TNI Journal of Engineering and Technology
Vol.3 No.1 January-June 2015

การรู้จำเสียงพูดภาษาไทยในสภาพแวดล้อมเสียงรบกวนโดยใช้ PocketSphinx : กรณีศึกษาการนับสินค้าคงคลัง

An Experiment of Thai Speech Recognition in a Noisy Environment using PocketSphinx : A Case of Inventory Counting

อุดม ไได้พร้อม[1], วีรวุฒิ ทัพหิกรรม[2], ดัชกรณ์ ตันเจริญ[3]
คณะวิศวกรรมศาสตร์และเทคโนโลยี, สถาบันการจัดการปัญญาภิวัฒน์
85/1 หมู่ 2 ถ.แจ้งวัฒนะ ต.บางตลาด อ.ปากเกร็ด จังหวัด นนทบุรี 11120
[1] u.daiphorm@gmail.com
[2] weerawuttha@pim.ac.th
[3] datchakorntan@pim.ac.th

บทคัดย่อ

ปัจจุบันเทคโนโลยีการรู้จำเสียงพูด (Speech Recognition) ได้เข้ามามีบทบาทในการดำเนินชีวิตเป็นอย่างมาก โดยประโยชน์ของการใช้เทคโนโลยีการรู้จำเสียงพูดก่อให้เกิดความสะดวกสบาย เช่น การใช้คำสั่งเสียง (Voice Command) ในการสั่งเปิดหรือปิดอุปกรณ์ต่างๆ หรือการใช้โทรศัพท์มือสั่งการเพื่อเข้าสู่โปรแกรมต่างๆ นอกจากนั้นเทคโนโลยีนี้ยังมีความสำคัญต่อคนพิการทางสายตาเป็นอย่างมาก เนื่องจากบุคคลเหล่านี้ไม่สามารถมองเห็นได้ จึงจำเป็นต้องใช้คำสั่งเสียง รวมทั้งคนปกติก็ได้รับประโยชน์เช่นเดียวกัน คือ การดำเนินกิจกรรมต่างๆ ได้ด้วยความสะดวกรวดเร็ว

โครงงานนี้มีวัตถุประสงค์เพื่อประยุกต์เทคโนโลยีการรู้จำเสียงพูดโดยนำมาใช้ในการนับของเพื่อช่วยอำนวยความสะดวกรวดเร็ว โดยไม่ต้องใช้มือในการจดบันทึกสินค้าและจำนวนลงบนกระดาษหรือป้อนข้อมูลด้วยการพิมพ์ ซึ่งโปรแกรมที่พัฒนาจะทำงานบนโทรศัพท์มือถือในระบบปฏิบัติการแอนดรอยด์ ซึ่งเป็นระบบปฏิบัติการที่ใช้กับโทรศัพท์มือถือที่มีจำนวนผู้ใช้มากที่สุดในปัจจุบัน ซึ่งจากผลการทดลองในสภาพแวดล้อมต่างๆ หลังการปรับโมเดลเสียง มีอัตราความผิดพลาดในการรู้จำเสียงพูดลดลง

คำสำคัญ: การรู้จำเสียงพูด, สภาพแวดล้อมเสียงรบกวน

Abstract

Recently, speech recognition technologies play an increasingly important role in everyday life. Speech recognition applications have become very popular and increased convenience of using Smartphone such as voice commands, turn on/off hardware and access the different apps and features on Smartphone. Speech recognition technology can be very helpful for users who are blind or visually impaired because operating Smartphone with difficulties seeing is complicated task. In addition, general Smartphone users can also benefit from this technology such as the rapidity and ease of operation.

