รายงานวิจัยฉบับสมบูรณ์

โครงการ การรู้จำอารมณ์จากเสียงพูดที่แสดงความรู้สึกด้วยวิธีการแบ่งกลุ่มผสม (Emotion Recognition of Affective Speech Based on Hybrid Classifiers)

คณะผู้วิจัย

นายภูสิต กุลเกษม	หัวหน้าโครงการวิจัย
นางสาวสุวรรณา รัศมีขวัญ	ผู้ร่วมวิจัย
นางสาวเบญจภรณ์ จันทรกองกุล	ผู้ร่วมวิจัย
นางสาวสุนิสา ริมเจริญ	ผู้ร่วมวิจัย
นายกฤษณะ ชินสาร	ผู้ร่วมวิจัย
นายปิยตระกูล บุญทอง	ผู้ช่วยนักวิจัย
นายมานิต ชาญสุภาพ	ผู้ช่วยนักวิจัย

โครงการวิจัยประเภทงบประมาณเงินรายได้
จากเงินอุดหนุนรัฐบาล (งบประมาณแผ่นดิน)
ปีงบประมาณ พ.ศ. ๒๕๕๗
มหาวิทยาลัยบูรพา

รหัสโครงการ 2557A10802243
เลขที่สัญญา 31/2557

รายงานวิจัยฉบับสมบูรณ์
โครงการ การรู้จำอารมณ์จากเสียงพูดที่แสดงความรู้สึกด้วยวิธีการแบ่งกลุ่มผสม (Emotion Recognition of Affective Speech Based on Hybrid Classifiers)

คณะผู้วิจัย

นายภูสิต กุลเกษม	หัวหน้าโครงการวิจัย
นางสาวสุวรรณา รัศมีขวัญ	ผู้ร่วมวิจัย
นางสาวเบญจภรณ์ จันทรกองกุล	ผู้ร่วมวิจัย
นางสาวสุนิสา ริมเจริญ	ผู้ร่วมวิจัย
นายกฤษณะ ชินสาร	ผู้ร่วมวิจัย
นายปิยตระกูล บุญทอง	ผู้ช่วยนักวิจัย
นายมานิต ชาญสุภาพ	ผู้ช่วยนักวิจัย

คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา

กันยายน 2558

กิจกรรมประกาศ

งานวิจัยนี้ได้รับการสนับสนุนการวิจัยจากงบประมาณเงินรายได้จากเงินอุดหนุนรัฐบาล (งบประมาณแผ่นดิน) ประจำปีงบประมาณ พ.ศ. 2557 มหาวิทยาลัยบูรพา ผ่านสำนักงานคณะกรรมการการวิจัยแห่งชาติ เลขที่สัญญา 31/2557

คณะผู้วิจัย
กันยายน 2558

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาและค้นคว้าอัลกอริทึมในการรู้จำอารมณ์จากเสียงพูดภาษาไทย งานวิจัยประเภทนี้รู้จักกันโดยทั่วไปว่า “การคณนาเชิงอารมณ์” ซึ่งสามารถลดช่องว่างในการสื่อสารระหว่างผู้ใช้กับคอมพิวเตอร์หรือช่วยพัฒนาความฉลาดทางด้านอารมณ์ให้กับคอมพิวเตอร์ เพื่อให้คอมพิวเตอร์สามารถเลือกตอบสนองกับมนุษย์ได้อย่างเหมาะสมยิ่งขึ้น

ในงานวิจัยนี้ได้นำเสนอวิธีคัดเลือกคุณลักษณะแบบฟิชเชอร์สกอร์ (Fisher’s Score) สำหรับการจำแนกอารมณ์ 4 อารมณ์ ได้แก่ อารมณ์เศร้า, โกรธ, มีความสุข และ กลัว ผู้วิจัยได้เลือกใช้เสียงพูดภาษาไทยเนื่องด้วยว่าในภาษาไทย ระดับเสียงพูดที่ใช้ จะมีผลต่อความหมายที่เปลี่ยนไปของคำนั้นๆ ซึ่งถือว่ามีความท้าทายและน่าสนใจอย่างมากในการจำแนกอารมณ์ วิธีที่นำเสนอจะแบ่งออกเป็น 2 ส่วนด้วยกัน ในส่วนแรก เสียงพูดภาษาไทยจะถูกสกัดเพื่อดึงเอา 14 คุณลักษณะเด่นของสัญญาณเสียงออกมา แล้วจึงนำมาคัดเลือกเฉพาะคุณลักษณะที่เหมาะสมกับการรู้จำเสียงภาษาไทยโดยใช้วิธีคัดเลือก Fisher’s Score ส่วนสุดท้าย คุณลักษณะที่คัดเลือกแล้วจะผ่านโครงข่ายการเรียนรู้ 2 แบบเพื่อเปรียบเทียบประสิทธิภาพในการจำแนก จากผลการทดลองที่ได้แสดงให้เห็นว่า การคัดเลือกคุณลักษณะแบบฟิชเชอร์สกอร์กับวิธีจำแนกผ่านโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น ให้อัตราการรู้จำอารมณ์จากเสียงพูดภาษาไทยสูงถึง 95%

Abstract

This research is aimed at studying and developing an algorithm to recognize human feeling or emotion from Thai speech. This type of research is commonly known as “affective computing”. Affective Computing is intended to reduce the communication gap between human and machine or to increase the intelligence to the computer. This type of research is done to raise the efficiency of human and computer interaction.

In this research, we propose Fisher Feature Selection for Emotion Recognition of Thai Speech to classify 4 different emotions of human speech: Sad, Angry, Happy and Fear. The essence of our work lies on the inherit difficulty on different tones of the sound made different meanings in Thai Language. The approach has been divided into two steps. For the first step, the human sound is extracted to get the 14 dominant features using Fisher Feature Selection. Then in step two, two different structures of learning networks are used to compare the classification performance. The results showed that with the use of Fisher Feature Selection as a feature selection method combines with Multi Layers Perceptron as a learning network offers a distinctive recognition emotion of Thai Speech at the rate of 95%.

สารบัญ

บทที่ 1 บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบัน เทคโนโลยีเริ่มเข้ามามีบทบาทสำคัญในการดำเนินชีวิตของมนุษย์มากขึ้น และมากขึ้นเรื่อยๆ เทคโนโลยีนั้นมีความสัมพันธ์กับการดำรงชีวิตของมนุษย์มาอย่างยาวนาน ตัวอย่างเช่น โทรศัพท์ วิทยุ เทคโนโลยีเหล่านี้เข้ามาช่วยอำนวยความสะดวก และเริ่มเปลี่ยนแปลงรูปแบบการใช้ชีวิตของมนุษย์ จวบจนทุกวันนี้มนุษย์เริ่มต้องพึ่งพาเทคโนโลยี ในเกือบทุกกิจกรรม และช่วงอายุวัย ตั้งแต่แรกเกิด จนแก่ชรา ด้วยความสำคัญที่กล่าวมานี้ เทคโนโลยีจึงถูกพัฒนาให้ใกล้ชิดกับมนุษย์มากขึ้น มีความเข้าใจมนุษย์มากขึ้นเพื่อที่จะได้ลดช่องว่างในการสื่อสารระหว่างกันลง หากคอมพิวเตอร์สามารถรับรู้อารมณ์ของมนุษย์ได้ ก็จะทำให้สามารถเลือกตอบสนองต่อมนุษย์ได้อย่างถูกต้องเหมาะสมขึ้น การที่คอมพิวเตอร์สามารถโต้ตอบโดยคำนึงถึงอารมณ์ด้วย ส่งผลให้มนุษย์รู้สึกเป็นมิตร และ รู้สึกสนุกสนานมากมากขึ้นไปด้วย

แต่ยังคงเป็นการยากที่คอมพิวเตอร์จะสามารถทำความเข้าใจมนุษย์ซึ่งมีความละเอียดซับซ้อนมาก สิ่งที่คอมพิวเตอร์จะทำได้จึงเป็นเพียงความเข้าใจแบบเทียมๆผ่านปัญญาประดิษฐ์ (*Artificial Intelligence)* และการรู้จำ (*Recognition)* การรับรู้ของคอมพิวเตอร์กำลังเป็นที่สนใจและได้รับการพัฒนามาอย่างต่อเนื่อง อาทิเช่น การรู้ถึงการเคลื่อนไหวของมนุษย์ (*Human motion recognition),* การจดจำใบหน้า (*Facial recognition)* และ การรู้จำเสียงพูด (*Speech recognition)* แต่ส่วนที่มีความยากคือการรับรู้ถึงอารมณ์ของมนุษย์ (*Emotion recognition)* ซึ่งมีหลากหลายและแสดงออกได้ในหลายทางซ้ำยังมีรูปแบบคล้ายคลึงกันในบางอารมณ์ โดยในงานวิจัยนี้เลือกใช้เสียงพูดเพื่อจำแนกอารมณ์ของผู้พูด (*Speech emotion recognition)* จำแนกอารมณ์ต่างๆด้วยคุณลักษณะของเสียง เช่น ระดับเสียง,ความถี่ งานชิ้นนี้เลือกใช้เสียงพูดภาษาไทย เนื่องด้วยว่า งานวิจัยเกี่ยวกับการจำแนกอารมณ์จากเสียง ในรอบหลายปีที่ผ่านมานั้น มีการจำแนกอารมณ์จากเสียงภาษาไทยน้อยมาก นอกจากนั้น คำๆเดียวกันในภาษาไทย หากใช้ระดับเสียงที่ต่างกัน ความหมายที่ได้ก็แตกต่างออกไป ซึ่งถือว่ามีความท้าทายและน่าสนใจอย่างมากสำหรับการจำแนกอารมณ์ งานวิจัยชิ้นนี้จะแบ่งออกเป็น 2 ส่วนหลักด้วยกัน ส่วนที่หนึ่ง เสียงพูดภาษาไทยจะถูกนำมาสกัดหา 14 คุณลักษณะเด่นของเสียง

(*Features)* แล้วจึงผ่านกระบวนการคัดเลือกคุณลักษณะแบบฟิชเชอร์ (*Fisher score or F-Score)* ในส่วนสุดท้าย เป็นส่วนการจำแนกอารมณ์และการรับรู้ โดยใช้โปรแกรม *WEKA* ข้อมูลคุณลักษณะที่ได้จะถูกนำเข้า โครงข่ายประสาทเทียมชนิดแพร่กลับ (*BPNN)* และ โครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน (*RBF)* เพื่อวัดประสิทธิภาพและเปรียบเทียบเป็นร้อยล่ะความถูกต้องในการจำแนก

1.2 *วัตถุประสงค์ของโครงการวิจัย*

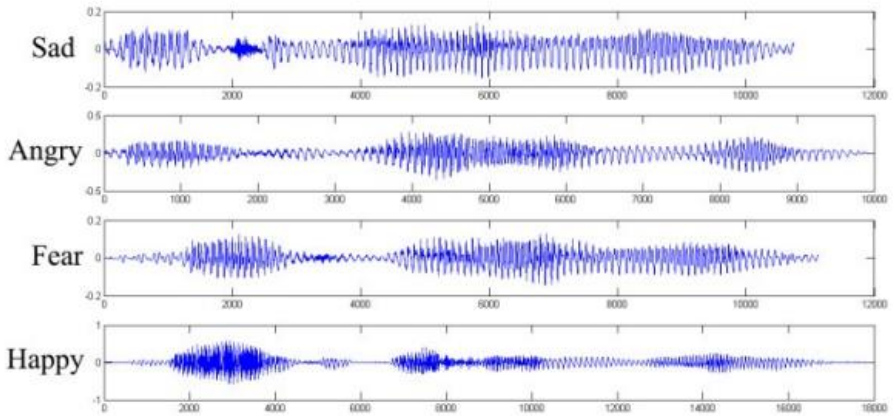
1. เพื่อศึกษาเทคนิคการวิเคราะห์เสียงพูดที่มีอารมณ์และความรู้สึก สำหรับใช้ในการสกัดคุณลักษณะของข้อมูลสำหรับการนำเข้า
2. พัฒนาซอฟท์แวร์ที่สามารถจำแนกอารมณ์ของเสียงพูดมนุษย์ ซึ่งจะเป็นประโยชน์ในการพัฒนาเทคโนโลยีที่มีปฏิสัมพันธ์กับมนุษย์ *(Affective Computing)*
3. เพื่อให้ผู้ที่สนใจสามารถนำแนวความคิดที่นำเสนอ ไปศึกษาเพื่อทำการพัฒนาหรือประยุกต์ใช้ในงานวิจัยของตนเองต่อไป

1.3 *ขอบเขตของโครงการวิจัย*

การวิจัยครั้งนี้มุ่งเน้นที่จะพัฒนาซอฟท์แวร์จำแนกอารมณ์ เศร้า, โกรธ, มีความสุข และ กลัว จากเสียงพูดภาษาไทย งานวิจัยชิ้นนี้จะประกอบไปด้วยค้นคว้าและพัฒนาหลากหลายอัลกอริธึมการประมวลผลเสียง ได้แก่ ขั้นตอนการเตรียมข้อมูลเสียงก่อนการประมวลผล โดยทำการแบ่งข้อมูลออกเป็นส่วนๆแล้วใช้ กรอบสัญญาณแฮมมิงจะลดความไม่ราบรื่นของสัญญาณเสียงเพื่อหลีกเลี่ยงการรั่วของสเปกตรัม การค้นหาตัวสกัดคุณลักษณะที่เหมาะสมกับเสียงพูดภาษาไทยและการคัดเลือกคุณลักษณะที่มีความซ้ำซ้อนออกเพื่อลดเวลาในการประมวลผลและคงประสิทธิภาพไว้เสียงพูดที่นำมาใช้ในงานวิจัยชิ้นนี้ ได้ถูกรวบรวมจากกลุ่มตัวอย่างจำนวน 6 คน ในแต่ละคำจะประกอบไปด้วย 1-7 พยางค์พูด ซึ่งถูกนำมาใช้ในงานวิจัยจำนวน 800 แฟ้มเสียง

1.4 วิธีการดำเนินงานวิจัยโดยสรุป

ในงานวิจัยนี้ ได้กำหนดขั้นตอนหลักของการทำงานวิจัยไว้ 3 ขั้นตอนหลัก ได้แก่ ขั้นตอนการเตรียมข้อมูลก่อนการประมวลผล (*pre-processing)* ซึ่งจะเป็นการแบ่งข้อมูลเสียงออกเป็นส่วนๆแล้วจึงนำวิธี กรอบสัญญาณแฮมมิง (*Hamming window)* มาใช้ลดความไม่ราบรื่นของสัญญาณข้อมูลเสียง เพื่อหลีกเลี่ยงการรั่วของสเปกตรัม (*Spectral Leakage)* จากนั้นข้อมูลเสียงจะเข้าสู่กระบวนการที่สอง คือกระบวนการสกัดคุณลักษณะและคัดเลือกคุณลักษณะ คุณลักษณะที่ได้จะมีด้วยกัน 14 คุณลักษณะ เช่น พลังงานของเสียง,ค่าอัตราการตัดผ่านศูนย์ แต่ละคุณลักษณะเหล่านี้จะถูกนำมาคำนวณหาค่าทางสถิติ 5 ค่า ได้แก่ ค่าเฉลี่ย(*Mean),* มัธยฐาน(*Median),*ค่าสูงสุด(*Max),* ค่าต่ำสุด(*Min),* ความแปรปรวน(*Variance)* ดังนั้น ในหนึ่งข้อมูลเสียงก็จะประกอบด้วย 70 ค่าคุณลักษณะ ซึ่งจะถูกนำมาคัดเลือก โดยวิธีการทางสถิติ คือคัดเลือกคุณลักษณะแบบ *F-Score* เพื่อช่วยกำจัดคุณลักษณะที่มีความซ้ำซ้อนกับคุณลักษณะอื่นๆ และไม่มีส่วนพัฒนาผลในกระบวนการเรียนรู้ ข้อมูลที่ได้จากการคัดเลือก และ ข้อมูลที่ไม่ได้ถูกกำจัดคุณลักษณะใดออก จะถูกนำเข้าสู่ วิธีการจำแนก 2 แบบ ได้แก่โครงข่ายประสาทเทียม *BPNN* และ *RBF* เพื่อหาประสิทธิภาพในการจำแนก จากนั้นผลที่ได้จากทั้ง 2 วิธี จะนำมาเปรียบเทียบอีกครั้ง ด้วยการตรวจสอบความถูกต้องแบบข้อผิดพลาดรากค่าเฉลี่ยกำลังสอง (*Root mean square errors RMSE)*



รูปที่ 1-1 คลื่นเสียง 4 อารมณ์ของคำว่า “โดยเฉพาะอย่างยิ่ง”

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ซอฟต์แวร์ ที่สามารถจำแนกอารมณ์ ไฟล์เสียงพูดคำภาษาไทย โดยจำแนกได้ 4 อารมณ์ เศร้า, โกรธ, มีความสุข, กลัว
2. สามารถนำไปพัฒนาระบบช่วยตัดสินใจสำหรับปัญหาการรู้จำเสียงพูดที่มีอารมณ์ ซึ่งเกิดขึ้นเป็นประจำในชีวิตประจำวัน เช่น ระบบการรู้จำเสียงในห้องประชุม เป็นต้น
3. ซอฟต์แวร์นี้สามารถนำไปประยุกต์ใช้ กับเทคโนโลยีทางด้านการคณนาเชิงอารมณ์ *(Affective Computing)* เพื่อพัฒนาให้คอมพิวเตอร์มีความฉลาดทางด้านอารมณ์มากขึ้น

