วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏศรีสะเกษ
ปีที่ 2 ฉบับที่ 2 กรกฎาคม – ธันวาคม 2565

การจำแนกอารมณ์ของมนุษย์จากการรู้จำเสียงพูดโดยใช้การเรียนรู้เชิงลึก

Classification of Human Emotion from Speech Recognition Using Deep Learning

ศรัญญา กาญจนวัฒนา[1*] อัษฎายุธ จารัตน์[1] และ ปัญญ์ชลี ปราณีตพลกรัง[2]

Sarunya Kanjanawattana[1*] Atsadayoot Jarat[1] and Panchalee Praneetpholkrang[2]

สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี[1]

สาขาวิชาเทคโนโลยีการจัดการ สำนักวิชาเทคโนโลยีสังคม มหาวิทยาลัยเทคโนโลยีสุรนารี[2]

School of Computer Engineering, Institute of Engineering, Suranaree University of Technology[1]

School of Management Technology, Institute of Social Technology, Suranaree University of Technology[2]

*Corresponding Author: Sarunya.k@sut.ac.th

ข้อมูลบทความ

ประวัติบทความ:

รับเพื่อพิจารณา: 29 ธันวาคม 2564

แก้ไข: 10 กรกฎาคม 2565

ตอบรับ: 20 กรกฎาคม 2565

คำสำคัญ:

อารมณ์ของมนุษย์/การจำแนกประเภท/Long Short-Term Memory/Convolution Neuron Networks/การรู้จำเสียงพูด

บทคัดย่อ

อารมณ์ของมนุษย์เป็นกระบวนการทางจิตที่ตอบสนองต่อสิ่งเร้าที่เกิดขึ้นรอบตัวและมีความซับซ้อนสูง ซึ่งเป็นกลไกทำให้มนุษย์ปรับตัวและการแสดงออกทางอารมณ์ในสถานการณ์ต่าง ๆ อย่างไรก็ตามในสถานการณ์เดียวกันนั้นมนุษย์มีการแสดงอารมณ์แตกต่างกัน ทำให้การที่จะเข้าถึงและจับความรู้สึกของผู้อื่นได้อย่างถูกต้องนั้นเป็นเรื่องยาก การคาดเดาอารมณ์ของคู่สนทนา ทำให้เกิดการตัดสินใจและการกระทำที่เหมาะสมต่อสถานการณ์ เช่น การรักษาผู้ป่วยที่เป็นโรคซึมเศร้าหรือผู้ที่ต้องการได้รับการบำบัดทางจิต การศึกษานี้มีวัตถุประสงค์คือ เพื่อเปรียบเทียบประสิทธิภาพระหว่างโมเดล Convolution Neuron Networks (CNN) และ Long Short-Term Memory (LSTM) และเพื่อหาโครงสร้างโมเดลที่เหมาะสมในการจำแนกอารมณ์จากเสียง โดยได้มีการดำเนินการทดลองกับการปรับแต่งค่าต่างๆ ของ LSTM และ CNN ซึ่งผลการทดลองพบว่า LSTM ที่มีระดับชั้น 4 เหมาะสมกับการจำแนกอารมณ์จากเสียงพูดของมนุษย์ในงานวิจัยนี้ได้พัฒนาแบบจำลองโดยใช้เทคนิคการเรียนรู้เชิงลึกในการจำแนกอารมณ์จากเสียงของมนุษย์ มี 5 อารมณ์ ประกอบด้วย ปกติ โกรธ ประหลาดใจ มีความสุขและเศร้า

Article Info

Article History:

Received: December 29, 2021

Revised: July 10, 2022

Accepted: July 20, 2022

Abstract

Human emotions are complex mental processes that respond to surrounding stimulators. It is a mechanism that allows humans to adjust themselves and express their emotions in various situations. In a particular situation, humans can manifest their emotions diversely. Therefore, it is difficult to

Keywords:
Human Emotions/ Classification/Long Short-Term Memory/Convolution Neuron Networks/Speech Recognition

catch and understand the actual emotions. Predicting the interlocutors' emotions help to decide proper actions for specific situations such as for treating patients with depression or those who need psychotherapy. This study develops deep learning models to classify human emotions by using human speech. Then, humans' voices are classified based on five emotional types including normal emotion, anger, surprise, happiness, and sadness. The objectives of this study are to 1) compare the performances of two classifying models i.e., Convolution Neuron Networks (CNN), and Long Short-Term Memory (LSTM), and 2) propose the most appropriate model for classifying humans' emotions from speech recognition. It reveals that classification results generated by LSTM outperform CNN. With LSTM, there are four classes to recognize humans' speech emotions such as normal, angry, surprised, happy, and sad.