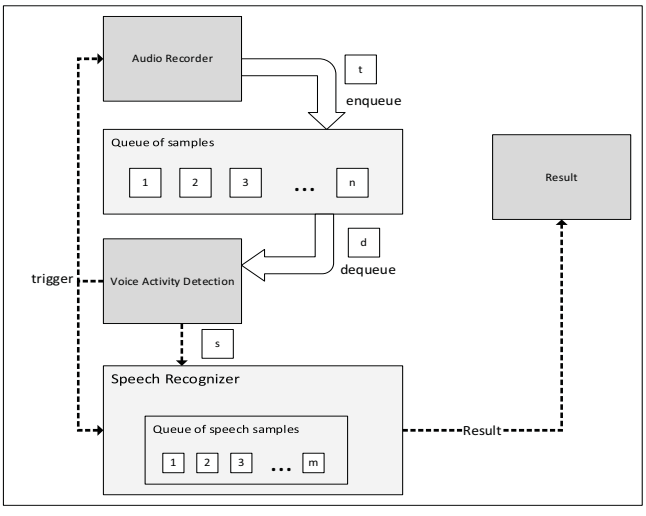
The purpose of developing this project is to apply technology of speech recognition to items counting system. Speech recognition not only makes it easier and quicker to use Smartphone, it also allows hands-free use in situations where our hands are busy such as counting stock items in retail. Speech recognition can improve the working speed by counting items and memorizing quantities without any paper note or typing on devices.

Goal of this project is to develop items counting system by speech that can run on smart phones with the Android operating system, which has increasingly number of users. Furthermore, this system can operate under noisy environment such as office and convenience store. The simulation results show that improvement of speech recognition model can reduce error rate.

Keywords: Speech Recognition, Noisy Environment

1. บทนำ

ปัจจุบันร้านสะดวกซื้อหลายแห่งมีสินค้าสำหรับจำหน่ายเป็นจำนวนมาก เนื่องจากความต้องการของลูกค้าในการซื้อสินค้าต่างกัน ส่งผลให้ร้านสะดวกซื้อจำเป็นต้องมีสินค้าเป็นจำนวนมากเพื่อรองรับความต้องการของลูกค้า ในขณะเดียวกันผลกระทบที่เกิดจากการมีจำนวนสินค้ามาก คือ ความยุ่งยากสำหรับการตรวจนับสินค้า ผู้จัดทำขอยกตัวอย่างเป็น ร้านสะดวกซื้อเซเว่น อีเลฟเว่น ซึ่งเป็นร้านสะดวกซื้อขนาดใหญ่ ที่มีสาขามากกว่า 8,000 สาขาทั่วประเทศไทย สำหรับการตรวจนับสินค้าภายในร้านเซเว่น อีเลฟเว่น จะมีการตรวจนับ 1 ครั้ง ต่อ 1 เดือน โดยจะมีพนักงานสำหรับตรวจนับสินค้าโดยเฉพาะ พนักงานตรวจนับสินค้าจะมีเครื่องคอมพิวเตอร์พกพาเป็นเครื่องมือสำหรับบันทึกการตรวจนับ ในขั้นตอนการบันทึกการตรวจนับสินค้า พนักงานจะทำการสแกนรหัสแท่งของสินค้า แล้วกรอกจำนวนโดยการกดปุ่มบนเครื่องคอมพิวเตอร์พกพาเพื่อบันทึกจำนวนสินค้า ตามกระบวนการทำงานของพนักงานตรวจนับสินค้านี้ จะมีปัญหาในการตรวจนับสินค้าที่มีความล่าช้า เนื่องจากการกดปุ่มตัวเลข หากกดตัวเลขเร็วขึ้น ก็มีโอกาสในการใส่ข้อมูลผิดพลาดมากขึ้นเช่นกัน



รูปที่ 1 ระบบการทำงานส่วนการรู้จำเสียงพูด

ดังนั้นงานวิจัยนี้ จึงพัฒนาระบบนับของด้วยเสียงภายใต้สภาพแวดล้อมเสียงรบกวน เพื่อแก้ปัญหาดังกล่าว โดยระบบนับของด้วยเสียงจะกรอกจำนวนของสินค้าโดยการพูดตัวเลข ทำให้เกิดความรวดเร็วในการนับสินค้า เปรียบเสมือนการมีพนักงาน 2 คน ทั้งคนที่นับสินค้า และคนที่จดบันทึกจำนวน ซึ่งเพิ่มประสิทธิภาพกับร้านสะดวกซื้อในการตรวจนับสินค้าได้