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การรู้จำอารมณ์จากเสียงพูดที่แสดงความรู้สึก

อารมณ์ *(Emotion)* เป็นปัจจัยพื้นฐานสำคัญอย่างหนึ่งต่อประสิทธิภาพหรือความสำเร็จของการรู้จำในงานทางปัญญาประดิษฐ์ ทั้งนี้เพราะ เมื่ออารมณ์เปลี่ยนก็จะทำให้คุณลักษณะของเสียง *(Feature)* เปลี่ยนไปตาม ซึ่งผลการรู้จำก็จะเปลี่ยนตามไปด้วย ดังนั้น ปัญหาการรู้จำเสียงซึ่งมีอารมณ์เป็นองค์ประกอบจึงเป็นปัญหาที่มียากตามมาโดยหลีกเลี่ยงไม่ได้ ซึ่งถ้านักวิจัยสามารถพัฒนาขั้นตอนวิธีสำหรับการรู้จำเสียงที่มีอารมณ์ *(Emotion Recognition)* ได้ประสบความสำเร็จจะทำให้การพัฒนาของระบบ *Human-Machine Interactions (HMI)* จะเป็นไปอย่างก้าวกระโดดอย่างแน่นอน ซึ่งการที่จะทำให้ระบบการรู้จำเสียงที่มีอารมณ์มีประสิทธิภาพนั้น จะต้องมีการพัฒนาความสามารถให้ระบบคอมพิวเตอร์มีความสามารถในการตรวจจับคุณลักษณะซึ่งมีความแปรปรวนสูง *(Detect or Feature Extraction)* จากนั้นก็จะเข้าสู่ขั้นตอนการรู้จำซึ่งต้องมีความยืดหยุ่นเช่นเดียวกัน

2.2 การประมวลผลความรู้สึก

การประมวลผลความรู้สึก (*Affective Computing*) หมายถึง การประมวลผลที่เกี่ยวข้อง เกิดจาก หรือ มีผลกระทบกับอารมณ์ ซึ่งเกิดขึ้นมาจากแนวคิดที่ต้องการการลดช่องว่างในการสื่อสารระหว่างมนุษย์ซึ่งมีอารมณ์ในการตัดสินใจกับคอมพิวเตอร์ โดยการพัฒนาระบบคอมพิวเตอร์ให้สามารถรู้จำและตอบสนองต่อสถานะของความรู้สึกและอารมณ์ของมนุษย์แบบเวลาจริง โดยในระหว่างการโต้ตอบกันระหว่างมนุษย์กับคอมพิวเตอร์นั้น คอมพิวเตอร์สามารถที่จะเรียนรู้เพื่อเพิ่มประสิทธิภาพของการปฏิสัมพันธ์ได้ นั้นคือ จะทำให้การติดต่อกับคอมพิวเตอร์มีความน่าใช้มากยิ่งขึ้น มีความร่าเริงและสนุกสนานมากขึ้น และ มีประสิทธิภาพมากขึ้น

2.3 การประมวลผลเสียงพูด

เสียงพูดถูกสร้างจากเวลาที่เปลี่ยนแปลงเป็นตัวกระตุ้น ซึ่งตามลักษณะผลลัพธ์ของสัญญาณเสียงพูดจึงไม่คงที่ *(Non-stationary)* โดยเกือบทั้งหมดของเครื่องมือที่ใช้ใน

การประมวลผลสัญญาณทำการศึกษาเกี่ยวกับ รูปแบบสัญญาณ และ การประมวลผลสัญญาณ โดยมีสมมุติฐานของระบบเวลาคงที่ การกระตุ้นเวลาคงที่และสัญญาณคงที่ จากกรณีตัวอย่าง เครื่องมือสำหรับการคำนวณพลังงานทั้งหมดของความสัมพันธ์ขั้นปฐมภูมิในการประมวลผลเสียงพูด ดังนี้

$$E_T = \sum_{n=-\infty}^{\infty} S^2(n) \qquad (1)$$

ความสัมพันธ์ของสัญญาณคงที่จะมีพลังงานจำกัดสมมุติว่าใช้เครื่องมือนี้สำหรับการคำนวณพลังงานทั้งหมดของสัญญาณเสียงพูด ซึ่งจะเห็นได้ว่าพลังงานทั้งหมดก็อยู่ในสัญญาณเสียงพูด อย่างไรก็ตามพลังงานทั้งหมดก็ยังไม่ได้ถูกใช้ไป เพราะว่าเสียงพูดตามลักษณะนั้น เรารู้ว่าเสียงพูดมีทั้งเวลาที่เปลี่ยนแปลงของแอมปิจูดและพลังงาน เช่นนั้นอะไรคือสิ่งสำคัญของการสร้างเสียงพูดของเครื่องมือ ที่ให้ข้อมูลเกี่ยวกับเวลาที่เปลี่ยนแปลงตามพลังงาน ดังนั้นแล้วจึงต้องมีแนวทางการประมวลผลสัญญาณเสียงพูดที่แตกต่างกัน

กระบวนการทางวิศวกรรมได้นำเสนอแนวทางสำหรับการประมวลผลเสียงพูดจากการใช้สัญญาณที่ยังคงอยู่ในการประมวลผลของเครื่องมือทีมีการปรับแต่งรูปแบบที่เฉพาะเจาะจงเครื่องมือนี้ก็ยังคงอยู่ในสมมุติฐานการประมวลผลสัญญาณคงที่ สัญญาณเสียงพูดอาจจะคงที่เมื่ออยู่ช่วงเวลาที่จำกัด ประมาณ *10-30* มิลลิวินาที ดังนั้นกระบวนการของเสียงพูดจะใช้เครื่องมือในการประมวลผลสัญญาณที่แตกต่างกัน ซึ่งเรียกกระบวนการนี้ว่า *“การประมวลผลระยะสั้น (Short Time Processing: STP)”* การประมวลผลระยะสั้นของเสียงพูดสามารถแสดงทั้งในขอบเขตของเวลาหรือขอบเขตของความถี่ ขอบเขตของกระบวนการขึ้นอยู่กับข้อมูลจากเสียงพูดที่เราสนใจ สำหรับตัวอย่าง, พารามิเตอร์ เหมือนกันทั้ง พลังงานระยะสั้น (*Short Time energy*), อัตราการผ่านค่าศูนย์ระยะสั้น (*Short Time zero crossing rate*) และความสัมพันธ์คลาดเคลื่อนระยะสั้น *(Short Time autocorrelation)* สามารถนำมาคำนวณขอบเขตเวลาการประมวลผลเสียงพูด อีกหนึ่งทางเลือก คือ *Short Time Fourier transform* สามารถนำมาคำนวณขอบเขตความถี่การประมวลผลเสียงพูด ในแต่ละพารามิเตอร์นั้นให้ข้อมูลแตกต่างกันเกี่ยวกับเสียงพูดถูกนำมาใช้สำหรับการประมวลผลอัตโนมัติ

2.4 การเตรียมข้อมูลก่อนการประมวลผล

การเตรียมข้อมูลก่อนการประมวลผล (*Pre-processing*) เริ่มแรกสัญญาณเสียงพูดที่รับเข้ามาจะผ่านตัวกรองเน้นล่วงหน้า *(Pre-emphasis filtering)* เพื่อลดสัญญาณรบกวน กระบวนการนี้สัญญาณเสียงจะถูกบีบอัดในช่วงพิสัยพลวัต *(Dynamic range)* ส่งผลให้อัตราส่วนระหว่างสัญญาณเสียงและสัญญาณรบกวน *(Signal to Noise Ratio)* ปรับตัวสูงขึ้น โดยผ่านการนำสัญญาณเสียงพูดเข้าสู่วงจรกรองสัญญาณลำดับที่หนึ่ง *(first-order filter)* วงจรนี้มีฟังก์ชั่นการถ่ายโอนดังสมการที่ 2 และ สมการที่ 3

$$H(z) = 1 - az^{-1} \quad , \quad 0.9375 \leq a \leq 1 \qquad (2)$$

$$\tilde{s}(n) = s(n) - as(n-1) \qquad (3)$$

กำหนดให้

a - คือค่าสัมประสิทธิ์ของวงจรกรอง

$s(n)$ - คือสัญญาณเสียงพูดที่รับเข้ามา

$\tilde{s}(n)$ - คือสัญญาณเสียงที่ผ่านตัวกรองเน้นล่วงหน้าแล้ว

หลังจากสัญญาณเสียงผ่านตัวกรองเน้นล่วงหน้าแล้วก็จะถูกแบ่งเป็นกรอบสัญญาณด้วยวิธีกรอบสัญญาณแฮมมิง*(Hamming window)* โดยทั่วไปจะแบ่งสัญญาณเสียงเป็นกรอบ ขนาด 10-30มิลลิวินาที และในกรอบถัดๆไปควรมีส่วนเหลื่อมกันคิดเป็น 1/2 หรือ 1/3 ของขนาดกรอบสัญญาณเพื่อให้สัญญาณมีความต่อเนื่องกัน การแบ่งกรอบสัญญาณแบบแฮมมิ่งมีสมการ ดังนี้

$$w(n) = \begin{cases} 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) & , 0 \leq n \leq N-1 \\ 0, & \end{cases} \qquad (4)$$

กำหนดให้

$w(n)$ - คือผลลัพธ์ของฟังก์ชั่นกรอบสัญญาณตำแหน่งที่ n

N - คือขนาดของกรอบ

n - คือข้อมูลในกรอบมีค่าตั้ง 0 ถึง $N-1$

2.5 การสกัดคุณลักษณะเด่น

ในงานวิจัยนี้สัญญาณเสียงพูด จะถูกสกัดดึงลักษณะเฉพาะ (*Feature extraction*) ของหน่วยเสียงแต่ละหน่วยที่แตกต่างกันออกมา กระบวนการเรียนรู้จะทำการจดจำคุณลักษณะของเสียงในแต่ละกลุ่มไว้จนหมด เพื่อเปรียบเทียบเมื่อเริ่มทำการแบ่งกลุ่ม หากสัญญาณเสียงมีคุณลักษณะต่างๆเหมือนหรือใกล้เคียงกับกลุ่มใดก็จะสามารถระบุได้ว่าสัญญาณนั้นเป็นสัญญาณในกลุ่มใด การสกัดคุณลักษณะยังช่วยในการลดปริมาณข้อมูลที่ต้องนำมาวิคราะห์ ประมวลผล โดยไม่จำเป็นต้องใช้ข้อมูลสัญญาณเสียงทั้งหมดมาเปรียบเทียบกัน สิ่งที่นำมาวิเคราะห์จะเป็นเพียงคุณลักษณะที่สกัดออกมาเท่านั้น และยังคงคุณสมบัติที่สำคัญๆของข้อมูลไว้ได้ งานวิจัยนี้จะสกัดคุณลักษณะของเสียงพูดภาษาไทย ออกมาด้วยกัน 14 คุณลักษณะ ดังนี้

2.5.1 Energy Entropy Block

Energy Entropy Block ($\boldsymbol{E_e}$) เป็นค่าที่แทนถึงคุณสมบัติของอุณหพลศาสตร์ที่สามารถนำมาใช้ประเมินปริมาณพลังงานที่พร้อมใช้งานที่เป็นประโยชน์ในกระบวนการ *(Entropy)* ของข้อมูลภายในแต่ละกลุ่มย่อยที่แบ่ง ซึ่งสามารถคำนวณได้จากสัญญาณเสียงที่รับเข้ามา โดยเริ่มแรกสัญญาณจะถูกแบ่งเป็นส่วนๆ จำนวน f ส่วน แล้วจึงทำการลดความซ้ำซ้อน *(Normalize)* ค่าพลังงานของแต่ละส่วน เพื่อง่ายต่อการคำนวณหาค่าเอนโทรปีในแต่ละส่วนย่อยที่ถูกแบ่ง ค่าเอนโทรปีสามารถคำนวณได้ตามสมการ ดังนี้

$$E_e = -\sum_{k=0}^{f-1} \mu^2 . log_2(\mu^2) \qquad (5)$$

$$\mu^2 = \sum_{b=1}^{N} \frac{(N * \frac{W_1}{S_b})}{F} \qquad (6)$$

กำหนดให้

μ^2 – คือ ค่าพลังงานที่ถูกนอร์มัลไลซ์แล้ว

N – คือ จำนวนของกลุ่มที่แบ่ง

W_1 – คือ ความยาวของส่วนที่แบ่ง

S_b – คือ จำนวนของกลุ่มย่อย

F – คือ ความถี่

2.5.2 Short Time Energy

พลังงานที่เกี่ยวข้องกับเสียงพูด (*Short Time Energy:* S_e) คือ ระยะเวลาที่เปลี่ยนแปลงตามลักษณะ ดังนั้น เป็นที่น่าสนใจสำหรับการประมวลผลเสียงพูดแบบอัตโนมัติ จึงทราบได้ว่าพลังงานมีการเปลี่ยนแปลงกับระยะเวลาและปัจจัยอื่นๆ พลังงานที่เกี่ยวข้องกับขอบเขตของเสียงพูดระยะสั้น (*Short Time region of speech*) โดยลักษณะสร้างเสียงพูดที่มีสัญญาณประกอบไปด้วยการเปล่งเสียง ไม่ออกเสียง และขอบเขตความเงียบพลังงานที่เกี่ยวข้องกับขอบเขตการเปล่งเสียงถูกนำไปเปรียบเทียบอย่างแพร่หลายกับขอบเขตการไม่ออกเสียงและขอบเขตความเงียบจะไม่เป็นอย่างน้อย หรือ พลังงานที่ไม่สำคัญ *(negligible energy)* ดังนั้นพลังงานระยะสั้นสามารถใช้สำหรับการจัดหมวดหมู่ของเสียงพูดแบบออกเสียง ไม่ออกเสียง และเสียงเงียบ

ในกรณีของการคำนวณพลังงานระยะสั้น จะพิจารณาจากเสียงพูดที่อยู่ในเทอมระยะเวลา *10-30* มิลลิวินาทีตัวอย่างเช่นเฟรมของเสียงพูดจะเป็น *“n=0 to n=N-1”*, ที่ *“N”* คือ ความยาวของเฟรม *(ตัวอย่าง)* ดังนั้นสำหรับการคำนวณพลังงานเสียงพูดจะเป็นศูนย์ *(0)* ที่อยู่นอกความยาวเฟรม สำหรับการคำนวณพลังงานแอมปิจูดของเสียงพูดตัวอย่างจะเป็นศูนย์ *(0)* ที่อยู่นอกความยาวเฟรม เพราะฉะนั้นเราสมารถเขียนความสัมพันธ์ดังกล่าวได้ ดังนี้

$$E_T = \sum_{n=-\infty}^{\infty} S^2(n) + \sum_{n=0}^{N-1} S^2(n) + \sum_{n=N}^{\infty} S^2(n)$$

$$E_T = \sum_{n=0}^{N-1} S^2(n) \qquad (7)$$

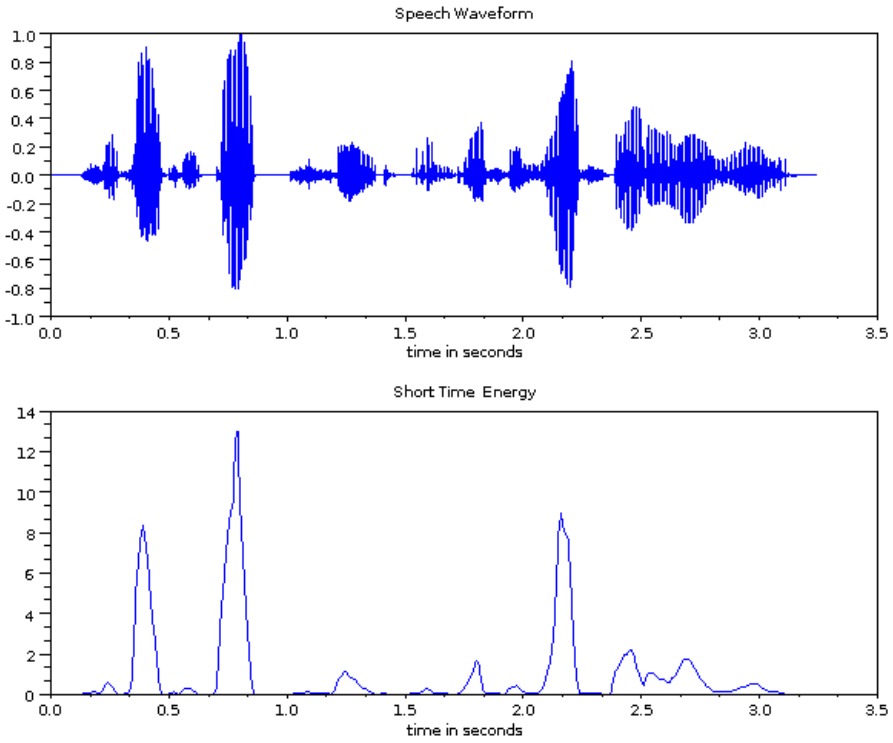
ความสัมพันธ์นี้ จะให้แสดงค่าพลังงานทั้งหมดที่อยู่ในเฟรมของเสียงพูด จาก *n=0 to n=N-1* การแสดงแบบเจาะจง เพียงแค่ *1* เฟรมของเสียงพูด เราให้ความสัมพันธ์ดังนี้

$$S_w(n) = S(m).w(n-m) \qquad (8)$$

เมื่อ $w(n)$ แสดงถึงของช่วงเวลาสุดท้าย ซึ่งมีหลายช่วงที่แสดงอยู่ในประมวลผลสัญญาณ ส่วนใหญ่แล้วจะใช้รวมกับ *rectangular, hanning and hamming* สำหรับการประมาณการขอบเขตเวลาของพารามิเตอร์ เราใช้ *rectangular window* เพื่อให้มีความใช้งานง่าย สามารถเขียนสูตรของความสัมพันธ์ของพลังงานระยะสั้น ได้ดังนี้

$$e(n) = \sum_{m=-\infty}^{\infty} (s(m).w(\mathrm{n}-\mathrm{m}))^2 \qquad (9)$$

