รายงานการวิจัย

การรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้ของผู้สูงอายุเพื่อควบคุมอุปกรณ์ไฟฟ้า

Elderly Southern Thai Dialect Speech Recognition Based Electrical Device Control

ปฏิมากร จันทร์พริ้ม

ธีรพงษ์ ฉิมเพชร

กีรติ อินทวิเศษ

คณะวิศวกรรมศาสตร์

มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย

ได้รับการสนับสนุนทุนวิจัยจากมหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย

งบประมาณรายได้ ประจำปี พ.ศ. 2558

คำนำ

รายงานการวิจัย เรื่อง การรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้ของผู้สูงอายุเพื่อควบคุมอุปกรณ์ไฟฟ้า ได้รับการสนับสนุนทุนอุดหนุนวิจัยจากงบประมาณรายได้ ประจำปี พ.ศ. 2558 คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย เป็นรายงานโครงการวิจัยใหม่ที่สอดคล้องกับยุทธศาสตร์การพัฒนาคนสู่สังคมแห่งการเรียนรู้ตลอดชีวิตอย่างยั่งยืน และการสร้างศักยภาพและความสามารถเพื่อการพัฒนาทางสังคม มีเป้าประสงค์การวิจัยเพื่อสร้างเสริมองค์ความรู้ให้เป็นพื้นฐานเพื่อความมั่นคงของประเทศโดยการสร้างความเข้มแข็งของสังคม การพัฒนาและยกระดับคุณภาพชีวิตและความผาสุกของประชาชน เนื้อหาภายในประกอบด้วยรายละเอียดส่วนหลักของโครงการ ทฤษฎีและการทบทวนวรรณกรรมวิจัยที่เกี่ยวข้อง วิธีดำเนินงาน ผลการทดลอง สรุปผลการทดลองและข้อเสนอแนะ คณะผู้จัดทำหวังเป็นอย่างยิ่งว่ารายงานการวิจัยนี้จะเป็นประโยชน์กับคณะวิศวกรรมศาสตร์และสามารถใช้เป็นแนวทางสำหรับงานวิจัยทางด้านวิศวกรรมศาสตร์ต่อไป

ปฏิมากร จันทร์พริ้ม
ธีรพงษ์ ฉิมเพชร
กีรติ อินทวิเศษ

บทคัดย่อ

งานวิจัยนี้ทำการศึกษาการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้ของผู้สูงอายุเพื่อควบคุมอุปกรณ์ไฟฟ้า ใช้เสียงพูดสำเนียงภาคใต้ของจังหวัดสงขลา พัทลุง ตรัง และนครศรีธรรมราชจากผู้สูงอายุที่อาศัยอยู่ในท้องถิ่นนั้นทั้งเพศชายและหญิง เพศละ 2 คนต่อ 1 จังหวัด รวมเป็น 16 คน เสียงพูดมีทั้งสิ้น 5 คำ ได้แก่ คำว่า เปิด ปิด พัดลม หลอดไฟ และ ทีวี โดยให้พูดซ้ำคำสั่งละ 4 ครั้ง ทำการบันทึกเสียงจากโปรแกรมที่เขียนขึ้นจาก LabVIEW และทำการวิเคราะห์ด้วยโปรแกรมที่เขียนขึ้นจาก MATLAB ทำการศึกษา 2 การทดลอง คือ 1) ศึกษาเพื่อดูว่าคุณลักษณะเด่นเชิงเวลาร่วมกับคุณลักษณะเด่นเชิงความถี่สามารถช่วยในการรู้จำเสียงพูดได้ 2) ศึกษาเพื่อดูว่าการใช้ Discrete Wavelet Transform (DWT) สามารถช่วยเพิ่มประสิทธิภาพการรู้จำเสียงพูดจากการใช้คุณลักษณะเด่นจำนวนพยางค์และสัมประสิทธิ์การประมาณพันธะเชิงเส้นที่ผ่าน DWT ผลการทดสอบการรู้จำด้วยโครงข่ายประสาทเทียมชนิด Backpropagation กับข้อมูลทดสอบ 160 ข้อมูล พบว่า 1) ระยะเวลาของเสียงพูด และจำนวนพยางค์ ซึ่งเป็นคุณลักษณะเด่นเชิงเวลา ร่วมกับจำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ ซึ่งเป็นคุณลักษณะเด่นเชิงความถี่สามารถใช้เป็นคุณลักษณะเด่นในการรู้จำได้ โดยให้ความถูกต้องของการรู้จำเท่ากับ 80.2 เปอร์เซ็นต์ 2) การใช้กระบวนการ DWT ร่วมด้วยช่วยเพิ่มประสิทธิภาพการรู้จำสำหรับคุณลักษณะเด่นจำนวนพยางค์ร่วมกับสัมประสิทธิ์การประมาณพันธะเชิงเส้น โดยเฉพาะจำนวนพยางค์ร่วมกับสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 จากสัญญาณองค์ประกอบความถี่ต่ำของ Wavelet Decomposition ที่ระดับ 2 ให้เปอร์เซ็นต์ความถูกต้องของการรู้จำสูงสุดเท่ากับ 83.75 เปอร์เซ็นต์ อย่างไรก็ตามการรู้จำเสียงพูดที่ใช้จำนวนคำสั่ง และจำนวนผู้พูดมากขึ้นควรได้รับการวิจัยต่อไปเพื่อเป็นการยืนยันประสิทธิภาพของวิธีการเพิ่มเติม

คำสำคัญ รู้จำเสียงพูด, เสียงพูดภาษาไทยสำเนียงภาคใต้, เวฟเลตทรานฟอร์ม

Abstract

This research describes a study on southern Thai dialect speech recognition based on electrical device control for the elderly. Elderly speech is southern Thai dialect including Songkhla, Phatthalung, Trang and Nakhon Si Thammarat. 16 subjects, 2 subjects/sex/province, stay for those local areas. The study focused on 5 commands for control electrical device. Those were 'turn on', 'turn off', 'lamp', 'fan', and 'TV'. Each command was repeated 4 times. The speech was recorded by created programs from LabVIEW and analyzed by created programs from MATLAB. Two experiments were studied. 1) To find that time-domain features together with frequency domain feature can recognize speech. 2) To find that Linear Predictive Coefficient from Discrete Wavelet Transform and number of syllable can improve the speech recognition performance. For 160 data, the backpropagation results showed that 1) duration of speech and number of syllable, time-domain features, and number of formant frequency peak for each frequency group, frequency domain feature, utilized for speech recognition. 2) Linear Predictive Coefficient from Discrete Wavelet Transform could improve the speech recognition performance. Especially, the 3rd order Linear Predictive coefficients from low pass filter of wavelet decomposition at level 2 and number of syllable offered the highest percent recognition at 83.75. However, data of more commands and subjects should be ongoing studied to verify this method.

Keywords: speech recognition, southern Thai dialect speech, Wavelet Transform

กิตติกรรมประกาศ

งานวิจัยนี้ได้รับการสนับสนุนทุนอุดหนุนวิจัยจากงบประมาณรายได้ ประจำปี พ.ศ. 2558 คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย ขอขอบคุณครอบครัว เพื่อนคณาจารย์ และเจ้าหน้าที่ทั้งในสาขาวิศวกรรมไฟฟ้าและคณะวิศวกรรมศาสตร์ที่คอยช่วยเหลือ ให้กำลังใจ และอำนวยความสะดวกในเรื่องต่างๆ จนทำให้งานวิจัยนี้สำเร็จลุล่วงไปได้ด้วยดี

ปฏิมากร จันทร์พริ้ม
ธีรพงษ์ ฉิมเพชร
กีรติ อินทวิเศษ

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

บทนี้จะกล่าวถึงความสำคัญและที่มาของปัญหาที่ทำการวิจัย วัตถุประสงค์ ขอบเขต และประโยชน์ที่จะได้รับ

1.1 ความสำคัญและที่มาของปัญหาที่ทำการวิจัย

ตัวเลขจากการคาดประมาณประชากรของสำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติ ได้ประมาณการแนวโน้มการเปลี่ยนแปลงประชากรผู้สูงอายุว่า ในปี 2566 ประชากรผู้สูงอายุในประเทศไทยจะมีจำนวนเพิ่มขึ้นเป็น 14.1 ล้านคน คิดเป็นร้อยละ 21 ของประชากรทั้งหมด และในปี 2576 ประเทศไทยจะมีประชากรผู้สูงอายุมากถึง 18.7 ล้านคน หรือคิดเป็นร้อยละ 29 ของประชากรทั้งหมด เท่ากับว่า ประเทศไทยจะกลายเป็น "สังคมสูงวัยอย่างสมบูรณ์" [1]

การสูงวัยของร่างกายย่อมส่งผลต่อความคล่องตัวในการเคลื่อนไหวเพื่อควบคุมอุปกรณ์ไฟฟ้าเพื่ออำนวยความสะดวกในการดำรงชีวิตของผู้สูงอายุ อีกทั้งผู้สูงอายุส่วนใหญ่ในปัจจุบันมีแนวโน้มที่จะมีช่วงเวลาที่อยู่เพียงลำพังมากขึ้น อาจเนื่องจากผู้ดูแลมีความจำเป็นต้องออกไปทำงานนอกบ้านหรือความต้องการความเป็นส่วนตัวของผู้สูงอายุเอง ทำให้ผู้สูงอายุจำเป็นต้องพึ่งพาตนเองมากขึ้น การควบคุมอุปกรณ์ไฟฟ้าด้วยเสียงพูดจึงเป็นทางเลือกหนึ่งที่ช่วยอำนวยความสะดวกให้กับผู้สูงอายุ ทำให้สามารถแบ่งเบาภาระของผู้ดูแล ช่วยให้ผู้สูงอายุมีคุณภาพชีวิตดีขึ้น และสามารถอยู่ในสังคมได้อย่างมีความสุขมากขึ้น

โครงการวิจัยนี้จึงทำการศึกษาการรู้จำเสียงพูดของผู้สูงอายุเพื่อควบคุมอุปกรณ์ไฟฟ้าและเพื่อให้รองรับสำหรับผู้สูงอายุในจังหวัดสงขลาและจังหวัดใกล้เคียงซึ่งผู้สูงอายุส่วนใหญ่มีความคุ้นเคยกับการใช้ภาษาไทยสำเนียงท้องถิ่นตนเอง จึงได้มุ่งศึกษาการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้ของผู้สูงอายุเพื่อควบคุมอุปกรณ์ไฟฟ้าสำหรับจังหวัดสงขลา พัทลุง ตรัง และนครศรีธรรมราช

1.2 วัตถุประสงค์

1. เพื่อศึกษาการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้ของผู้สูงอายุ

2. เพื่อช่วยยกระดับคุณภาพชีวิตของผู้สูงอายุเนื่องจากสามารถควบคุมอุปกรณ์ไฟฟ้าได้ด้วยตนเอง

3. เพื่อช่วยลดภาระของผู้ดูแลผู้สูงอายุ

1.3 ขอบเขต

1. เป็นโปรแกรมวิเคราะห์และแสดงผลการรู้จำ

2. ใช้เสียงพูดจากผู้สูงอายุตั้งแต่ 60 ปีขึ้นไปที่เป็นคนท้องถิ่นจังหวัดสงขลา พัทลุง ตรัง และนครศรีธรรมราช

3. ผู้พูดอาศัยอยู่ในท้องถิ่นของทั้ง 4 จังหวัดดังกล่าวยาวนานมากกว่า 10 ปี และใช้สำเนียงภาคใต้ของท้องถิ่นนั้นๆ เป็นหลัก

4. ผู้พูดมีทั้งเพศชายและหญิง เพศละอย่างน้อย 2 คนต่อ 1 จังหวัด

5. เสียงพูดมีทั้งสิ้น 5 คำ ได้แก่ คำว่า เปิด ปิด พัดลม หลอดไฟ และ ทีวี

6. บันทึกเสียงพูดในสภาพแวดล้อมค่อนข้างเงียบ

1.4 ประโยชน์ที่จะได้รับ

1. ได้อัลกอริทึมในการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้ของผู้สูงอายุ

2. สามารถนำไปพัฒนาต่อเป็นระบบรู้จำเสียงพูดเพื่อควบคุมอุปกรณ์ไฟฟ้าจริงได้

3. ช่วยยกระดับคุณภาพชีวิตของผู้สูงอายุและลดภาระของผู้ดูแลได้

4. ได้รับการตีพิมพ์ผลงานในการประชุมวิชาการเครือข่ายวิศวกรรมไฟฟ้า ครั้งที่ 7 (EENET 2015) [11]

5. ได้รับการตีพิมพ์ผลงานในการประชุมวิชาการงานวิจัยและพัฒนาเชิงประยุกต์ ครั้งที่ 7 (ECTI-CARD 2015) [12]

บทที่ 2
ทฤษฎีและการทบทวนวรรณกรรมวิจัยที่เกี่ยวข้อง

บทนี้จะกล่าวถึงกรอบแนวความคิดของการทำวิจัย จากนั้นจะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง ทั้งนี้จะมุ่งเน้นกล่าวถึงคุณลักษณะเด่นของสัญญาณเสียงพูดและวิธีการตัดสินใจรู้จำหรือการรู้จำในงานวิจัยที่ผ่านมา

2.1 กรอบแนวความคิด

กรอบแนวความคิดของโครงการวิจัยเกิดจากการสังเกตเห็นว่า เสียงพูดของผู้สูงอายุให้คุณลักษณะเฉพาะที่แตกต่างจากเสียงพูดของหนุ่ม-สาว และ เสียงพูดสำเนียงภาคใต้ให้โทนเสียงที่แตกต่างออกไปจากสำเนียงภาคกลาง ขณะที่การทบทวนวรรณกรรมเกี่ยวกับการรู้จำเสียงพูดภาษาไทยที่ผ่านมาส่วนใหญ่ [2]-[7] มุ่งศึกษาและทดสอบเฉพาะเสียงพูดของหนุ่ม-สาวและเป็นเสียงพูดสำเนียงภาคกลางแทบทั้งสิ้น ซึ่งระบบรู้จำตามกรอบการทดสอบดังกล่าวนั้นอาจไม่สามารถรองรับสำหรับเสียงพูดของผู้สูงอายุที่พูดสำเนียงภาคใต้ได้ สมมติฐานเบื้องต้นจึงกล่าวว่า "สำหรับคำพูดเดียวกันแล้วเสียงพูดภาษาไทยสำเนียงภาคใต้ของผู้สูงอายุให้ความถี่เด่น (formant frequency) ที่แตกต่างจากเสียงพูดสำเนียงภาคกลางของหนุ่ม-สาว" จึงได้ทำการศึกษาและวิเคราะห์เสียงพูดสำเนียงภาคใต้ของผู้สูงอายุ โดยจะทำการศึกษาคุณลักษณะของสัญญาณทั้งใน time domain และ frequency domain เพื่อหาคุณลักษณะเด่น (feature) ที่เหมาะสมสำหรับการรู้จำต่อไป

2.2 ทฤษฎีและการทบทวนวรรณกรรมที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับการรู้จำเสียงพูดภาษาไทยที่ผ่านมาส่วนใหญ่ใช้ภาษาไทยมาตรฐานสำเนียงภาคกลางแทบทั้งสิ้น [2]-[7] เมื่อเร็วๆ นี้ S. Aunkaew [8] ได้พัฒนาชุดข้อมูล (corpus) สำหรับเสียงพูดภาษาไทยสำเนียงภาคใต้ขึ้นแต่ไม่ได้เสนองานวิจัยในเชิงของการรู้จำ มีเพียงบทความวิชาการล่าสุดของผู้วิจัยเอง [9] ที่เสนอการรู้จำเสียงพูดคำสั่งควบคุมอุปกรณ์ไฟฟ้าทั้งภาษาไทยมาตรฐานสำเนียงภาคกลางและภาคใต้ แต่ก็เป็นเพียงการศึกษาเบื้องต้นที่ใช้

เสียงพูดจากผู้หญิงวัยกลางคนเพียงคนเดียว และคุณลักษณะเด่นที่ใช้ก็มีเพียงความยาวของเสียงพูด และสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 เท่านั้น

2.2.1 คุณลักษณะเด่นของสัญญาณเสียงพูด

คุณลักษณะเด่นของสัญญาณเสียงพูดสามารถพิจารณาได้ 3 กลุ่มหลัก [7]

1. กลุ่มค่าคุณลักษณะเด่นระดับสูง (High level feature) ได้แก่ สำเนียงการพูด รูปแบบในการพูด และความเร็วในการพูด เป็นต้น

