วิธีการรู้จำเสียงพูดภาษาไทยแบบทนทานต่อเสียงรบกวนภายนอก

Robust Thai Speech Recognition System against Real Environmental Noises

มนตรี โพธิโสโนทัย* และ เฉลิมภัณฑ์ ฟองสมุทร**

บทคัดย่อ

บทความนี้เสนอวิธีการรู้จำเสียงพูดภาษาไทยแบบทนทานต่อสัญญาณรบกวนจริงจากสภาพแวดล้อมภายนอก โดยวิธีการรู้จำเสียงนี้จะใช้ช่องสัญญาณจำนวนสองช่อง เพื่อใช้สำหรับรับเสียงพูดและใช้สำหรับรับเสียงรบกวน โดยเสียงรบกวนจะถูกกำหนดให้เป็นเสียงรบกวนที่เกิดขึ้น จากนั้นจะใช้วิธีการลบเชิงสเปกตรัม (Spectral subtraction: SS) เพื่อกำจัดเสียงรบกวนออก ขั้นตอนการทดลองได้ใช้เสียงพูดตัวเลขภาษาไทยจำนวน 10 คำ เป็นเสียงทดสอบ และได้เปรียบเทียบผลการทดลองกับสภาพแวดล้อมจริง โดยใช้ทดสอบกับสภาพแวดล้อมที่ไม่มีเสียงรบกวน และสภาพแวดล้อมที่มีเสียงรบกวนจริงที่ระดับความดังประมาณ 40–50 เดซิเบล ผลที่ได้พบว่าระบบที่ได้นำเสนอสามารถแยกเสียงตัวเลขได้ถูกต้องด้วยอัตราแยกเฉลี่ยประมาณ 98.0 เปอร์เซ็นต์ในกรณีที่ไม่มีเสียงรบกวน และอัตราแยกเฉลี่ยประมาณ 83.6 เปอร์เซ็นต์ในกรณีที่มีเสียงรบกวน

คำสำคัญ: การรู้จำเสียงพูด เสียงพูดภาษาไทย การประมวลผลเสียงพูด สัญญาณรบกวนภายนอก

Abstract

This paper proposes a robust speech recognition method against real environmental noises. The proposed method is on the basis of double-channel system in which original speech and noise are recorded separately. To reduce the noise, in this paper, the spectral subtraction (SS) method has been employed. During experiment, we used numerical Thai speech as the input data. Noiseless and noisy sound level of 40-50 dB conditions were compared to assess the performance of the proposed method. Based on the experimental results, the average recognition rates are about 98.0% and 83.6% for noiseless and noisy environments, respectively.

Keyword: speech recognition, Thai speech, speech processing, environmental noise.

1. บทนำ

เทคโนโลยีการรู้จำเสียงพูด (speech recognition) ได้ถูกพัฒนาก้าวหน้าอย่างมาก ซึ่งในปัจจุบันสามารถนำเอาระบบดังกล่าวมาใช้ประโยชน์ได้อย่างมีประสิทธิภาพและหลากหลายมากขึ้น การรู้จำเสียงนั้นสามารถพัฒนาเป็นโปรแกรมอำนวยความสะดวกสำหรับใช้งานจริงได้ โดยพบว่าจะมีประโยชน์อย่างยิ่งเมื่อนำไปใช้กับผู้พิการ จากการศึกษาพบว่าขณะนี้ทั่วโลกมีคนพิการประมาณ 600 ล้านคน ในส่วนของประเทศไทยจากการสำรวจของสำนักงานสถิติแห่งชาติล่าสุดเมื่อ พ.ศ.2544 ทั่วประเทศมี ผู้พิการจำนวน 1.1 ล้านคน หรือเกือบ 2% ของประชากรประเทศ โดยพิการทางกาย แขน-ขาขาด หรือด้วน เท้ากุดมากที่สุด 47% ตาบอด 11% [1] และสถานการณ์คนพิการทั่วโลกมีแนวโน้มเพิ่มขึ้น เราสามารถใช้ระบบรู้จำเสียงซึ่งจะมีบทบาทสำคัญที่จะอำนวยความสะดวกให้กับคนพิการ เพื่อต้องการเพิ่มคุณภาพชีวิตให้ดีขึ้นและเพิ่มความสะดวกใน การดำรงชีวิต ด้วยเหตุนี้จึงทำให้การพัฒนางานด้านระบบอัตโนมัติสำหรับโต้ตอบกับมนุษย์เป็นสิ่งที่จำเป็น ตัวอย่างของกลุ่มงานวิจัย [2] – [4] ที่ใช้ระบบรู้จำเสียงมีดังนี้

* วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา คณะวิศวกรรมศาสตร์ มหาวิทยาลัยบูรพา

** ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยบูรพา

1) การสั่งงานอุปกรณ์ไฟฟ้าภายในบ้านด้วยเสียงหรืออุปกรณ์รถเข็นด้วยเสียง

2) การสั่งงานหรือพิมพ์เอกสารด้วยเสียงบนเครื่องคอมพิวเตอร์

3) การสั่งงานบนอุปกรณ์พกพาด้วยเสียง (เช่น โทรศัพท์เคลื่อนที่ เครื่องเล่น MP3 และ PDA) ปัจจุบันเป็นอุปกรณ์สำคัญที่ตลาดกำลังเติบโตเป็นอย่างมาก

4) การช่วยเหลือผู้ที่ปฏิบัติหน้าที่ในสถานการณ์ที่ไม่สามารถใช้วิธีการกดหรือพิมพ์เพื่อสั่งงานได้ หรือใช้ได้ไม่สะดวก (เช่น ในโรงงานอุตสาหกรรม และในขณะขับรถ)

5) ใช้สำหรับตอบรับอัตโนมัติในระบบโต้ตอบด้วยเสียง (Interactive Voice Response : IVR) ในระบบศูนย์บริการข้อมูลให้ความช่วยเหลือ สอบถามปัญหา ข้อแนะนำ บริการต่างๆ (Call center หรือ Operator) เป็นต้น

สำหรับวิธีการรู้จำเสียงในภาษาไทยนั้นยังต้องใช้ในสภาวะแวดล้อมที่เงียบไร้เสียงรบกวน เพื่อให้การทำงานของระบบการรู้จำเสียงมีการประมวลผลที่ถูกต้อง แต่ในความเป็นจริงแล้วเสียงรบกวนจะทำให้ประสิทธิภาพของการประมวลผลสัญญาณลดลงเนื่องจากเสียงรบกวนจากสภาพแวดล้อมนั้นเกิดขึ้นได้ง่าย ทำให้การใช้งานจริงของสัญญาณเสียงที่เป็นอินพุตส่วนใหญ่จะเป็นสัญญาณเสียงพูดผสมเสียงรบกวน ซึ่งเกิดมาจากสภาพแวดล้อมที่ใช้งานอยู่ เช่น เสียงคนกำลังสนทนา แอร์ พัดลม เสียงรถยนต์ เป็นต้น ดังนั้นการรู้จำเสียงพูดแบบเดิมอาจให้ประโยชน์แก่ผู้ใช้งานได้ไม่เต็มที่เนื่องจากมีข้อจำกัดทั้งสถานที่และเสียงรบกวนภายนอกต่างๆ จากงานวิจัยที่ผ่านมาพบว่ากระบวนการรู้จำเสียงที่ทนทานต่อสภาพแวดล้อมจริง [5-6] ได้มีนำเสนอวิธีการลดเสียงรบกวนออกจากเสียงอินพุตด้วยวิธีการแปลงเวฟเลต (wavelet transform) และการแปลงโคไซน์ (cosine transform) อย่างไรก็ตามพบว่าวิธีการเหล่านี้มีวิธีการลดเสียงรบกวนที่ซับซ้อนและไม่เหมาะต่อการนำมาใช้งานในเวลาจริง

ดังนั้นงานวิจัยนี้จึงเน้นการศึกษาและพัฒนาวิธีการรู้จำเสียงภาษาไทยแบบทนทานต่อเสียงรบกวนจากภายนอกอย่างง่ายด้วยการวิเคราะห์เชิงความถี่ของเสียง ขั้นตอนการทดลองของงานวิจัยนี้จะใช้ตัวอย่างเสียงพูดตัวเลขภาษาไทยเป็นกรณีศึกษา