เมื่อ *"n"* คือ การเปลี่ยนแปลง (*Shift*) / จำนวนอัตราของตัวอย่าง ซึ่งเป็นสิ่งที่เราสนใจ เกี่ยวกับ *Short Time energy* การเปลี่ยนแปลงอาจมีขนาดเล็ก ตัวอย่างหนึ่งหรือ เปลี่ยนแปลงความยาวของขนาดเฟรม *Short Time energy* ถูกให้คำนวณในทุกการเปลี่ยนแปลงตัวอย่าง หรืออาจจะไม่ใช้ตั้งแต่แรกของการเปลี่ยนแปลงของพลังงานเสียงที่ค่อนข้างช้า ด้วยเหตุนี้การเปลี่ยนแปลงจะถูกเก็บไว้ที่มีขนาดใหญ่กว่าตัวอย่าง ปกติแล้วจะเป็นขนาดครึ่งหนึ่งของขนาดเฟรม



รูปที่ 2-1 รูปร่างของสัญญาณเสียงพูดแบบ *Short Time energy*

จุดสิ้นสุดของพลังงานระยะสั้น คือ ค่าขนาดเฟรม เนื่องมาจากสมมุติฐานคงที่ของกรณีเสียงพูดที่ถูกต้องในช่วง *10* ถึง *30* มิลลิวินาที ตามตัวอย่างค่าของขนาดเฟรม คือ *20* มิลลิวินาที อีกทางเลือกหนึ่งสำหรับขนาดเฟรมที่ใหญ่ เราจะต้องใช้ช่วงของพลังงานที่สม่ำเสมอและจะต้องไม่พบการเปลี่ยนแปลงเวลาตามลักษณะของพลังงานระยะสั้น จากรูปที่ 2-1 แสดงรูปทรงของพลังงานที่เป็นสัญญาณเสียงพูด ที่นำมาใช้ในการศึกษา

2.5.3 Zero Crossing Rate

Zero Crossing Rate (zcr) หมายถึงจำนวนครั้งของแกน *zero* ที่ถูกตัดผ่านในแต่ละเฟรม ซึ่งอัตราการผ่านศูนย์ (*zero*) จะให้ข้อมูลเกี่ยวกับจำนวนของการผ่านศูนย์ของสัญญาณ การคาดการณ์ กล่าวคือ เมื่อสัญญาณเปลี่ยนแปลงอย่างรวดเร็ว แล้วจำนวน

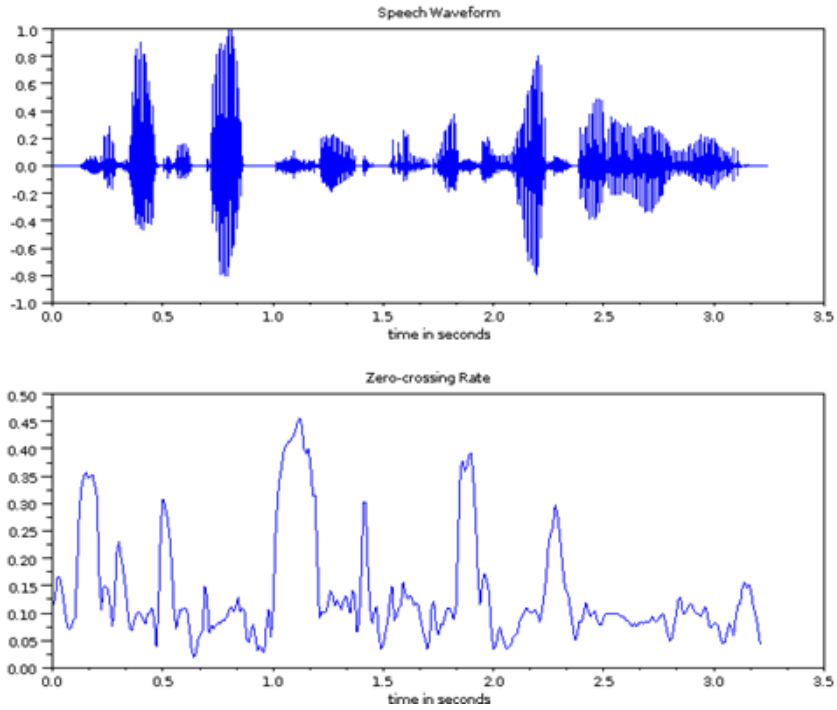
ของการผ่านศูนย์มีมากขึ้นในการรับสัญญาณ นั้นคือ สัญญาณนั้นจะต้องมีข้อมูลความถี่สูงอยู่ ในทำนองเดียวกันสัญญาณที่มีการเปลี่ยนแปลงช้า (หรืออัตราผ่านศูนย์น้อย) จะหมายความว่า สัญญาณนั้นจะต้องมีข้อมูลความถี่ต่ำอยู่ เช่นนั้นแล้ววิธีการ *ZCR* จะให้ข้อมูลทางอ้อมเกี่ยวกับสัญญาณความถี่ที่มีอยู่ของสัญญาณ กรณีที่ *ZCR* เป็นสัญญาณคงที่จะกำหนดให้เป็น

$$z = \sum_{n=-\infty}^{\infty} \left|\mathrm{sgn}(s(n)) - \mathrm{sgn}(s(n-1))\right|$$
$$where\ \ \mathrm{sgn}(\mathrm{s}(n)) = 1 if\ \ s(n) \geq 0$$
$$= -1 if\ \ s(n) < 0 \qquad (10)$$

ความสัมพันธ์จะถูกปรับปรุงสำหรับสัญญาณที่ไม่คงที่ ที่มีความคล้ายกับเสียงพูดและเวลาที่กำหนดเรียกว่า *ZCR* จะกำหนดให้เป็น

$$z(n) = \frac{1}{2N} \sum_{m=0}^{N-1} s(m).w(n-m) \qquad (11)$$

เมื่อ ปัจจัย *“2”* มาเป็นตัวหารที่จากข้อเท็จจริงจะต้องมี 2 ค่าผ่านศูนย์ต่อหนึ่งรอบสัญญาณ



รูปที่ 1-2 สัญญาณเสียงพูดแบบ *Zero Crossing Rate*

ในกรณีเสียงพูดตามลักษณะของสัญญาณเปลี่ยนแปลงกับเวลาที่เกินประมาณ *2-3* มิลลิวินาที จากตัวอย่างเริ่มต้นที่การเปล่งเสียงพูด การไม่ออกเสียงและกลับมาที่การเปล่งเสียง และแบบอื่นๆ จะมีข้อมูลทีเป็นประโยชน์มาก *ZCR* จำเป็นต้องมีการคำนวณโดยใช้ขนาดเฟรมทั่วไปอยู่ที่ *10-30* มิลลิวินาที กับครึ่งของขนาดเฟรมที่เปลี่ยนแปลงสัญญาณเสียงพูดจะแสดงข้อความ *"she had your suit in your greasy wash water all year" และ ZCR* ถูกคำนวณดังแสดงตามรูปที่ *2-2* จะสังเกตได้ว่า กรณีของการไม่ออกเสียงพูดค่าของ *ZCR* อยู่ในระดับสูงอย่างมีนัยสำคัญ ขอบเขตของเสียง ดังนั้น จึงสามารถใช้ *ZCR* ในการแยกความแตกต่างของขอบเขตการเปล่งเสียงและไม่ออกเสียงได้

2.5.4 Spectral-Roll-Off

Spectral-Roll-Off นั้นแสดงถึง ผลรวมของค่าพลังงานสเปกตรัมส่วนที่เบ้ขวา *(right skewness)* เมื่อดูจากกราฟ ส่วนที่เบ้ขวาหรือส่วนความถี่สูงนั้นจะเป็นส่วนที่มีค่า

สเปกตรอล โรล ออฟสูง ซึ่งถูกกำหนดให้เป็น องค์ประกอบทางความถี่รอง *(second frequency bin)* $M_C^R(J)$ ภายใต้ c เปอร์เซ็นของการกระจายขนาดของ *DFT* X_r สมการในการหาค่าสเปกตรอล โรล ออฟสูง มีดังนี้

$$\sum_{k=0}^{M_C^R(j)} \left| X_{jk} \right| = \frac{C}{100} \sum_{k=0}^{S-1} \left| X_{jk} \right| \qquad (12)$$

2.5.5 Spectral Centroid

Spectral Centroid คือ ลักษณะหรือคุณสมบัติของสเปกตรัม ซึ่งคำนวณได้จากการประเมิน จุดศูนย์ถ่วง *(*Center of gravity*)* ซึ่งก็คือจุดที่ทำให้เกิดความสมดุลของทั้งสองด้าน โดยใช้การแปลงฟูเรียร์ความถี่ *(Fourier Transform's Frequency)*ด้วยขนาดของมัน ในแต่ละจุดศูนย์กลาง *(centroid point)* ของส่วนสเปกตรัมที่ถูกแบ่งออกจะหมายถึงความถี่เฉลี่ย ที่คำนวณจากความสูงคลื่นสัญญาณ *(Amplitudes)* และนำมาหารด้วยผลรวมของความสูงคลื่นสัญญาณอีกที ดังที่แสดงในสมการนี้

$$Spectral\ Centroid = \frac{\sum_{k=1}^{N} kF[k]}{\sum_{k=1}^{N} F[k]} \qquad (13)$$

กำหนดให้

$F[k]$ คือ ความสูงคลื่นสัญญาณที่สอดคล้องกับ องค์ประกอบ k ใน *DFT* สเปกตรัม

2.5.6 Fundamental Frequency

ความถี่พื้นฐาน *(Fundamental frequency or Pitch: F_0)* คือความถี่ ที่เกิดขึ้นจากการสั่นของสายเสียงในกล่องเสียงมนุษย์เพื่อทำให้เกิดเสียงออกมาในกาประมวลผลหาค่าความถี่พื้นฐานนั้นมีอุปสรรคอยู่หลายประการ เนื่องจากการคำนวณจำเป็นต้องใช้

ทัพยากรคอมพิวเตอร์มาก และยังยากที่จะกำหนดลักษณะปฏิสัมพันธ์ที่ซับซ้อนระหว่าง F_0 และ *supra-segmental phenomenon*ได้ แต่ F_0 นั้นถือเป็นคุณลักษณะที่มีประโยชน์อย่างมากในภาษาที่ใช้เสียงวรรณยุกต์ *(Tonal language)* เช่นภาษา จีน,ไทย อันเนื่องมาจากภาษาเหล่านี้ ระดับเสียงที่ใช้มีผลต่อการให้ความหมายของคำนั้นๆ การประมวลผล F_0 นั้นจะอยู่บนลอการิทึมสเกลเพื่อให้ผลที่ได้สอดคล้องกับระบบการได้ยินของมนุษย์ ในการคำนวณ F_0 มีด้วยกันหลายวิธีในงานวิจัยชิ้นนี้เลือกใช้วิธี *zero-crossing*

2.5.7 Mel Frequency Cepstral Coefficient

Mel Frequency Cepstral Coefficient (MFCCs) คือค่าพลังงานสเปกตรัมของสัญญาณเสียงในช่วงสั้นๆ *MFCCs* ถูกนำมาใช้ในงานรู้จำเสียงพูดมนุษย์อย่างกว้างขวาง โดยผู้คิดค้นวิธีนี้คือ *Davis* และ *Mermelstein* ในปี ค.ศ. 1980 *MFCCs* นั้นมีกระบวนการทำงานสอดรับกับระบบการได้ยินของมนุษย์หมายความว่ากระบวนการนี้รองรับสัญญาณเสียงตั้งแต่ 1 กิโลเฮิร์ทซ์ ขึ้นไปเท่านั้น ซึ่งเป็นช่วงความถี่ที่มนุษย์จะสามารถได้ยินได้ ทำให้วิธี *MFCCs* นั้นเหมาะที่จะนำมาใช้ในงานรู้จำเสียงพูด การคำนวณค่าสัมประสิทธิ์ *Mel Frequency Cepstral* มีขั้นตอนดังนี้

1. ใช้วินโดว์ฟังก์ชัน
2. คำนวณหาค่าพลังงานสเปกตรัม โดยใช้วิธีแปลงฟูเรียร์แบบเร็ว (FFT)
3. ใช้ *mel-filter bank*
4. ใช้การแปลงโคไซน์แบบไม่ต่อเนื่อง *(DCT)*
5. *MFCCs* คือความสูงของสเปกตรัมที่ได้

$$MFCC_i = \sum_{k=1}^{N} X_k \cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{N}\right], i = 1,2,\dots,M \qquad (14)$$

เมื่อ M คือจำนวนค่าสัมประสิทธิ์ของสเปกตรัม , X_k ,$k = 1,2,\dots,N,$ คือผล *log-energy* ของ k^{th} *filter และ* N *คือจำนวนของ triangular band pass filters* โดยตามมาตรฐานจะมีค่าอยู่ที่ราวๆ 20

2.5.8 Linear Predictive Coding

Linear Predictive Coding (LPC) หรือ การประมาณพันธะเชิงเส้น นั้นอยู่บนพื้นฐานของ *source-filter model* ซึ่งมีด้วยกันสองระบบ คือ ต้นกำเนิดเสียง *(Voice source)* และส่วนที่ใช้ปรับเสียงที่เกิดจากต้นกำเนิดให้เป็นคำ *(Vocal tract transfer)* ซึ่งส่วนกำทอนเสียงนี้ถูกนำมาจำลองเป็น *all-pole filter* โดยมีสมการดังนี้

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \qquad (15)$$

เมื่อ a_i คือ ค่าสัมประสิทธิ์

สัญญาณเสียง s_n จะถูกทำนายจากปัจจุบันจนถึงอดีต p มีสมการดังนี้

$$\hat{s}_n = \sum_{i=1}^{p} a_i s_{n-i} \qquad (16)$$

ในสมการ (12) a_i คำนวณได้จากการ *minimizing the mean square filter prediction error* ระหว่าง $\hat{s}_n$ และ s_n

2.5.9 Formant Frequencies

ค่าความถี่ฟอร์แมนท์ *(F)* คือ ยอดของสเปกตรัมเสียง *(Spectral peak)* คุณลักษณะนี้ถูกเสนอขึ้น โดย *Gunnar Fant* ในปี ค.ศ.1960 ค่าความถี่นี้ใช้แสดงถึงความถี่ที่เกิดขึ้นจากการสั่นพ้องเสียงในช่องทางเดินเสียงมนุษย์ *(Vocal tract)* ซึ่งผลที่ได้จะทำให้เห็นถึงความแตกต่างของความถี่ในเสียงพูด ในเสียงพูดจะมีค่าความถี่ฟอร์แมนท์ได้มากกว่าหนึ่งค่า อาจประกอบด้วยความถี่ฟอร์แมนท์ *F1, F2* และ *F3* แต่เพียงค่าความถี่สองค่าแรก *F1, F2* นั้นก็เพียงพอสำหรับการแสดงถึงเสียงสระในเสียงพูดมนุษย์ ค่าความถี่ฟอร์แมนท์สามารถคำนวณได้จาก

$$F = \frac{Fs}{2\pi} arctan \frac{im(s)}{re(s)} \qquad (17)$$

โดย $im(s),\ re(s)$ คือ ความถี่สุ่ม ตามลำดับ และ s คือสัญญาณเสียงที่รับเข้ามา

2.5.10 Perceptual Linear Predictive

Perceptual Linear Predictive หรือ การประมาณพันธะแบบอิงการรับฟังของมนุษย์ มีกระบวนการทำงานที่อยู่บนพื้นฐานของ 3 กฎจิตฟิสิกส์ในระบบการได้ยิน ได้แก่ 1. *the critical-band spectral resolution 2. the equal-loudness curve* และ *3. the intensity-loudness power law* *PLP* ถูกคิดค้นขึ้นโดย *Hermansky* *PLP* นั้นใช้หาแบบจำลอง *all-pole* ในสเปกตรัมของสัญญาณเสียงพูดช่วงเวลาสั้นๆ *(Short-Time Spectrum)* การจะหาค่าสัมประสิทธิ์ *PLP* สามารถหาได้จาก ค่าสัมประสิทธิ์ตัวกรองสัญญาณ *(filter bank coefficients)* โดยมีขั้นตอนดังนี้

1. ใช้วิธีเน้นล่วงหน้า *(Pre-emphasis)*โดยใช้การจำลองโค้งความดังเทียบเท่า *(simulated equal-loudness)* และ บีบอัดแอมพลิจูด
2. โค้งความดังเทียบเท่าจะสอดคล้องกับความถี่ f_k สำหรับตัวกรองสัญญาณ k^{th} คำนวณได้ดังนี้

$$L_k = \left(\frac{f_k^2}{f_k^2+1.6e5}\right)^2 \left(\frac{f_k^2+1.44e6}{f_k^2+9.61e6}\right) \qquad (18)$$