2. กลุ่มค่าคุณลักษณะเด่นทางฉันทลักษณ์ (Prosodic feature) เช่น ค่าความถี่มูลฐาน (Fundamental frequency) และ ความถี่ฟอร์แมนต์ (Formant frequency) เป็นต้น

3. กลุ่มค่าคุณลักษณะเด่นแบบเอนเวโลปของสเปกตรัม (Spectrum envelop feature) เช่น สัมประสิทธิ์การประมาณพันธะเชิงเส้น (Linear prediction coefficients: LPC) และ สัมประสิทธิ์เซปสตรัล (Cepstral coefficient) เป็นต้น

สำหรับคุณลักษณะเด่นของสัญญาณเสียงพูดในงานวิจัยที่ผ่านมา ได้แก่ สัมประสิทธิ์การประมาณพันธะเชิงเส้น (Linear prediction coefficients: LPC) [2],[3],[9], Line Spectral Pairs Coefficients (LSP) [2] และ ความถี่ฟอร์แมนต์ (Formant frequency) [3] เป็นต้น

2.2.2 วิธีการตัดสินใจรู้จำหรือการรู้จำ (recognition)

การตัดสินใจรู้จำเสียงพูดในงานวิจัยที่ผ่านมาจะใช้วิธีการหรือเทคนิคหลายชนิดเข้ามาช่วย อาทิเช่น Dynamic Time Warping (DTW) [2],[5],[6], Artificial Neural Networks (ANNs) [3] และ Hidden Markov Models (HMMs) [4],[7] เป็นต้น

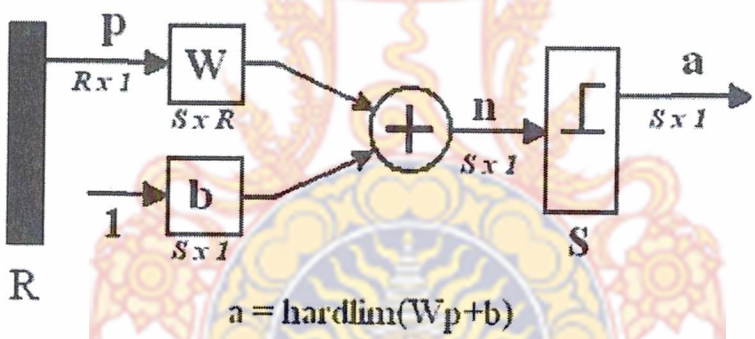
2.2.2.1 Dynamic Time Warping (DTW)

DTW [10] เป็นขั้นตอนวิธีสำหรับการเปรียบเทียบความคล้ายของลำดับที่มีความแตกต่างกันในด้านเวลาหรือความเร็ว สามารถใช้ DTW เพื่อจัดการกับคำพูดที่มีความเร็วไม่เท่ากันแม้จะสื่อความหมายเดียวกัน โดยทั่วไป DTW เป็นวิธีที่สามารถหาการจับคู่ที่เหมาะสมของลำดับสองชุดได้ภายใต้ข้อจำกัด ลำดับเหล่านั้นจะถูกบิดเบือน (warp) แบบไม่คงที่ในหน่วยของเวลา เพื่อที่จะพิจารณาความคล้ายจากการกระจายแบบไม่คงที่ในหน่วยของเวลา โดยจะ

ให้ผลลัพธ์ออกมาเป็นระยะทางและวิถีการปรับแนว (alignment) ที่ดีสุด งานวิจัยของเกรียงไกร เหลืองอำพล [2], P. Thong-in [5] และ R.Boonsin [6] ใช้ DTW ในการตัดสินใจรู้จำ

2.2.2.2 Artificial Neural Networks (ANNs)

ANNs หรือ โครงข่ายประสาทเทียม [11] คือ แบบจำลองทางคณิตศาสตร์สำหรับประมวลผลข้อมูลด้วยการคำนวณแบบคอนเนคชันนิสต์ (connectionist) แนวคิดเริ่มต้นของเทคนิคนี้ได้มาจากการศึกษาโครงข่ายไฟฟ้าชีวภาพ (bioelectric network) ในสมอง ซึ่งประกอบด้วย เซลล์ประสาท (neurons) และ จุดประสานประสาท (synapses) ตามแบบจำลองนี้ข่ายงานประสาทเกิดจากการเชื่อมต่อระหว่างเซลล์ประสาทจนเป็นเครือข่ายที่ทำงานร่วมกัน ตัวอย่างของโครงข่ายประสาทเทียมอย่างง่าย ได้แก่ Perceptron Neural Network ซึ่งเป็นโครงข่ายประสาทเทียมที่ประกอบด้วยชั้นซ่อนอย่างน้อยหนึ่งชั้นซ่อน ดังรูปที่ 2.1



รูปที่ 2.1 Perceptron Neural Network [12]

การเรียนรู้แบบ supervised learning ของโครงข่ายชนิดนี้ เกิดจากข้อมูลสอน (p) แต่ละชุดที่ป้อนให้โครงข่ายจะถูกคำนวณกับค่าน้ำหนัก (W) และไบอัส (b) ที่กำหนดไว้ก่อนตาม transfer function ที่กำหนด จากนั้นผลลัพธ์ที่ได้จากข้อมูลสอนแต่ละชุดถูกนำมาเปรียบเทียบกับผลลัพธ์ที่ต้องการ เมื่อมีความต่าง (error) ของผลลัพธ์ทั้งสองเกินค่าที่ยอมรับได้ ระบบจะปรับค่าน้ำหนักไปเรื่อยๆ ตามผลที่ได้จากข้อมูลสอนแต่ละชุดจนกระทั่งได้ความต่างที่ยอมรับได้หรือครบจำนวนรอบการทำซ้ำที่กำหนดไว้ ผลที่ได้จากการเรียนรู้จะเป็นฟังก์ชันการตัดสินที่ใช้ทำนายผลลัพธ์ที่ถูกต้อง งานวิจัยของปฏิมากร กิ้มสวัสดิ์[3] ใช้ ANNs ในการตัดสินใจรู้จำ

2.2.2.3 Hidden Markov Models (HMMs)

HMMs [7] เป็นวิธีการจำแนกรูปแบบโดยอาศัยวิธีการทางสถิติในการรู้จำ ระบบจะทำการเก็บรวบรวมรายละเอียดทางสถิติเกี่ยวกับเสียงพูด โดยเก็บข้อมูล การกระจายที่สมบูรณ์ของลักษณะสำคัญของเสียงไว้ในข้อมูลฝึกฝน เมื่อมีการทดสอบกับชุดทดสอบก็จะสามารถจำแนกความแตกต่างระหว่างเสียงพูดได้เป็นอย่างดี งานวิจัยของ M. Karnjanadecha [4] และนลินรัตน์ วิศวกิตติ [7] ใช้ HMMs ในการตัดสินใจรู้จำ

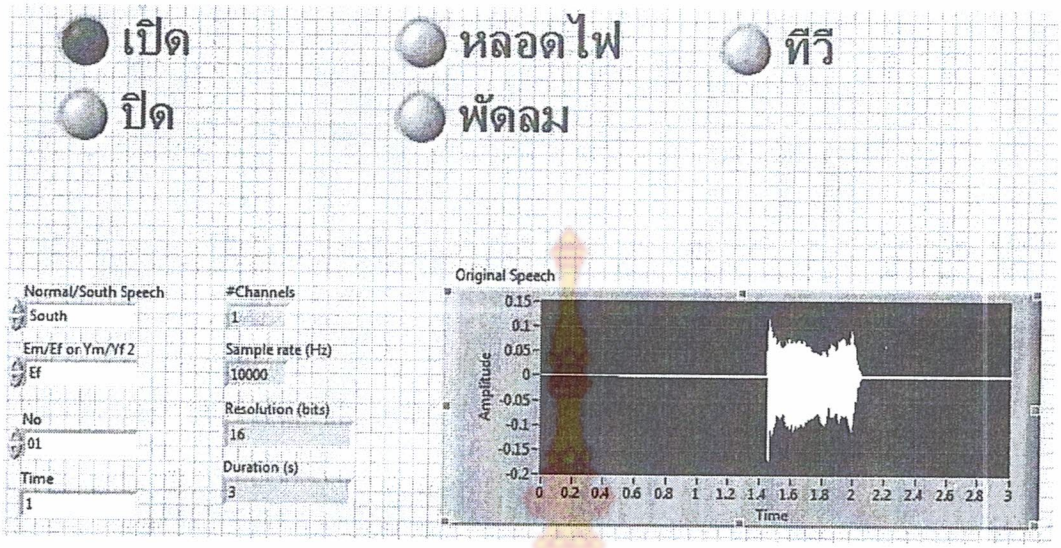
บทที่ 3
วิธีการดำเนินงาน

บทนี้จะกล่าวถึงวิธีการดำเนินงานเริ่มตั้งแต่การบันทึกเสียง วิธีการวิเคราะห์และคุณลักษณะเด่นของเสียงพูด กระบวนการ Discrete Wavelet Transform วิธีการตัดสินใจรู้จำโดยใช้โครงข่ายประสาทเทียม และการคิดคำนวณประสิทธิภาพการรู้จำ จากนั้นจะกล่าวถึงการทดลองการรู้จำที่แบ่งศึกษาเป็น 2 การทดลอง คือ การทดลอง 1 เป็นการศึกษาเพื่อดูว่าคุณลักษณะเด่นเชิงเวลาร่วมกับคุณลักษณะเด่นเชิงความถี่สามารถช่วยในการรู้จำเสียงพูดได้และการทดลองที่ 2 เป็นศึกษาเพื่อดูว่าการใช้กระบวนการ DWT สามารถช่วยเพิ่มประสิทธิภาพการรู้จำเสียงพูดจากการใช้คุณลักษณะเด่นจำนวนพยางค์และสัมประสิทธิ์การประมาณพันธะเชิงเส้นที่ผ่าน DWT

3.1 การบันทึกเสียง

บันทึกเสียงพูดของผู้สูงอายุดังแสดงตัวอย่างในรูปที่ 3.1 ด้วยโปรแกรมที่เขียนขึ้นจากโปรแกรม LabVIEW ดังรูปที่ 3.2 ทำการบันทึกเสียงพูดสำเนียงภาคใต้แบบ mono ด้วยอัตราการสุ่มเท่ากับ 10 kHz เป็นเวลาคำสั่งละ 3 s ความละเอียด 16 bits ผู้พูดมีทั้งเพศชายและหญิงเพศละ 2 คนต่อ 1 จังหวัด ได้แก่ สงขลา พัทลุง ตรัง และนครศรีธรรมราช ทำให้ได้จำนวนผู้พูด 16 คน คำสั่งที่บันทึกมีทั้งหมด 5 คำสั่ง ได้แก่ เปิด ปิด หลอดไฟ พัดลม และทีวี พูดซ้ำคำสั่งละ 4 ครั้ง ทำให้ได้ข้อมูลเสียงพูดทั้งหมด 320 ข้อมูล (16 คน x 5 คำสั่ง x 4 ครั้ง)

รูปที่ 3.1 ตัวอย่างผู้สูงอายุบันทึกเสียงพูดในสภาพแวดล้อมที่พักอาศัย



รูปที่ 3.2 โปรแกรมบันทึกเสียงพูด

3.2 วิธีการวิเคราะห์และคุณลักษณะเด่นของเสียงพูด

เสียงที่บันทึกถูกนำมาผ่านกระบวนการหาจุดเริ่มต้นและสิ้นสุดของสัญญาณเสียงด้วยวิธีการคอนโวลูชั่น (Convolution) แล้วนำมาวิเคราะห์ทั้งใน time domain และ frequency domain เพื่อหาคุณลักษณะเด่นของเสียงพูดแต่ละคำด้วยโปรแกรมที่เขียนขึ้นจาก MATLAB

3.2.1 Convolution

การคอนโวลูชั่น (Convolution) เป็นกระบวนการที่สามารถแสดงเค้าโครงรูปร่าง (envelop) หรือขอบเขตเฉพาะเสียงพูด ทำให้สามารถระบุจุดเริ่มต้นและสิ้นสุดของเสียงพูดได้ มีวิธีคำนวณตามสมการที่ 1

$$c(n) = \sum_{i=0}^{n} |s(i)|.w(n-i) \qquad (1)$$

เมื่อ

$c(n)$ คือ ค่าคอนโวลูชั่น

$s(i)$ คือ ข้อมูลสัญญาณเสียงพูดลำดับที่ i

w คือ หน้าต่างชนิดสี่เหลี่ยมขนาด 40 ms

n คือ จำนวนค่าคอนโวลูชั่นทั้งหมด

ทั้งนี้จุดเริ่มต้นและจุดสิ้นสุดของเสียงพูดมีค่าเท่ากับจุดแรกที่ให้ค่าคอนโวลูชั่นมากกว่า 0.1 เท่าของค่าสูงสุดนับจากต้นเสียงและนับจากปลายเสียง ตามลำดับ

3.2.2 คุณลักษณะเด่นเชิงเวลา

คุณลักษณะเด่นเชิงเวลาที่ใช้ในงานวิจัยได้แก่ ระยะเวลาของเสียงพูด (Duration of Sound: DS) จำนวนพยางค์ (Number of Syllable: NS) และการประมาณพันธะเชิงเส้น (Linear Predictive Coefficient: LPC)

ระยะเวลาของเสียงพูด

คือ ผลต่างระหว่างจุดเริ่มต้นและจุดสิ้นสุดของสัญญาณเสียง

จำนวนพยางค์

จากระยะเวลาของเสียงพูดโดยแบ่งระยะเวลาออกเป็น 3 เฟรมเท่าๆ กัน พิจารณาเฟรมตรงกลาง หากเสียงพูดที่เฟรมตรงกลางมีค่าคอนโวลูชั่นน้อยกว่า 20% ของค่าคอนโวลูชั่นสูงสุดเป็นระยะเวลามากกว่า 5% ของระยะเวลาของเสียงพูด กำหนดให้สัญญาณเสียงนั้นมี 2 พยางค์

การประมาณพันธะเชิงเส้น (Linear Predictive Coefficient: LPC) เป็นกระบวนการหาค่าสัมประสิทธิ์ของ forward linear predictor โดยพิจารณาว่าเสียงเกิดจากผลรวมเชิงเส้นของสัญญาณที่ทราบค่าแล้วก่อนหน้าจำนวน p ค่า ดังสมการที่ 2[10] งานวิจัยนี้จะใช้การประมาณพันธะเชิงเส้นอันดับ 3 ดังสมการที่ 2

$$\hat{x}(n) = -a(2)x(n-1) - a(3)x(n-2) - \ldots - a(p+1)x(n-p) \qquad (2)$$

เมื่อ

$\hat{x}(n)$ คือ สัญญาณค่าถัดไปที่ทำนาย

$x(n)$ คือ สัญญาณที่ทราบค่าแล้ว

a คือ ค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น

3.2.3 คุณลักษณะเด่นเชิงความถี่

คุณลักษณะเด่นเชิงความถี่ที่ใช้ในงานวิจัย ได้แก่ จำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ (FG)

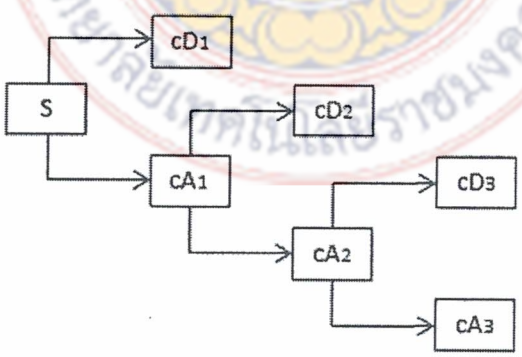
จำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ (FG)

จากเสียงพูดที่ผ่านวิธีการ Fast Fourier Transform (FFT) ทำให้ได้ความถี่ของเสียงที่นำมาพิจารณาในช่วง 0-5000 Hz (ความถี่สุ่มเท่ากับ 10 kHz) ช่วงความถี่ถูกแบ่งออกเป็น 10 กลุ่ม (G) โดยที่ G1 มีความถี่ช่วง (0,500] , G2 มีความถี่ช่วง (500,1000], G3 มีความถี่ช่วง (1000,1500], ..., และ G10 มีความถี่ช่วง (4500,5000]