2. ระเบียบวิธีวิจัย

ระบบโดยรวมที่นำเสนอของงานวิจัยแสดงดังภาพที่ 1 ประกอบด้วยขั้นตอนย่อยดังต่อไปนี้

2.1 การรับสัญญาณเสียง

ขั้นตอนการรับสัญญาณเสียงสำหรับงานวิจัยนี้ได้เลือกใช้คอนเดนเซอร์ไมโครโฟนจำนวน 2 ช่องสัญญาณ โดยช่องสัญญาณแรกจะถูกใช้สำหรับรับเสียงพูดของผู้ใช้งาน ส่วนช่องสัญญาณที่สองจะถูกใช้สำหรับรับเสียงรบกวนที่เกิดขึ้นบริเวณที่ผู้ใช้งานอยู่ ซึ่งสามารถเขียนเป็นรูปสมการได้ดังนี้

$$Y_i = S_i + N_i \quad (1)$$

โดยที่ Y_i เป็นสัญญาณอินพุตของระบบ S_i เป็นเสียงพูด และ N_i เป็นเสียงรบกวนที่เกิดขึ้นที่ตำแหน่ง i

2.2 วิธีการลบเชิงสเปกตรัม (Spectral Subtraction : SS)

วิธีการลบเชิงสเปกตัม หรือเรียกย่อๆ ว่า วิธี SS เสนอโดย Boll [7] แนวความคิดพื้นฐานของวิธีนี้คือเราสามารถแยกสัญญาณสองสัญญาณที่รวมกันได้โดยนำเอาสัญญาณที่ประมาณได้ลบออก โดยวิธี SS เลือกการลบการทางโดเมนความถี่ จากสมการที่ 1 นำมาทำการแปลงฟูเรียร์ (Fourier transform) จะได้ผลดังสมการต่อไปนี้

$$Y(\omega) = S(\omega) + N(\omega) \quad (2)$$

จากนั้นเราสามารถประมาณเสียงรบกวนทางโดเมนความถี่ได้ เราสามารถแยกเสียงรบกวนออกจากเสียงพูดด้วยสมการต่อไปนี้

$$\hat{Y}(\omega) = Y(\omega) + \hat{N}(\omega) \quad (3)$$

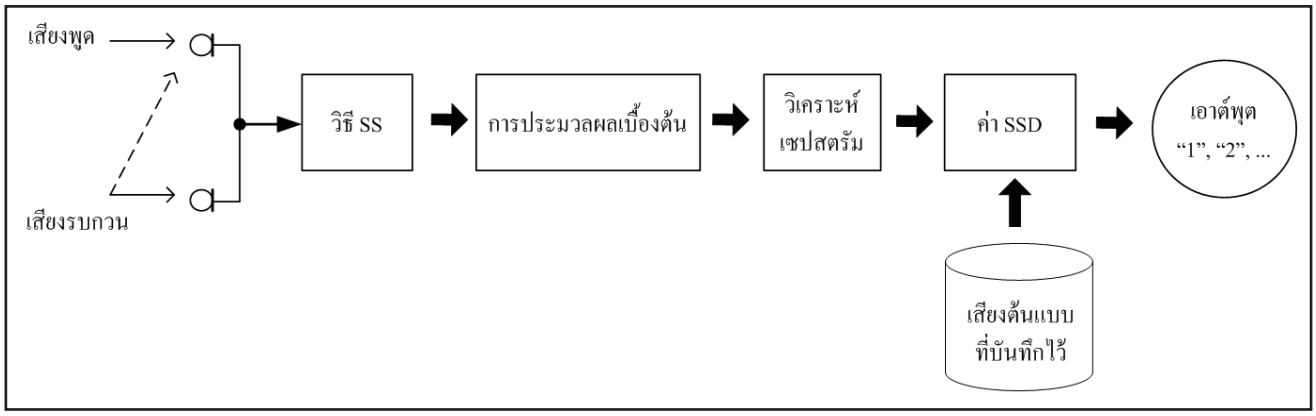
โดยที่ $\hat{Y}(\omega)$ คือเสียงพูดที่ประมาณได้ทางโดเมนความถี่ และ $\hat{N}(\omega)$ คือเสียงรบกวนที่ประมาณได้ทางโดเมนความถี่