2. วัตถุประสงค์ของการวิจัย

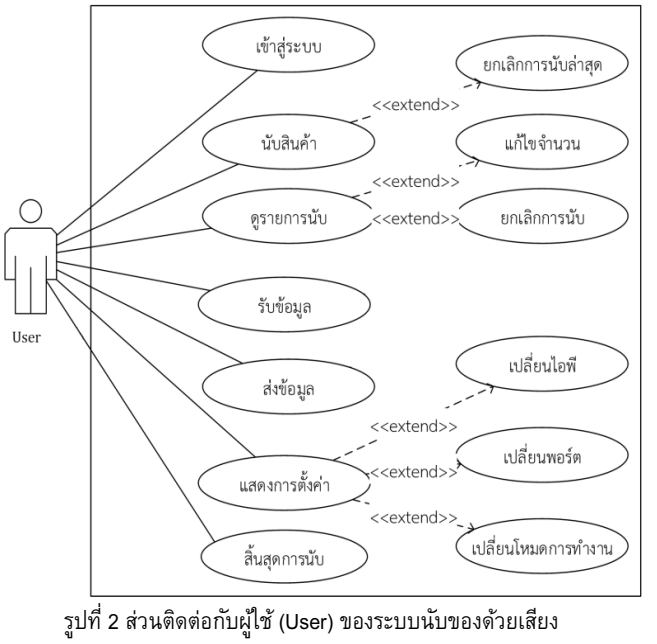
1. เพื่อพัฒนาโปรแกรมสำหรับบันทึกการตรวจนับของด้วยเสียงพูดบนโทรศัพท์มือถือระบบปฏิบัติการแอนดรอยด์
2. เพื่อความสะดวกรวดเร็วในการตรวจนับของภายในร้านสะดวกซื้อ
3. เพื่อเพิ่มศักยภาพในการทำงานของพนักงานตรวจนับของ
4. เพื่อการศึกษาการรู้จำเสียงพูด

3. วิธีการดำเนินการวิจัย

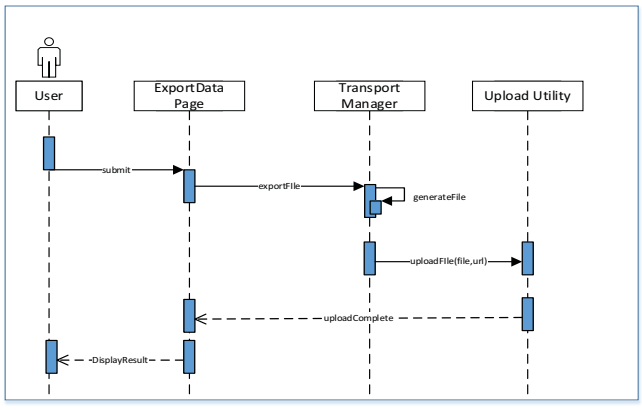
ในการพัฒนาโครงงานระบบนับของด้วยเสียง เป็นการประยุกต์ใช้เทคโนโลยีรู้จำเสียงพูดให้เกิดประโยชน์ โดยการพัฒนาส่วนของการรู้จำเสียงพูดได้ใช้ PocketSphinx [1] เป็นเครื่องมือในการพัฒนา ซึ่งเป็นโปรแกรม Open Source และมี Library สำหรับระบบปฏิบัติการแอนดรอยด์ ซึ่งเหมาะสำหรับการนำมาใช้เป็นเครื่องมือในการพัฒนาระบบสำหรับการรู้จำเสียงพูด [2] สิ่งสำคัญที่สุด คือ โมเดลเสียง ในการฝึกฝนโมเดลเสียงจะต้องใช้ข้อมูลเสียงเป็นจำนวนมากเพื่อประสิทธิภาพที่ดีของโมเดลเสียงและการรู้จำเสียงพูด โดยการฝึกฝนโมเดลเสียงให้มีประสิทธิภาพสูงจะต้องใช้เสียงที่บันทึกจากสภาพแวดล้อมจริง และใช้วิธีการปรับโมเดลเสียงซึ่งรวบรวมข้อมูลเสียงจากการบันทึกเสียงในร้านสะดวกซื้อแล้วนำมาใช้ปรับโมเดลเสียง ส่วนข้อมูลเสียงที่นำมาฝึกฝนโมเดลใช้ฐานข้อมูลเสียง LOTUS (Large vOcabulary Thai continoUos Speech recognition Corpus) [3] ซึ่งเป็นฐานข้อมูลเสียงขนาดใหญ่ที่สร้างขึ้นโดยศูนย์อิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ซึ่งเป็นฐานข้อมูลเสียงที่บันทึกจากสภาพแวดล้อมที่เงียบ เหมาะที่จะนำมาฝึกฝนและเป็นพื้นฐานในการปรับโมเดลเสียง [4] ซึ่งจะเน้นที่การปรับโมเดลเสียงเพื่อการนับสินค้าเป็นหลัก