3. ใช้ โค้งความดังเทียบเท่า และ บีบอัดแอมพลิจูด :

$$\hat{m}_k = (L_k M_k)^{\beta}, (\beta = comp.factor)$$

ทำการ *inverse DFT* และ *Linear Prediction* ก็จะได้ ค่าสัมประสิทธิ์ *LP*

4. ใช้ *inverse DFT* เพื่อกรอง *pre-emphasised bank* ก็จะได้ผลลัพธ์ของค่าสัมประสิทธิ์ *auto-correlation*
5. ใช้ *Durbin algorithm* เพื่อหาค่าสัมประสิทธิ์ *LP*
6. แปลง ***LP coeff.*** (a_i) เป็น ***cepstral coef*** (c_n)***:***

$$c_n = -(a_n + \frac{1}{n}\sum_{i=1}^{n-1}(n-i)\,a_i c_{n-i} \qquad (19)$$

2.5.11 Harmonic Product Spectrum

Harmonic Product Spectrum (HPS) คือค่ามัชฌิมเรขาคณิตของแอมพลิจูดองค์ประกอบทางฮาร์โมนิคที่ความถี่ที่เป็นจำนวนเต็มเท่าของความถี่มูลฐาน ในแอมพลิจูดสเปกตรัมของสัญญาณเสียงนั้นจะประกอบไปด้วยจุดยอดหลายๆจุด และในจุดยอดที่แหลมก็มีหลายความถี่มูลฐานประกอบด้วย เมื่อทำการบีบอัดสเปกตรัมด้วยการหารด้วยจำนวนเต็มเท่า (*Down sampling)* แล้วเมื่อนำมาเทียบกับสเปกตรัมก่อนการบีบอัดจะพบว่าจุดยอดสามารถเห็นได้เด่นชัดมากขึ้น *HPS* $P(n)$ เป็นผลของ R *frequency-shrunken* จำลองแอมพลิจูดสเปกตรัม

$$\left|X\left(e^{j\frac{2\pi}{N}n}\right)\right|$$

$$P(n) = \sqrt[R]{\prod_{r=1}^{R}\left|X\left(e^{j\frac{2\pi}{N}nr}\right)\right|} \qquad (20)$$

โดย N คือจำนวนของจุด *FFT* และ $R = [N/2n]$ คือ ค่าที่น้อยลงสูงสุดของแอมพลิจูดสเปกตรัม ที่ยังคงมีค่าแอมพลิจูดบนความถี่ไม่ต่อเนื่อง n

2.5.12 Autocorrelation

Autocorrelation (AC) คุณลักษณะนี้มีคุณสมบัติที่ช่วยในการแบ่งแยกสัญญาณรบกวนได้เป็นอย่างดี *AC* นั้นใช้หาช่วงของเวลาที่ความยาว T ในเวลาโดเมน ซึ่งจะแสดงได้ถึงความคล้ายคลึงกันระหว่าง สัญญาณ $x(t)$ และส่วนที่คัดลอกที่เปลี่ยนโดย t

$$R(t) = \frac{1}{T-t}\sum_{\tau=0}^{T-t-1} x(\tau)x(\tau+t) \qquad (21)$$

จะได้ค่าสูงสุดของAC ที่ $t = 0$ $(|R(t)| \leq R(0)$ สำหรับทุกๆ $t)$

2.5.13 Spectral Flux

Spectral Flux (SF) หรือ การวิเคราะห์ความแตกต่างสเปกตรอล คือค่าที่สเปกตรอลเปลี่ยนแปลงในสองส่วนที่ต่อเนื่องกัน หรือก็คือส่วนต่างระหว่างค่าสเปกตรัมเชิงขนาดในช่วงปัจจุบัน กับ ค่าสเปกตรัมเชิงขนาดในช่วงก่อน รู้จักกันในอีกชื่อว่าระยะห่างยูคริเดียน *(Euclidean distance)* ระหว่างสองสเปกตร้าที่นอร์มัลไลซ์แล้ว *SF* หาได้จากสมการดังนี้

$$SF_i = \sum_{k=1}^{N/2} (|X_i(k)| - |X_i(k-1)|)^2 \qquad (22)$$

2.5.14 Harmonic Ratio

Harmonic Ratio (HR) คือ สัดส่วนขององค์ประกอบทางฮาร์โมนิคในค่าพลังงานสเปกตรัม ค่าสูงสุดของ *autocorrelation function* จะถูกคำนวณในทุกส่วนที่แบ่ง และจุดยอดของสัญญาณเสียงเสียงจะเป็น ค่าความล่าช้า M *(m* เป็นดัชนีล่าช้าของ *autocorrelation)* ค่าความล่าช้าที่สูงที่สุดนั้น จะสอดคล้องกับ ความถี่มูลฐานที่ต่ำที่สุด ซึ่งสามารถประมาณได้โดย

$$M = \frac{F_s}{f_0^{min}} \qquad (23)$$

และ อัตราส่วนฮาร์โมนิคคำนวณได้ตามสมการนี้

$$HR = \max_{M_0 \leq m \leq M}\{T_i(m)\} \qquad (24)$$

ค่า HR มีค่าเข้าใกล้ 1 หมายถึงสัญญาณฮาร์โมนิค หากเป็น 0 หมายถึง สัญญาณรบกวน

2.6 การคัดเลือกคุณลักษณะ

งานวิจัยชิ้นนี้ใช้การคัดเลือกคุณลักษณะแบบฟิชเชอร์ *(Fisher score or F-Score)* ซึ่งเป็นวิธีคัดเลือกคุณลักษณะที่มีความสัมพันธ์สำหรับการแบ่งกลุ่ม และทำการจัดอันดับคุณลักษณะตามความสัมพันธ์จากมากไปหาน้อยเพื่อเลือกคุณลักษณะที่ซ้ำซ้อนออก *F-Score* นั้นนำวิธีการจำแนกและสร้างแบบจำลองทางสถิติ โดยจะกำหนดให้คะแนนสูงสำหรับคุณลักษณะที่มีจุดข้อมูลห่างกับกลุ่มอื่นๆมากและใขณะเดียวกันจุดข้อมูลภายในกลุ่มนั้นๆก็ต้องใกล้เคียงกันด้วย *F-Score* สามารถคำนวณได้ตามสมการนี้

$$F_r = \frac{\sum_{i=1}^{c} n_i(\mu_r^i - \mu_r)^2}{\sum_{i=1}^{c} n_i(\sigma_r^i)^2} \quad ------ \qquad (25)$$

กำหนดให้

n_i - คือจำนวนของข้อมูลในกลุ่ม i

μ_r^i และ $(\sigma_r^i)^2$ - คือค่าเฉลี่ยและความแปรปรวนของกลุ่ม i ตามลำดับ $i = 1,\ldots,c$

2.7 ทบทวนวรรณกรรม/สารสนเทศ ที่เกี่ยวข้อง

Yun Jin (A feature selection and feature fusion combination method for speaker-independent speech emotion recognition) 2014 ได้นำเสนอวิธีที่สามารถเพิ่มอัตราการรู้จำจากเสียงพูดอิสระโดยใช้การคัดเลือกคุณลักษณะและวิธีผสมผสานคุณลักษณะที่อยู่บนพื้นฐานของ *multiple kernel learning (MKL)* เริ่มแรกใช้ *MKL* ทำการเลือกคุณลักษณะ ในลำดับที่สองคุณลักษณะที่คัดเลือกมาแล้วเหล่านี้ จะถูกนำมาผสมผสานกันในระดับเคอร์เนล ในขั้นตอนสุดท้ายเคอร์เนลทั้งหมดจะรวมเข้าด้วยกันผลลัพธ์ที่ได้จะเป็นเคอร์เนลผสาน จากผลการทดลองกับฐานข้อมูลเบอร์ลินซึ่งประกอบไปด้วย 7 อารมณ์ อัตราการรู้จำสูงสุดได้อยู่ที่ *83.10%*

Jun-Seok Park and Soo-Hong Kim (Emotion Recognition from Speech Signals using Fractal Features) 2014 ในงานวิจัยเลือกใช้คุณลักษณะแฟร็กทัลในระบบรู้จำเสียงพูด แฟร็กทัลถูกใช้แสดงถึงความไม่เป็นเชิงเส้นและคุณสมบัติความคล้ายตนเอง *(self-similarity)* ของสัญญาณเสียงพูด เทคนิคซัพพอร์ตเวกเตอร์แมชชีนถูกนำมาใช้เป็นตัวจำแนกและการรู้จำ ส่วนฐานข้อมูลเสียงที่นำมาใช้เป็ฐานข้อมูลมาตรฐานเบอร์ลิน ผลการที่ได้ อัตราการรู้จำอยู่ที่ราวๆ *77%*

Dipti D. Joshi and M.B. Zalte (Recognition of Emotion from Marathi Speech using MFCC and DWT algorithms) 2013 ได้เสนอระบบรู้จำอารมณ์จากเสียงพูดภาษามราฐีซึ่งเป็นหนึ่งในภาษาอินเดีย ภายในงานวิจัยคุณลักษณะที่สกัดจากเสียงพูดจะประกอบไปด้วยพลังงานเสียง, ระดับเสียง*(Pitch),* ความทุ้มของเสียง (*formant), MFCC* และ การแปลงเวฟเล็ตแบบเต็มหน่วย *(Discrete Wavelet Transform)* จะใช้สกัดคุณลักษณะเชิงเวกเตอร์ ตัวจำแนกใช้ *SVM* ในการจำแนกอารมณ์ เช่น โกรธ, เศร้า, มีความสุข และ สถานะอารมณ์ปกติ

Yan-You Chen (Emotion Aware System Based on Acoustic and Textual Features from Speech) 2010 ได้ศึกษาวิจัยระบบการรู้จำอารมณ์โดยรวมเอาคุณลักษณะที่ได้จากเสียงและเนื้อหาของบทพูดมาตรวจจับสถานะทางอารมณ์ทั้งสิ้น 7 ประการ ได้แก่ ดีใจ เสียใจ โกรธ กลัว ตกใจ กังวล และขยะแขยง ระบบการรู้จำอารมณ์แบ่งออกเป็นสองขั้นตอนได้แก่ขั้นตอนการสอน และการรู้จำ ทั้งสองขั้นตอนเริ่มต้นจากการสกัดคุณลักษณะเสียงและเนื้อหาของบทพูด วิธีการที่ใช้เหมือนกันเกือบทั้งหมดยกเว้นเนื้อหาที่ใช้ในกระบวนการสอนนั้นมาจากบทพูด แต่ในกระบวนการรู้จำถูกสร้างขึ้นจาก *ASR* หลังจากสกัดและรวมคุณลักษณะของเสียงและเนื้อหาแล้วจะนำ *Ada-Boost* อัลกอริทึมมาใช้เพื่อจำแนกคุณลักษณะทางอารมณ์ต่อไป ในการทดลอง ผู้วิจัยได้รวบรวมบทพูดที่เกี่ยวข้องกับอารมณ์ทั้งสิ้น 400 ประโยค ทุกประโยคในบทพูดถูกเชื่อมเข้ากับคลาสแบบแมนนวล จากผลการทดลองพบว่าค่าเฉลี่ยของอัตราการรู้จำของระบบมีค่าสูงกว่าวิธีการรู้จำอารมณ์จากคุณลักษณะอย่างใดอย่างหนึ่ง *5.15%* และ *3.73%* ตามลำดับ

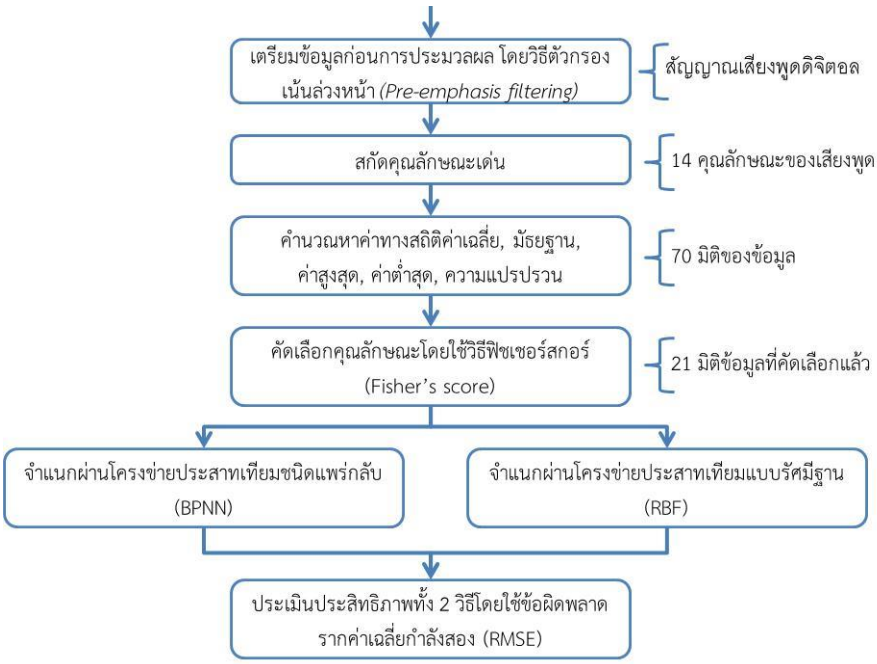
Chung-Hsien Wu (Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Lables) 2011 นำเสนอวิธีการรู้จำอารมณ์จากเสียงพูดที่มีอารมณ์ (*Affective Speech)* ด้วยวิธีการแบ่งหลายๆวิธี จากข้อมูลจังหวะของเสียง (*Acoustic-prosodic:AP)* และจากความหมายของคำที่ปรากฏ (*Semantic labels:SLs)* การรู้จำจาก *AP* จะทำการสกัดคุณลักษณะของเสียงได้แก่ *Spectrum,* ความทุ้มของเสียง (*formant)* และ ระดับเสียง *(pitch)* โดยเบื้องต้นจะใช้ตัวแบบสามประเภทได้แก่ *GMM, SVM* และ *MLP* จากนั้นจะใช้ *Meta Decision Tree (MDT)* เพื่อวัดระดับความเชื่อมั่น *(Confidence)* การรู้จำอารมณ์ที่ได้จากเสียงนั้นๆ การรู้จำจาก *SLs* จะนำข้อมูลที่มีอยู่แล้วในฐานความรู้ภาษาจีนที่เรียกว่า *HowNet* มาใช้ในการสกัด *Emotion Association Rules (EARs)* จากชุดของคำที่รู้จำได้จากเสียงพูดที่มีอารมณ์ โดยใช้ตัวแบบระดับเอนโทรปีที่มากที่สุด *(Maximum Entropy : MaxEnt)* ในการแสดงความสัมพันธ์ระหว่างสถานะของอารมณ์กับกฎ *EARs* ที่ใช้ในการรู้จำอารมณ์ หลังจากการรู้จำจากข้อมูล *AP* และ *SL* แล้ว ขั้นตอนสุดท้ายได้ใช้วิธี *weighted product fusion* ในการรวมผลการรู้จำจากข้อมูลทั้งสองประเภทข้างต้นเข้าด้วยกันเพื่อใช้ในการตัดสินใจขั้นสุดท้ายว่าเป็นเสียงพูดนั้นมีอารมณ์อยู่ในสถานะใด การประเมินผลทำจากข้อมูลเสียงพูด 2033 เสียงโดยไม่เจาะจงผู้พูด ข้อมูลที่ใช้มีสถานะอารมณ์ 4 สถานะได้แก่ อารมณ์ปกติ, มีความสุข, โกรธ, และเสียใจ พบว่าประสิทธิภาพการรู้จำอารมณ์โดยใช้ *MDT* สูงถึง 80% ซึ่งดีกว่าการแบ่งกลุ่มวิธีอื่นๆ และการรู้จำจาก *SL* ก็มีความแม่นยำโดยเฉลี่ยที่ 80.92% เมื่อรวมวิธีการรู้จำจากข้อมูลทั้งสองประเภทเข้าด้วยกันจึงมีประสิทธิภาพถึง 83.55% ซึ่งดีกว่าการรู้จำจากข้อมูลประเภทใดประเภทหนึ่งเพียงอย่างเดียว และหากมีการพิจารณาถึงคุณลักษณะส่วนบุคคลของผู้พูดด้วยจะยิ่งเพิ่มประสิทธิภาพการรู้จำมากขึ้นเป็น 85.79%

Santiago Planet(Spontaneous Children's Emotion Recognition by Categorical Classification of Acoustic Features) 2011 เสนอวิธีการแบ่งกลุ่มเพื่อรู้จำอารมณ์ตามธรรมชาติของเด็กจากคุณลักษณะของเสียงพูด จากข้อมูลที่ได้จาก *FAU Aibo Corpus* แบ่งเป็นข้อมูลที่ใช้ในการสอน 9,959 ชุด และข้อมูลที่ใช้ในการทดสอบ 8,257 ชุดโดยอารมณ์ที่ต้องการจำแนกในบทความนี้มี 5 อารมณ์ด้วยกันได้แก่ โกรธ *(Angry), Emphatics,* ปกติ *(Neutral),* อารมณ์ดี *(Positive),* อารมณ์อื่นๆนอกเหนือจากข้างต้น (*Rest)* วิธีการแบ่งงกลุ่มที่ใช้ได้แก่วิธีที่ 1 ใช้ *Naïve-Bayes,* วิธีที่ 2 ใช้ *Support Vector Machine* กับข้อมูลต้นฉบับที่ผ่านการนอร์มัลไลซ์แล้ว วิธีที่ 3 ใช้ *Support*