2.3 การประมวลผลเบื้องต้น

หลังจากผ่านกระบวนการกำจัดเสียงรบกวนออกด้วยวิธี SS แล้ว สัญญาณเอาต์พุตที่ได้จะถูกนำไปผ่านการประมวลผลเบื้องต้นดังนี้

(1) การปรับค่าเฉลี่ยเป็นศูนย์ (Zero Mean) : สัญญาณเสียงที่ได้มาแปลงจากสัญญาณอะนาลอกเป็นดิจิตอลจากนั้นนำสัญญาณเสียงที่ได้มาปรับค่าแกนกลางให้เป็นศูนย์เพื่อให้ง่ายต่อการนำไปประมวลผล

(2) รูปแบบบรรทัดฐาน (Normalization) : การทำรูปแบบบรรทัดฐานจะทำให้สัญญาณเสียงจะมีค่าอยู่ระหว่าง -1.0 ถึง 1.0 เสมอ เพื่อให้สัญญาณเสียงแต่ละครั้งมีลักษณะของสัญญาณอยู่ในช่วงที่กำหนด



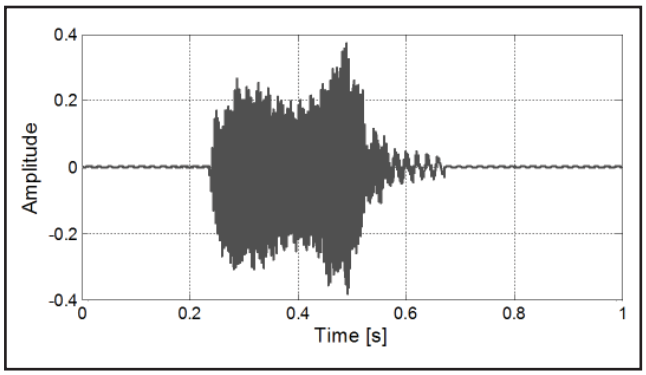
ภาพที่ 1 *ขั้นตอนโดยรวมของงานวิจัยที่นำเสนอ*

2.4 การวิเคราะห์เซปสตรัม (Cepstrum Analysis)

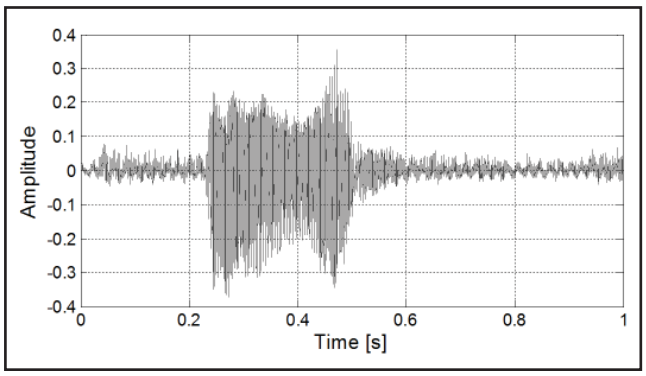
เซปสตรัมเป็นวิธีการวิเคราะห์สัญญาณในโดเมนความถี่ ซึ่งคำว่าเซปสตรัมมาจากการสลับคำจากคำว่า สเปกตรัม (Spectrum) โดยการวิเคราะห์เซปสตรัมนั้นจะนำผลการแปลงฟูเรียร์มาผ่านฟังก์ชันลอการิธึมเพื่อขยายรายละเอียดทางความถี่ต่ำด้วยสมการดังนี้

$$P = \left(\log\left\{ \left| \hat{Y}(\omega) \right|^2 \right\} \right)^2 \qquad (4)$$

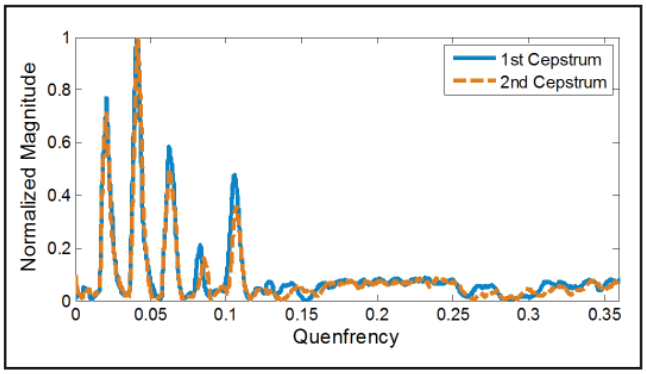
โดยที่ P คือค่าพลังงานของเซปสตรัม ตัวอย่างการเปรียบเทียบผลลัพธ์การวิเคราะห์เซปสตรัมของคำว่า ศูนย์ จำนวนสองครั้งดังภาพที่ 4