สำหรับการบันทึกเสียงในร้านสะดวกซื้อเพื่อการนับสินค้า จะทำการบันทึกจากผู้พูดจำนวน 15 คน โดยแบ่งเป็นเพศชาย 8 คน และเพศหญิง 7 คน ซึ่งแต่ละคนจะพูดคนละ 40 ประโยค โดยใช้สำหรับปรับโมเดลเสียงโดยเฉพาะ ซึ่งจะแตกต่างกับคำที่ใช้ในส่วนผลวิจัยในหัวข้อถัดไป ซึ่งประโยคสำหรับปรับโมเดลเสียงจะทำการบันทึกเสียงผ่านโทรศัพท์มือถือระบบปฏิบัติการแอนดรอยด์ เมื่อได้ทำการรวบรวมข้อมูลเรียบร้อยแล้ว จึงทำการฝึกฝนโมเดลเสียงและปรับโมเดลเสียง 3 แบบ ประกอบด้วย เสียงแบบไม่แบ่งเพศและผู้พูด เสียงแบบแบ่งตามเพศ และเสียงแบบแบ่งตามผู้พูด

การพัฒนาระบบจะแบ่งออกเป็น 2 ส่วนหลัก ได้แก่ ส่วนรู้จำเสียงพูดและส่วนติดต่อกับผู้ใช้ ซึ่งมีรายละเอียดประกอบด้วย การพัฒนาส่วนรู้จำเสียงพูดด้วยเครื่องมือ PocketSphinx ซึ่งสามารถทำงานแบบ Real Time และมีการรู้จำเสียงพูดที่แม่นยำ โดยมีกระบวนการกรองเสียงพูดก่อนนำไปประมวลผล เรียกว่าการตรวจจับพฤติกรรมเสียงพูดโดยมีการตรวจจับเสียงว่าเป็นเสียงคนพูดหรือเสียงรบกวน [5] ซึ่งกระบวนการนี้สามารถตรวจจับการเริ่มพูดและการหยุดพูดได้ ส่งผลให้สามารถทราบถึงผลลัพธ์ของการรู้จำเสียงพูดโดยอัตโนมัติ [6] การทำงานในส่วนของการรู้จำเสียงพูดแสดงดังรูปที่ 1 ซึ่งระบบโดยภาพรวมจะประกอบด้วยตัวบันทึกเสียง (Audio Recorder) ซึ่งเสียงจะถูกนำไปเรียงและคัดเฉพาะส่วนที่มีเสียงพูดโดยตัดส่วนที่เงียบออกเพื่อตรวจจับการเริ่มพูดและการหยุด (Queue of Samples, Voice Activity Detection) หลังจากได้เสียงพูดแล้วจึงนำเข้าสู่กระบวนการรู้จำเสียงพูดต่อไป (Speech Recognizer)