Vector Machine กับข้อมูลต้นฉบับที่ผ่านการนอร์มัลไลซ์แล้วนำมาทำการสุ่มตัวอย่างซ้ำเพื่อกระจายกลุ่มข้อมูลให้อยู่ในรูปการกระจายตัวแบบยูนิฟอร์มก่อน และวิธีที่ 4 เป็นการรวมวิธีที่ 1 และ 3 เข้าด้วยกัน ในการวัดประสิทธิภาพวิธีการแบ่งกลุ่มแต่ละวิธีจะวัดจากค่า *Unweighted Average Recall (UAR)* ซึ่งจากผลการทดลองพบว่าวิธีที่ 3 ให้ค่าผลลัพธ์ที่มีประสิทธิภาพดีกว่าวิธีที่ 2 อยู่ 11.38%

บทที่ 3 วิธีดำเนินการวิจัย

ในงานวิจัยนี้ ผู้วิจัยได้ศึกษาการรู้จำอารมณ์จากเสียงพูดภาษาไทยโดยประกอบไปด้วย อารมณ์ เศร้า,โกรธ,มีความสุข และ กลัว ซึ่งผู้วิจัยได้เสนอวิธีการรู้จำโดยมีขั้นตอนการทำงานตามแผนภาพแสดงขั้นตอนใน รูปที่ 3-1



รูปที่ 3-1 แผนภาพแสดงขั้นตอนการรู้จำอารมณ์เสียงพูดภาษาไทย

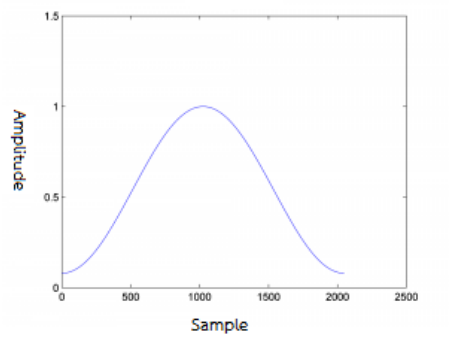
เสียงที่นำมาใช้ในงานวิจัยประกอบไปด้วยไฟล์เสียงพูดภาษาไทยจำนวน 800ไฟล์ ส่วนรายละเอียดได้กล่าวไปแล้วในตอนต้น กระบวนการรู้จำอารมณ์จากเสียงพูดที่เสนอนี้ จะแบ่งออกได้เป็น 3 ขั้นตอนหลักด้วยกันคือ

- ขั้นตอนเตรียมข้อมูลก่อนการประมวลผล
- ขั้นตอนสกัดคุณลักษณะและคัดเลือกคุณลักษณะ
- ขั้นตอนจำแนกอารมณ์

3.1 ขั้นตอนเตรียมข้อมูลก่อนการประมวลผล

สัญญาณเสียงพูดภาษาไทยที่นำมาใช้แม้จะมีการบันทึกเสียงในสถานที่ปิดแล้วแต่ก็ยังคงมีสัญญาณรบกวนซึ่งอาจเกิดได้จากหลายปัจจัยไม่ว่าจะเป็นเสียงลมที่เกิดจากการเคลื่อนไหว ทั้งนี้จึงต้องผ่านกระบวนการตัวกรองเน้นล่วงหน้า เพื่อลดสัญญาณรบกวนที่ยังหลงเหลืออยู่ โดยการปรับอัตราส่วนระหว่างสัญญาณเสียงและสัญญาณรบกวนให้สูงขึ้น ผ่านการนำสัญญาณเสียงพูดเข้าสู่วงจรกรองสัญญาณลำดับที่หนึ่ง สัญญาณที่ผ่านตัวกรองนี้แล้ว ส่วนสัญญาณรบกวนที่มีความถี่ต่ำจะถูกขจัดออก ในงานวิจัยนี้ ได้กำหนดให้ค่าสัมประสิทธิ์ของวงจรตัวกรอง มีค่าเท่ากับ 0.9375 เนื่องจากเป็นค่าที่เหมาะสมให้อัตราการรู้จำที่สูง

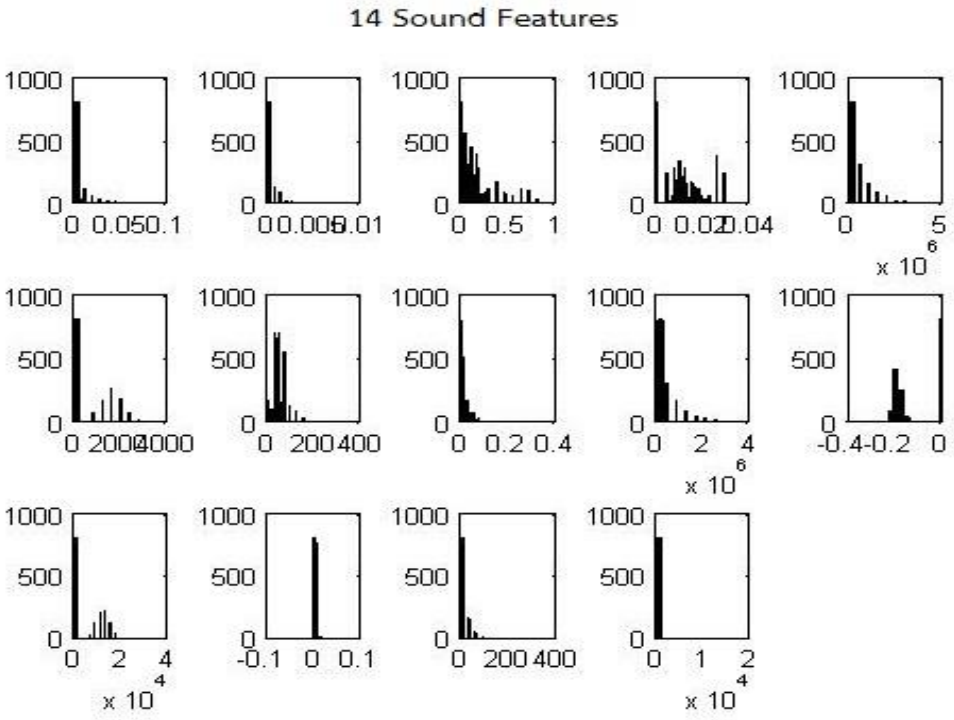
เนื่องจากสัญญาณเสียงพูดจะมีลักษณะไม่คงที่ ไม่แน่นอนและมีการเปลี่ยนแปลงทีละน้อยอย่างช้าๆผันแปรไปตามเวลา จึงจำเป็นต้องมีการวิเคราะห์ทีละช่วงสั้นๆ ประกอบกันเพื่อให้เกิดความแม่นยำ สัญญาณเสียงที่ผ่านตัวกรองเน้นล่วงหน้าแล้วจะถูกแบ่งเป็นกรอบสัญญาณด้วยวิธีกรอบสัญญาณแฮมมิง ในงานวิจัยนี้กำหนดขนาดกรอบสัญญาณอยู่ที่ 30 มิลลิวินาที และให้แต่ละกรอบมีส่วนเหลื่อมล้ำกัน 15 มิลลิวินาที วิธีนี้จะลดทอนแอมปลิจูดลงอย่างช้าๆ ที่บริเวณปลายแต่ละข้างของกรอบข้อมูลเสียงพูด เพื่อป้องกันการเปลี่ยนแปลงที่ไม่ต่อเนื่องอย่างกระทันหันที่ส่วนปลายกรอบสัญญาณเพื่อให้สัญญาณมีความต่อเนื่องกัน และ ยังช่วยหลีกเลี่ยงการรั่วของสเปกตรัม สัญญาณเสียงที่ผ่านการวางกรอบสัญญาณจะยังคงข้อมูลไว้ครบถ้วน



รูปที่ 3-2 ฟังก์ชั่นกรอบสัญญาณแฮมมิง

3.2 ขั้นตอนสกัดคุณลักษณะและคัดเลือกคุณลักษณะ

ในเบื้องต้นผู้วิจัยได้ตั้งสมมุติฐานว่าหากความรู้สึก หรือ อารมณ์ ของผู้พูดเปลี่ยน ลักษณะของคลื่นเสียงที่เปล่งออกมาย่อมจะเปลี่ยนแปลงตามไปด้วย ในงานวิจัยนี้จึงทำการดึงเอาคุณลักษณะของเสียงพูดออกมาเพื่อใช้เป็นตัวบ่งชี้ถึงอารมณ์ โดยหลัง จากสัญญาณเสียงผ่านการเตรียมข้อมูลแล้ว ก็จะเข้าสู่ขั้นตอนการสกัดเอาคุณลักษณะเด่นของเสียงออกมา ซึ่งเสียงพูดจะถูกดึงเอา 14 คุณลักษณะซึ่งมาจากการศึกษาคุณลักษณะ ที่มีประสิทธิภาพและนิยมนำมาใช้ในงานประมวลผลสัญญาณเสียง *"A Large Set of Audio Features for Sound Description (similarity and classification) in the CUIDADO project" Geotfroy Peeters* เมื่อได้คุณลักษณะทั้ง 14 แล้ว จึงนำแต่ละคุณลักษณะมาคำนวณหาค่าทางสถิติ เนื่องจาก คุณลักษณะที่สกัดออกมาจะอยู่ในรูปของเมทริกซ์ขนาดต่างๆกัน การนำข้อมูลทั้งหมดมาวิเคราะห์จึงทำได้ยาก ค่าทางสถิติจึงช่วยกำจัดความซ้ำซ้อนของข้อมูลและเป็นตัวแทนของข้อมูลที่มีประสิทธิภาพ ช่วยให้การวิเคราะห์ทำได้รวดเร็วขึ้นอย่างมาก ค่าทางสถิติที่นำมาใช้ได้แก่ ค่าเฉลี่ย,มัธยฐาน, ค่าสูงสุด,ค่าต่ำสุด และ ความแปรปรวน แต่ละคุณลักษณะก็จะถูกเพิ่มมิติของข้อมูลขึ้น สรุปได้ว่าหนึ่งเสียงพูดจะประกอบไปด้วย 70 มิติข้อมูล ข้อมูลทั้งหมดเหล่านี้ จะถูกนำมาคัดเลือกเฉพาะข้อมูลที่มีความสัมพันธ์เพื่อลดความซ้ำซ้อนของข้อมูลลง ตามดังรูปที่ 3-3 จะเห็นได้ว่าในคุณลักษณะทั้งหมด มีบางคุณลักษณะที่มีความคล้ายคลึงกันของข้อมูลกับคุณลักษณะอื่นๆ วิธีที่นำมาใช้คัดเลือกคุณลักษณะในงานวิจัยนี้ คือ การคัดเลือกคุณลักษณะแบบ *Fisher's Score* เพื่อ เลือกเฉพาะคุณลักษณะที่เหมาะสมที่สุดในกระบวนการเรียนรู้ อย่างที่ได้กล่าวไปแล้วในบทที่ 2 *Fisher's Score* เป็นวิธีทางสถิติที่สามารถประเมิน และ แบ่งแยกความแตกต่างของแต่ละคุณลักษณะ โดยหลีกเลี่ยงการทับซ้อนกันของข้อมูลในคุณลักษณะต่างๆ เท่าที่จะเป็นไปได้ จากการทดลองนี้ *Fisher's Score* ได้เลือก 22 มิติข้อมูลจาก 7 คุณลักษณะให้เป็นคุณลักษณะที่เหมาะสมที่สุดต่อการจำแนก

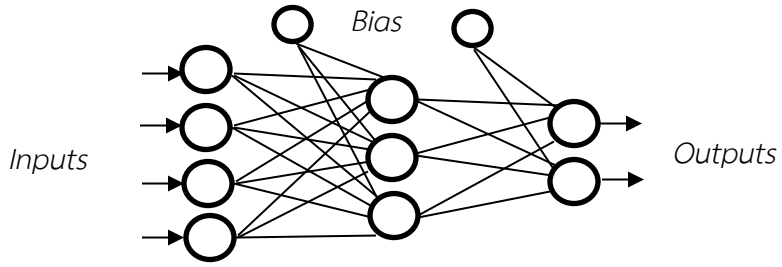


รูปที่ 3-3 ลักษณะข้อมูลของทั้ง 14 คุณลักษณะ

3.3 ขั้นตอนจำแนกอารมณ์

ในขั้นตอนการจำแนกอารมณ์ มิติข้อมูลของคุณลักษณะทั้งหมดและที่ได้จากการคัดเลือก จะเข้าสู่กระบวนการเรียนรู้ เพื่อจำแนกอารมณ์ ซึ่งผู้วิจัยได้เลือกใช้ 2 วิธีจำแนก อันได้แก่โครงข่ายประสารทเทียมแบบ *BPNN* กับ *RBF* และมีการแบ่งกลุ่มข้อมูลเพื่อทดสอบ เป็นแบบ *10 folds cross validation* แบ่งเป็นข้อมูลชุดทดสอบและชุดเทรนนิ่ง โดยใช้โปรแกรมจักรกลเรียนรู้ *WEKA* ภายในโครงข่ายแบบ *BPNN* จะประกอบไปด้วย 3 ชั้น ได้แก่ ชั้นอินท์พุทและชั้นเอาท์พุทอย่างละ 1 ชั้น ชั้นซ่อนซึ่งอยู่ระหว่างชั้นอินท์พุทกับชั้นเอาท์พุทสามารถมีได้หลายชั้น และยังมีหน่วยที่เชื่อมโยงกับทุกๆโหนดในแต่ละชั้น เรียกว่า หน่วยไบอัสจะมีค่าเป็น 1 เสมอ เริ่มแรกมิติข้อมูลที่ได้จากสัญญาณเสียงพูดจะเข้าสู่ชั้นอินพุท ในชั้นนี้จะไม่มีการประมวลผลใดๆมีเพียงการส่งสัญญาณต่อไปยังทุกๆโหนดในชั้นซ่อนชั้นแรก เมื่อชั้นซ่อนประมวลผลเสร็จก็จะมีการส่งสัญญาณแบบนี้ไปยังชั้น

ซ่อนอื่นเรื่อยๆ จนกระทั่งชั้นซ่อนชั้นสุดท้ายส่งสัญญาณไปยังทุกโหนดของชั้นเอาท์พุท ผลลัพธ์ที่ได้จากชั้นนี้ จะเป็นค่าเอาท์พุทของโครงข่าย



รูปที่ 3-4 สถาปัตยกรรมโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

ส่วนพารามิเตอร์ที่เหมาะสมสำหรับโครงข่ายประสาทเทียมแบบ *BPNN* ที่ใช้ในงานวิจัยได้กำหนดไว้ดังนี้

- *BPNN*

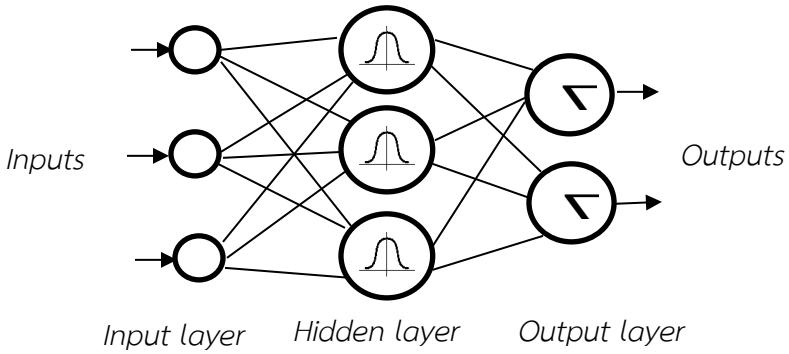
  Learning rate 0.1

  Momentum 0.1

  Number of epoch 500

  Number of Hidden layers 12

ในส่วนของโครงข่ายประสาทเทียมแบบ *RBF* จะมีชั้นคล้ายคลึงกับแบบ *BPNN* ประกอบด้วย ชั้นอินท์พุท,เอาท์พุท และ ชั้นซ่อน โครงข่ายมีลักษณะการทำงานแบบไปข้างหน้า *(feed forward networks)* ทุกโหนดภายในชั้นซ่อนจะมีเรเดียลเบสิส-ฟังก์ชั่น



รูปที่ 3-5 สถาปัตยกรรมโครงข่ายประสาทเทียมแบบเรเดียลเบสิส

ส่วนพารามิเตอร์ที่เหมาะสมสำหรับโครงข่ายประสาทเทียมแบบ *RBF* ที่ใช้ในงานวิจัยได้กำหนดไว้ดังนี้

- *RBF*

  Gaussian function
  clusteringDees 1
  maxIts -1
  minStdDev 0.1
  numCluster 2
  ridge 1.00E-08

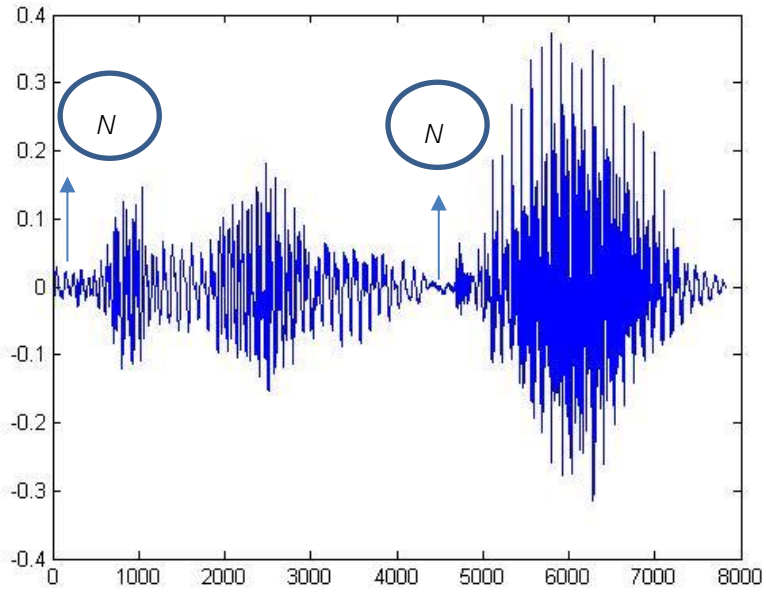
ผลลัพธ์ที่ได้จากทั้ง 2 วิธีจำแนกจะเป็นร้อยล่ะความถูกต้องในการจำแนก ซึ่งในส่วนสุดท้ายเป็นการประเมินประสิทธิภาพ การรู้จำเสียงพูดภาษาไทย ของทั้ง 2 วิธี โดยใช้วิธีหาข้อผิดพลาดรากค่าเฉลี่ยกำลังสอง *(RMSE)*