ภาพที่ 2 *ตัวอย่างเสียงพูด “ศูนย์”*



ภาพที่ 3 *ตัวอย่างเสียงพูดที่มีเสียงรบกวน*



ภาพที่ 4 *ตัวอย่างการเปรียบเทียบผลลัพธ์การวิเคราะห์เซปสตรัมของคำว่า “ศูนย์” จำนวนสองครั้ง*

2.5 ค่ารวมของผลต่างกำลังสอง (Sum of Squared Difference: SSD)

การทำให้ระบบสามารถแยกแยะเสียงพูดได้นั้นจะต้องมีการเปรียบเทียบกับเสียงต้นแบบ (template matching) ที่ได้เก็บไว้ โดยงานวิจัยนี้จะใช้วิธีหาค่ารวมของผลต่างกำลังสองซึ่งนิยามได้จากสมการ

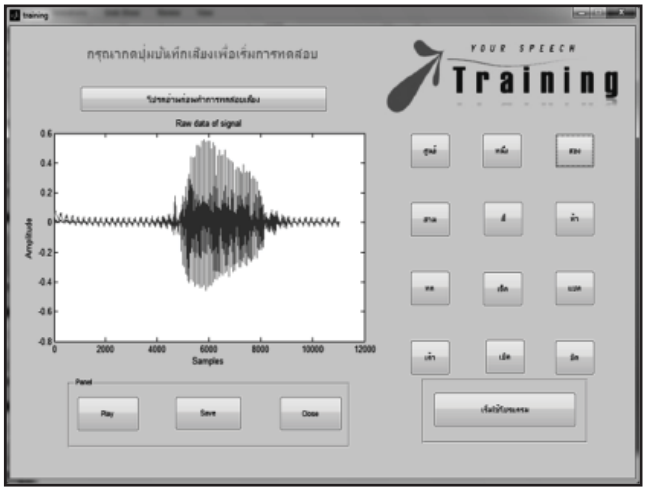
$$E_{SSD} = \sum_{k=1,2,\ldots,N} (P_k - P_k^t)^2 \qquad (5)$$

โดยค่า E_{SSD} คือค่ารวมของผลต่างกำลังสอง หรือค่า *SSD*, P_k คือค่าเซปสตรัมของเสียงอินพุต และ P_k^t คือเซปสตรัมของเสียงแม่แบบ (จากสมการที่ 4) ซึ่งค่าพารามิเตอร์นี้จะให้ค่าเข้าใกล้ศูนย์เมื่อเสียงพูดกับเสียงต้นแบบมีลักษณะของข้อมูลที่ใกล้เคียงกัน (จะมีค่าเป็นศูนย์หากนำเสียงพูดชุดเดียวกันมาวิเคราะห์) ดังนั้นวิธีการคัดแยกนั้นจะใช้เงื่อนไขการคัดแยกค่า *SSD* ที่ต่ำที่สุดเมื่อเทียบเทียบเสียงอินพุตกับเสียงต้นแบบนิยามได้จาก

$$Z_{out} = \min_{T_{SSD} \in \{0,1,2,\ldots,9\}} (E_{SSD}, T_{SSD}) \qquad (6)$$

3. การทดลองและผลการทดลอง

งานวิจัยนี้ได้ออกแบบระบบด้วยโปรแกรม MATLAB™, MathWorks, Inc. [8] ขั้นตอนการบันทึกเสียงนั้นได้เลือกใช้อัตราสุ่มข้อมูลที่ 11,025 เฮิร์ต ด้วยความละเอียดของเสียง 16 บิต กำหนดความยาวของเสียงเท่ากับ 1 วินาทีตลอดการทดลอง การออกแบบระบบนั้นได้ใช้ Graphical User Interface (GUI) เพื่อให้ง่ายต่อการใช้งาน ดังแสดงในภาพที่ 5