1. บทนำ

อารมณ์เป็นกระบวนการทางจิตของมนุษย์ที่ตอบสนองต่อสิ่งเร้าที่เกิดขึ้นรอบตัว ซึ่งมีความซับซ้อนและเป็นประสบการณ์ส่วนบุคคล หนึ่งในทฤษฎีที่มีการอธิบายการแสดงออกทางอารมณ์ของมนุษย์ คือ ทฤษฎีวิวัฒนาการ [1] ได้กล่าวถึงอารมณ์ว่า เป็นส่วนหนึ่งที่ทำให้สิ่งมีชีวิตสามารถดำรงเผ่าพันธุ์ได้ผ่านกระบวนการปรับตัวและการแสดงออกทางอารมณ์เป็นการช่วยให้มนุษย์สามารถทำนายหรือคาดเดาพฤติกรรม ซึ่งทำให้ง่ายต่อการแสดงพฤติกรรมตอบโต้สิ่งนั้นๆ แต่พบว่าในสถานการณ์เดียวกันนั้นมนุษย์มีการแสดงอารมณ์แตกต่างกัน ทำให้การที่จะเข้าถึงและจับความรู้สึกของผู้อื่นได้อย่างถูกต้องนั้นเป็นเรื่องยาก รวมทั้งมีกรณีที่เมื่อมนุษย์อยู่ในอารมณ์ความรู้สึกเดียวกันแต่มีการแสดงออกที่แตกต่างกัน และในบางสถานการณ์หากเรารับรู้อารมณ์ของผู้ที่สนทนาได้ช้าเกินไป อาจนำไปสู่เหตุการณ์ที่ไม่คาดคิดได้ เช่น การรักษาผู้ป่วยที่เป็นโรคซึมเศร้าหรือผู้ที่ต้องการได้รับการบำบัดทางจิต [2-3] การแสดงออกทางอารมณ์ของผู้ได้รับการบำบัดจะช่วยทำให้ผู้บำบัดหรือแพทย์ผู้รักษาสามารถดำเนินวิธีการรักษาได้อย่างแม่นยำและมีประสิทธิภาพมากขึ้น ในทางกลับกันถ้าหากเราสามารถรับรู้อารมณ์ได้อย่างรวดเร็วและแม่นยำมากขึ้น จะทำให้เราสามารถตอบสนองต่อความต้องการที่แตกต่างกันของมนุษย์ได้อย่างมีประสิทธิภาพมากขึ้น เช่น การประยุกต์ใช้ในการตรวจจับอารมณ์ของลูกค้าที่มารับบริการ เพื่อปรับปรุงแนวทางการให้บริการต่อไป เป็นต้น

ในปัจจุบันปัญญาประดิษฐ์ (Artificial Intelligence: AI) เป็นเทคโนโลยีที่เข้ามามีบทบาทและความสำคัญต่อชีวิตมนุษย์ในยุคดิจิทัลมากขึ้น เพื่ออำนวยความสะดวกสบายในการใช้ชีวิตประจำวันของมนุษย์ โดยทั่วไปแล้วการตรวจจับอารมณ์ของมนุษย์สามารถวิเคราะห์ได้จากการแสดงออกผ่านทางใบหน้า มีงานวิจัยทางด้านปัญญาประดิษฐ์จำนวนมากศึกษากระบวนการตรวจจับอารมณ์ของมนุษย์ผ่านทางใบหน้า ซึ่งได้มีความพยายามในการเพิ่มความแม่นยำและประสิทธิภาพให้สูงยิ่งขึ้น Piyapong, et al. [4] ได้นำเสนอวิธีในการตรวจจับใบหน้าและ

ทำการจำแนกอารมณ์โดยใช้การเรียนรู้เชิงลึก มีตรวจจับวัตถุด้วย Object Detection API และการจำแนกประเภทด้วยสถาปัตยกรรม mini-Xception ได้มุ่งเน้นในการวิเคราะห์อารมณ์ 4 กลุ่ม คือ อารมณ์โกรธ อารมณ์เศร้า มีความสุข และอารมณ์ปกติ โดยมีความแม่นยำอยู่ที่ 76.23% ในอีกงานวิจัยหนึ่ง Song, et al. [5] พัฒนาแอพพลิเคชั่นในการตรวจจับอารมณ์บนใบหน้าผ่านสมาร์ทโฟน ซึ่งได้ใช้ 65,000 นิวรอน 5 ระดับชั้นในโครงข่ายของ Artificial Neuron Networks และได้ใช้วิธีที่ชื่อว่า Dropout ในการแก้ปัญหา Overfitting

อย่างไรก็ตามจากการวิเคราะห์งานวิจัยในอดีตที่ผ่านมา [4, 6] พบว่าในการแสดงออกบางอารมณ์ เช่น โกรธ และ เศร้า บนใบหน้าของมนุษย์ มีการแสดงออกที่คล้ายคลึงกันและส่งผลให้ความแม่นยำในการจำแนกอารมณ์ด้วยใบหน้าของมนุษย์ลดลง อีกทั้งการแสดงออกทางใบหน้าของแต่ละคนไม่เหมือนกันแม้จะอยู่ในอารมณ์เดียวกัน ดังนั้นเพื่อลดช่องว่างของการจดจำอารมณ์จากใบหน้ามนุษย์เพียงอย่างเดียว การใช้เสียงของมนุษย์จึงเป็นอีกทางเลือกหนึ่งในการตรวจจับอารมณ์แทนที่จะเป็นใบหน้า [7]

