาคำสำคัญซึ่งทำให้ประสิทธิภาพในการค้นหาดียิ่งขึ้น