ภาพที่ 5 *หน้าต่างโปรแกรมสำหรับการฝึกให้ระบบรู้จำเสียง*

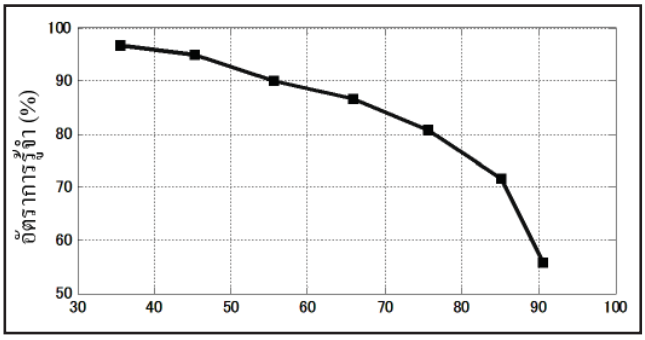
ภาพที่ 6 *หน้าต่างโปรแกรมที่ได้พัฒนาขึ้นสำหรับงานวิจัยนี้*

การทดสอบได้ทดสอบกับผู้ใช้งานจำนวน 10 คน (เพศหญิง 5 คน เพศชาย 5 คน) อายุระหว่าง 20 – 22 ปี โดยให้ผู้ทดสอบพูดเสียงตัวเลขภาษาไทยตั้งแต่ศูนย์ถึงเก้าคำละ 10 ครั้ง ขั้นตอนการทดสอบนั้นแบ่งได้ดังนี้

1) ขั้นตอนการฝึก (training process) เพื่อให้ระบบจำจดเสียงต้นแบบไว้จะให้ผู้ใช้งานทำการบันทึกเสียงพูดทุกคำก่อน โดยหลังจากผ่านขั้นตอนการฝึกแล้วเสียงพูดของผู้ใช้งานจะถูกนำไปประมวลผลเพื่อเก็บไว้เปรียบเทียบกับเสียงพูดจริงอีกครั้ง

2) ขั้นตอนการรู้จำ (recognition process) จะเป็นขั้นตอนการทดสอบแยกเสียงจริงโดยให้ผู้ใช้งานพูดผ่านไมโครโฟน จากนั้นระบบจะแสดงตัวเลขผ่านทางหน้าจออัตโนมัติ

จากนั้นจะทำการเก็บผลการทดลองโดยทำการประเมินค่าอัตราการแยกเสียงที่ถูกต้องของระบบกับเสียงพูดจริงของผู้ใช้งาน แบ่งเป็นสภาพแวดล้อมที่ไม่มีเสียงรบกวน (ระดับสียงต่ำกว่า 10 dB) และได้ทดสอบระบบกับสภาพแวดล้อมจริงบริเวณข้างถนนที่มีรถสัญจรไปมาตลอด (ระดับเสียงดังมากกว่า 40 dB) ผลการทดลองแสดงในตารางที่ 1 นอกจากนี้งานวิจัยนี้ยังได้ใช้เครื่องวัดระดับเสียง (sound level meter) รุ่น DT-8820 หน่วยวัดเดซิเบลเพื่อทดสอบวิธีการที่นำเสนอกับสภาพแวดล้อมจริงที่มีระดับเสียงรบกวนแตกต่างกันผลที่ได้แสดงดังภาพที่ 7



ภาพที่ 7 *กราฟแสดงอัตราการรู้จำของวิธีการที่ได้นำเสนอ โดยเปรียบเทียบกับระดับเสียงรบกวนแตกต่างกัน*

4. อภิปรายผลการทดลอง

จากผลการทดลอง วิธีการที่นำเสนอสามารถแยกเสียงพูดได้ถูกต้องเฉลี่ย 98.0% และ 83.6% อย่างไรก็ตามพบว่าวิธีการนี้ยังมีข้อจำกัดอยู่ดังนี้

1) ปัญหาจากผู้ใช้งานบางคนออกเสียงพูดไม่ชัดเจนระหว่างขั้นตอนการฝึกและขั้นตอนการรู้จำ ทำให้ระบบเกิดความผิดพลาด

2) ปัญหาจากตำแหน่งการวางไมโครโฟนเพื่อรับเสียงรบกวน พบว่าหากตำแหน่งของช่องเสียงรบกวนไม่ตรงกับเสียงรบกวนจริงจะทำให้ผลของการวิเคราะห์ผิดพลาดได้