บทที่ 4 ผลการทดลอง

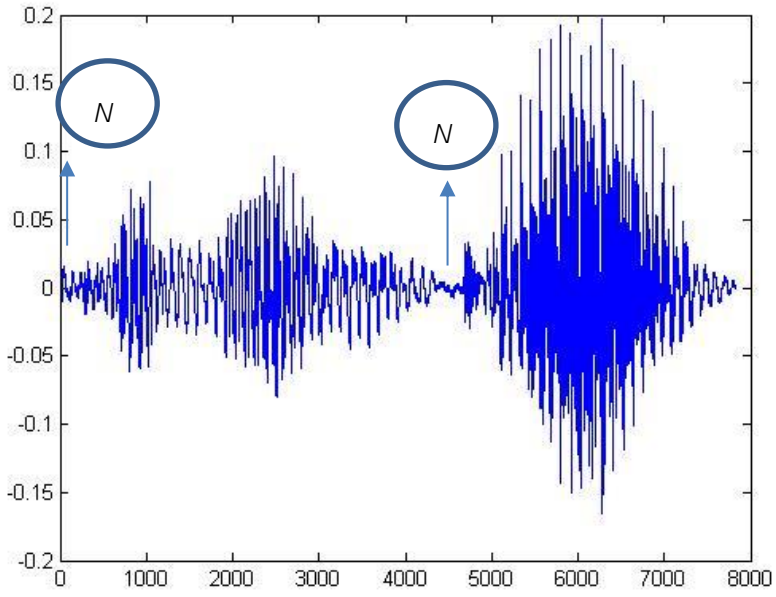
ในบทนี้ แสดงผลการทดลองที่ได้จากการใช้วิธีที่นำเสนอในบทที่ 3 โดยจะแสดงตัวอย่างผลที่ได้จากทั้ง 3 ขั้นตอน ก่อนจะนำมาเปรียบเทียบเพื่อหาวิธีที่เหมาะสมที่สุดในการรู้จำเสียงพูดภาษาไทย ผลการทดลองที่ได้มีรายละเอียดดังต่อไปนี้

4.1 ขั้นตอนเตรียมข้อมูลก่อนการประมวลผล

สัญญาณเสียงพูดทั้งหมด 800 แฟ้มเสียง ถูกนำมาผ่านตัวกรองเน้นล่วงหน้าเป็นขั้นตอนแรกเพื่อลดสัญญาณรบกวน จากรูปที่ 4-1 (ก) สัญญาณเสียงพูดก่อนเข้าสู่ตัวกรองเน้นล่วงหน้า อักษร *N* ในรูปหมายถึงบริเวณสัญญาณรบกวนความถี่ต่ำ ซึ่งหลังผ่านตัวกรองเน้นล่วงหน้าแล้ว สัญญาณจะถูกปรับอัตราส่วนระหว่างสัญญาณเสียงและสัญญาณรบกวนส่งผลให้ สัญญาณรบกวนบริเวณ *N* ถูกปรับให้เท่ากันตลอดย่านความถี่เสียง ซึ่งแสดงยังภาพ (ข) สัญญาณหลังผ่านตัวกรองเน้นล่วงหน้า



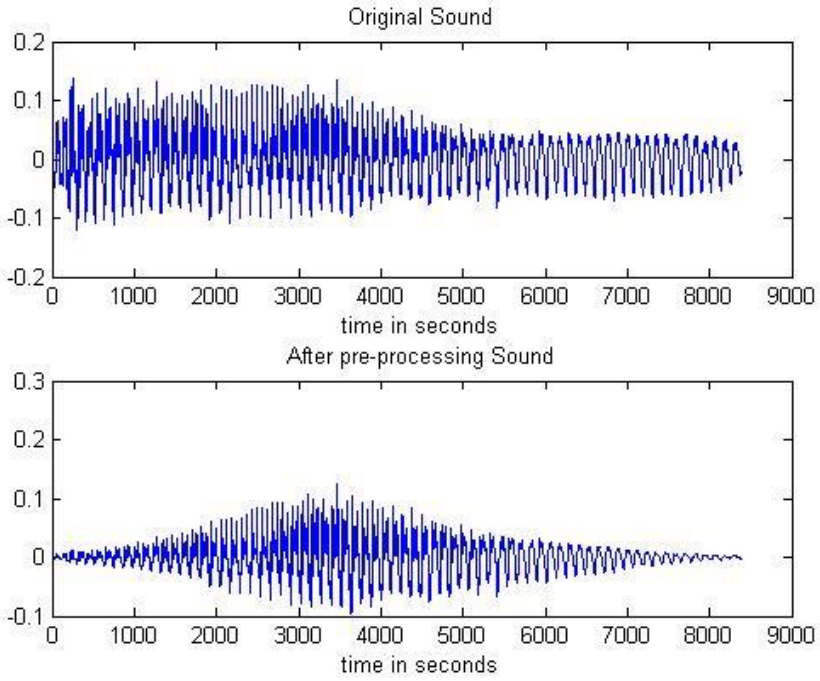
(ก)



(ข)

รูปที่ 4-1 สัญญาณเสียงพูดภาษาไทย (ก)ก่อน และ (ข)หลัง ตัวกรองเน้นล่วงหน้า

หลังจากสัญญาณเสียงพูดผ่านตัวกรองเน้นล่วงหน้าแล้ว ก็จะถูกนำมาแบ่งออกเป็นกรอบสัญญาณช่วงสั้นๆด้วยวิธีกรอบสัญญาณแฮมมิงเพื่อให้การวิเคราะห์สัญญาณเสียงเป็นไปอย่างถูกต้อง กรอบสัญญาณที่แบ่งจะเหลื่อมทับกัน ผลที่ได้หลังจากสัญญาณเสียงผ่านตัวกรองเน้นล่วงหน้า และ การแบ่งกรอบสัญญาณแบบแฮมมิงแล้วจะปรากฏยังรูปที่ 4-2



รูปที่ 4-2 แสดงสัญญาณเสียงพูดภาษาไทย ก่อน และ หลัง กระบวนการเตรียมข้อมูลก่อนการประมวลผล

4.2 ขั้นตอนสกัดคุณลักษณะและคัดเลือกคุณลักษณะ

หลังจากสัญญาณเสียงผ่านการเตรียมข้อมูลก่อนการประมวลผลแล้ว สัญญาณเสียงทั้งหมดจะเข้าสู่กระบวนการสกัดเอาคุณลักษณะเด่นทั้ง 14 ออกมา ข้อมูลคุณลักษณะที่ได้จาก 1 เสียงพูด จะอยู่ในรูปเมทริกซ์ขนาดต่างกันออกไป ดังในตารางที่ 4-1 โดยจะแสดงตัวอย่างการคำนวณเฉพาะค่า *Energy Entropy Block, Zero Crossing Rate, Mel Frequency Cepstral Coefficient, Spectral Flux* และ *Spectral-Roll-Off* เท่านั้น

ตารางที่ 4-1. แสดงเมทริกซ์ผลลัพธ์ที่ได้จาก การสกัดคุณลักษณะ เสียงพูดภาษาไทย 1 เสียง

Energy Entropy Block

0.0045	0.0053	0.0083	0.015	0.014	0.01	0.036	....	0.0047

Zero Crossing Rate

0.52	0.75	0.85	0.84	0.85	0.68	0.37	0.11	0.096	...	0.021

Mel Frequency Cepstral Coefficient

32.99	*32.43*	*34.47*	*35.67*	*38.14*	*38.29*	*45.24*	*49.2*	...	*32.02*
-5.62	*-7.61*	*-10.82*	*-12.57*	*-9.04*	*-8.05*	*0.084*	*5.37*	...	-2.01
-1.26	*-0.75*	*1.66*	*-0.25*	*-1.0*	*-1.47*	*1.25*	*-3.69*	...	*0.28*
0.13	*-2.28*	*-3.47*	*-1.6*	*-4.03*	*-3.8*	*-4.48*	*-1.9*	...	*-3.43*
1.9	*1.65*	*-1.2*	*-0.037*	*5.53*	*0.078*	*-3.4*	*-7.53*	...	*4.69*
2.79	*-0.54*	*-4.97*	*-4.54*	*-0.95*	*2.1*	*-2.17*	*2.41*	...	*-6.75*

Spectral Flux

5.25	5.46	4.53	3.47	2.71	10.63	1.83	2.59	1.26	0.93	...	*0.20*

Spectral-Roll-Off

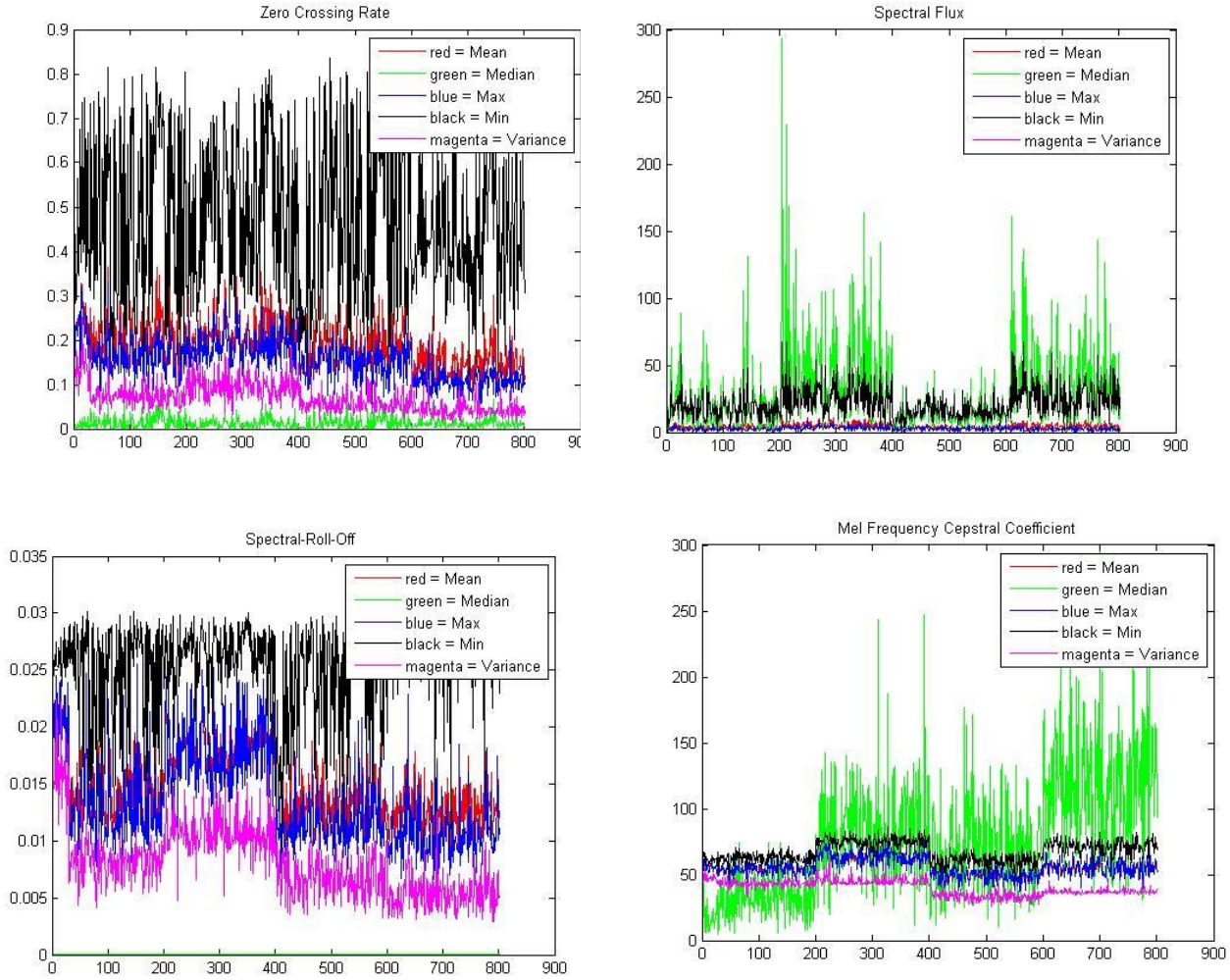
0.031	*0.0309*	*0.0309*	*0.0309*	*0.0309*	*0.0309*	*0.024*	*0.014*	...	*0.008*

จากเมทริกซ์ข้อมูลของแต่ละคุณลักษณะเหล่านี้จะถูกนำมาคำนวณหาค่าทางสถิติ *5* ค่าอันได้แก่ ค่าเฉลี่ย *(Mean),* มัธยฐาน *(Median),* ค่าสูงสุด *(Max),* ค่าต่ำสุด *(Min),* ความแปรปรวน *(Variance)* เพื่อใช้เป็นตัวแทนข้อมูล ดังในตารางที่ 4-2

ตารางที่ 4-*2* ข้อมูล 70 มิติที่ได้จากการคำนวณค่าทางสถิติในแต่ละคุณลักษณะของเสียงพูดภาษาไทย 1 เสียง

Features	*Mean*	*Median*	*Max*	*Min*	*Variance*
1. *Energy Entropy Block*	*0.0003*	*4.95E-08*	*0.000244*	*0.000653*	*5.30E-05*
2. *Short Time Energy*	*2.82E-05*	*4.72E-10*	*2.17E-05*	*6.30E-05*	*4.18E-06*
3. *Zero Crossing Rate*	*0.17642*	*0.001238*	*0.176042*	*0.244792*	*0.105208*
4. *Spectral-Roll-Off*	*0.017023*	*4.70E-06*	*0.016625*	*0.021*	*0.013125*
5. *Spectral Centroid*	*1158.946*	*1309804*	*2060*	*2422.66*	*0*
6. *Fundamental Frequency*	*119.1919*	*1738.215*	*133.3333*	*200*	*33.33333*
7. *MFCC*	*59.21113*	*38.585*	*60.95077*	*66.66609*	*49.51607*
8. *Linear Predictive Coding*	*0.015611*	*0*	*0.015611*	*0.015611*	*0.015611*
9. *Formant Frequencies*	*4350.735*	*405925.4*	*4193.978*	*6722.965*	*3889.503*
10. *Perceptual Linear Predictive*	*-0.21572*	*0*	*-0.21572*	*-0.21572*	*-0.21572*
11.*Harmonic Product Spectrum*	*285.9493*	*8302.731*	*300.4219*	*362.0469*	*15.40625*
12. *Autocorrelation*	*0.001164*	*0*	*0.001164*	*0.001164*	*0.001164*
13. *Spectral Flux*	*2.203583*	*4.105226*	*2.084765*	*9.094121*	*0*
14. *Harmonic Ratio*	*0.999937*	*2.36E-05*	*1.000119*	*1.074211*	*0.990924*

หลังจากการคำนวณหาค่าทางสถิติแล้ว หนึ่งไฟล์เสียงพูดจะประกอบด้วย 70 มิติข้อมูลที่ได้มาจากค่าทางสถิติทั้ง 5 ของทั้ง 14 คุณลักษณะ ซึ่งในบางคุณลักษณะผลลัพธ์ที่ได้ อาจไม่ใช่เมทริกซ์แต่เป็นเพียงค่า ค่าเดียว เมื่อคำนวณมัธยฐาน จึงมีค่าเป็น 0 เมื่อได้มิติข้อมูลจากไฟล์เสียงพูดภาษาไทยทั้ง 800 ไฟล์แล้ว กระบวนการต่อไปจะทำการคัดเลือกบางคุณลักษณะและบางค่าทางสถิติที่มีความสัมพันธ์



รูปที่ 4-3 มิติข้อมูลทั้ง 5 *(Mean,Median,Max,Min,Variance)* ของคุณลักษณะที่สกัดจาก 800 เสียงพูด (ก) *Zero Crossing Rate* (ข) *Spectral Flux* (ค) *Spectral-Roll-Off* (ง) *Mel Frequency Cepstral Coefficient*

มิติข้อมูลทั้งหมด จะถูกคัดเลือกเฉพาะข้อมูลที่มีความสัมพันธ์ โดยการคัดเลือกแบบ *Fisher's Score* ซึ่งผลจากการคัดเลือกที่ได้ในการทดลองครั้งนี้ประกอบไปด้วย 22 มิติข้อมูล รายละเอียดแสดงยังตารางที่ 4-3

ตารางที่ 4-3 แสดงมิติข้อมูลทั้ง 22 ที่ถูกคัดเลือกโดยวิธี *Fisher's Score*

คุณลักษณะที่ถูกคัดเลือก	มิติข้อมูลทางสถิติที่ถูกคัดเลือก
1. *Energy Entropy Block*	*Mean,Max,Min*
2. *Short-time energy*	*Mean,Min*
3. *Zero crossing rate*	*Mean,Max,Variance*
4. *Spectral-roll-off*	*Mean,Median,Max,Variance*
5. *Spectral Centroid*	*Mean,Max*
6. *Formants frequency*	*Mean,Median,Max,Min,Variance*
7. *MFCC*	*Median,Min,Variance*