รูปที่ 2 ส่วนติดต่อกับผู้ใช้ (User) ของระบบนับของด้วยเสียง



รูปที่ 3 ลำดับการทำงานของการส่งข้อมูล

เมื่อทำการสร้างโมเดลในส่วนการรู้จำเสียงพูดสำเร็จแล้ว [7] ขั้นตอนต่อไปจะเป็นการวิเคราะห์และออกแบบระบบอีกส่วนหนึ่ง คือ การพัฒนาส่วนติดต่อกับผู้ใช้ ผู้จัดทำจะทำการวิเคราะห์ภาพรวมของกระบวนการทำงานโดยวิเคราะห์จากฟังก์ชันการทำงานต่าง ๆ ที่ จำเป็นในการใช้งานระบบ ซึ่งจะใช้ Use case Diagram ดังรูปที่ 2 ในการอธิบายเพื่อให้เกิดความเข้าใจง่ายมากขึ้น โดยขั้นตอนแรกในการใช้งานต้องทำการเข้าสู่ระบบ เพื่อใช้งานฟังก์ชันอื่นๆเช่น การนับสินค้าและดูรายการนับ ซึ่งสามารถแก้ไขหรือยกเลิกการนับได้ หรือการรับส่งข้อมูลการนับสินค้าไปยังเครื่องเซิร์ฟเวอร์ ซึ่งแสดงตัวอย่างดังรูปที่ 3 การทำงานของการส่งข้อมูลได้ออกแบบไว้ในกรณีที่ผู้ใช้ไม่สามารถเชื่อมต่อกับเซิร์ฟเวอร์ได้ เมื่อผู้ใช้เชื่อมต่อเซิร์ฟเวอร์ได้ตามปกติ จะมีการแจ้งเตือนผู้ใช้หลังจากทำการอัพโหลดข้อมูลเรียบร้อยแล้ว (Display Result)

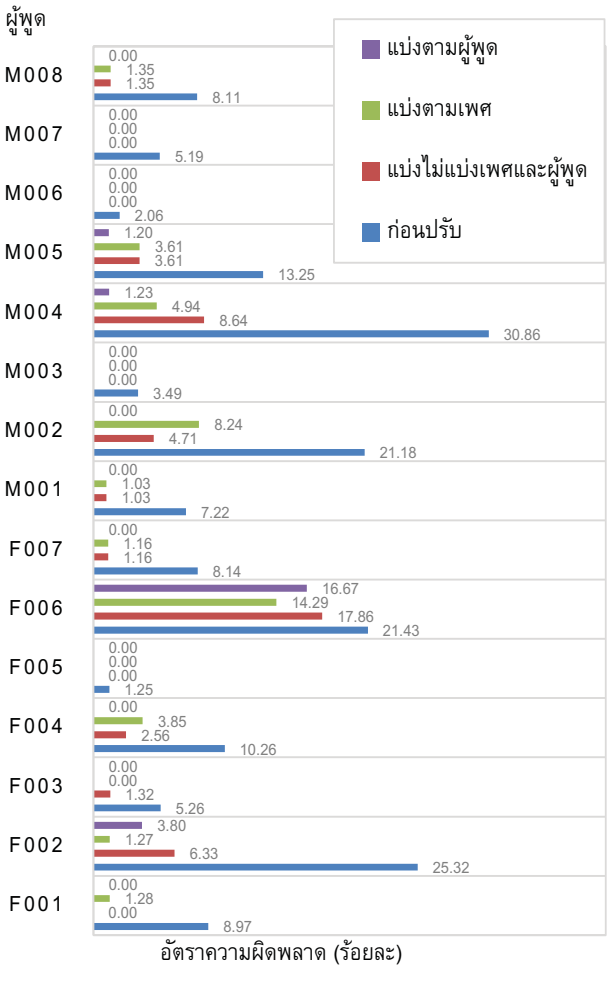
4. ผลการวิจัยและอภิปรายผล

จากการทดสอบการฝึกฝนและปรับโมเดลเสียง ได้ผลการรู้จำเสียงพูดโดยใช้โมเดลที่ผ่านการปรับทั้งสามแบบดังนี้ การปรับโมเดลเสียงที่ส่งผลให้การรู้จำเสียงพูดมีอัตราความแม่นยำมากที่สุด คือ การปรับโมเดลเสียงแบบแบ่งตามผู้พูด โดยมีอัตราความผิดพลาดร้อยละ 11.08 รองลงมา คือ การปรับโมเดลเสียงแบบแบ่งตามเพศ และการปรับโมเดลเสียงแบบไม่แบ่งตามเพศ ซึ่งมีอัตราความผิดพลาดร้อยละ 2.80 และ 3.30 ตามลำดับ ในการปรับโมเดลเสียงทั้งสามแบบมีชุดทดสอบจากผู้พูดจำนวน 4 คนที่มีอัตราความแม่นยำสูงสุด ซึ่งมีอัตราความผิดพลาดร้อยละ 0 ได้แก่ ผู้พูด F005 M003 M006 และ M007 ดังแสดงในรูปที่ 4