จุดยอดของความถี่เด่น พิจารณาจาก จุดยอดของขนาดความถี่ภายในหน้าต่างสี่เหลี่ยมขนาดเท่าช่วงความถี่ 500 Hz โดยที่จุดยอดนั้นจะต้องที่มีค่ามากที่สุดภายในหน้าต่างและมากกว่าค่าเฉลี่ยของขนาดความถี่ตลอดช่วงความถี่ทั้งหมด 0-5000 Hz หน้าต่างนี้จะถูกกำหนดให้เริ่มต้นจากความถี่ 0 Hz แล้วเลื่อนไปตลอดช่วงความถี่ทั้งหมดทีละ 250 Hz (หน้าต่างซ้อนทับกัน 50%)

3.3 กระบวนการ Discrete Wavelet Transform (DWT)

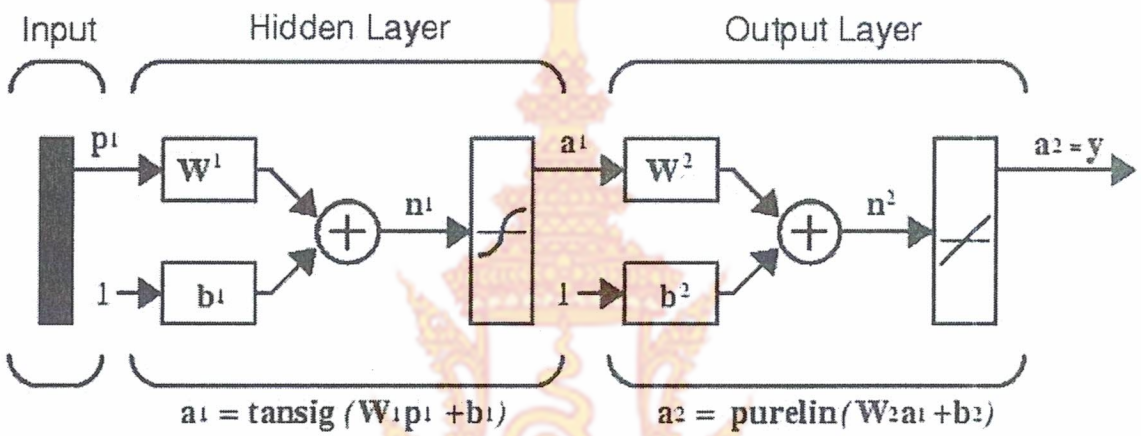
DWT เป็นกระบวนการที่ให้สัญญาณผ่าน filter 2 ชนิด คือ Digital low-pass filter และ Digital high-pass filter แล้ว Down Sampling ลง 2 เท่า ทำให้ได้องค์ประกอบสัญญาณความถี่ต่ำ (Approximation: cA1) และ องค์ประกอบสัญญาณความถี่สูง (Detail: cD1) ของ Wavelet Decomposition ในระดับที่ 1 ต่อมาองค์ประกอบสัญญาณความถี่ต่ำ (Approximation) ยังสามารถถูกแยกในระดับต่อๆไปด้วยกระบวนการเดิม ทำให้ได้องค์ประกอบสัญญาณความถี่ต่ำและสูงในระดับต่างๆ ดังแสดงในรูปที่ 3.3 งานวิจัยนี้ใช้ mother wavelet แบบ Daubechies [10]



รูปที่ 3.3 แผนภาพ Wavelet Decomposition Tree[12]

3.4 โครงข่ายประสาทเทียม (ANNs)

โครงข่ายประสาทเทียมที่ใช้ในงานวิจัยนี้เป็นชนิด Backpropagation จำนวน 1 ชั้นซ่อน โครงสร้างเป็นแบบ Multilayer Feedforward มีจำนวนอินพุตขึ้นอยู่กับแต่ละการทดลอง จำนวนโหนดในชั้นซ่อนเท่ากับ 5 โหนด ทรานเฟอร์ฟังก์ชันในชั้นซ่อนเป็น tansig และชั้นเอาต์พุตเป็น purelin มีโครงสร้างดังรูปที่ 3.4



รูปที่ 3.4 โครงสร้างของ Multilayer Feedforward Network[11]

3.5 ประสิทธิภาพการรู้จำ

ประสิทธิภาพการรู้จำหาได้จากเปอร์เซ็นต์ความถูกต้องของการรู้จำ (Percent Recognition: PR) ตามสมการที่ 3

$$PR = \frac{TR}{TR + FR} * 100 \qquad (3)$$

เมื่อ

PR (Percent Recognition)	คือ เปอร์เซ็นต์ความถูกต้องของการรู้จำ
TR (TrueRecognition)	คือ จำนวนการรู้จำที่ถูกต้อง
FR (False Recognition)	คือ จำนวนการรู้จำที่ไม่ถูกต้อง

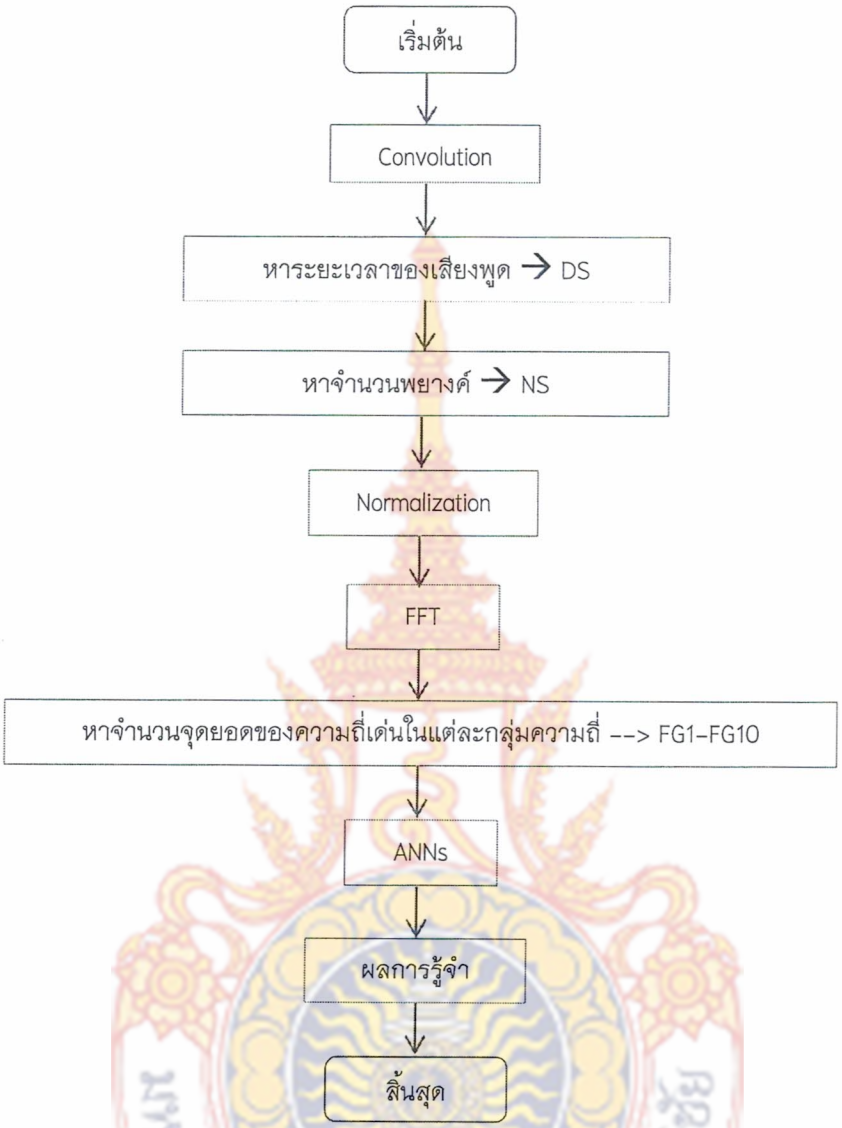
3.6 การทดลอง

ทำการทดลองเพื่อศึกษาคุณลักษณะเด่นที่ใช้ในการรู้จำเสียงพูดและผลของการใช้กระบวนการ DWT ร่วมกับคุณลักษณะเด่นที่ใช้รู้จำเสียงพูด ทำให้แบ่งการทดลองหลักออกเป็น 2 การทดลอง แต่ใช้วิธีการตัดสินใจรู้จำชนิดเดียวกัน คือ ใช้โครงข่ายประสาทเทียมดังที่กล่าวในหัวข้อ 3.4

3.6.1 การทดลองที่ 1

เป็นการทดลองเพื่อศึกษาว่า "*คุณลักษณะเด่นเชิงเวลาร่วมกับคุณลักษณะเด่นเชิงความถี่สามารถช่วยในการรู้จำเสียงพูดได้*" คุณลักษณะเด่นเชิงเวลาที่ใช้ คือ ระยะเวลาของเสียงพูด (DS) และจำนวนพยางค์ (NS) ส่วนคุณลักษณะเด่นเชิงความถี่ที่ใช้ คือ จำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ (FG) การทดลองมีกระบวนการทำงานดังแสดงในแผนภาพรูปที่ 3.5[11]

กระบวนการทำงานเริ่มจากการคอนโวลูชั่นสัญญาณเสียงพูดตลอดช่วงที่บันทึกกับหน้าต่างชนิดสี่เหลี่ยมขนาด 40 ms เพื่อหาขอบเขตเฉพาะเสียงพูด ค่าคอนโวลูชั่นจะถูกกำหนดจุดเริ่มต้นและสิ้นสุดเพื่อหาระยะเวลาของเสียงพูด (DS) เสียงพูดที่ได้จะผ่านกระบวนการพิจารณาเฟรมตรงกลางเพื่อหาจำนวนพยางค์ (NS) เป็นลำดับต่อไป จากนั้นเสียงพูดจะถูกนอร์มอลไลซ์ (normalize) แล้วเข้าสู่กระบวนการ FFT และหาจำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ (FG1–FG10) ทำให้ได้คุณลักษณะเด่นแทนเสียงพูด (feature) หรือจำนวนอินพุตสำหรับโครงข่ายประสาทเทียมทั้งหมด 12 ค่า (DS, NS, FG1–FG10) อินพุตทั้งหมดถูกนำไปฝึกฝน (train) หรือทดสอบ (test) กับโครงข่ายประสาทเทียมแล้วแสดงผลการรู้จำต่อไป ส่วนหนึ่งของการทดลองนี้ได้รับการตีพิมพ์ในการประชุมวิชาการเครือข่ายวิศวกรรมไฟฟ้า ครั้งที่ 7 (EENET 2015) [11]



รูปที่ 3.5 แผนภาพกระบวนการทำงาน

3.6.2 การทดลองที่ 2

เป็นการทดลองเพื่อศึกษาว่า "*การใช้กระบวนการ Discrete Wavelet Transform (DWT) สามารถช่วยเพิ่มประสิทธิภาพการรู้จำเสียงพูดจากการใช้คุณลักษณะเด่นจำนวนพยางค์และสัมประสิทธิ์การประมาณพันธะเชิงเส้นที่ผ่าน DWT*" กระบวนการทำงานเริ่มจากคอนโวลูชั่นสัญญาณเสียงพูดตลอดช่วงที่บันทึกกับหน้าต่างชนิดสี่เหลี่ยมเพื่อหาขอบเขตเฉพาะเสียงพูด

แล้วเข้าสู่กระบวนการหาจำนวนพยางค์ จากนั้นเสียงพูดจะถูกนอร์มอลไลซ์ (normalize) แล้วหาคุณลักษณะเด่น 3 ขั้นตอนหลัก คือ

1. หาสัมประสิทธิ์การประมาณพันธะเชิงเส้น (LPC)

2. ผ่านกระบวนการ DWT 3 ระดับ ทำให้ได้สัญญาณที่เป็นองค์ประกอบความถี่ต่ำ cA1, cA2 และ cA3 และได้สัญญาณที่เป็นองค์ประกอบความถี่สูง cD1, cD2 และ cD3

3. หาสัมประสิทธิ์การประมาณพันธะเชิงเส้น (LPC) จากสัญญาณองค์ประกอบความถี่ต่ำและองค์ประกอบความถี่สูงที่ผ่านกระบวนการ DWT ที่ระดับต่างๆ 3 ระดับ (cA1_LPC, cA2_LPC, cA3_LPC, cD1_LPC, cD2_LPC, cD3_LPC)

อินพุตที่ทดสอบเป็นคุณลักษณะเด่นที่ได้จากจำนวนพยางค์ (NS) ร่วมกับสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 ของสัญญาณ (LPC) หรือสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 ขององค์ประกอบความถี่ต่ำและองค์ประกอบความถี่สูงที่ผ่านกระบวนการ DWT ที่ระดับต่างๆ ทำให้แบ่งคุณลักษณะเด่นออกได้เป็น 7 กรณี คือ

1) NS + LPC
2) NS + cA1_LPC
3) NS + cA2_LPC
4) NS + cA3_LPC
5) NS + cD1_LPC
6) NS + cD2_LPC
7) NS + cD3_LPC

จากจำนวนคุณลักษณะเด่นที่ทดสอบในแต่ละกรณี ทำให้ได้จำนวนอินพุตสำหรับโครงข่ายประสาทเทียมกรณีละ 2 ค่า อินพุตทั้งหมดถูกนำไปฝึกฝน (train) หรือทดสอบ (test) กับโครงข่ายประสาทเทียมแล้วแสดงผลการรู้จำต่อไป ส่วนหนึ่งของการทดลองนี้ได้รับการตีพิมพ์ผลงานในการประชุมวิชาการงานวิจัยและพัฒนาเชิงประยุกต์ ครั้งที่ 7 (ECTI-CARD 2015) [12]

ในการทดสอบการรู้จำด้วยโครงข่ายประสาทเทียมสำหรับทั้งสองการทดลอง ข้อมูลจะถูกแบ่งออกเป็น 2 ส่วนเท่ากัน คือ สัญญาณเสียงพูดครั้งที่ 1-2 เป็นข้อมูลฝึกฝนและสัญญาณเสียงพูดครั้งที่ 3-4 เป็นข้อมูลทดสอบ ทำให้ได้ข้อมูลสำหรับฝึกฝน 160 ข้อมูล (จำนวนผู้พูด 16

คน x 5 คำสั่ง x 2 ครั้ง) และข้อมูลทดสอบ 160 ข้อมูล (จำนวนผู้พูด 16 คน x 5 คำสั่ง x 1 ครั้ง) เปอร์เซ็นต์ความถูกต้องของการรู้จำคำนวณจากค่าเฉลี่ยจำนวน 3 ครั้ง เนื่องจากค่าน้ำหนักที่ลู่เข้าของโครงข่ายประสาทเทียมในการฝึกฝนแต่ละครั้งมีค่าไม่เท่ากัน

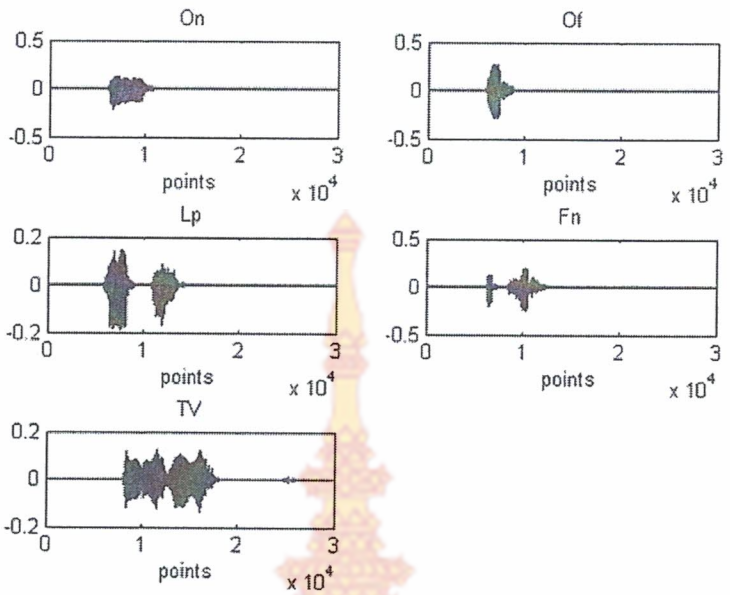
บทที่ 4

ผลการทดลอง

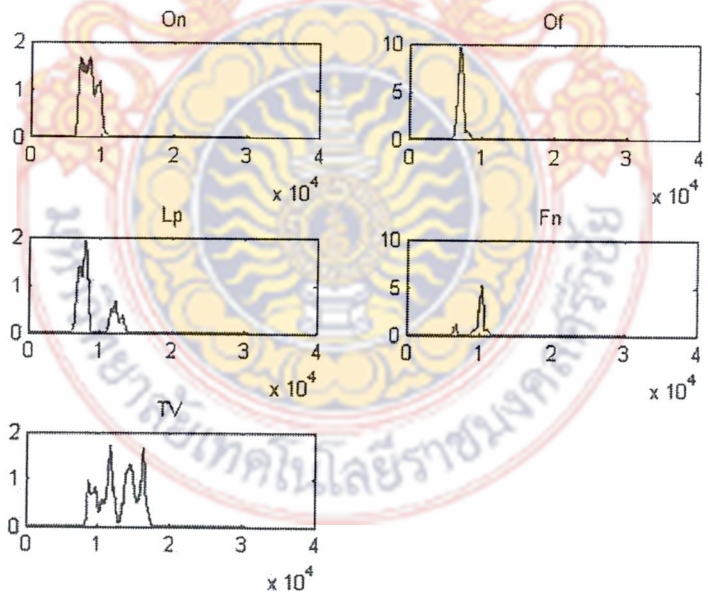
บทนี้จะแสดงผลของรูปแบบสัญญาณเสียงพูดแต่ละคำ ค่าคอนโวลูชั่น ตำแหน่งเริ่มต้นและสิ้นสุดของสัญญาณเสียงพูด รวมทั้งผลการพิจารณาจำนวนพยางค์ จากนั้นจึงแสดงผลการทดลองที่ 1 ที่ศึกษาว่าคุณลักษณะเด่นเชิงเวลาร่วมกับคุณลักษณะเด่นเชิงความถี่สามารถช่วยในการรู้จำเสียงพูดได้ และผลการทดลองที่ 2 ที่ศึกษาว่าการใช้กระบวนการ DWT สามารถช่วยเพิ่มประสิทธิภาพการรู้จำเสียงพูดจากการใช้คุณลักษณะเด่นจำนวนพยางค์และสัมประสิทธิ์การประมาณพันธะเชิงเส้นที่ผ่าน DWT