4.3 ขั้นตอนจำแนกอารมณ์ และ เปรียบเทียบผลการทดลอง

ขั้นตอนจำแนกอารมณ์ มิติข้อมูลที่ได้จากการสกัดสัญญาณเสียงจะถูกกำกับอารมณ์ของเสียงนั้นๆไว้ในทุกเสียงพูดทั้งหมด ก่อนเข้าสู่กระบวนการเรียนรู้ เพื่อคำนวณหาอัตราการรู้จำ ผลการทดลองการรู้จำอารมณ์จากเสียงพูดจะแสดงในรูปของตาราง ตั้งแต่ตารางที่ *4*-4 ถึง *4*-8 ซึ่งแบ่งออกได้เป็น 3 ส่วน ในส่วนแรก (ตารางที่ 4-4 และ 4-5) แสดง ผลการทดลองกับข้อมูลที่ได้จากการสกัดคุณลักษณะทั้งหมด 14 คุณลักษณะซึ่งในแต่ละเสียงพูดจะประกอบไปด้วย 70 มิติข้อมูล กับส่วนที่สอง (ตารางที่ 4-6 และ 4-7) เป็นผลการทดลองจากข้อมูลที่ได้รับการคัดเลือกนำมาเฉพาะคุณลักษณะที่มีความสัมพันธ์โดยใช้วิธีการคัดเลือกแบบ *Fisher's Score* ประกอบไปด้วย 22 มิติข้อมูลจาก 7 คุณลักษณะ อันได้แก่ *(Energy Entropy Block, Short-time energy, Zero crossing rate, Spectral-roll-off, Spectral Centroid, Formants frequency, MFCC)* ผลการทดลองทั้งสองส่วนถูกวัดประสิทธิภาพในการจำแนกอารมณ์โดยใช้โครงข่ายประสาทเทียมแบบ *BPNN* และ *RBF* ในส่วนสุดท้าย (ตารางที่ 4-8) แสดงประสิทธิภาพเปรียบเทียบระหว่าง ทั้ง 2 วิธีจำแนก *BPNN* และ *RBF* โดยประเมินจากข้อผิดพลาดรากค่าเฉลี่ยกำลังสอง *(RMSE)*

ตารางที่ 4-4 แสดงผลการทดลองการรู้จำเสียงพูดภาษาไทย 70 มิติข้อมูลจาก 14 คุณลักษณะ โดยใช้วิธีจำแนก ***BPNN***

อารมณ์(%)	เศร้า	โกรธ	กลัว	มีความสุข
เศร้า	*92.5*	*2.5*	*4.5*	*0.5*
โกรธ	*3*	*96*	0	1
กลัว	6	0	93.5	0.5
มีความสุข	*0.5*	2.5	1	96

ตารางที่ 4-5 แสดงผลการทดลองการรู้จำเสียงพูดภาษาไทย 70 มิติข้อมูลจาก 14 คุณลักษณะ โดยใช้วิธีจำแนก ***RBF***

อารมณ์ (%)	เศร้า	โกรธ	กลัว	มีความสุข
เศร้า	*84.5*	*1*	*14.5*	*0*
โกรธ	*8*	*89.5*	*0.5*	*2*
กลัว	*9*	*1*	*88.5*	*1.5*
มีความสุข	*1.5*	*5*	*5.5*	*88*

ตารางที่ 4-6 แสดงผลการทดลองการรู้จำเสียงพูดภาษาไทย 22 มิติข้อมูลจาก 7 คุณลักษณะที่คัดเลือกจาก *Fisher's Score* โดยใช้วิธีจำแนก ***BPNN***

อารมณ์(%)	เศร้า	โกรธ	กลัว	มีความสุข
เศร้า	*92.5*	*1.5*	*6*	*0*
โกรธ	*3.5*	*96*	*0.5*	*0*
กลัว	*5*	*0*	*95*	*0*
มีความสุข	*0*	*2.5*	*0.5*	*97*

ตารางที่ 4-7 แสดงผลการทดลองการรู้จำเสียงพูดภาษาไทย 22 มิติข้อมูลจาก 7 คุณลักษณะที่คัดเลือกจาก *Fisher's Score* โดยใช้วิธีจำแนก ***RBF***

อารมณ์(%)	เศร้า	โกรธ	กลัว	มีความสุข
เศร้า	*91*	*0.5*	*8.5*	*0*
โกรธ	*5.5*	*93.5*	*0.5*	*0.5*
กลัว	*7*	*0*	*91.5*	*1.5*
มีความสุข	*0.5*	*3*	*3*	*93.5*

จากตารางที่ *4*-1 ถึงตารางที่ 4-7 ผลการทดลองระหว่างข้อมูลทั้ง 70 มิติข้อมูล กับ 22 มิติข้อมูลที่ถูกคัดเลือก เด่นชัดว่า ข้อมูลที่ผ่านการคัดเลือกคุณลักษณะ (ตารางที่ 4-6, 4-7) มีประสิทธิภาพในการจำแนกอารมณ์สูงกว่า ไม่ว่าจะเป็นวิธีจำแนกแบบ *BPNN* หรือ *RBF*สรุปได้ว่า ในงานวิจัยนี้ การรู้จำอารมณ์จากเสียงพูดภาษาไทยตามวิธีที่เสนอโดยใช้คุณลักษณะทั้ง 14 วิธีจำแนกใช้แบบ *RBF* (จากตารางที่ 4-5) ให้อัตราการจำแนกถูกต้องต่ำที่สุด และการคัดเลือกคุณลักษณะด้วยวิธี *Fisher's Score* 7 คุณลักษณะจำแนกด้วย *BPNN* (จากตารางที่ 4-6) มีประสิทธิภาพและเหมาะสมกับการรู้จำอารมณ์จากเสียงพูดภาษาไทยมากที่สุด

ตารางที่ *4-8* ประเมินประสิทธิภาพวิธีจำแนก ด้วย ข้อผิดพลาดรากค่าเฉลี่ยกำลังสอง

อัตราความแม่นยำ (%)	14 คุณลักษณะ (70 มิติข้อมูล)	7 คุณลักษณะ (22 มิติข้อมูล)
BPNN	*0.1525*	*0.1413*
RBF	*0.2317*	*0.1854*

จากในตารางที่ 4-8 วิธีจำแนกแบบ *BPNN* มีข้อผิดพลาดน้อยกว่า *RBF* ดังนั้นในงานวิจัยนี้ วิธีจำแนกอารมณ์จากเสียงพูดภาษาไทยที่เหมาะสมที่สุดคือ วิธี *BPNN*

บทที่ 5 สรุปผลการทดลอง

5.1 สรุปผลการทดลอง

จากการศึกษาและค้นคว้าการประมวลผลเสียง การรู้จำอารมณ์ งานวิจัยนี้ได้เสนออัลกอริทึมที่สามารถรู้จำอารมณ์จากเสียงพูดภาษาไทย โดยใช้วิธีการคัดเลือกคุณลักษณะแบบฟิชเชอร์สกอ *(Fisher score or F- score)* ซึ่งช่วยลดคุณลักษณะของเสียงที่ซ้ำซ้อนกันลง จากการทดลองพบว่าสามารถลดคุณลักษณะลงได้ถึง 70% จากมิติของข้อมูลทั้งหมดที่นำมาใช้ จึงส่งผลให้ปริมาณของข้อมูล และ เวลา ในการประมวลผลลดลงอย่างมาก ภายในการทดลองนี้คุณลักษณะที่ถูกเลือกว่าเหมาะสมกับเสียงพูดภาษาไทยมีด้วยกัน 7 คุณลักษณะ *(Energy Entropy Block, Short-time energy, Zero crossing rate, Spectral-roll-off, Spectral Centroid, Formants frequency, MFCC)* จากทั้งหมด 14 คุณลักษณะ ซึ่งคุณลักษณะที่ผ่านการคัดเลือกเหล่านี้พิสูจน์ให้เห็นถึงประสิทธิภาพในการแยกแยะความแตกต่างของอารมณ์ในเสียงพูดภาษาไทย ด้วยอัตราการรู้จำที่สูง *0.1525, 0.1413, 0.2317* และ *0.185* ตามลำดับ

นอกจากนี้ผู้วิจัย ได้ทำการทดลองถึงกระบวนการเรียนรู้ที่เหมาะสมกับเสียงพูดภาษาไทย โดยเปรียบเทียบระหว่าง 2 วิธี จากผลสรุปได้ว่าจำแนกด้วยวิธี โครงข่ายประสาทเทียมชนิดแพร่กลับ *(BPNN)* ให้ผลลัพธ์ที่ดีกว่า

บรรณานุกรม

Akshay S. Utane and S. L. Nalbalwar, 2013, "Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine," International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013.

Amita Dev and Poonam Bansal, 2010, "Robust Features for Noisy Speech Recognition using MFCC Computation from Magnitude Spectrum of Higher Order Autocorrelation Coefficients". International Journal of Computer Applications, Volume 10– No.8, November 2010, pp.36-38.

Bidoor Noori Ishaq and Bharti W. Gawali, 2014, "Comparative Analysis of MFCC, DTW&ANN for Arabic Speech Recognition," International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349- 2163, Vol.1 Issue 11 (November 2014), pp.56-61.

Catherine J Nereveettil, M.Kalamani, Dr.S.Valarmathy, 2014, "Feature Selection Algorithm for Automatic Speech Recognition Based On Fuzzy Logic", Proceedings of the 2014 International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering, Vol.3 Issue 1.

Chung-Hsien Wu, and Wei-Bin Liang, 2011, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels", IEEE Transactions on Affective Computing, Vol.2, pp. 10-21, 2011.

Dipen Nath and Sanjib Kr. Kalita, 2014, "An effective age detection method based on Short Time Energy and Zero Crossing Rate," Proceedings of the 2014 2nd International Conference on Business and Information Management (ICBIM), pp.99- 103.

Dipti D. Joshi and M.B. Zalte, 2013, "Recognition of Emotion from Marathi Speech using MFCC and DWT algorithms," International Journal of

Advanced Computer Engineering and Communication Technology (IJACECT), Vol.2, Issue (2013).

Jun-Seok Park, Soo-Hong Kim, 2014, "Emotion Recognition from Speech Signals using Fractal Features". International Journal of Software Engineering and Its Applications, Vol.8, No.5 (2014), pp.15-22.

Muhammad Sanaullah and Masud H. Chowdhury, 2014, "Neural network based classification of stressed speech using nonlinear spectral and cepstral features", Proceedings of the 2014 IEEE 12th International New Circuits and Systems Conference (NEWCAS), pp.33-36.

Namrata Dave, 2013, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", International Journal for Advance Research in Engineering and Technology, July 2013, Volume 1 Issue 6 pp.1-4.

S. Sultana, C. Shahnaz, S.A. Fattah, I. Ahmmed and W.- P. Zhu and M.O. Ahmad, 2014, "Speech Emotion Recognition Based on Entropy of Enhanced Wavelet Coefficients," Proceedings of the 2014 IEEE International Symposium on Circuits and Systems (ISCAS), pp.137-140.

Santiago Planet, 2011,"Spontaneous Children's Emotion Recognition by Categorical Classification of Acoustic Features", Information Systems and Technologies (CISTI), pp.1 – 6, 2011.

Stankovic, I., Karnjanadecha, M., and Delic, V., 2011, "Improvement of Thai speech emotion recognition by using face feature analysis", Proceedings of the Nineteenth IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2011), Chiang Mai, Thailand, December 7-9, pp. 87, 2011.

Tran Huy Dat, Cuntai Guan, "Feature selection based on fisher ratio and mutual information analyses for robust brain computer interface". Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on (Volume:1), pp.337-339.

Yan-You Chen, Bo-Wei Chen, Jhing-Fa Wang, and Yi-Cheng Chen, 2010, "Emotion Aware System Based on Acoustic and Textual Features from Speech", Aware Computing (ISAC), pp.92-96, 2010.

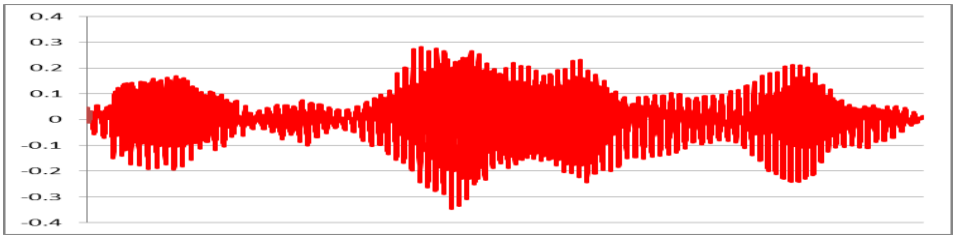
Yun Jin, Peng Song, Wenming Zheng, Li Zhao, 2014, "A feature selection and feature fusion combination method for speaker-independent speech emotion recognition". Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May-2014.

ภาคผนวก ก

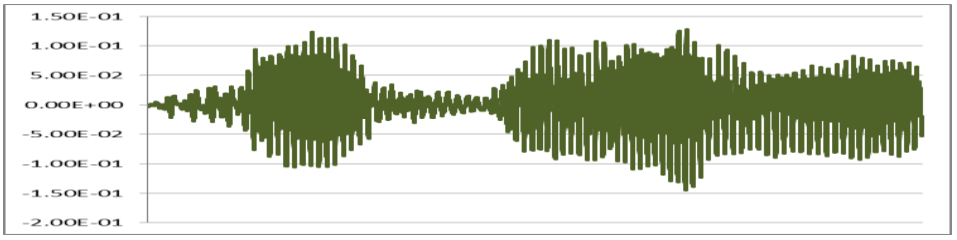
คลื่นเสียงพูด และ คุณลักษณะ

คลื่นเสียง 4 อารมณ์ของคำว่า “โดยเฉพาะอย่างยิ่ง”

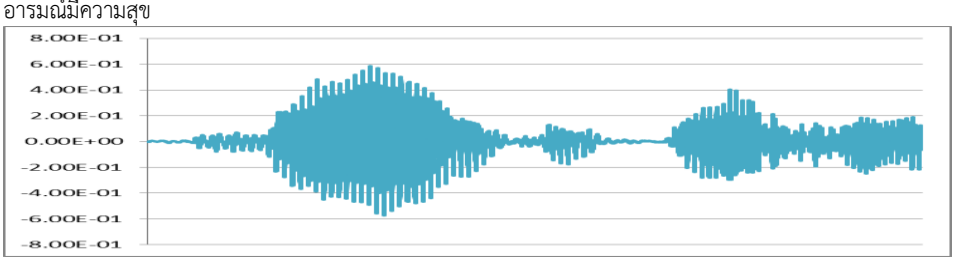
อารมณ์โกรธ



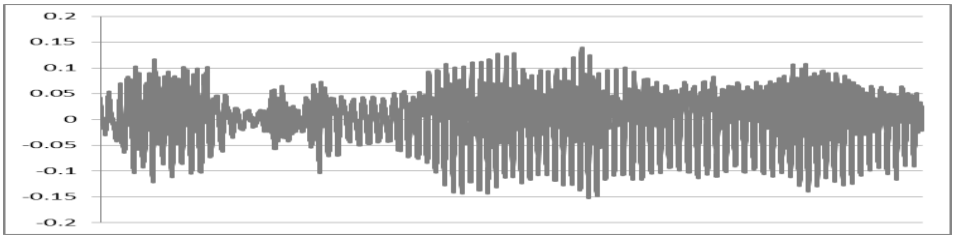
อารมณ์กลัว



อารมณ์มีความสุข



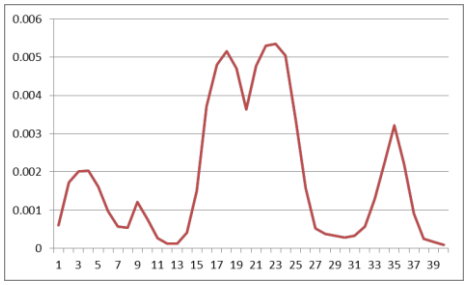
อารมณ์เศร้า



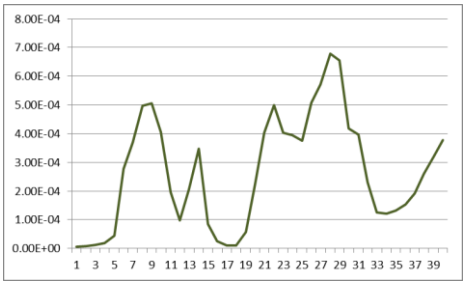
คุณลักษณะตัวอย่างที่ได้จากการสกัดสัญญาณเสียงพูดคำว่า “โดยเฉพาะอย่างยิ่ง” ในอารมณ์ต่าง ๆ ที่นำเสนอในงานวิจัย