ในงานวิจัยนี้ ผู้วิจัยและคณะได้พัฒนาแบบจำลองโดยใช้เทคนิคการเรียนรู้เชิงลึกในการจำแนกอารมณ์จากเสียงของมนุษย์ มี 5 อารมณ์ประกอบด้วย ปกติ โกรธ ประหลาดใจ มีความสุขและเศร้า จุดประสงค์ของงานวิจัยคือเพื่อเปรียบเทียบประสิทธิภาพระหว่างโมเดล Convolution Neural Networks (CNN) และ Long Short-Term Memory (LSTM) และเพื่อหาโครงสร้างโมเดลที่เหมาะสมในการจำแนกอารมณ์จากเสียง ทั้ง 2 อัลกอริทึมมีโครงสร้างโมเดลที่เหมาะสมกับการเรียนรู้ข้อมูลที่มีความซับซ้อนได้ดี CNN มักจะเป็นโมเดลที่ทำงานกับข้อมูลประเภทรูปภาพแต่ก็สามารถทำงานกับข้อมูลเสียงได้เช่นกัน ยกตัวอย่างเช่น ในงานวิจัยของ Abdul Malik Badshah, et al. [8] ได้นำเสนอวิธีการเรียนรู้จดจำอารมณ์จากเสียงพูดโดยใช้ spectrogram ร่วมกับ CNN ข้อมูลที่ใช้ คือชุดข้อมูลของ Berlin emotion จากนั้นได้ทำการฝึกฝนกับ AlexNet ซึ่งได้ผลปรากฏว่า แบบจำลองที่ถูกพัฒนาขึ้นให้ผลลัพธ์ที่ดีกว่าแบบจำลองที่ได้มีการปรับแต่ง Sathit Prasomphan [9] ได้พัฒนาอัลกอริทึมใหม่ที่ใช้สำหรับในการตรวจจับอารมณ์ของมนุษย์โดยใช้คำพูด ที่ถูกแปลงให้อยู่ในรูปของ spectrogram แบบจำลองที่ใช้คือ Neural network แหล่งข้อมูลคือ ฐานข้อมูล Berlin emotion และ Audiovisual Thai Emotion ผลของการทดลองแสดงให้เห็นว่า แบบจำลองที่พัฒนาขึ้นให้ประสิทธิภาพที่ดี งานวิจัยทั้ง 2 งานได้ใช้ CNN ร่วมกับข้อมูลเสียงเพื่อทำการจำแนกอารมณ์ แต่มีวิธีการจัดการข้อมูลเสียงโดยแปลงให้เป็นสเปกโตรแกรมก่อนนำเข้าสู่โมเดล ซึ่งจะเป็นการนำจุดแข็งของ CNN มาใช้กับข้อมูลเสียง และ LSTM เป็นโมเดลที่ทำงานร่วมกับข้อมูลที่เป็น Time-series เช่น คลื่นสัญญาณต่าง ๆ ข้อมูลหุ้นหรือข้อมูลที่เป็นลักษณะ Transaction ที่ขึ้นกับเวลา เป็นต้น ในงานวิจัยของ [8] ได้ใช้ข้อมูลเสียงร่วมกับ LSTM เพื่อจำแนกอารมณ์ โดยวิเคราะห์ถึงระดับเฟรมของเสียงพูด ซึ่งจะเห็นถึงความสัมพันธ์ระหว่างช่วงเวลาและลำดับของเฟรม จากเหตุผลดังกล่าว ในงานวิจัยนี้จึงได้เลือก CNN และ LSTM ในการประมวลผลเนื่องจากอัลกอริทึมมีกระบวนการที่ใช้ในการจัดการข้อมูลอย่างเหมาะสม

CNN และ LSTM เป็นอัลกอลิทึมที่มีพื้นฐานมาจาก Artificial Neural Network (ANN) ดังนั้นโดยหลักการแล้วทั้ง 3 อัลกอริทึมมีการทำงานที่คล้ายกันมาก CNN หรือโครงข่ายประสาทแบบคอนโวลูชันมีความสามารถจำลองการมองเห็นของมนุษย์ที่มองพื้นที่ที่สนใจเป็นส่วนย่อย ๆ และนำกลุ่มของพื้นที่เหล่านั้นมาผสานกัน เพื่อพิจารณาว่าสิ่งที่มองอยู่คืออะไร ดังนั้นในส่วนการทำงานของ CNN จึงต้องอาศัยหลักการทางคณิตศาสตร์มารองรับเพื่อให้ CNN สามารถทำงานตามแนวคิดของมนุษย์ได้ CNN มีโครงสร้างสำคัญ 2 ส่วน คือ Feature Extraction ที่ทำหน้าที่ดึงคุณลักษณะที่ใช้ในการรู้จำวัตถุออกมาโดยใช้ Filter และ Classification ทำหน้าที่จำแนกประเภทของการรู้จำวัตถุ ซึ่งเป็นการทำงานของ ANN ในส่วน LSTM ได้ถูกพัฒนามาจาก Recurrent Neural Network (RNN) การทำงานของ RNN และ LSTM ใช้หลักการเดียวกันกับ ANN ต่างกันที่ระหว่างการ

ฝึกสอนข้อมูลที่เข้าไปใน Neuron สามารถนำเอาความรู้ที่เกิดขึ้นก่อนหน้ามาบวกกับข้อมูลเข้าตัวใหม่ที่เข้ามาเพื่อเกิดการเรียนรู้จากลำดับก่อนหน้า

2. วิธีดำเนินการวิจัย

2.1 เทคโนโลยีที่เลือกใช้

ภาษาหลักที่ใช้ คือ Python3 โดยมี Library หลัก ๆ คือ TensorFlow เป็นเครื่องมือในการสร้างโมเดล Librosa ใช้สำหรับการสกัดคุณลักษณะของข้อมูล Pandas ใช้สำหรับการสร้างชุดข้อมูล Numpy เป็นเครื่องมือในการจัดการค่าต่าง ๆ ทางคณิตศาสตร์และ Matplotlib เป็นเครื่องมือในการช่วยแสดงผลลัพธ์ให้เข้าใจง่ายมากขึ้น

2.2 แหล่งข้อมูลเพื่อการวิจัย

ในงานวิจัยนี้ได้รวมรวมข้อมูล 2 ส่วนคือ ส่วนที่ 1 คือ ข้อมูลเสียงพูดที่ผู้วิจัยได้รวมรวบมาเอง แบ่งเป็น 5 อารมณ์ ประกอบด้วย โกรธ ปกติ ประหลาดใจ มีความสุข และเศร้า มีจำนวนราว 500 ข้อมูล เป็นเฉพาะเสียงพูดภาษาไทยเท่านั้น ส่วนที่ 2 คือ ข้อมูลจาก AIResearch.in.th มีจำนวนข้อมูลทั้งหมด 27,854 ข้อมูล แบ่งเป็นอารมณ์ปกติ 7,360 ข้อมูล อารมณ์โกรธ 3,112 ข้อมูล อารมณ์สุข 4,071 ข้อมูล อารมณ์เศร้า 2,781 ข้อมูล อารมณ์หงุดหงิด 7,861 ข้อมูล อารมณ์อื่น ๆ 5 ข้อมูล และไม่มีความเห็นพ้อง 2664 ข้อมูล