กำหนดให้เสียงพูดคำสั่งต่างๆ แทนด้วยสัญลักษณ์ต่อไปนี้ เปิด (On), ปิด (Of), หลอดไฟ (Lp), พัดลม (Fn) และทีวี (TV) รูปที่ 4.1 แสดงตัวอย่างสัญญาณเสียงพูดเปิด, ปิด, หลอดไฟ, พัดลมและทีวี ที่ถูกบันทึกนาน 3 s (อัตราการสุ่ม 10 kHz, 30000 points) จากรูปจะเห็นได้ว่าเสียงพูดแต่ละคำสั่งให้ลักษณะรูปร่างสัญญาณเสียงใน time domain ที่แตกต่างกัน เสียงพูดที่มี 1 พยางค์จะให้สัญญาณเสียงที่มองเห็นเป็น 1 กลุ่มอย่างชัดเจน ขณะที่เสียงพูดที่มี 2 พยางค์จะให้สัญญาณเสียงที่มองเห็นเป็น 2 กลุ่มที่ติดหรือแยกกันแม้ว่าเสียงพูดคำว่าพัดลมอาจเห็นเป็นเพียงกลุ่มขนาดเล็กในช่วงแรก

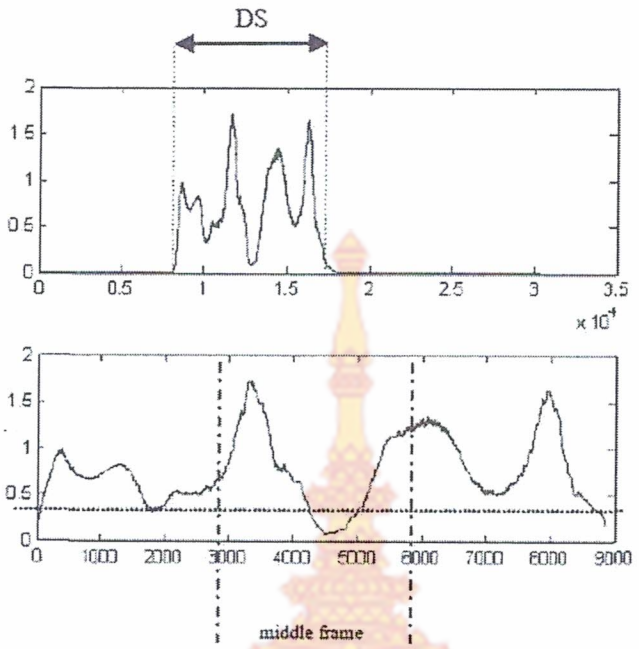
รูปที่ 4.2 แสดงค่าคอนโวลูชั่นของเสียงพูดที่สอดคล้องกับสัญญาณในรูปที่ 4.1 จะเห็นได้ว่าค่าคอนโวลูชั่นสามารถแสดงขอบเขตเฉพาะเสียงพูดได้ จุดเริ่มต้นและจุดสิ้นสุดของสัญญาณเสียงสามารถคำนวณได้จากการตรึงค่ากำหนด (threshold) ตามเงื่อนไขในข้อ 3.2.1 ทำให้ได้ระยะเวลาของเสียงพูด (DS) ซึ่งเป็นผลต่างระหว่างจุดสิ้นสุดและจุดเริ่มต้น ดังแสดงในรูปที่ 4.3a และเมื่อพิจารณาเฟรมตรงกลางของเสียงพูดที่มีค่าคอนโวลูชั่นน้อยกว่า 20% ของค่าคอนโวลูชั่นสูงสุดเป็นระยะเวลามากกว่า 5% ของระยะเวลาของเสียงพูด ทำให้พิจารณาได้ว่าสัญญาณเสียงนั้นมี 2 พยางค์ แสดงเฟรมตรงกลางของเสียงพูดและเส้นแสดงค่าคอนโวลูชั่น 20% ของค่าคอนโวลูชั่นสูงสุดในรูปที่ 4.3b



รูปที่ 4.1 สัญญาณเสียงพูดคำว่า เปิด, ปิด, หลอดไฟ, พัดลม และทีวี



รูปที่ 4.2 ค่าคอนโวลูชั่นของเสียงพูดเปิด, ปิด, หลอดไฟ, พัดลม และทีวี

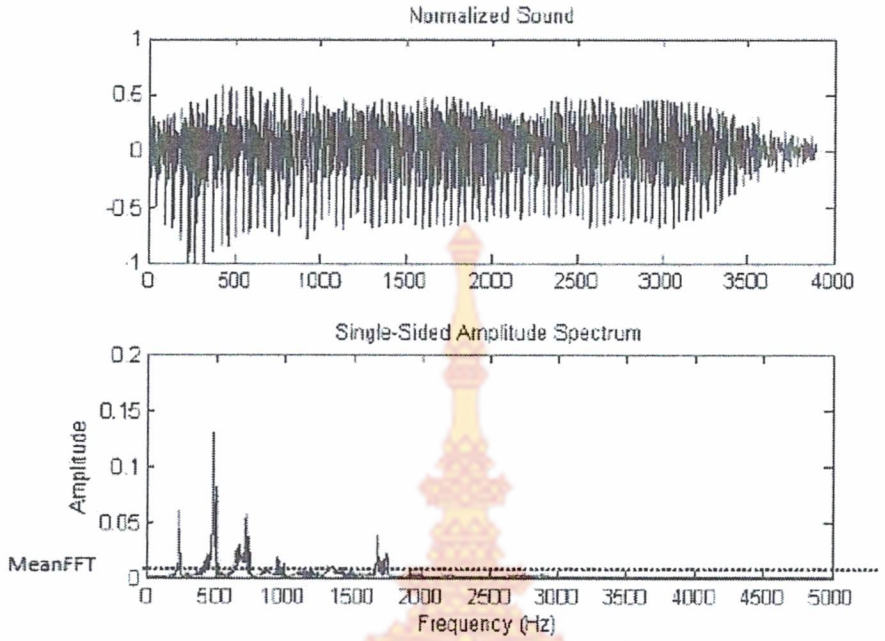


รูปที่ 4.3 a) ระยะเวลาของเสียงพูดของคำว่าทีวี

b) เฟรมตรงกลาง

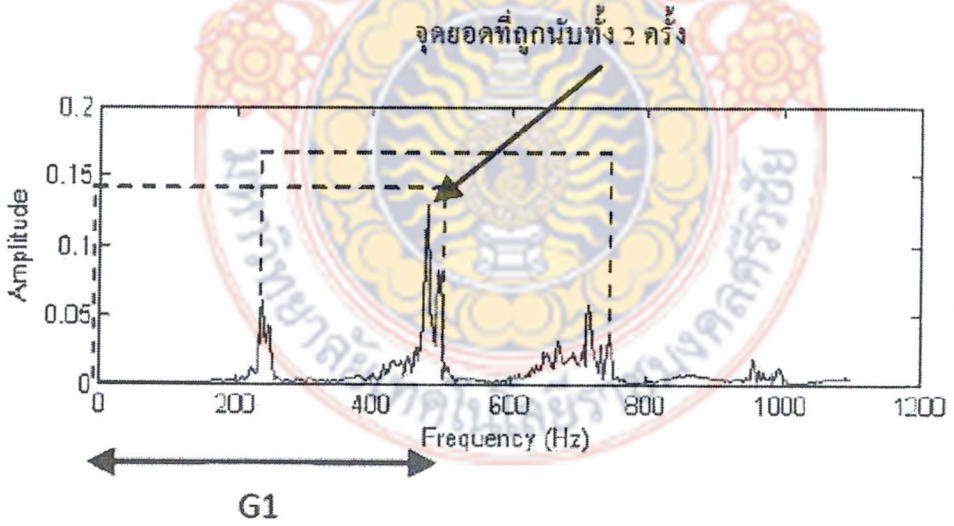
4.1 ผลการทดลองที่ 1

การทดลองที่ 1 เป็นการทดลองเพื่อศึกษาว่า "คุณลักษณะเด่นเชิงเวลาร่วมกับคุณลักษณะเด่นเชิงความถี่สามารถช่วยในการรู้จำเสียงพูดได้" คุณลักษณะเด่นเชิงเวลาที่ใช้ คือ ระยะเวลาของเสียงพูด (DS) และจำนวนพยางค์ (NS) มีวิธีการหาค่าและผลที่ได้ดังที่ได้กล่าวข้างต้น ส่วนคุณลักษณะเด่นเชิงความถี่ที่ใช้ คือ จำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ (FG) จากเสียงพูดที่ถูกนอร์มอลไลซ์ (normalize) แล้วเข้าสู่กระบวนการ FFT และหาจำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ (FG1-FG10) ทำให้ได้ตัวอย่างผลการทดลองดังรูปที่ 4.4 และ 4.5



รูปที่ 4.4 a) ตัวอย่างเสียงพูดที่ถูกนอร์มอลไลซ์ของคำว่าเปิด

b) ผลตอบสนองเชิงความถี่



รูปที่ 4.5 ผลตอบสนองเชิงความถี่ของคำว่าเปิดในช่วง 0–1100 Hz และหน้าต่างที่เลื่อนผ่านความถี่ในกลุ่มที่ 1

รูปที่ 4.4a แสดงตัวอย่างเสียงพูดที่ถูกนอร์มอลไลซ์ของคำว่าเปิด ส่วนรูปที่ 4.4b แสดงผลตอบสนองเชิงความถี่ของสัญญาณในรูปที่ 4.4a และค่าเฉลี่ย (MeanFFT) ของขนาดความถี่ตลอดช่วงความถี่ทั้งหมดมีค่าเท่ากับ 0.0029 ส่วนรูปที่ 4.5 แสดงตัวอย่างจำนวนจุดยอดของความถี่เด่นภายในหน้าต่างที่เลื่อนผ่านความถี่ในกลุ่มที่ 1 โดยที่หน้าต่างซ้อนทับกัน 50% ทำให้นับจุดยอดของความถี่เด่นในกลุ่มที่ 1 ได้ 2 ครั้งหรือจำนวนจุดยอดของความถี่เด่นในกลุ่มที่ 1 เท่ากับ 2 คุณลักษณะเด่นทั้ง 12 ค่า (DS, NS, FG1–FG10) ถูกนำไปใช้เป็นอินพุตสำหรับโครงข่ายประสาทเทียมเพื่อฝึกฝน (train) หรือทดสอบ (test) แล้วแสดงผลการรู้จำต่อไป

ผลการทดสอบเสียงพูดครั้งที่ 3–4 กับโครงข่ายประสาทเทียมสำหรับคุณลักษณะเด่นในเชิงเวลาและความถี่ดังกล่าว พบว่า ให้เปอร์เซ็นต์ความถูกต้องของการรู้จำ 3 ครั้งดังนี้

ครั้งที่ 1 ถูกต้อง 125 ข้อมูลจาก 160 ข้อมูล คิดเป็น 78.12 เปอร์เซ็นต์

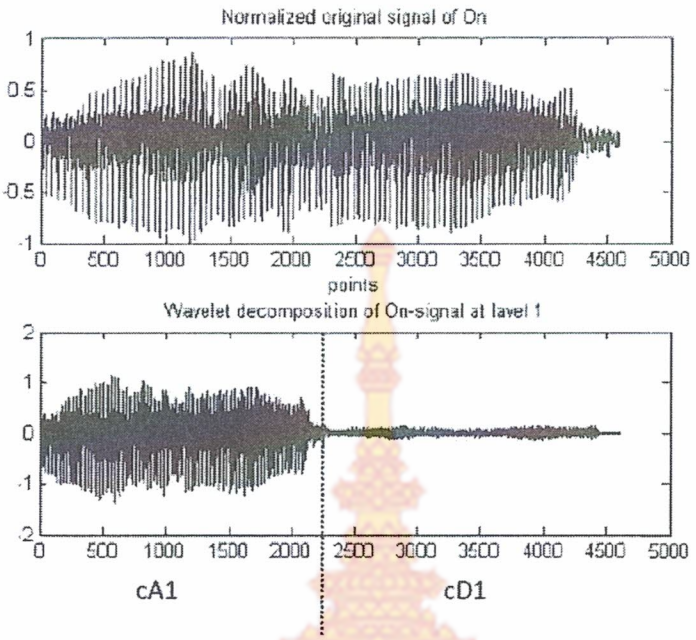
ครั้งที่ 2 ถูกต้อง 131 ข้อมูลจาก 160 ข้อมูล คิดเป็น 81.86 เปอร์เซ็นต์

ครั้งที่ 3 ถูกต้อง 129 ข้อมูลจาก 160 ข้อมูล คิดเป็น 80.63 เปอร์เซ็นต์

ทำให้ได้เปอร์เซ็นต์ความถูกต้องของการรู้จำเฉลี่ยสำหรับคุณลักษณะเด่นในเชิงเวลาและความถี่ดังกล่าว เท่ากับ 80.2 เปอร์เซ็นต์ เสียงพูดที่ผิดพลาดส่วนใหญ่เป็นคำว่า พัดลม และหลอดไฟ

4.2 ผลการทดลองที่ 2

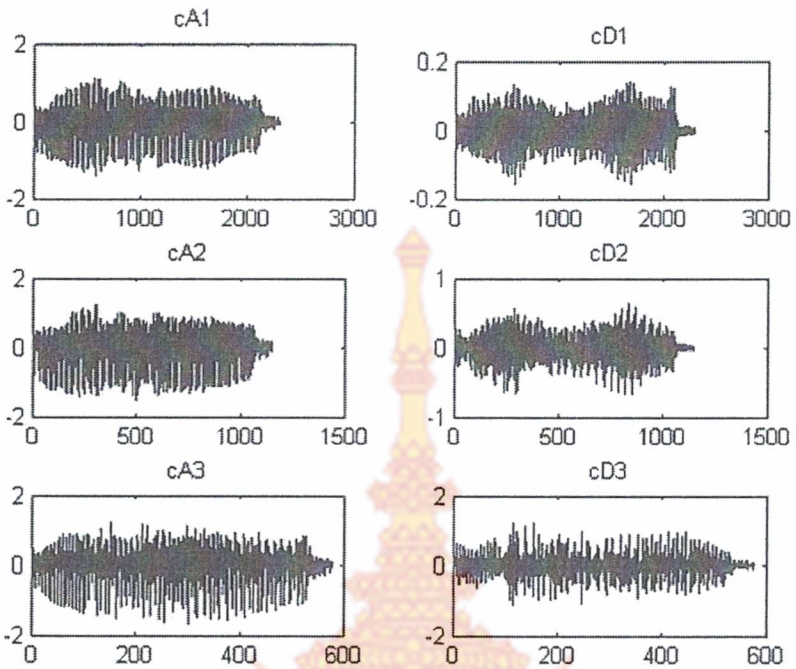
การทดลองที่ 2 เป็นการทดลองเพื่อศึกษาว่า “การใช้กระบวนการ DWT สามารถช่วยเพิ่มประสิทธิภาพการรู้จำเสียงพูดจากการใช้คุณลักษณะเด่นจำนวนพยางค์และสัมประสิทธิ์การประมาณพันธะเชิงเส้นที่ผ่าน DWT”