เนื่องจากวิธีการกำจัดเสียงรบกวนของงานวิจัยนี้ใช้

ตารางที่ 1 *อัตราการแยกเสียงของระบบที่นำเสนอรบกวนแตกต่างกัน*

เสียงพูด	ไม่มีเสียงรบกวน	มีเสียงรบกวน
ศูนย์	99.6 %	82.4 %
หนึ่ง	99.6 %	82.2 %
สอง	99.0 %	80.2 %
สาม	98.2 %	86.0 %
สี่	97.6 %	84.8 %
ห้า	99.6 %	83.8 %
หก	99.2 %	86.6 %
เจ็ด	90.0 %	79.6 %
แปด	99.4 %	90.4 %
เก้า	97.8 %	80.2 %
เฉลี่ย	**98.0 %**	**83.6 %**

สมมุติฐานที่ว่าเสียงรบกวนที่เข้ามากับเสียงพูดนั้นเกิดขึ้นจากสภาพแวดล้อมของผู้ทดสอบ การออกแบบติดตั้งตำแหน่งของไมโครโฟนที่ดีในการรับเสียงรบกวนนั้นสามารถที่จะหลีกเลี่ยงข้อผิดพลาดนี้ลงได้

5. สรุปผลการทดลอง

งานวิจัยนี้ได้นำเสนอวิธีการรู้จำเสียงภาษาไทยแบบทนทานต่อเสียงรบกวนภายนอก โดยมีขั้นตอนการประมวลผลที่ไม่ซับซ้อนมาก จากผลการทดลองพบว่าวิธีการที่ได้นำเสนอนี้สามารถนำไปใช้กับสภาพแวดล้อมที่มีเสียงรบกวนที่เกิดขึ้นจริงได้ขณะที่ระบบยังสามารถแยกแยะเสียงพูดตัวเลขภาษาไทยได้ งานวิจัยในอนาคตคาดว่าจะนำวิธีการที่ได้นำเสนอนี้ทดสอบกับสภาพแวดล้อมที่หลากหลายมากขึ้นและระดับเสียงรบกวนที่แตกต่างกัน

6. เอกสารอ้างอิง

[1] สำนักงานสารนิเทศและประชาสัมพันธ์กระทรวงสาธารณสุข, *สถิติผู้พิการ*, Available online at http://www.moph.go.th

[2] สกล ลิโมทัย, "ระบบรู้จำเสียงพูดสำหรับควบคุมอุปกรณ์ไฟฟ้าภายในบ้าน" วิทยานิพนธ์ปริญญาวิศวกรรมศาสตรมหาบัณฑิต, สาขาวิชาวิศวกรรมคอมพิวเตอร์, สถาบันมหาวิทยาลัยเกษตรศาสตร์, พ.ศ. 2550.

[3] N. Hataoka, H. Kokzibo, Y. Obuchi, and A. Amano. "Development of Robust Speech Recognition Middleware on Microprocessor," *IEEE Int. Conf. ASSP*'98, pp. 837-840, 1998.

[4] ชัย วุฒิวิวัฒน์ชัย, "การพัฒนาระบบสนทนาภาษาไทย" *Int. Symp. Natural Lang. Process.* (SNLP), พ.ศ. 2548.

[5] M. Phothisonothai, P. Kumhom, and K. Chamnongthai, "Single-Channel Noise Reduction for Multiple Background Noises using Perceptual Wavelet Package Transform and Fuzzy Logic," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol.8, no.6, pp.613-620, 2004.

[6] O. Jarungpornsawas, P. Nawarat Na Ayudaya, P. Kumhom, and K. Chamnongthai, "Noise Estimation for Low Distortion Speech Enhancement in Real Environment," *Int. Symposium on Communications and Information Technology* (ISCIT), November 14-16, Chiang Mai, Thailand, pp. 733-736, 2001.

[7] S. F. Boll, "Suppressing of Acoustic Noise in Speech using Spectral Subtraction," *IEEE Trans. Acoust. Speech Signal Processing*, vol.27, pp. 133-120, 1979.

[8] Matlab, MathWorks, Inc., Available online at http://www.mathworks.com