2.3 ระเบียบวิธีการวิจัย

2.3.1 การจัดเตรียมข้อมูล

ก่อนนำไปฝึกสอน ข้อมูลเข้าถูกดึงค่าคุณลักษณะด้วย Mel Frequency Cepstral Coefficients (MFCCs) และตัดข้อมูลที่มีความช่วงคำยาวน้อยเกินไปและข้อมูลที่ผู้วิจัยพิจารณาแล้วไม่สามารถใช้ได้ออกไปจากชุดข้อมูล MFCCs เป็นวิธีการที่เป็นที่นิยมอย่างมากในการดึงคุณสมบัติสำคัญที่อยู่ในข้อมูลเสียงออกมาให้แสดงอยู่ในรูปของเวกเตอร์ MFCCs สามารถแสดงถึงเสียงความถี่ต่ำได้ดี ซึ่งมนุษย์นั้นได้ยินเสียงความถี่ต่ำได้ดีกว่าเสียงที่มีความถี่สูง เมื่อได้ค่าคุณลักษณะที่สำคัญและเหมาะสมกับการนำมาเรียนรู้แล้ว ได้ทำการแบ่งข้อมูลเป็น 3 ส่วนแบบสุ่ม ประกอบด้วย ชุดข้อมูลสำหรับการฝึก เป็นสัดส่วนร้อยละ 70 ของชุดข้อมูล โดยแต่ละคลาสมีจำนวนข้อมูลประมาณ 3,000 ข้อมูล ชุดข้อมูลสำหรับการตรวจสอบ เป็นสัดส่วนร้อยละ 15 ของชุดข้อมูล และชุดข้อมูลสำหรับการทดสอบเป็นสัดส่วนร้อยละ 15 ของชุดข้อมูล โดยข้อมูลที่ถูกแบ่งนี้จะทำการแบ่งโดยคำนึงถึงความสมดุลของคลาสอารมณ์เป้าหมายด้วย (Balanced data)

2.3.2 การฝึกสอนแบบจำลอง

ในงานวิจัยนี้ ได้เลือก 2 อัลกอริทึม คือ CNN และ LSTM มาทำการเปรียบเทียบประสิทธิภาพกันเพื่อค้นหาโมเดลที่เหมาะสมสำหรับการจำแนกอารมณ์จากเสียงพูด โดยผู้วิจัยได้นำชุดข้อมูลสำหรับการฝึกมาสอนโมเดลของทั้ง 2 อัลกอริทึมนี้ ดังนั้นโมเดลที่ได้จะมีการเรียนรู้มาจากข้อมูลชุดเดียวกัน คลาสเป้าหมายที่เหมือน จากนั้นจึงนำชุดข้อมูลสำหรับการทดสอบมาประเมินโมเดลที่ได้เพื่อพิจารณาเปรียบเทียบความน่าเชื่อถือของโมเดลทั้ง 2 ตัว

3. การทดลองและผลการทดลอง

3.1 การทดลองเพื่อเปรียบเทียบประสิทธิภาพระหว่าง CNN และ LSTM

กำหนดให้โครงสร้างของ CNN และ LSTM ที่ใช้ในการทดลองนี้แสดงในรูปที่ 1 และ 2 ตามลำดับ เมื่อนำข้อมูลที่เตรียมไว้ฝึกสอนและทดสอบโมเดลได้ผลปรากฏว่า การฝึกโมเดล CNN มีค่าการสูญเสียของโมเดลที่

ด้วยชุดข้อมูลฝึกฝนใน epoch ที่ 500 มีค่า 1.334 และค่าการสูญเสีย (Loss Function) ของโมเดลที่ด้วยชุดข้อมูลทดสอบใน epoch ที่ 500 มีค่า 1.117 ดังที่แสดงในรูปที่ 3 และมีความแม่นยำอยู่ที่ร้อยละ 42.06 สำหรับการฝึกโมเดล LSTM มีค่าการสูญเสียของโมเดลที่ด้วยชุดข้อมูลฝึกฝนใน epoch ที่ 500 มีค่า 0.142 และค่าการสูญเสียของโมเดลที่ด้วยชุดข้อมูลตรวจสอบความถูกต้องใน epoch ที่ 500 มีค่า 0.160 ดังที่แสดงในรูปที่ 4 และมีความแม่นยำอยู่ที่ร้อยละ 40.80 ในการทดลอง ผู้วิจัยเลือก epoch ที่ 500 เนื่องจากเป็นจุดที่ค่าการสูญเสียต่ำลงอย่างต่อเนื่องและให้ค่าความแม่นยำดีที่สุด

ในการทดลอง ผู้วิจัยได้ทำการกำหนดค่า Hyperparameter ของแบบจำลองเป็นค่าปกติที่ทางไลบรารี่ Keras ได้ทำการกำหนดไว้ให้แล้ว การดำหนดค่า epoch เป็นการกำหนดโดยพิจารณาจากค่าความสูญเสีย (Loss) ขณะการฝึกฝน โดยได้เลือก epoch ที่ได้รับค่า Loss น้อยที่สุดที่ผู้วิจัยรับได้ ในที่นี้จึงใช้ค่า epoch ที่ 500 ตามที่แสดงในรูปที่ 3 และ 4

ค่าการสูญเสีย (Loss Function) เป็นค่าที่แสดงถึงเป้าหมายของการฝึกสอนของโมเดล โดยถ้าหากมีค่าสูญเสียสูง หมายความว่า การฝึกสอนของโมเดลยังไม่ถึงเป้าหมายที่ตั้งไว้ การนำโมเดลที่มีค่าสูญเสียสูงมาใช้งานอาจได้รับผลที่ไม่ถูกต้องหรือมีความแม่นยำต่ำได้ โดยการลดค่าสูญเสียสามารถทำได้โดยการเพิ่มจำนวน Epoch ซึ่งเป็นจำนวนรอบของโมเดลเพื่อฝึกสอน ในทางกลับกันถ้าค่าสูญเสียต่ำมากจนถึงจุดที่เป็นเป้าหมายหรือผู้วิจัยสามารถรับได้แล้ว สามารถนำเอาโมเดลดังกล่าวไปใช้งานได้ ดังนั้นในงานวิจัยนี้ ผู้วิจัยได้ใช้ค่าสูญเสียและค่าความแม่นยำเป็นตัววัดประสิทธิภาพของแบบจำลอง