รูปที่ 4.6 a) เสียงพูดที่ถูกนอร์มอลไลซ์ของคำว่าเปิด

b) Wavelet Decomposition ที่ระดับ 1 ของสัญญาณใน 4.6a

รูปที่ 4.6a แสดงเสียงพูดที่ถูกนอร์มอลไลซ์ของคำว่าเปิด ส่วนรูปที่ 4.6b แสดง Wavelet Decomposition ที่ระดับ 1 ของสัญญาณในรูปที่ 4.6a ซึ่งประกอบไปด้วยองค์ประกอบความถี่ต่ำ (cA1) และองค์ประกอบความถี่สูง (cD1) การหา Wavelet Decomposition ทั้ง 3 ระดับของสัญญาณดังกล่าวแสดงองค์ประกอบความถี่ต่ำและสูงที่ระดับต่างๆ ได้ในรูปที่ 4.7 ส่วนผลการรู้จำที่เกิดจากค่าเฉลี่ยจำนวน 3 ครั้ง ของการทดสอบ 7 คุณลักษณะเด่นของเสียงพูดครั้งที่ 3–4 กับโครงข่ายประสาทเทียม พบว่า เปอร์เซ็นต์ความถูกต้องของการรู้จำแสดงในตารางที่ 1



รูปที่ 4.7 องค์ประกอบสัญญาณความถี่ต่ำ cA1, cA2, cA3 และองค์ประกอบสัญญาณความถี่สูง cD1, cD2, cD3 ของสัญญาณในรูปที่ 4.6a

ตารางที่ 4.1 เปอร์เซ็นต์ความถูกต้องของการรู้จำ (PR)

คุณลักษณะเด่น	จำนวนข้อมูลถูกต้อง			เปอร์เซนต์ความถูกต้อง			เปอร์เซนต์ความถูกต้องเฉลี่ย
	ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	
NS + LPC	128	124	127	80	77.5	79.375	78.96
NS + cA1_LPC	128	126	126	80	78.75	78.75	79.17
NS + cA2_LPC	132	136	134	82.5	85	83.75	83.75
NS + cA3_LPC	125	126	127	78.125	78.75	79.375	78.75
NS + cD1_LPC	128	128	129	80	80	80.625	80.21
NS + cD2_LPC	132	129	127	82.5	80.625	79.375	80.83
NS + cD3_LPC	128	132	130	80	82.5	81.25	81.25

จากผลการทดลองในตารางที่ 4.1 จะเห็นว่าสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 ที่ได้จากสัญญาณที่ผ่านกระบวนการ DWT ร่วมกับจำนวนพยางค์ (NS) .ให้ความถูกต้องในการรู้จำสูงกว่ากรณีสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 ที่ไม่ผ่านกระบวนการ DWT ดังนั้นการใช้กระบวนการ DWT ร่วมด้วยจึงช่วยเพิ่มประสิทธิภาพการรู้จำ

บทที่ 5

สรุปผลและข้อเสนอแนะ

บทนี้จะกล่าวถึงสรุปผลการทดลองทั้ง 2 การทดลอง สรุปผลรวม และข้อเสนอแนะ

5.1 สรุปผลการทดลองที่ 1

การทดลองที่ 1 เป็นการทดลองเพื่อศึกษาว่า "คุณลักษณะเด่นเชิงเวลาร่วมกับคุณลักษณะเด่นเชิงความถี่สามารถช่วยในการรู้จำเสียงพูดได้"

ระยะเวลาของเสียงพูด (DS) และจำนวนพยางค์ (NS) ซึ่งเป็นคุณลักษณะเด่นเชิงเวลาร่วมกับจำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ (FG) ซึ่งเป็นคุณลักษณะเด่นเชิงความถี่สามารถใช้เป็นคุณลักษณะเด่นในการรู้จำเสียงพูดสำเนียงภาคใต้ของคำสั่งเปิด ปิด หลอดไฟ พัดลม และทีวีได้ โดยคุณลักษณะเด่นดังกล่าวให้เปอร์เซ็นต์ความถูกต้องของการรู้จำเท่ากับ 80.2 เมื่อเปรียบเทียบกับงานวิจัยที่ใช้วิธีการเดียวกันก่อนหน้านี้[11] ซึ่งใช้เสียงจากผู้พูดเพียงคนเดียวซึ่งได้ความถูกต้องของการรู้จำเท่ากับ 86 เปอร์เซนต์ จะเห็นว่าเปอร์เซ็นต์ความถูกต้องของการรู้จำกรณีไม่ขึ้นกับผู้พูด (16 คน) ให้ค่าต่ำกว่าเปอร์เซ็นต์ความถูกต้องของการรู้จำกรณีขึ้นกับผู้พูด (1 คน) ทั้งนี้เป็นเพราะความหลากหลายของเสียงพูดที่เกิดจากต่างบุคคล นอกจากนี้ยังพบว่าข้อผิดพลาดของการรู้จำส่วนใหญ่อยู่ที่คำว่า หลอดไฟ และพัดลม

5.2 สรุปผลการทดลองที่ 2

การทดลองที่ 2 เป็นการทดลองเพื่อศึกษาว่า "การใช้กระบวนการ DWT สามารถช่วยเพิ่มประสิทธิภาพการรู้จำเสียงพูดจากการใช้คุณลักษณะเด่นจำนวนพยางค์และสัมประสิทธิ์การประมาณพันธะเชิงเส้นที่ผ่าน DWT"

สัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 ที่ได้จากสัญญาณเสียงพูดหรือจากองค์ประกอบความถี่ต่ำหรือสูงของ Wavelet Decomposition ที่ระดับ 1-3 ร่วมกับจำนวนพยางค์สามารถใช้เป็นคุณลักษณะเด่นสำหรับการรู้จำเสียงพูดสำเนียงภาคใต้ของคำสั่งเปิด ปิด หลอดไฟ พัดลม และทีวีได้ โดยที่สัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 ที่ได้จาก

สัญญาณที่ผ่านกระบวนการ DWT ร่วมกับจำนวนพยางค์ให้ความถูกต้องในการรู้จำสูงกว่ากรณีสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 ที่ไม่ผ่านกระบวนการ DWT ดังนั้นการใช้กระบวนการ DWT ร่วมด้วยจึงช่วยเพิ่มประสิทธิภาพการรู้จำ และสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 จากสัญญาณองค์ประกอบความถี่ต่ำของ Wavelet Decomposition ที่ระดับ 2 ร่วมกับจำนวนพยางค์ให้เปอร์เซ็นต์ความถูกต้องของการรู้จำสูงสุดมีค่าเท่ากับ 83.75 เมื่อเปรียบเทียบกับงานวิจัยที่ใช้วิธีการเดียวกันก่อนหน้านี้[12] ซึ่งใช้เสียงจากผู้พูดเพียงคนเดียว ซึ่งได้ความถูกต้องของการรู้จำเท่ากับ 92 เปอร์เซนต์ จะเห็นว่าเปอร์เซ็นต์ความถูกต้องของการรู้จำกรณีไม่ขึ้นกับผู้พูด (16 คน) ให้ค่าต่ำกว่าเปอร์เซ็นต์ความถูกต้องของการรู้จำกรณีขึ้นกับผู้พูด (1 คน) ทั้งนี้เป็นเพราะความหลากหลายของเสียงพูดที่เกิดจากต่างบุคคล ข้อผิดพลาดของการรู้จำส่วนใหญ่ยังคงอยู่ที่คำว่าหลอดไฟ และพัดลม

5.3 สรุปผล

ในการศึกษาการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้ของผู้สูงอายุเพื่อควบคุมอุปกรณ์ไฟฟ้าด้วยคำสั่ง เปิด ปิด พัดลม หลอดไฟ และ ทีวี พบว่า

1. สามารถใช้ระยะเวลาของเสียงพูด (DS) และจำนวนพยางค์ (NS) ซึ่งเป็นคุณลักษณะเด่นเชิงเวลา ร่วมกับจำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ (FG) เป็นคุณลักษณะเด่นในการรู้จำได้ โดยให้ความถูกต้องของการรู้จำเท่ากับ 80.2 เปอร์เซ็นต์
2. สามารถใช้สัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 ที่ได้จากสัญญาณเสียงพูดหรือจากองค์ประกอบความถี่ต่ำหรือสูงของ Wavelet Decomposition ที่ระดับ 1-3 ร่วมกับจำนวนพยางค์เป็นคุณลักษณะเด่นในการรู้จำได้
3. การใช้กระบวนการ DWT ร่วมด้วยสำหรับสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 ช่วยเพิ่มประสิทธิภาพการรู้จำ
4. สัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 จากสัญญาณองค์ประกอบความถี่ต่ำของ Wavelet Decomposition ที่ระดับ 2 ร่วมกับจำนวนพยางค์ให้เปอร์เซ็นต์ความถูกต้องของการรู้จำสูงสุดเท่ากับ 83.75 เปอร์เซ็นต์
5. เปอร์เซ็นต์ความถูกต้องของการรู้จำกรณีไม่ขึ้นกับผู้พูด (16 คน) ให้ค่าต่ำกว่าเปอร์เซ็นต์ความถูกต้องของการรู้จำกรณีขึ้นกับผู้พูด (1 คน)

6. ข้อผิดพลาดของการรู้จำส่วนใหญ่อยู่ที่คำว่าหลอดไฟ และพัดลม

5.4 ข้อเสนอแนะ

1. เนื่องจากข้อผิดพลาดของการรู้จำส่วนใหญ่อยู่ที่คำว่าหลอดไฟ และพัดลม ในการวิจัยต่อไปจึงควรหาวิธีการลดความผิดพลาดสำหรับทั้งสองคำสั่งด้วย
2. การรู้จำเสียงพูดที่ใช้จำนวนคำสั่ง และจำนวนผู้พูดมากขึ้นควรได้รับการวิจัยต่อไปเพื่อเป็นการยืนยันประสิทธิภาพของวิธีการเพิ่มเติม
3. เปอร์เซ็นต์ความถูกต้องของการรู้จำกรณีไม่ขึ้นกับผู้พูดยังคงต่ำกว่ากรณีขึ้นกับผู้พูด จึงควรทำการวิจัยเพิ่มเติมเพื่อหาวิธีเพิ่มประสิทธิภาพสำหรับการรู้จำเสียงพูดของหลากหลายบุคคล

บรรณานุกรม

1. ยุทธศาสตร์การวิจัยรายประเด็นด้านผู้สูงอายุและสังคมสูงอายุ (พ.ศ. 2556-2559), สำนักงานคณะกรรมการวิจัยแห่งชาติ.
2. เกรียงไกร เหลืองอำพล, "การพัฒนาเทคนิคการรู้จำเสียงพูดด้วย DTW กับ LPC และ LSP," วิทยานิพนธ์ปริญญามหาบัณฑิตสาขาวิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี, 2553.
3. ปฏิมากร กิ้มสวัสดิ์, "การศึกษาการรู้จำเสียงพูดตัวเลขภาษาไทยแบบแยกคำชนิดไม่ขึ้นกับผู้พูดโดยใช้โครงข่ายประสาทที่มีการเรียนรู้แบบแพร่กลับ," วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์, 2544.
4. M. Karnjanadecha, P. Kimsawad, and P. Tanthanakit, "HMM Based Speech Recognition of Continuous Thai Digits," Proceedings of the 2001 International Symposium on Communications and Information Technology, pp. 271–274, Chiang Mai, Thailand, Nov. 14–16, 2001.
5. P. Thong-in and S. Wongthanavasu, "Microcontroller – Base Thai speech recognition," in the 13th National Computer Science and Engineering Conference, Bangkok, Thailand, vol.13(1), pp.10–15, Nov. 4–6, 2009.
6. R. Boonsin and C. Jaruskulchai, "Thai Voice Command and Control for PocketPC," Kasetsart University Conference, 2010.
7. นลินรัตน์ วิศวกิตติ และ พกิจ สุวัตถิ์ "รายงานการวิจัย เรื่องการประยุกต์ใช้แบบจำลองฮิดเดนมาร์คอฟในการรู้จำเสียงพยัญชนะต้นภาษาไทย," มหาวิทยาลัยสยาม.
8. S. Aunkaew, M. Karnjanadecha, and C. Wutiwiwatchai, "Development of a Corpus for Southern Thai Dialect Speech Recognition: Design and Text Preparation," in the 10th international symposium on natural language processing, Phuket, Thailand, pp. 147–152, Oct. 28–30, 2013.
9. P. Jantaraprim, T. Chimphet, and C. Jantaraprim, "Speech recognition based electrical device control for central and southern Thai dialects," in the 6th Conference of Electrical Engineering Network of Rajamangala University of

Technology 2014 (EENET 2014), Krabi, Thailand, pp. 909-912, Mar. 26-28, 2014.

10. User's Guide, MATLAB.
11. Jantaraprim, P., Chimphet, T., Inthavisas, K., "Applying Time and Frequency Domain Features to Improve Recognition of Southern Thai Dialect Speech," in the 7th Conference of Electrical Engineering Network of Rajamangala University of Technology 2015 (EENET 2015), Phattaya, Thailand, pp. 230-243, May. 27-29, 2015.
12. Jantaraprim, P., Chimphet, T., Inthavisas, K., "Improving Southern Thai Dialect Speech Recognition using Wavelet Transform," in the 7th ECTI-CARD 2015, Trang, Thailand, pp.610-613, July. 8-10, 2015.

ภาคผนวก บทความที่ตีพิมพ์

บทความวิจัย
การประชุมวิชาการเครือข่ายวิศวกรรมไฟฟ้ามหาวิทยาลัยเทคโนโลยีราชมงคล ครั้งที่ 7
Proceedings of the 7th Conference of Electrical Engineering Network of Rajamangala University of Technology 2015 (EENET 2015)

การประยุกต์ใช้คุณลักษณะเด่นเชิงเวลาและความถี่เพื่อพัฒนาการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้

Applying Time and Frequency Domain Features to Improve Recognition of Southern Thai Dialect Speech

ปฏิมากร จันทร์พริ้ม [1] ธีรพงษ์ ฉิมเพชร [1] และ กีรติ อินทวิเศษ [2]
[1]สาขาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย
[2]สาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย
เลขที่ 1 ถนนราชดำเนินนอก ตำบลบ่อยาง อำเภอเมืองสงขลา จังหวัดสงขลา E-mail: patimakorn.j@hyac.in.th, patimakorn.j@rmutsv.ac.th

บทคัดย่อ

บทความนี้เสนอการศึกษาเบื้องต้นของการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้ที่ใช้ทั้งคุณลักษณะเด่นเชิงเวลาและความถี่ โดยใช้เสียงพูดของผู้สูงอายุและมุ่งศึกษาเสียงพูดของคำสั่งในการควบคุมอุปกรณ์ไฟฟ้า ได้แก่ เปิด ปิด หลอดไฟ พัดลม และทีวี คุณลักษณะเด่นที่ใช้พิจารณา ได้แก่ ระยะเวลาของเสียงพูด จำนวนพยางค์ และจำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ โครงข่ายประสาทเทียมชนิด Backpropagation จำนวน 1 ชั้นซ่อน มีจำนวนอินพุตเท่ากับ 12-5-1 ถูกนำมาใช้ในการตัดสินใจรู้จำ จากข้อมูลทดลอง 20 ชุด (100 ข้อมูล) พบว่าคุณลักษณะเด่นดังกล่าวสามารถให้เปอร์เซ็นต์ความถูกต้องของการรู้จำเท่ากับ 86 เปอร์เซ็นต์ อย่างไรก็ตาม จำนวนผู้พูดที่มากขึ้นควรได้รับการวิจัยต่อไปเพื่อเป็นการยืนยันประสิทธิภาพของวิธีการเพิ่มเติม

คำสำคัญ: รู้จำเสียงพูด, เสียงพูดภาษาไทยสำเนียงภาคใต้

Abstract

This article presents a preliminary study of speech recognition for southern Thai dialect using features in time and frequency domains. Elderly speech is used in the experiment. The study focuses on 5 commands for control electrical device, that is 'turn on', 'turn off', 'lamp', 'fan', and 'TV'. A duration of sound, a number of syllable and peak number of formant frequency for frequency group are 3 features for speech recognition. A backpropagation of a Neural Network, consisting of 1 hidden layer and inputs of 12-5-1, is used for making decision. From 20 data sets (100 data), the result shows the percent recognition of 86. However, data of more subjects should be ongoing studied to verify this method.