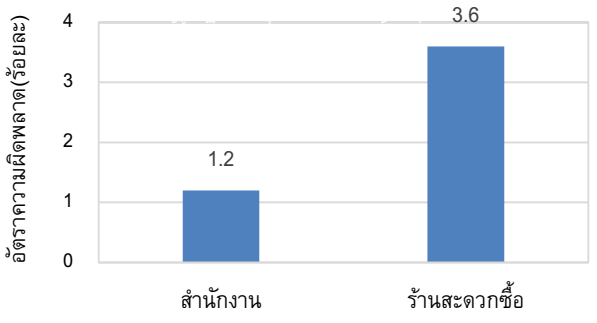
โดยการเลือกใช้โมเดลในการรู้จำเสียงพูด สำหรับพัฒนาระบบนับของด้วยเสียง คือ โมเดลที่ผ่านการปรับแบบไม่แบ่งเพศและผู้พูด เนื่องจากในการนับสินค้านั้นพนักงานนับสินค้าจะมีทั้งเพศชายและเพศหญิง จึงเลือกใช้โมเดลเสียงดังกล่าว

ระบบนับของด้วยเสียงภายใต้สภาพแวดล้อมเสียงรบกวน พัฒนาขึ้นให้มีส่วนติดต่อกับผู้ใช้ให้สะดวกน่าใช้ และนำไปใช้ประโยชน์ได้จริง [8] เช่น เพิ่มความสะดวกในการนับสินค้า เพิ่มศักยภาพในการทำงานให้สามารถนับสินค้าได้อย่างรวดเร็วขึ้น เป็นต้น นอกจากนั้นถ้าต้องการเพิ่มประสิทธิภาพในการรู้จำเสียงพูดจะต้องใช้โมเดลเสียงที่มีคุณภาพดีในกรณีที่แต่ละอุปกรณ์มีผู้ใช้เฉพาะเพียงหนึ่งคน สามารถใช้โมเดลเสียงที่มีการปรับแบบแบ่งตามผู้พูดได้ โดยจะให้ผลการทดสอบที่มีความแม่นยำสูงที่สุด จากการพัฒนาระบบนับของด้วยเสียงพูด ได้ทดลองวัดประสิทธิภาพของระบบ 2 ลักษณะสภาพแวดล้อม ได้แก่ สภาพแวดล้อมในสำนักงาน และสภาพแวดล้อมในร้านสะดวกซื้อ โดยใช้คำทั้งหมด 50 คำ และได้ผลดังแสดงในรูปที่ 5

จากรูปที่ 5 เห็นผลการทดสอบการรู้จำเสียงพูดใน 2 สภาพแวดล้อม โดยประสิทธิภาพการรู้จำเสียงในสำนักงานจะดีกว่าในร้านสะดวกซื้อ ซึ่งในสำนักงานและร้านสะดวกซื้อมีอัตราความผิดพลาดร้อยละ 1.2 และ 3.6 ตามลำดับ ซึ่งอัตราความผิดพลาดในร้านสะดวกซื้อจะสูงกว่าสองเท่า สาเหตุที่อัตราความผิดพลาดการรู้จำเสียงพูดในร้านสะดวกซื้อสูงกว่าเนื่องจากภายในร้านมีเสียงรบกวนมากซึ่งมีผลกระทบต่อความแม่นยำในการรู้จำเสียงพูด



รูปที่ 4 การเปรียบเทียบความผิดพลาดของการรู้จำเสียงพูดก่อนและหลังปรับโมเดลเสียง



รูปที่ 5 แสดงการเปรียบเทียบอัตราความผิดพลาดในการรู้จำเสียงพูดของสภาพแวดล้อมในสำนักงานและร้านสะดวกซื้อ