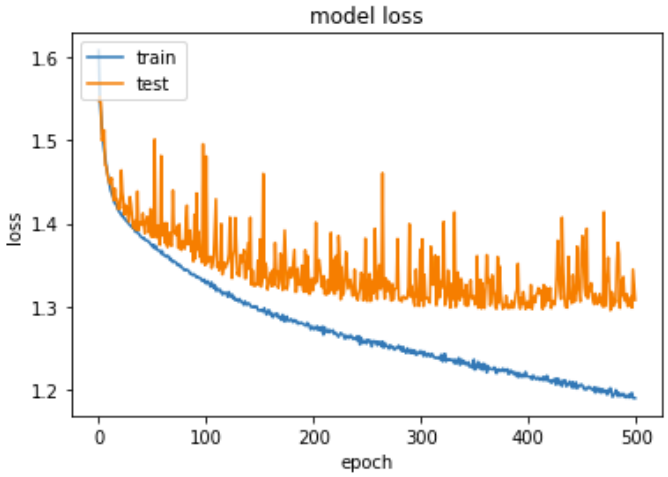
จากการพิจารณาค่าสูญเสียของข้อมูลทดสอบของโมเดล 2 ตัวที่ epoch ที่ 500 เห็นได้ว่า ถึงแม้ว่า CNN ให้ความแม่นยำสูงกว่า LSTM เล็กน้อยแต่โมเดล LSTM ให้ค่าสูญเสียน้อยกว่า CNN อย่างชัดเจน ดังนั้นเราจึงเลือกใช้ LSTM ในการจำแนกอารมณ์จากเสียง

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d (Conv1D)              (None, 216, 256)          1536
_________________________________________________________________
activation (Activation)      (None, 216, 256)          0
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 216, 128)          163968
_________________________________________________________________
activation_1 (Activation)    (None, 216, 128)          0
_________________________________________________________________
dropout (Dropout)            (None, 216, 128)          0
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 27, 128)           0
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 27, 128)           82048
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 27, 128)           82048
_________________________________________________________________
activation_2 (Activation)    (None, 27, 128)           0
_________________________________________________________________
flatten (Flatten)            (None, 3456)              0
_________________________________________________________________
dense (Dense)                (None, 20)                69140
_________________________________________________________________
dense_1 (Dense)              (None, 5)                 105
_________________________________________________________________
activation_3 (Activation)    (None, 5)                 0
=================================================================
```

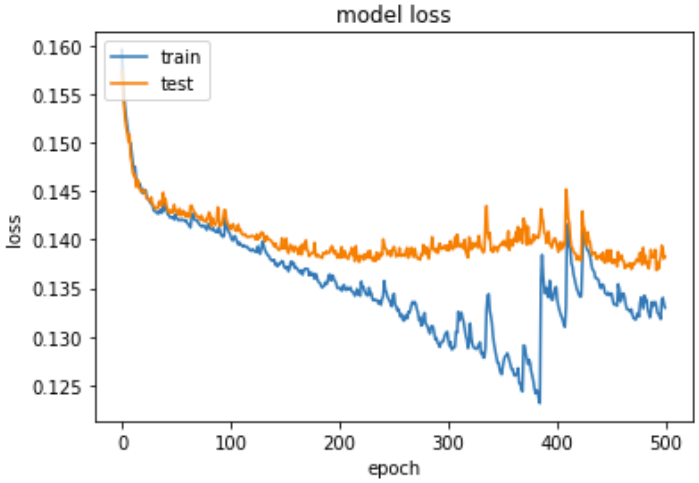
ภาพที่ 1 โครงสร้างของ CNN ที่ใช้ในงานวิจัยนี้

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_5 (LSTM)                (None, 216, 128)          66560
_________________________________________________________________
lstm_6 (LSTM)                (None, 216, 64)           49408
_________________________________________________________________
lstm_7 (LSTM)                (None, 216, 32)           12416
_________________________________________________________________
lstm_8 (LSTM)                (None, 216, 16)           3136
_________________________________________________________________
lstm_9 (LSTM)                (None, 8)                 800
_________________________________________________________________
dense_1 (Dense)              (None, 5)                 45
=================================================================
```

ภาพที่ 2 โครงสร้างของ LSTM ที่ใช้ในงานวิจัยนี้



ภาพที่ 3 ค่าสูญเสีย (Loss) การฝึกสอนและทดสอบโมเดล CNN



ภาพที่ 4 ค่าสูญเสีย (Loss) การฝึกสอนและทดสอบโมเดล LSTM

3.2 การทดลองเพื่อหาโครงสร้างโมเดลที่เหมาะสมในการจำแนกอารมณ์จากเสียง

จากผลการทดลองที่ 3.1 ได้ผลว่า LSTM เหมาะสมกับการจำแนกอารมณ์จากเสียงมากกว่า CNN เนื่องจากค่าสูญเสียที่ได้ต่ำกว่า ดังนั้นในการทดลองนี้จึง ทำการปรับโครงสร้าง LSTM เป็นรูปแบบต่างกัน 3 รูปแบบ ดังภาพที่ 5ก-5ค

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 216, 128)	66560
lstm_1 (LSTM)	(None, 216, 64)	49408
lstm_2 (LSTM)	(None, 216, 32)	12416
lstm_3 (LSTM)	(None, 216, 16)	3136
lstm_4 (LSTM)	(None, 8)	800
dense (Dense)	(None, 3)	27

5ก

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 216, 128)	66560
lstm_1 (LSTM)	(None, 216, 64)	49408
lstm_2 (LSTM)	(None, 216, 32)	12416
lstm_3 (LSTM)	(None, 16)	3136
dense (Dense)	(None, 3)	51