Keywords: speech recognition, southern Thai dialect speech

1. บทนำ

เทคโนโลยีการรู้จำเสียงพูดเป็นทางเลือกหนึ่งที่ช่วยอำนวยความสะดวกในการดำรงชีวิตของผู้คนในยุคปัจจุบัน การควบคุมอุปกรณ์ไฟฟ้าด้วยเสียงพูดเป็นตัวอย่างหนึ่งที่แสดงให้เห็นความจำเป็นอย่างยิ่งของเทคโนโลยีนี้โดยเฉพาะสำหรับบุคคลบางกลุ่ม เช่น ผู้พิการที่ยังสามารถพูดได้ ผู้สูงอายุที่ไม่สามารถเคลื่อนไหวได้สะดวก เป็นต้น งานวิจัยที่เกี่ยวข้องกับการรู้จำเสียงพูดภาษาไทยที่ผ่านมาส่วนใหญ่ใช้ภาษาไทยมาตรฐานสำเนียงภาคกลางแทบทั้งสิ้น [1]-[3] อย่างไรก็ตาม บุคคลที่อาศัยในบางท้องถิ่นจะคุ้นเคยกับการใช้ภาษาไทยสำเนียงของท้องถิ่นนั้นๆ มากกว่าสำเนียงภาคกลาง และบางคนอาจไม่สามารถพูดสำเนียงภาคกลางได้ เช่น ผู้สูงอายุส่วนใหญ่ที่อาศัยในภาคใต้ ระบบรู้จำเสียงพูดที่ได้รับการวิจัยและทดสอบด้วยสำเนียงภาคกลางอาจไม่สามารถรองรับหรือให้ประสิทธิภาพการรู้จำที่ดีสำหรับเสียงพูดของสำเนียงภาคอื่นๆ ได้

เมื่อเร็วๆ นี้ S. Aunkaew [4] ได้พัฒนาชุดข้อมูล (corpus) สำหรับเสียงพูดภาษาไทยสำเนียงภาคใต้ แต่ก็ไม่ได้รายงานการวิจัยในเชิงการรู้จำ ส่วน P. Jantaraprim [5] เสนอการศึกษาเบื้องต้นของการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้เพื่อควบคุมอุปกรณ์ไฟฟ้าโดยใช้คุณลักษณะเด่นเชิงเวลา ได้แก่ ระยะเวลาของเสียงพูด และสัมประสิทธิ์การประมาณพันธะเชิงเส้น อย่างไรก็ตามคุณลักษณะเด่นดังกล่าวยังมีแนวโน้มให้ความผิดพลาดสำหรับเสียงพูดคำว่า 'หลอดไฟ' และ 'พัดลม' อีกทั้งยังไม่ได้วิจัยไปจนกระทั่งถึงการตัดสินใจรู้จำ

บทความนี้จึงเสนอการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้ที่ใช้ทั้งคุณสมบัติเชิงเวลาและความถี่ร่วมกันเพื่อเพิ่มประสิทธิภาพการรู้จำ ทั้งนี้จะใช้เสียงของผู้สูงอายุและมุ่งศึกษาเสียงพูดเกี่ยวกับคำสั่งในการควบคุมอุปกรณ์ไฟฟ้า ได้แก่ เปิด ปิด หลอดไฟ พัดลม และทีวี

2. หลักการ

2.1 คุณลักษณะเด่นเชิงเวลา

คุณลักษณะเด่นเชิงเวลาที่ใช้ในงานวิจัยได้แก่ ระยะเวลาของเสียงพูด (Duration of Sound: DS) และจำนวนพยางค์ (Number of Syllable: NS)

ระยะเวลาของเสียงพูด จากการคอนโวลูชั่นระหว่างสัญญาณเสียงตลอดช่วงที่บันทึกกับหน้าต่างชนิดสี่เหลี่ยมขนาด 40 ms ทำให้ได้ผลของคอนโวลูชั่นที่สามารถแสดงขอบเขตเฉพาะเสียงพูดได้ โดยที่จุดเริ่มต้นและจุดสิ้นสุดของเสียงพูดมีค่าเท่ากับจุดแรกที่ให้ค่าคอนโวลูชั่นมากกว่า 0.1 เท่าของค่าสูงสุดนับจากต้นเสียงและนับจากปลายเสียง ตามลำดับ การทำคอนโวลูชั่นแสดงในสมการที่ 1 [5]

$$c(n) = \sum_{i=0}^{n} |s(i)|.w(n-i) \qquad (1)$$

เมื่อ

$c(n)$ คือ ค่าคอนโวลูชั่น

$s(i)$ คือ ข้อมูลสัญญาณเสียงพูดลำดับที่ i

w คือ หน้าต่างชนิดสี่เหลี่ยม

n คือ จำนวนค่าคอนโวลูชั่นทั้งหมด

จำนวนพยางค์ จากระยะเวลาของเสียงพูดแบ่งระยะเวลาออกเป็น 3 เฟรมเท่าๆ กัน พิจารณาเฟรมตรงกลาง หากเสียงพูดที่เฟรมตรงกลางมีค่าคอนโวลูชั่นน้อยกว่า 20% ของค่าคอนโวลูชั่นสูงสุดเป็นระยะเวลามากกว่า 5% ของระยะเวลาของเสียงพูด กำหนดให้สัญญาณเสียงนั้นมี 2 พยางค์

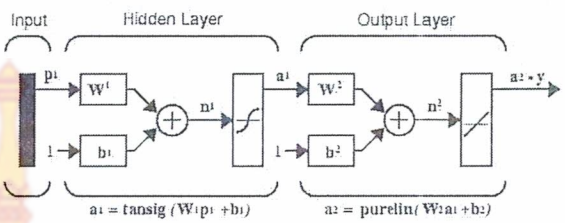
2.2 คุณลักษณะเด่นเชิงความถี่

คุณลักษณะเด่นเชิงความถี่ที่ใช้ในงานวิจัย ได้แก่ จำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ (FG) เสียงพูดที่ผ่านวิธีการ Fast Fourier Transform (FFT) ทำให้ได้ความถี่ของเสียงที่นำมาพิจารณาในช่วง 0-5000 Hz (ความถี่สุ่มเท่ากับ 10 kHz) ช่วงความถี่ถูกแบ่งออกเป็น 10 กลุ่ม (G) โดยที่ G1 มีความถี่ช่วง (0,500] , G2 มีความถี่ช่วง (500,1000], G3 มีความถี่ช่วง (1000,1500], ..., และ G10 มีความถี่ช่วง (4500,5000]

จุดยอดของความถี่เด่น พิจารณาจาก จุดยอดของขนาดความถี่ภายในหน้าต่างสี่เหลี่ยมขนาดเท่าช่วงความถี่ 500 Hz โดยที่จุดยอดนั้นจะต้องที่มีค่ามากที่สุดภายในหน้าต่างและมากกว่าค่าเฉลี่ยของขนาดความถี่ตลอดช่วงความถี่ทั้งหมด 0-5000 Hz หน้าต่างนี้จะถูกกำหนดให้เริ่มต้นจากความถี่ 0 Hz แล้วเลื่อนไปตลอดช่วงความถี่ทั้งหมดทีละ 250 Hz (หน้าต่างซ้อนทับกัน 50%)

2.3 โครงข่ายประสาทเทียม

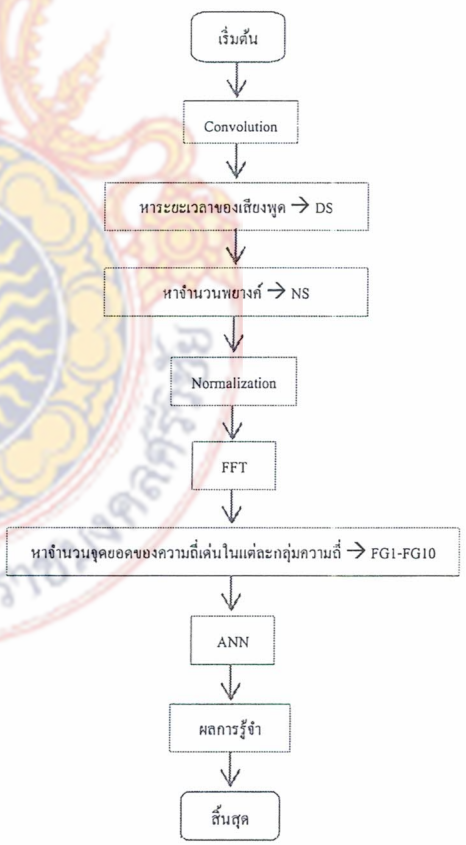
โครงข่ายประสาทเทียม (ANN) ที่ใช้ในงานวิจัยนี้ เป็นชนิด Backpropagation ที่มีโครงสร้างเป็นแบบ Multilayer Feedforward Network จำนวน 1 ชั้นซ่อน มีจำนวนอินพุตเท่ากับ 12-5-1 (input-hiden layer-output layer) และมีทรานเฟอร์ฟังก์ชันเป็น tansig/purelin ดังแสดงในรูปที่ 1



รูปที่ 1 โครงสร้างของ Multilayer Feedforward Network

3. วิธีการดำเนินงาน

บันทึกเสียงพูดสำเนียงภาคใต้แบบ mono ด้วยอัตราการสุ่มเท่ากับ 10 kHz เป็นเวลาคำสั่งละ 3 s ความละเอียด 16 bits ผู้พูดเป็นผู้สูงอายุเพศหญิง 1 คน อายุ 79 ปี และเป็นคนภาคใต้โดยกำเนิด คำสั่งที่บันทึกมีทั้งหมด 5 คำสั่ง ได้แก่ เปิด ปิด หลอดไฟ พัดลม และทีวี ทำการบันทึกคำสั่งละ 20 ครั้ง ทำให้ได้ข้อมูลเสียงพูดทั้งหมด 100 ข้อมูล



รูปที่ 2 แผนภาพกระบวนการทำงาน

รูปที่ 2 แสดงแผนภาพกระบวนการทำงานโดยเริ่มจากการคอนโวลูชั่นสัญญาณเสียงพูดตลอดช่วงที่บันทึกกับหน้าต่างชนิดสี่เหลี่ยมขนาด 40 ms เพื่อหาขอบเขตเฉพาะเสียงพูด ค่าคอนโวลูชั่นจะถูกกำหนดจุดเริ่มต้นและสิ้นสุดเพื่อหาระยะเวลาของเสียงพูด (DS) เสียงพูดที่ได้จะผ่านกระบวนการพิจารณาเฟรมตรงกลางเพื่อหาจำนวนพยางค์ (NS) เป็นลำดับต่อไป จากนั้นเสียงพูดจะถูกนอร์มอลไลซ์ (normalize) แล้วเข้าสู่กระบวนการ FFT และหาจำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ (FG1-FG10) ทำให้ได้คุณลักษณะเด่นแทนเสียงพูด (feature) หรือจำนวนอินพุตสำหรับโครงข่ายประสาทเทียมทั้งหมด 12 ค่า (DS, NS, FG1-FG10) อินพุตทั้งหมดถูกนำไปฝึกฝน (train) หรือทดสอบ (test) กับโครงข่ายประสาทเทียมแล้วแสดงผลการรู้จำต่อไป

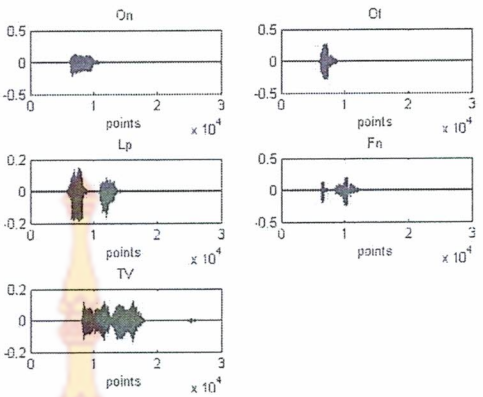
สัญญาณเสียงพูดครั้งที่ 1-10 ของแต่ละคำสั่งจะถูกนำมาใช้เป็นข้อมูลสำหรับฝึกฝน (หรือข้อมูลทดสอบสำหรับกรณี 2-fold cross validaton) กับโครงข่ายประสาทเทียม และสัญญาณเสียงพูดครั้งที่ 11-20 ของแต่ละคำสั่งจะถูกนำมาใช้เป็นข้อมูลสำหรับทดสอบ (หรือข้อมูลฝึกฝนสำหรับกรณี 2-fold cross validaton) ทำให้ได้ข้อมูลสำหรับฝึกฝนและทดสอบอย่างละ 50 ข้อมูล (5 คำสั่ง x 10 ครั้ง) คำนวณหาเปอร์เซ็นต์ความถูกต้องของการรู้จำ (Percent Recognition: PR) ตามสมการที่ 2 โดยที่ TR (True Recognition) คือ จำนวนการรู้จำที่ถูกต้อง และ FR (False Recognition) คือ จำนวนการรู้จำที่ *ไม่*ถูกต้อง

$$PR = \frac{TR}{TR+FR} * 100 \qquad (2)$$

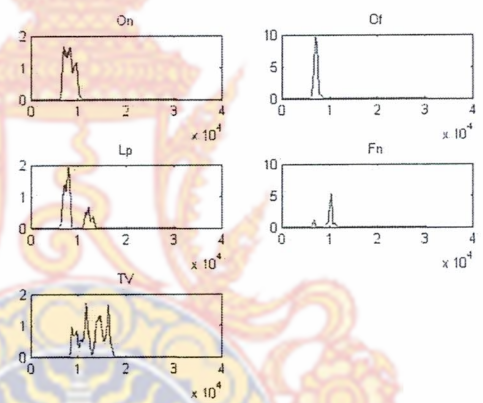
ผลการรู้จำที่ได้คิดจากค่าเฉลี่ยที่เกิดจากการฝึกฝนและทดสอบ 3 ครั้ง และสลับข้อมูลฝึกฝนและทดสอบด้วยการทำ 2-fold cross validation เพื่อความน่าเชื่อถือของผลการรู้จำ

4. ผลการทดลอง

กำหนดให้เสียงพูดคำสั่งต่างๆ แทนด้วยสัญลักษณ์ต่อไปนี้ เปิด (On), ปิด (Of), หลอดไฟ (Lp), พัดลม (Fn) และทีวี (TV) รูปที่ 3 แสดงตัวอย่างสัญญาณเสียงพูดเปิด, ปิด, หลอดไฟ, พัดลมและทีวี ที่ถูกบันทึกนาน 3 s (อัตราการสุ่ม 10 kHz, 30000 points) จากรูปจะเห็นได้ว่าเสียงพูดที่มี 1 พยางค์จะให้สัญญาณเสียงที่มองเห็นเป็น 1 กลุ่มอย่างชัดเจน ขณะที่เสียงพูดที่มี 2 พยางค์จะให้สัญญาณเสียงที่มองเห็นเป็น 2 กลุ่มที่ติดหรือแยกกันแม้ว่าเสียงพูดคำว่าพัดลมอาจเห็นเป็นเพียงกลุ่มขนาดเล็กในช่วงแรก