5. สรุป

การรู้จำเสียงพูดจะมีประสิทธิภาพมากที่สุด เมื่อใช้งานในสภาพแวดล้อมที่มีเสียงรบกวนน้อย การรู้จำเสียงพูดในสำนักงานซึ่งมีเสียงรบกวนน้อยจะมีอัตราความผิดพลาดร้อยละ 1.2 ซึ่งอัตราความผิดพลาดในร้านสะดวกซื้อซึ่งมีเสียงรบกวนมากจะสูงกว่าถึงสองเท่าคือ ร้อยละ 3.6 ตามลำดับ สำหรับการใช้งานในสภาพแวดล้อมที่มีเสียงรบกวนระดับร้านสะดวกซื้อได้ผลระดับพอใช้โดยที่ผู้พูดจำเป็นต้องพูดเสียงให้ดังกว่าเสียงรบกวนนั้น ระบบสามารถเพิ่มความถูกต้องได้โดยใช้ชุดหูฟังและไมโครโฟนที่มีคุณสมบัติในการกำจัดเสียงรบกวน โดยจะทำให้โปรแกรมรู้จำเสียงพูดมีความถูกต้องมาก

กิตติกรรมประกาศ

ขอขอบพระคุณห้องปฏิบัติการวิจัยเทคโนโลยีเสียง ศูนย์อิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ที่ให้ความอนุเคราะห์ในการให้การปรึกษาและฐานข้อมูลเสียง (Large vOcabulary Thai continoUos Speech recognition Corpus) สำหรับการฝึกฝนโมเดลเสียงในการทำวิจัย

ขอขอบพระคุณบริษัท โกซอฟท์ (ประเทศไทย) จำกัด ซึ่งเป็นบริษัทที่ดูแลระบบภายในร้านเซเว่น อีเลฟเว่น ที่ให้ความกรุณาให้ความรู้และข้อมูลในการพัฒนาโปรแกรมสำหรับใช้งานจริงในการนับสินค้าภายในร้านเซเว่น อีเลฟเว่น ตลอดจนการให้การสนับสนุนด้านต่างๆ เช่น เครื่องมือ อุปกรณ์ในการพัฒนาและการทดสอบระบบ

สุดท้ายนี้ ขอขอบพระคุณคณะวิศวกรรมศาสตร์และเทคโนโลยี สถาบันการจัดการปัญญาภิวัฒน์ ที่ให้โอกาสผู้จัดทำได้จัดทำวิจัยจนสำเร็จลุล่วงไปได้ด้วยดี

เอกสารอ้างอิง

[1] D. Huggins-Daines, M. Kumar, A. Chan, A.W. Black, M. Ravishankar & A.I. Rudnicky, “Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices,” in 2006 IEEE International Conference on Acoustics Speech and Signal Processing, Toulouse, France, 2006.

[2] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf and J. Woelfel, “Sphinx-4: A Flexible Open Source Framework for Speech Recognition,” in SMLI TR-2004-139, Nov. 2004.

[3] P. Cotsomrong, T. Sunpetchniyom, S. Kasuriya, N. Thatphithakkul & C. Wutiwiwatchai, “LOTUS: Large vOcabulary Thai continUous Speech Recognition Corpus,” in NAC2005, Nonthaburi, Thailand, 2005.

[4] บุญเสริม กิจศิริกุล, ณัฐกร ทับทอง, “การพัฒนาระบบรู้จำเสียงพูดภาษาไทย,”โครงการเชื่อมโยงการวิจัยภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย, 2546.

[5] Freeman D.K., Cosier G., Southcott C.B., Boyd., “The Voice Activity Detector for the PAN-European Digital Cellular Mobile Telephone Service,” International Conference on Acoustics, 1989.

[6] Jon P. Nedel, “Duration normalization for robust recognition of spontaneous speech via missing feature methods,” Ph.D. Thesis, Carnegie Mellon University, 2004.

[7] J. Baker, “Stochastic Modeling as a Means of Automatic Speech Recognition,” Ph.D. Thesis, Carnegie Mellon University, 1975.

[8] มนตรี โพธิโสโนทัย, เฉลิมภัณฑ์ ฟองสมุทร, “วิธีการรู้จำเสียงพูดภาษาไทยแบบทนทานต่อเสียงรบกวนภายนอก,” วารสารเทคโนโลยีสารสนเทศ, ฉบับที่ 13, มกราคม – มิถุนายน 2554.