5ข

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 216, 128)	66560
lstm_1 (LSTM)	(None, 216, 128)	131584
lstm_2 (LSTM)	(None, 216, 64)	49408
lstm_3 (LSTM)	(None, 216, 32)	12416
lstm_4 (LSTM)	(None, 216, 16)	3136
lstm_5 (LSTM)	(None, 216, 8)	800
lstm_6 (LSTM)	(None, 4)	208
dense (Dense)	(None, 3)	15

5ค

ภาพที่ 5 ก) โครงสร้าง LSTM แบบที่ 1 มีระดับชั้นอยู่ 4 ชั้น
ข) โครงสร้าง LSTM แบบที่ 2 มีระดับชั้นอยู่ 3 ชั้น
ค) โครงสร้าง LSTM แบบที่ 3 มีระดับชั้นอยู่ 6 ชั้น

จากการทดลองโดยนำข้อมูลฝึกสอนและทดสอบทำงานกับ LSTM แบบที่ 1-3 (ตารางที่ 1) พบว่าโครงสร้าง LSTM แบบที่ 1 มีค่าการสูญเสียของโมเดลที่ด้วยชุดข้อมูลฝึกฝนใน epoch ที่ 600 มีค่า 0.141 และค่าการสูญเสียของโมเดลที่ด้วยชุดข้อมูลตรวจสอบความถูกต้องใน epoch ที่ 600 มีค่า 0.140 (ภาพที่ 6) และมีค่าความแม่นยำ (Precision) อยู่ที่ 67.38%

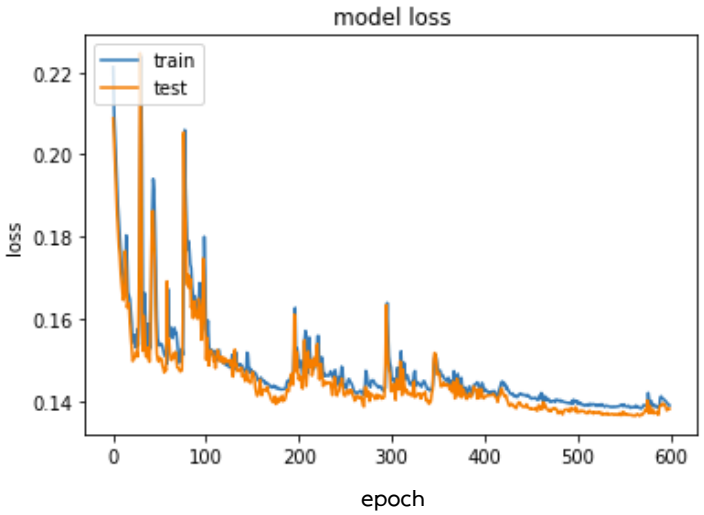
ในโครงสร้าง LSTM แบบที่ 2 มีค่าการสูญเสียของโมเดลที่ด้วยชุดข้อมูลฝึกฝนใน epoch ที่ 600 มีค่า 0.136 และค่าการสูญเสียของโมเดลที่ด้วยชุดข้อมูลตรวจสอบความถูกต้องใน epoch ที่ 600 มีค่า 0.140 (ภาพที่ 7) และมีค่าความแม่นยำอยู่ที่ 65.90%

ในโครงสร้าง LSTM แบบที่ 3 มีค่าการสูญเสียของโมเดลที่ด้วยชุดข้อมูลฝึกฝนใน epoch ที่ 600 มีค่า 0.133 และค่าการสูญเสียของโมเดลที่ด้วยชุดข้อมูลตรวจสอบความถูกต้องใน epoch ที่ 600 มีค่า 0.142 (ภาพที่ 8) และมีค่าความแม่นยำอยู่ที่ 66.13%

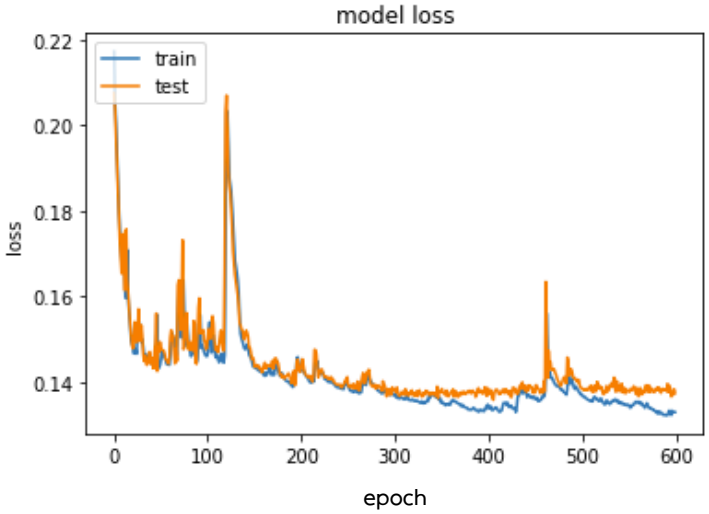
สรุปแล้ว จากการเปรียบเทียบประสิทธิภาพของโมเดล LSTM ทั้ง 3 โครงสร้าง โครงสร้างที่ให้ผลลัพธ์ที่ดีที่สุดคือ LSTM แบบที่ 1 ที่มีจำนวนระดับชั้นอยู่ 4 ระดับ เนื่องจากค่าการสูญเสียของโมเดลต่ำที่สุดและค่าความแม่นยำสูงที่สุด