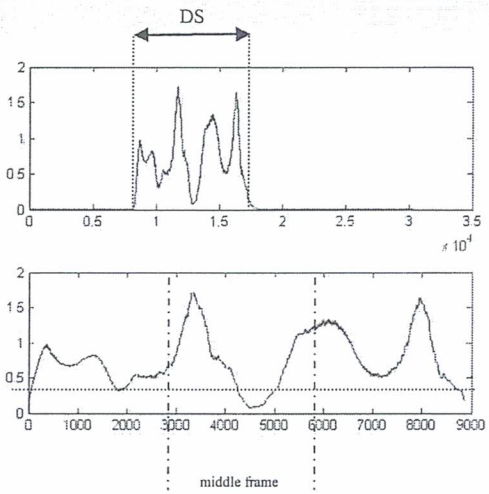


รูปที่ 3 สัญญาณเสียงพูดเปิด, ปิด, หลอดไฟ, พัดลม และทีวี

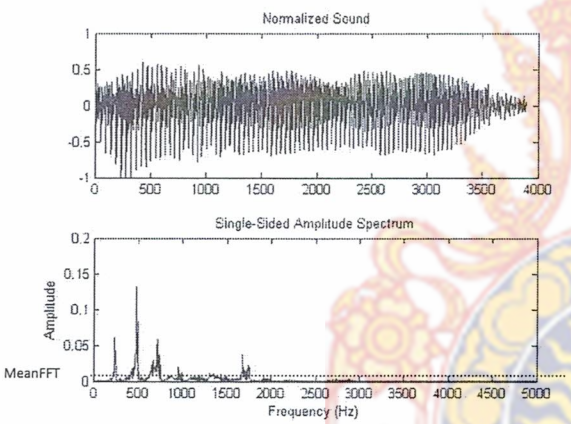


รูปที่ 4 ค่าคอนโวลูชั่นของเสียงพูดเปิด, ปิด, หลอดไฟ, พัดลม และทีวี

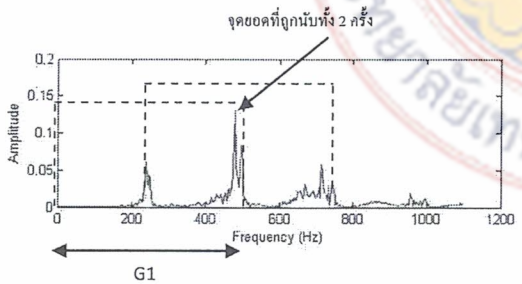
รูปที่ 4 แสดงค่าคอนโวลูชั่นของเสียงพูดที่สอดคล้องกับสัญญาณในรูปที่ 3 จะเห็นได้ว่าค่าคอนโวลูชั่นสามารถแสดงขอบเขตเฉพาะเสียงพูดได้ จุดเริ่มต้นและจุดสิ้นสุดของสัญญาณเสียงสามารถคำนวณได้จากการตรึงค่ากำหนด (threshold) ตามเงื่อนไขในข้อ 2.1 ทำให้ได้ระยะเวลาของเสียงพูด (DS) ซึ่งเป็นผลต่างระหว่างจุดสิ้นสุดและจุดเริ่มต้น ดังแสดงในรูปที่ 5a และเมื่อพิจารณาเฟรมตรงกลางของเสียงพูดที่มีค่าคอนโวลูชั่นน้อยกว่า 20% ของค่าคอนโวลูชั่นสูงสุดเป็นระยะเวลามากกว่า 5% ของระยะเวลาของเสียงพูด ทำให้พิจารณาได้ว่าสัญญาณเสียงนั้นมี 2 พยางค์ แสดงเฟรมตรงกลางของเสียงพูดและเส้นแสดงค่าคอนโวลูชั่น 20% ของค่าคอนโวลูชั่นสูงสุดในรูปที่ 5b



รูปที่ 5 a) ระยะเวลาของเสียงพูดของคำว่าทีวี
b) เฟรมตรงกลาง



รูปที่ 6 a) ตัวอย่างเสียงพูดที่ถูกนอร์มอลไลซ์ของคำว่าเปิด
b) ผลตอบสนองเชิงความถี่



รูปที่ 7 ผลตอบสนองเชิงความถี่ของคำว่าเปิดในช่วง 0-1100 Hz และหน้าต่างที่เลื่อนผ่านความถี่ในกลุ่มที่ 1

รูปที่ 6a แสดงตัวอย่างเสียงพูดที่ถูกนอร์มอลไลซ์ของคำว่าเปิด ส่วนรูปที่ 6b แสดงผลตอบสนองเชิงความถี่ของสัญญาณในรูปที่ 6a และค่าเฉลี่ย (MeanFFT) ของขนาดความถี่ตลอดช่วงความถี่ทั้งหมดมีค่าเท่ากับ 0.0029 ส่วนรูปที่ 7 แสดงตัวอย่างจำนวนจุดยอดของความถี่เด่นภายในหน้าต่างที่เลื่อนผ่านความถี่ในกลุ่มที่ 1 โดยที่หน้าต่างซ้อนทับกัน 50% ทำให้นับจุดยอดของความถี่เด่นในกลุ่มที่ 1 ได้ 2 ครั้งหรือจำนวนจุดยอดของความถี่เด่นในกลุ่มที่ 1 เท่ากับ 2

ผลการทดสอบการรู้จำเสียงพูดกับโครงข่ายประสาทเทียมที่เกิดจากค่าเฉลี่ยจำนวน 3 ครั้ง และทำ 2-fold cross validation พบว่าคุณลักษณะเด่นในเชิงเวลาและความถี่ดังกล่าวให้เปอร์เซ็นต์ความถูกต้องของการรู้จำเท่ากับ 86 เปอร์เซ็นต์ เมื่อเปรียบเทียบกับงานวิจัยก่อนหน้านี้ [5] พบว่า การใช้คุณลักษณะเด่นจำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ ร่วมกับระยะเวลาของเสียงพูดและจำนวนพยางค์ช่วยเพิ่มประสิทธิภาพการรู้จำ อย่างไรก็ตามข้อผิดพลาดของการรู้จำส่วนใหญ่ก็ยังคงอยู่ที่คำว่าหลอดไฟ และพัดลม จึงควรทำการวิจัยเพิ่มเติมเพื่อให้ได้ผลการรู้จำที่แม่นยำยิ่งขึ้น

5. สรุปและข้อเสนอแนะ

จำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ สามารถใช้เป็นคุณลักษณะเด่นเชิงความถี่สำหรับการรู้จำเสียงพูดสำเนียงภาคใต้ของคำสั่งเปิด ปิด หลอดไฟ พัดลม และทีวีได้ โดยที่เมื่อใช้ร่วมกับคุณลักษณะเด่นของระยะเวลาของเสียงพูดกับจำนวนพยางค์สามารถให้เปอร์เซ็นต์ความถูกต้องของการรู้จำเท่ากับ 86 เปอร์เซ็นต์ อย่างไรก็ตามการรู้จำเสียงพูดที่ใช้จำนวนผู้พูดที่มากขึ้นควรได้รับการวิจัยต่อไปเพื่อเป็นการยืนยันประสิทธิภาพของวิธีการเพิ่มเติม

6. กิตติกรรมประกาศ

ขอขอบคุณคณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัยที่ให้การสนับสนุนและส่งเสริมโครงการวิจัยเรื่องการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้ของผู้สูงอายุเพื่อควบคุมอุปกรณ์ไฟฟ้า

เอกสารอ้างอิง

[1] เกรียงไกร เหลืองอำพล, "การพัฒนาเทคนิคการรู้จำเสียงพูดด้วย DTW กับ LPC และ LSP," วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี, 2553

[2] P. Thong-in and S. Wongthanavasu, "Microcontroller - Base Thai speech recognition," in *the 13th National Computer Science and*

Engineering Conference, Bangkok, Thailand, vol.13(1), Nov. 4-6, 2009, pp.10-15.

[3] R. Boonsin and C. Jaruskulchai, "Thai Voice Command and Control for PocketPC," in *Kasetsart University Conference*, 2010.

[4] S. Aunkaew, M. Karnjanadecha, and C. Wutiwiwatchai, "Development of a Corpus for Southern Thai Dialect Speech Recognition: Design and Text Preparation," in *the 10th international symposium on natural language processing*, Phuket, Thailand, Oct. 28-30, 2013, pp. 147-152.

[5] P. Jantaraprim, T. Chimphet, and C. Jantaraprim, "Speech recognition based electrical device control for central and southern Thai dialects," in *the 6th Conference of Electrical Engineering Network of Rajamangala University of Technology 2014 (EENET 2014)*, Krabi, Thailand, Mar. 26-28, 2014, pp. 909-912.

ปฏิมากร จันทร์พริ้ม จบการศึกษาจากสาขาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์ในระดับปริญญาตรี พ.ศ. 2540 ระดับปริญญาโท พ.ศ. 2544 และระดับปริญญาเอก พ.ศ. 2555 ทำงานเป็นอาจารย์ประจำภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์ตั้งแต่ปี พ.ศ. 2544 – 2550 ปัจจุบันทำงานเป็นอาจารย์ประจำสาขาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย สนใจงานวิจัยด้าน Digital Signal Processing, Speech Recognition และ Pattern Recognition

ธีรพงษ์ ฉิมเพชร จบการศึกษาระดับปริญญาตรี พ.ศ.2554 จากสาขาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย จบการศึกษาระดับปริญญาโท พ.ศ. 2556 จากสาขาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์ ปัจจุบันทำงานเป็นอาจารย์ประจำสาขาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย

กีรติ อินทวิเศษ จบการศึกษาระดับปริญญาตรี พ.ศ. 2541 จากสาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์ จบการศึกษาระดับปริญญาโทและปริญญาเอก สาขาวิศวกรรมคอมพิวเตอร์จาก Lehigh University สหรัฐอเมริกา สนใจงานวิจัยด้าน Biometric Security, Speech Processing และ Computer Vision ปัจจุบันทำงานเป็นอาจารย์ประจำสาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย

การเพิ่มประสิทธิภาพการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้โดยใช้เวฟเลตทรานฟอร์ม

Improving Southern Thai Dialect Speech Recognition using Wavelet Transform

ปฏิมากร จันทร์พริ้ม [1] ธีรพงษ์ ฉิมเพชร [1] และ กีรติ อินทวิเศษ [2]
[1]สาขาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย
[2]สาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย
เลขที่ 1 ถนนราชดำเนินนอก ตำบลบ่อยาง อำเภอเมืองสงขลา จังหวัดสงขลา E-mail: patimakorn.j@hyac.in.th, patimakorn.j@rmutsv.ac.th

บทคัดย่อ

บทความนี้กล่าวเกี่ยวกับการศึกษาเบื้องต้นของการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้ที่ใช้จำนวนพยางค์ สัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 และ Wavelet Decomposition ที่ระดับ 1-3 ในการทดลองจะใช้เสียงพูดของผู้สูงอายุที่พูดคำสั่งสำหรับควบคุมอุปกรณ์ไฟฟ้า ได้แก่ เปิด ปิด หลอดไฟ พัดลม และทีวี โครงข่ายประสาทเทียมชนิด Backpropagation ถูกนำมาใช้ในกระบวนการตัดสินใจรู้จำ จากข้อมูลทดลอง 20 ชุด (100 ข้อมูล) พบว่า จำนวนพยางค์ร่วมกับสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 จากสัญญาณองค์ประกอบความถี่ต่ำของ Wavelet Decomposition ที่ระดับ 2 ให้เปอร์เซ็นต์ความถูกต้องของการรู้จำสูงสุดเท่ากับ 92.0 เปอร์เซ็นต์

คำสำคัญ: รู้จำเสียงพูด, เวฟเลตทรานฟอร์ม

Abstract

This article describes a preliminary study of speech recognition for southern Thai dialect using a number of syllable, the 3rd order of Linear Predictive Coefficients and wavelet decomposition at the 1st – 3rd level. Elderly speech is used in the experiment. The study focuses on 5 commands for control electrical device, that is 'turn on', 'turn off', 'lamp', 'fan', and 'TV'. A backpropagation of a Neural Network is used for making decision. From 20 data sets (100 data), the result shows that a number of syllable with the 3rd order of Linear Predictive Coefficients from approximation of wavelet decomposition offers the highest percent recognition at 92.0 percent.

Keywords: speech recognition, Wavelet Transform

1. บทนำ

การคาดประมาณประชากรของสำนักงานคณะกรรมการพัฒนาการเศรษฐกิจและสังคมแห่งชาติ ได้ประมาณการแนวโน้มการเปลี่ยนแปลงประชากรผู้สูงอายุว่า ในปี 2566 และ 2576 ประชากรผู้สูงอายุในประเทศไทยจะมีจำนวนเพิ่มขึ้นเป็น 14.1 ล้านคน และ 18.7 ล้านคน หรือคิดเป็นร้อยละ 21 และ 29 ของประชากรทั้งหมด ตามลำดับ [1] การควบคุมอุปกรณ์ไฟฟ้าด้วยเสียงพูดเป็นทางเลือกหนึ่งที่ช่วยอำนวยความสะดวกให้กับผู้สูงอายุในการใช้ชีวิตประจำวัน ผู้สูงอายุส่วนใหญ่ที่อาศัยในภาคใต้จะคุ้นเคยกับการใช้ภาษาไทยสำเนียงภาคใต้มากกว่าสำเนียงภาคกลาง และบางคนอาจไม่สามารถพูดสำเนียงภาคกลางได้ แม้ว่ามีงานวิจัยที่เกี่ยวข้องกับการรู้จำเสียงพูดภาษาไทยอยู่มาก แต่ส่วนใหญ่จะใช้ภาษาไทยมาตรฐานสำเนียงภาคกลางแทบทั้งสิ้น [2]-[4] ซึ่งอาจไม่สามารถให้ประสิทธิภาพการรู้จำที่ดีสำหรับเสียงพูดสำเนียงภาคใต้

ปี 2556 S. Aunkaew [5] ได้พัฒนาชุดข้อมูล (corpus) สำหรับเสียงพูดภาษาไทยสำเนียงภาคใต้ ต่อมา P. Jantaraprim [6] ศึกษาการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้โดยใช้ระยะเวลาของเสียงพูด และสัมประสิทธิ์การประมาณพันธะเชิงเส้น แต่ไม่ได้วิจัยไปจนกระทั่งถึงการตัดสินใจรู้จำ เมื่อเร็วๆ นี้ P. Jantaraprim [7] ยังคงพัฒนาการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้โดยได้เสนอคุณลักษณะเด่นจำนวนจุดยอดของความถี่เด่นในกลุ่มความถี่ต่างๆ ร่วมกับระยะเวลาของเสียงพูดกับจำนวนพยางค์ พบว่า สามารถให้เปอร์เซ็นต์ความถูกต้องของการรู้จำถึง 86 เปอร์เซ็นต์ อย่างไรก็ตามคุณลักษณะเด่นดังกล่าวยังคงให้ความผิดพลาดของการรู้จำสำหรับคำว่าหลอดไฟ และพัดลม

2. ทฤษฎีและหลักการ

2.1 จำนวนพยางค์

จากงานวิจัยก่อนหน้า [7] จำนวนพยางค์ของเสียงพูดหาได้จากการพิจารณาเฟรมตรงกลาง (เฟรมที่ 2 จากจำนวน 3 เฟรมเท่าๆกัน) ของค่าคอนโวลูชั่นระหว่างสัญญาณเสียงพูดกับหน้าต่างชนิดสี่เหลี่ยมขนาด 40 ms โดยที่หากค่าคอนโวลูชั่นของเสียงพูดที่เฟรมตรงกลางมีค่าน้อยกว่า 20% ของค่าคอนโวลูชั่นสูงสุดเป็นระยะเวลามากกว่า 5% ของระยะเวลาของเสียงพูด กำหนดให้สัญญาณเสียงนั้นมี 2 พยางค์ ทั้งนี้พิจารณาเฉพาะส่วนที่เป็นเสียงพูดเท่านั้น โดยกำหนดให้จุดเริ่มต้นและ

จุดสิ้นสุดของเสียงพูด คือ จุดแรกที่ให้ค่าคอนโวลูชั่นมากกว่า 0.1 เท่าของค่าสูงสุดนับจากต้นเสียงและนับจากปลายเสียง ตามลำดับ

สมการที่ 1 แสดงการหาค่าคอนโวลูชั่น (C) ระหว่างสัญญาณเสียงตลอดช่วงที่บันทึก (S) กับหน้าต่างชนิดสี่เหลี่ยมขนาด 40 ms (W) [7]

$$c(n) = \sum_{i=0}^{n} |s(i)|.w(n-i) \quad (1)$$

2.2 Discrete Wavelet Transform (DWT)

DWT เป็นกระบวนการที่ให้สัญญาณผ่าน filter 2 ชนิด คือ Digital low-pass filter และ Digital high-pass filter แล้ว Down Sampling ลง 2 เท่า ทำให้ได้องค์ประกอบสัญญาณความถี่ต่ำ (Approximation: cA1) และ องค์ประกอบสัญญาณความถี่สูง (Detail: cD1) ของ Wavelet Decomposition ในระดับที่ 1 ต่อมาองค์ประกอบสัญญาณความถี่ต่ำ (Approximation) ยังสามารถถูกแยกในระดับต่อๆไปด้วยกระบวนการเดิม ทำให้ได้องค์ประกอบสัญญาณความถี่ต่ำและสูงในระดับต่างๆ ดังแสดงในรูปที่ 1 งานวิจัยนี้ใช้ mother wavelet แบบ Daubechies