Energy Entropy Block

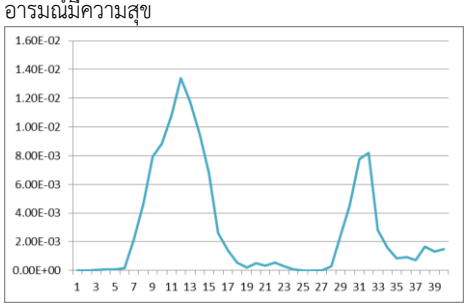
อารมณ์โกรธ



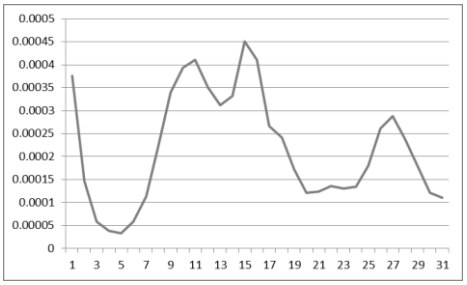
อารมณ์กลัว



อารมณ์มีความสุข

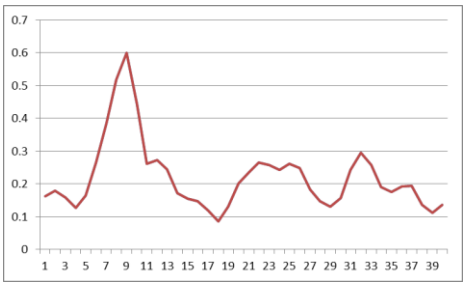


อารมณ์เศร้า

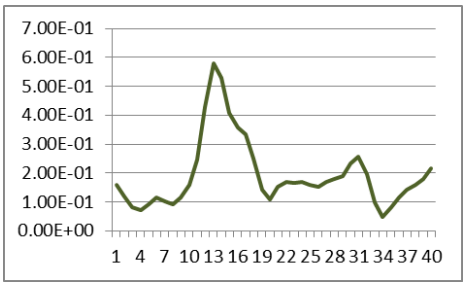


Zero Crossing Rate

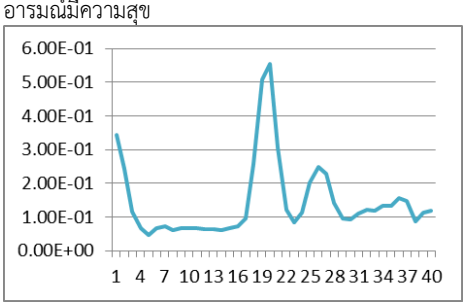
อารมณ์โกรธ



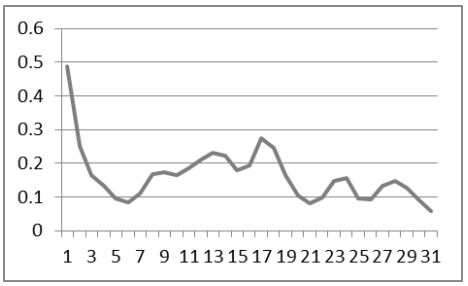
อารมณ์กลัว



อารมณ์มีความสุข

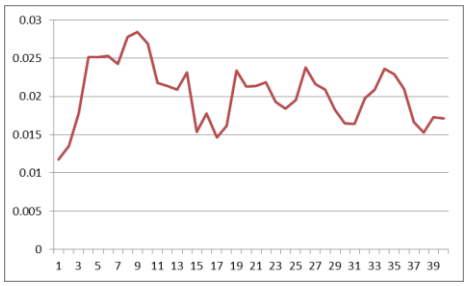


อารมณ์เศร้า

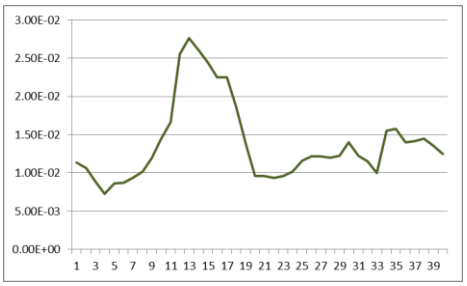


Spectral-Roll-Off

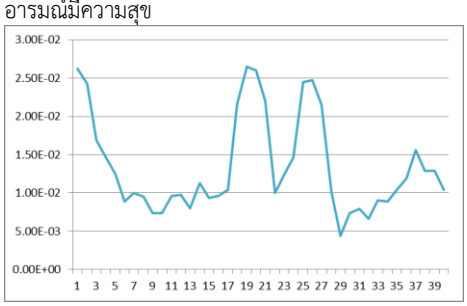
อารมณ์โกรธ



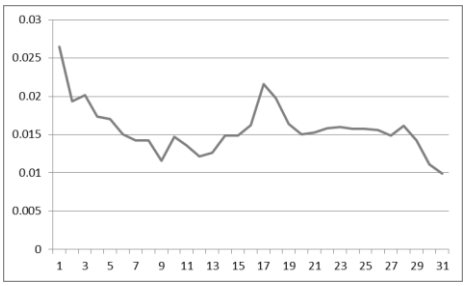
อารมณ์กลัว



อารมณ์มีความสุข

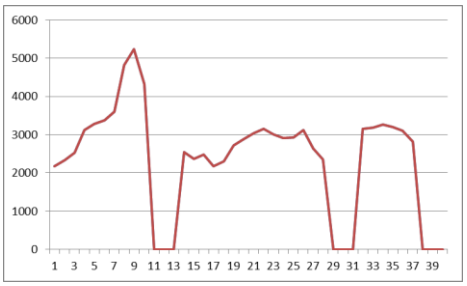


อารมณ์เศร้า

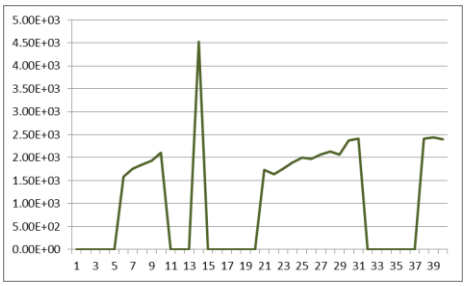


Spectral Centroid

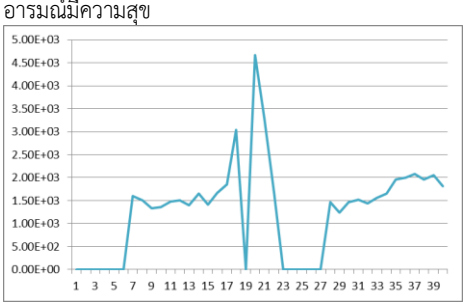
อารมณ์โกรธ



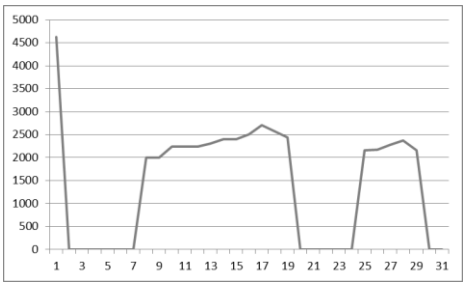
อารมณ์กลัว



อารมณ์มีความสุข



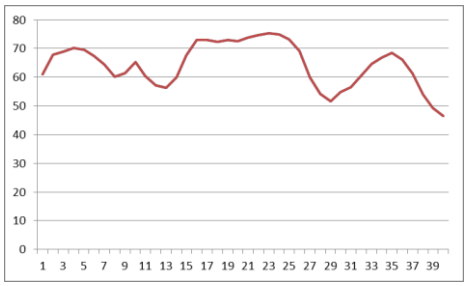
อารมณ์เศร้า



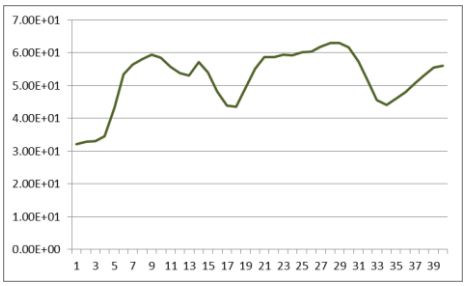
Mel Frequency Cepstral Coefficient (MFCCs)

(เลือกมาเฉพาะแถวที่ 1 ที่นำมาใช้ในงานวิจัย)

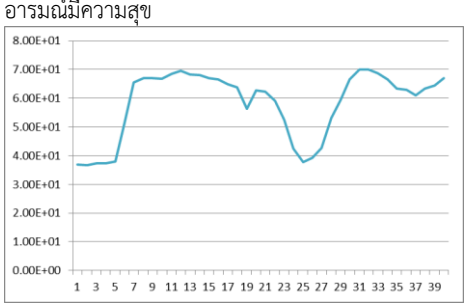
อารมณ์โกรธ



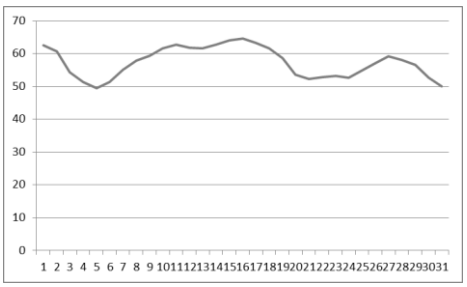
อารมณ์กลัว



อารมณ์มีความสุข

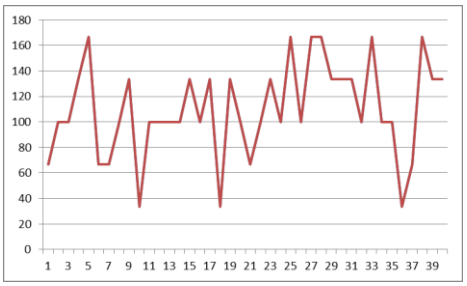


อารมณ์เศร้า

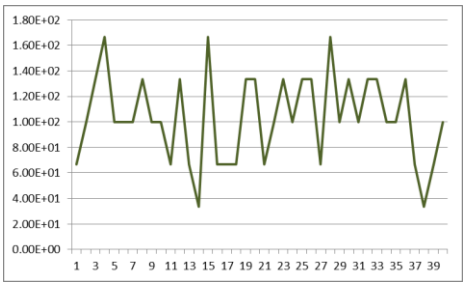


Fundamental Frequency: F0

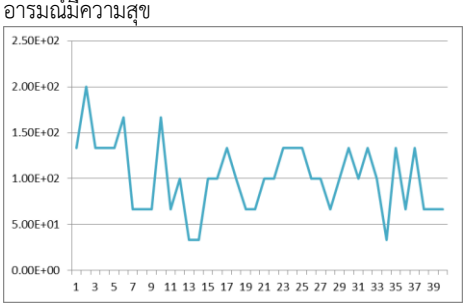
อารมณ์โกรธ



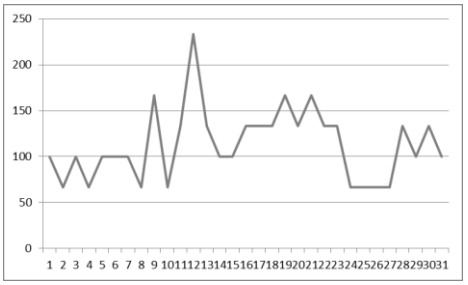
อารมณ์กลัว



อารมณ์มีความสุข

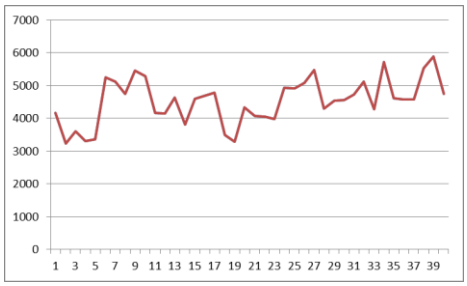


อารมณ์เศร้า

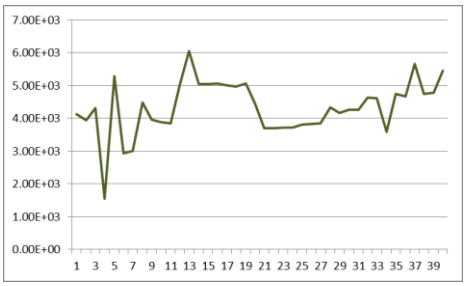


Formants

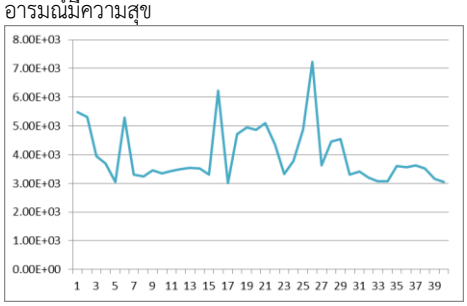
อารมณ์โกรธ



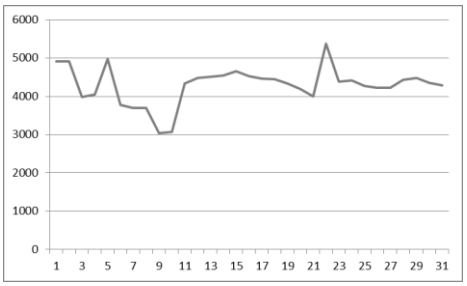
อารมณ์กลัว



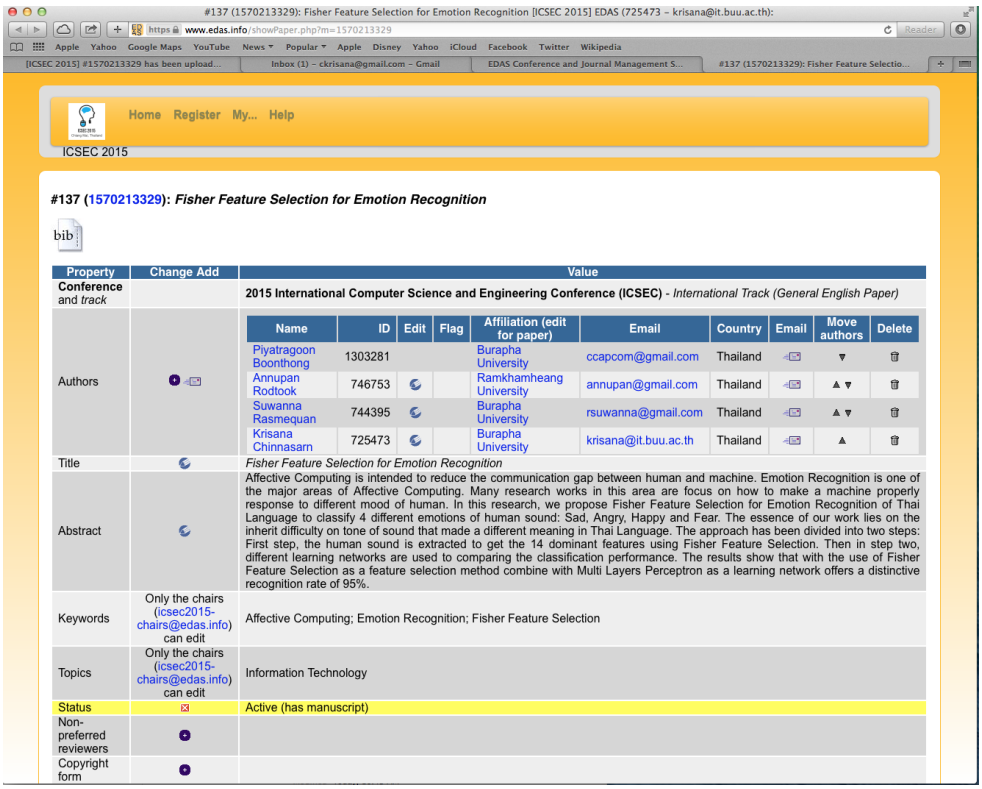
อารมณ์มีความสุข



อารมณ์เศร้า



ภาคผนวก ข



#137 (1570213329): Fisher Feature Selection for Emotion Recognition [ICSEC 2015] EDAS (725473 - krisana@it.buu.ac.th):

www.edas.info/showPaper.php?m=1570213329

Home Register My... Help

ICSEC 2015

#137 (1570213329): *Fisher Feature Selection for Emotion Recognition*

bib

Property	Change Add	Value
Conference and *track*		**2015 International Computer Science and Engineering Conference (ICSEC)** - *International Track (General English Paper)*
Authors		Name: Piyatragoon Boonthong; ID: 1303281; Affiliation: Burapha University; Email: ccapcom@gmail.com; Country: Thailand Name: Annupan Rodtook; ID: 746753; Affiliation: Ramkhamheang University; Email: annupan@gmail.com; Country: Thailand Name: Suwanna Rasmequan; ID: 744395; Affiliation: Burapha University; Email: rsuwanna@gmail.com; Country: Thailand Name: Krisana Chinnasarn; ID: 725473; Affiliation: Burapha University; Email: krisana@it.buu.ac.th; Country: Thailand
Title		*Fisher Feature Selection for Emotion Recognition*
Abstract		Affective Computing is intended to reduce the communication gap between human and machine. Emotion Recognition is one of the major areas of Affective Computing. Many research works in this area are focus on how to make a machine properly response to different mood of human. In this research, we propose Fisher Feature Selection for Emotion Recognition of Thai Language to classify 4 different emotions of human sound: Sad, Angry, Happy and Fear. The essence of our work lies on the inherit difficulty on tone of sound that made a different meaning in Thai Language. The approach has been divided into two steps: First step, the human sound is extracted to get the 14 dominant features using Fisher Feature Selection. Then in step two, different learning networks are used to comparing the classification performance. The results show that with the use of Fisher Feature Selection as a feature selection method combine with Multi Layers Perceptron as a learning network offers a distinctive recognition rate of 95%.
Keywords	Only the chairs (icsec2015-chairs@edas.info) can edit	Affective Computing; Emotion Recognition; Fisher Feature Selection
Topics	Only the chairs (icsec2015-chairs@edas.info) can edit	Information Technology
Status		Active (has manuscript)
Non-preferred reviewers		
Copyright form		

การส่งรายงานการวิจัยเพื่อตีพิมพ์