ตารางที่ 1 ผลการทดลองที่ 3.2

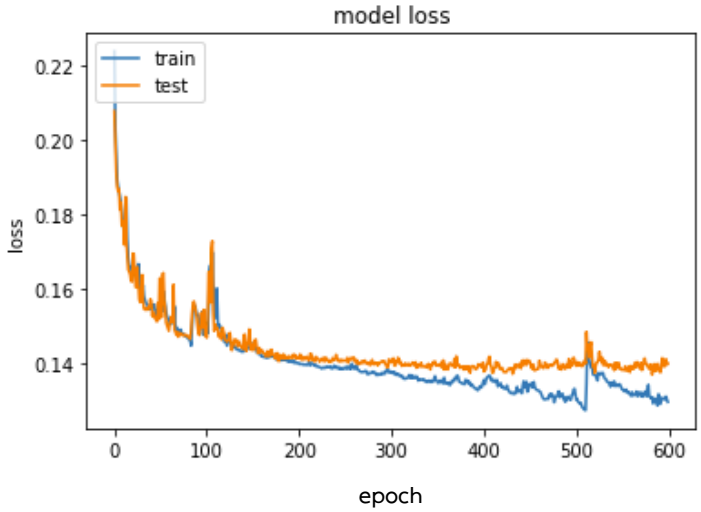
ระดับชั้นของ LSTM	ค่าสูญเสีย	ค่าความแม่นยำ
4 ระดับชั้น	0.14	67.38%
3 ระดับชั้น	0.14	65.90%
6 ระดับชั้น	0.142	66.13%



ภาพที่ 6 แสดงค่าความสูญเสียของโครงสร้าง LSTM แบบที่ 1



ภาพที่ 7 แสดงค่าความสูญเสียของโครงสร้าง LSTM แบบที่ 2



ภาพที่ 8 แสดงค่าความสูญเสียของโครงสร้าง LSTM แบบที่ 3

4. การอภิปรายผลวิจัย

ในการทดลองที่ 1 ผลปรากฏว่า ผลการทดลองกับ CNN ให้ความแม่นยำสูงกว่า LSTM เล็กน้อย แต่ LSTM ได้ให้ประสิทธิภาพสูงกว่า CNN เนื่องจากค่าการสูญเสียที่ได้น้อยกว่าในจำนวน epoch ที่เท่ากัน ซึ่งผลการทดลองที่ได้สอดคล้องกับงานวิจัย [9] ที่ได้เปรียบเทียบประสิทธิภาพของอัลกอริทึมประเภทอนุกรมเวลากับข้อมูลการรู้จำเสียง ซึ่งผลปรากฏว่า LSTM เป็นอัลกอริทึมที่ให้ประสิทธิภาพที่ดีที่สุด ในขณะที่ Gated Recurrent Unit (GRU) ใช้เวลาในการประมวลผลน้อยที่สุด

อย่างไรก็ตามเมื่อวิเคราะห์ค่าความสูญเสียของโมเดลทั้ง 2 ตัวที่ 500 epoch กับชุดข้อมูลทดสอบ CNN มีค่าความสูญเสียอยู่ที่ 1.117 ในขณะที่ LSTM มีค่าความสูญเสียอยู่ที่ 0.142 จากภาพที่ 3 และ 4 เส้นกราฟยังมีแนวโน้มลดลงถ้าหากมีการเพิ่มจำนวน epoch ดังนั้นจึงสามารถสรุปได้ว่า ถ้าหากผู้วิจัยได้ทำการ

เพิ่มจำนวน epoch มีความเป็นไปได้ว่า โมเดลทั้ง 2 ตัวนี้สามารถให้ค่าความแม่นยำที่สูงขึ้นและค่าความสูญเสียลดลง

เมื่อทำการเปรียบเทียบผลกับ CNN แล้ว สามารถวิเคราะห์ได้ว่า LSTM เป็นอัลกอลิทึมที่ถูกสร้างมาเพื่อจัดการกับข้อมูลอนุกรมเวลา (Time-series data) โดยเฉพาะ ลักษณะของข้อมูลอนุกรมเวลา คือ เป็นข้อมูลที่เก็บรวบรวมตามระยะเวลาเป็นช่วงอย่างต่อเนื่องกัน เมื่อข้อมูลที่นำมาวิเคราะห์ในงานวิจัยนี้เป็นข้อมูลคำพูด ทำให้ขนาดของข้อมูลแต่ละชุดมีความยาวไม่เท่ากัน ซึ่งจะแตกต่างจากข้อมูลประเภทรูปภาพที่ CNN สามารถทำงานด้วยดีที่สุด [10] เนื่องจากในอัลกอริทึม CNN มีกระบวนการเตรียมข้อมูลในระดับชั้น Convolution layer เพื่อสกัดปัจจัยสำคัญที่ใช้ในการจำแนกประเภทของรูปภาพ ดังนั้นเพื่อให้มีประสิทธิภาพในการทำงาน โดยทั่วไปแล้วผู้วิจัยจะทำการปรับขนาดของรูปภาพให้เท่ากันทั้งหมดก่อนนำเข้าสู่กระบวนการ ซึ่งขนาดของรูปภาพไม่ได้มีผลต่อปัจจัยสำคัญในการจำแนกรูปภาพ แต่ข้อมูลคำพูดที่ใช้ในการจำแนกอารมณ์ของผู้พูดไม่สามารถถูกตัดทอนออกไปได้ เนื่องจากอาจทำให้เสียปัจจัยสำคัญในการจำแนกอารมณ์ออกไป

ในการทดลองที่ 2 ผลปรากฏว่าโครงสร้างของ LSTM แบบที่ 1 ที่มีระดับชั้นอยู่ 4 ระดับชั้น มีค่าการสูญเสียของ Model ต่ำที่สุดและมีค่าความแม่นยำสูงที่สุด จากการวิเคราะห์ผลการทดลอง โครงสร้างของ LSTM แบบที่ 2 และ 3 มีค่าการสูญเสียและค่าความแม่นยำใกล้เคียงกัน แต่เมื่อมีระดับชั้นต่างกันก็มีผลต่อการปรับค่าน้ำหนักของ Neuron ที่อยู่ในแต่ละระดับชั้น ซึ่งถ้ามีระดับชั้นมากเกินไปอาจทำให้เกิดปัญหา ในระหว่างการฝึกสอน เมื่อ Gradients มีขนาดเล็กลงจนเท่ากับศูนย์ ทำให้ค่าน้ำหนักไม่ถูกอัพเดทอีกต่อไป ซึ่งเป็นสาเหตุทำให้โมเดลไม่สามารถฝึกสอนต่อไปได้ โดยปัญหาดังกล่าวเรียกว่า Vanishing Gradients และถ้าระดับชั้นน้อยเกินไปก็อาจส่งผลให้โมเดลมีการเรียนรู้ไม่เพียงพอต่อการวิเคราะห์ข้อมูลที่มีความซับซ้อนสูง