2.3 การประมาณพันธะเชิงเส้น

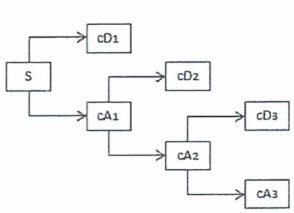
การประมาณพันธะเชิงเส้น (Linear Predictive Coefficients: LPC) เป็นกระบวนการหาค่าสัมประสิทธิ์ของ forward linear predictor โดยพิจารณาว่าเสียงเกิดจากผลรวมเชิงเส้นของสัญญาณที่ทราบค่าแล้วก่อนหน้าจำนวน p ค่า ดังสมการที่ 2 [8] งานวิจัยนี้ใช้การประมาณพันธะเชิงเส้นอันดับ 3

$$\hat{x}(n) = -a(2)x(n-1) - a(3)x(n-2) - \ldots - a(p+1)x(n-p) \quad (2)$$

เมื่อ $\hat{x}(n)$ คือ สัญญาณค่าถัดไปที่ทำนาย, $x(n)$ คือ สัญญาณที่ทราบค่าแล้ว และ a คือ ค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น

2.4 โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียมที่ใช้ในงานวิจัยนี้ เป็นชนิด Backpropagation จำนวน 1 ชั้นซ่อน มีโหนดในชั้นซ่อนเท่ากับ 5 ทรานเฟอร์ฟังก์ชันในชั้นซ่อนเป็น tansig และชั้นเอาต์พุตเป็น purelin



รูปที่ 1 แผนภาพ Wavelet Decomposition Tree

2.5 ประสิทธิภาพการรู้จำ

ประสิทธิภาพการรู้จำหาได้จากเปอร์เซ็นต์ความถูกต้องของการรู้จำ (Percent Recognition: PR) ตามสมการที่ 3 โดยที่ TR (True Recognition) คือ จำนวนการรู้จำที่ถูกต้อง และ FR (False Recognition) คือ จำนวนการรู้จำที่ไม่ถูกต้อง [7]

$$PR = \frac{TR}{TR + FR} * 100 \quad (3)$$

3. วิธีการดำเนินงาน

เสียงพูดสำเนียงภาคใต้ของผู้สูงอายุเพศหญิง 1 คน อายุ 79 ปี ถูกบันทึกแบบ mono ด้วยอัตราการสุ่มเท่ากับ 10 kHz ที่ความละเอียด 16 bits เป็นเวลาคำสั่งละ 3 s โดยมีคำสั่งที่บันทึกทั้งหมด 5 คำสั่ง ได้แก่ เปิด ปิด หลอดไฟ พัดลม และทีวี ทำการบันทึกคำสั่งละ 20 ครั้ง ทำให้ได้ข้อมูลเสียงพูดทั้งหมด 100 ข้อมูล [7]

กระบวนการทำงานเริ่มจากคอนโวลูชั่นสัญญาณเสียงพูดตลอดช่วงที่บันทึกกับหน้าต่างชนิดสี่เหลี่ยมเพื่อหาขอบเขตเฉพาะเสียงพูดแล้วเข้าสู่กระบวนการหาจำนวนพยางค์เป็นลำดับต่อไป จากนั้นเสียงพูดจะถูกนอร์มอลไลซ์ (normalize) แล้วหาคุณลักษณะเด่น 3 ขั้นตอนหลัก คือ

1. หาสัมประสิทธิ์การประมาณพันธะเชิงเส้น (LPC)
2. ผ่านกระบวนการ DWT 3 ระดับ ทำให้ได้สัญญาณที่เป็นองค์ประกอบความถี่ต่ำ cA1, cA2 และ cA3 และได้สัญญาณที่เป็นองค์ประกอบความถี่สูง cD1, cD2 และ cD3
3. หาสัมประสิทธิ์การประมาณพันธะเชิงเส้น (LPC) จากสัญญาณองค์ประกอบความถี่ต่ำและองค์ประกอบความถี่สูงที่ผ่านกระบวนการ DWT ที่ระดับต่างๆ 3 ระดับ (cA1_LPC, cA2_LPC, cA3_LPC, cD1_LPC, cD2_LPC, cD3_LPC)

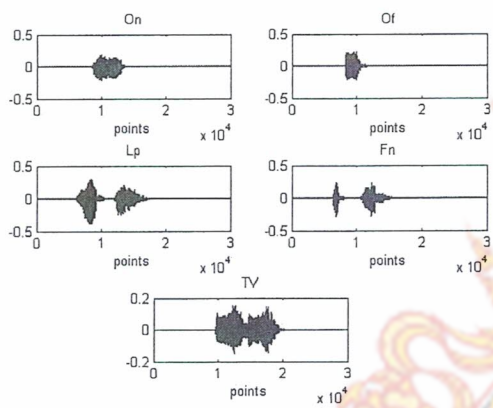
คุณลักษณะเด่นดังกล่าวเป็นอินพุตที่ถูกนำไปฝึกฝน (train) หรือทดสอบ (test) กับโครงข่ายประสาทเทียม (ANN) แล้วแสดงผลการรู้จำต่อไป อินพุตที่ทำการทดสอบเป็นคุณลักษณะเด่นที่ได้จากจำนวนพยางค์ (NS) ร่วมกับสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 ของสัญญาณ (LPC) หรือสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 ขององค์ประกอบความถี่ต่ำและองค์ประกอบความถี่สูงที่ผ่านกระบวนการ DWT ที่ระดับต่างๆ ทำให้แบ่งคุณลักษณะเด่นออกได้เป็น 7 กรณี คือ NS + LPC, NS + cA1_LPC, NS + cA2_LPC, NS + cA3_LPC, NS + cD1_LPC, NS + cD2_LPC และ NS + cD3_LPC

ในการทดสอบการรู้จำด้วยโครงข่ายประสาทเทียม ข้อมูลจะถูกแบ่งออกเป็น 2 ส่วนเท่ากัน คือ สัญญาณเสียงพูดครั้งที่ 1-10 เป็นข้อมูลฝึกฝนและสัญญาณเสียงพูดครั้งที่ 11-20 เป็นข้อมูลทดสอบ ทำให้ได้ข้อมูลสำหรับฝึกฝนและทดสอบอย่างละ 50 ข้อมูล (5 คำสั่ง x 10 ครั้ง)

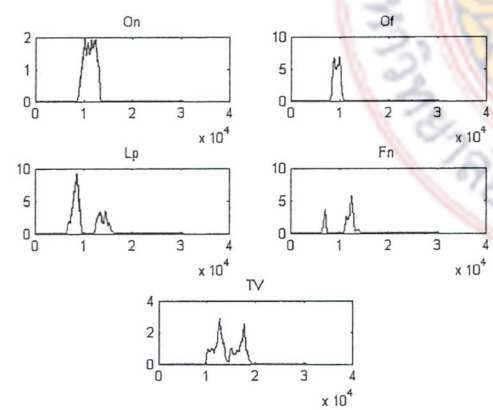
ทั้งนี้ได้สลับข้อมูลฝึกฝนและทดสอบโดยการทำ 2-fold cross validation เพื่อให้ได้ผลการรู้จำที่น่าเชื่อถือด้วย อีกทั้งทำการทดสอบในแต่ละกรณีซ้ำ 3 ครั้งเพื่อหาค่าเฉลี่ยของผลลัพธ์ แล้วแสดงประสิทธิภาพในรูปของเปอร์เซ็นต์ความถูกต้องของการรู้จำ

4. ผลการทดลอง

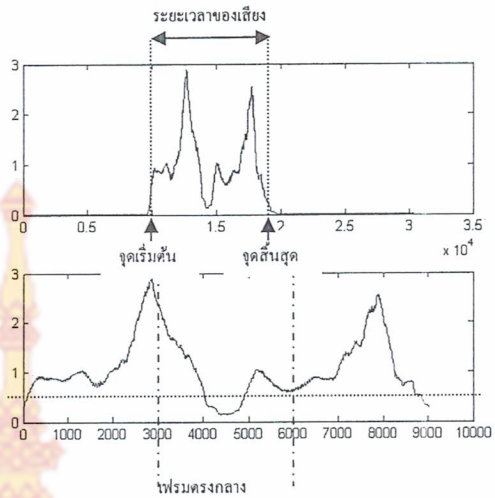
รูปที่ 2 แสดงตัวอย่างสัญญาณเสียงพูดคำว่า เปิด (On), ปิด (Of), หลอดไฟ (Lp), พัดลม (Fn) และทีวี (TV) ที่ถูกบันทึกนาน 3 s (อัตราการสุ่ม 10 kHz, 30000 points) ส่วนรูปที่ 3 แสดงค่าคอนโวลูชั่นของเสียงพูดที่สอดคล้องกับสัญญาณในรูปที่ 2 จะเห็นได้ว่าค่าคอนโวลูชั่นสามารถแสดงขอบเขตเฉพาะเสียงพูดได้



รูปที่ 2 สัญญาณเสียงพูดคำว่า เปิด, ปิด, หลอดไฟ, พัดลม และทีวี

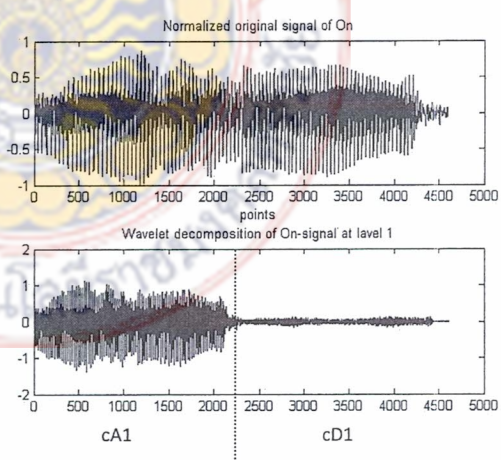


รูปที่ 3 ค่าคอนโวลูชั่นของเสียงพูดเปิด, ปิด, หลอดไฟ, พัดลม และทีวี

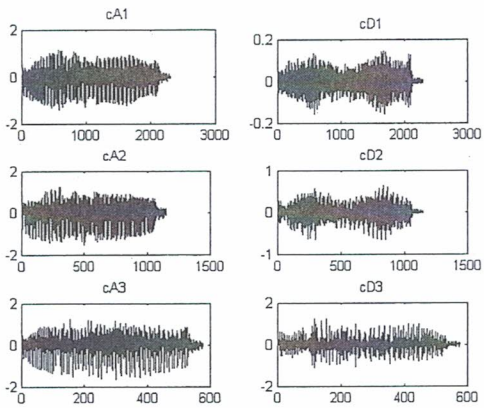


รูปที่ 4 a) ระยะเวลาของเสียงพูดของคำว่าทีวี b) เฟรมตรงกลาง

ค่าคอนโวลูชั่นของเสียงพูดคำว่าทีวีในรูปที่ 4a แสดงให้เห็นระยะเวลาของเสียงพูดซึ่งเป็นผลต่างระหว่างจุดเริ่มต้นและจุดสิ้นสุดของสัญญาณเสียง และเมื่อพิจารณาเฟรมตรงกลางของเสียงพูดในรูปที่ 4b ตามเงื่อนไขการหาจำนวนพยางค์ในข้อ 2.1 ทำให้พิจารณาได้ว่าสัญญาณเสียงนี้มี 2 พยางค์ รูปที่ 5a แสดงเสียงพูดที่ถูกนอร์มอลไลซ์ของคำว่าเปิด ส่วนรูปที่ 5b แสดง Wavelet Decomposition ที่ระดับ 1 ของสัญญาณในรูปที่ 5a ซึ่งประกอบไปด้วยองค์ประกอบความถี่ต่ำ (cA1) และองค์ประกอบความถี่สูง (cD1) การหา Wavelet Decomposition ทั้ง 3 ระดับของสัญญาณดังกล่าวแสดงองค์ประกอบความถี่ต่ำและสูงที่ระดับต่างๆ ได้ในรูปที่ 6



รูปที่ 5 a) เสียงพูดที่ถูกนอร์มอลไลซ์ของคำว่าเปิด
b) Wavelet Decomposition ที่ระดับ 1 ของสัญญาณใน 5a



รูปที่ 6 องค์ประกอบสัญญาณความถี่ต่ำ cA1, cA2, cA3 และ องค์ประกอบสัญญาณความถี่สูง cD1, cD2, cD3 ของสัญญาณในรูปที่ 5a

ผลการรู้จำที่เกิดจากค่าเฉลี่ยจำนวน 3 ครั้ง และทำ 2-fold cross validation ของการทดสอบ 7 คุณลักษณะเด่นกับโครงข่ายประสาทเทียม พบว่า เปอร์เซ็นต์ความถูกต้องของการรู้จำแสดงในตารางที่ 1

ตารางที่ 1 เปอร์เซ็นต์ความถูกต้องของการรู้จำ (PR)

คุณลักษณะเด่น	PR
NS + LPC	82.0
NS + cA1_LPC	82.3
NS + cA2_LPC	92.0
NS + cA3_LPC	81.7
NS + cD1_LPC	84.7
NS + cD2_LPC	86.0
NS + cD3_LPC	86.0

5. สรุป

สัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 ที่ได้จากสัญญาณเสียงพูดหรือจากองค์ประกอบความถี่ต่ำหรือสูงของ Wavelet Decomposition ที่ระดับ 1-3 ร่วมกับจำนวนพยางค์สามารถใช้เป็นคุณลักษณะเด่นสำหรับการรู้จำเสียงพูดสำเนียงภาคใต้ของคำสั่งเปิด ปิด หลอดไฟ พัดลม และทีวีได้ โดยที่สัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 จากสัญญาณองค์ประกอบความถี่ต่ำของ Wavelet Decomposition ที่ระดับ 2 ร่วมกับจำนวนพยางค์ให้เปอร์เซ็นต์ความถูกต้องของการรู้จำสูงสุดมีค่าเท่ากับ 92.0 เมื่อเปรียบเทียบกับผลการรู้จำของงานวิจัยก่อนหน้า [7] ที่ใช้ข้อมูลทดสอบเดียวกัน พบว่า คุณลักษณะเด่นสัมประสิทธิ์การประมาณพันธะเชิงเส้นอันดับ 3 จากสัญญาณองค์ประกอบความถี่ต่ำของ Wavelet Decomposition ที่ระดับ 2 ร่วมกับจำนวนพยางค์ช่วยเพิ่มประสิทธิภาพการรู้จำ และลดความผิดพลาดการรู้จำเสียงพูดคำว่า พัดลม และ หลอดไฟ ที่ทดสอบด้วยวิธีการของงานวิจัยก่อนหน้าได้ อย่างไรก็ตาม การรู้จำเสียงพูดที่ใช้จำนวนผู้พูดมากขึ้นควรได้รับการวิจัยต่อไปเพื่อเป็นการยืนยันประสิทธิภาพของวิธีการเพิ่มเติม

4. กิตติกรรมประกาศ

ขอขอบคุณคณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัยที่ให้การสนับสนุนและส่งเสริมโครงการวิจัยเรื่องการรู้จำเสียงพูดภาษาไทยสำเนียงภาคใต้ของผู้สูงอายุเพื่อควบคุมอุปกรณ์ไฟฟ้า

เอกสารอ้างอิง

[1] ยุทธศาสตร์การวิจัยรายประเด็นด้านผู้สูงอายุและสังคมสูงอายุ (พ.ศ. 2556-2559), สำนักงานคณะกรรมการวิจัยแห่งชาติ.

[2] เกรียงไกร เหลืองอำพล, "การพัฒนาเทคนิคการรู้จำเสียงพูดด้วย DTW กับ LPC และ LSP," วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี, 2553.

[3] P. Thong-in and S. Wongthanavasu, "Microcontroller - Base Thai speech recognition," in the 13th National Computer Science and Engineering Conference, Bangkok, Thailand, vol.13(1), Nov. 4-6, 2009, pp.10-15.

[4] R. Boonsin and C. Jaruskulchai, "Thai Voice Command and Control for PocketPC," in Kasetsart University Conference, 2010.

[5] S. Aunkaew, M. Karnjanadecha, and C. Wutiwiwatchai, "Development of a Corpus for Southern Thai Dialect Speech Recognition: Design and Text Preparation," in the 10th international symposium on natural language processing, Phuket, Thailand, Oct. 28-30, 2013, pp. 147-152.

[6] P. Jantaraprim, T. Chimphet, and C. Jantaraprim, "Speech recognition based electrical device control for central and southern Thai dialects," in the 6th Conference of Electrical Engineering Network of Rajamangala University of Technology 2014 (EENET 2014), Krabi, Thailand, Mar. 26-28, 2014, pp. 909-912.

[7] P. Jantaraprim, T. Chimphet, and K. Inthavisas, "Applying Time and Frequency Domain Features to Improve Recognition of Southern Thai Dialect Speech," in the 7th Conference of Electrical Engineering Network of Rajamangala University of Technology 2015 (EENET 2015), Phattaya, Thailand, May. 27-29, 2015, pp. 230-234.

[8] User's Guide, MATLAB.