เมื่อได้ทำการเปรียบเทียบการงานวิจัยในอดีตที่มีการดำเนินการคล้ายกัน Etienne, et al. [11] ได้ทำการออกแบบ neural network สำหรับการจดจำอารมณ์จากเสียงพูด แหล่งข้อมูลคือ IEMOCAP ซึ่งได้ผลปรากฏว่า มีความแม่นยำอยู่ที่ 64% เมื่อเทียบกับงานวิจัยนี้ ความแม่นยำที่ดีที่สุดที่ได้อยู่ที่ 67% มันแสดงให้เห็นว่าแบบจำลองที่ถูกพัฒนาขึ้นในงานวิจัยนี้ให้ผลลัพธ์ที่ดีขึ้น อย่างไรก็ตามเมื่อนำมาเปรียบเทียบกับ [12] ได้ทำการพัฒนาอัลกอริทึมใหม่ชื่อว่า 2D CNN LSTM และให้ความแม่นยำสูงถึง 95% เมื่อทำการวิเคราะห์การทดลองและผลลัพธ์ที่ได้ พบว่า อัลกอริทึม LSTM พื้นฐานอาจไม่เหมาะสมกับข้อมูลคำพูดของมนุษย์ เนื่องจากมีความยาวของข้อมูลไม่สม่ำเสมออีกครั้งเกิดโอกาสในการสูญเสียข้อมูลระหว่างการฝึกสอนอีกด้วย ดังนั้นเพื่อแก้ปัญหาดังกล่าวการใช้อัลกอริทึมอื่นเช่น BI-LSTM อาจเป็นทางเลือกที่ดีกว่า เนื่องจากมีการประมวลผลข้อมูล 2 ทิศทางต่างจาก LSTM พื้นฐานที่จะทำการวิเคราะห์และประมวลผลไปในทิศทางเดียว ยิ่งไปกว่านั้น เมื่อมีการกำหนดค่า epoch ที่สูงเกินไปอาจทำให้เกิดปัญหา Overfitting ได้

5. สรุปผลการวิจัย

วัตถุประสงค์งานวิจัยคือ เพื่อเปรียบเทียบประสิทธิภาพระหว่างโมเดล CNN และ LSTM และเพื่อหาโครงสร้างโมเดลที่เหมาะสมในการจำแนกอารมณ์จากเสียง ผู้วิจัยจึงได้ออกแบบการทดลองเพื่อตอบวัตถุประสงค์งานวิจัยออกเป็น 2 การทดลอง โดยในการทดลองที่ 1 ได้ทำการเปรียบเทียบการทำงานของ 2 อัลกอริทึม คือ CNN และ LSTM กับข้อมูลเสียงพูดที่ผู้วิจัยได้รวมรวบมาเองและข้อมูลจาก AIResearch.in.th มีจำนวนข้อมูลทั้งหมด 27,854 ข้อมูล แบ่งเป็น 5 อารมณ์ ประกอบด้วยโกรธ ปกติ ประหลาดใจ มีความสุข และเศร้า จากผลการทดลองสรุปได้ว่า LSTM เป็นอัลกอริทึมที่เหมาะสมกับการจำแนกอารมณ์จากเสียงพูดมากกว่า CNN เนื่องจากค่าสูญเสีย

ของ LSTM มีค่า 0.16 ขณะที่ CNN มีค่าความสูญเสียอยู่ที่ 1.117 และในผลการทดลองที่ 2 คือ ควรใช้ 4 ระดับชั้น LSTM ในการฝึกสอนเนื่องจากมีค่าความแม่นยำอยู่ที่ 67.38% สูงกว่าผลลัพธ์จากการทดลองอื่น

เอกสารอ้างอิง

[1] Hammond, M. (2006). Evolutionary theory and emotions. In Stets E.J. and Turnur J.H. Handbook of the Sociology of Emotions, New York: Springer, 368–385.

[2] Tokuno, S., Tsumatori, G., Shono, S., Takei, E., Yamamoto, T., Suzuki, G., Mituyoshi, S. and Shimura M. (2001). Usage of emotion recognition in military health care. In 2011 Defense Science Research Conference and Expo (DSR), 1–5.

[3] Yamashita, Y., Onodera, M., Shimoda, K. and Tobe, Y. (2019). Visualizing health with emotion polarity history using voice. In Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, 1210–1213.

[4] Kittichaiwatthana, P., Praneetpholkrang, P., and Kanjanawattana, S. (2020). Facial Expression Recognition using Deep Learning. SUT International Virtual Conference on Science and Technology, 41.

[5] Song, I., Kim, HJ. and Jeon, P. (2014). Deep learning for real-time robust facial expression recognition on a smartphone. In 2014 IEEE International Conference on Consumer Electronics (ICCE), 564–567.

[6] Dagar, D., Hudait, A., Tripathy, HK. and Das, MN. (2016). Automatic emotion detection model from facial expression. In 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 77–85.

[7] Lugović, S., Dunder, I. and Horvat, M. (2016). Techniques and applications of emotion recognition in speech. In 2016 39th international convention on information and communication technology, electronics and microelectronics (mipro), 278–1283.

[8] Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C. and Schuller, B. (2019). Speech emotion classification using attention-based LSTM. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(11), 1675–1685.

[9] Shewalkar, AN. (2018). Comparison of rnn, lstm and gru on speech recognition data. In Partial Fulfillment of the Requirements for the Degree of Master of Science. North Dakota State University of Agriculture and Applied Science.

[10] Rawat, W. & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. Neural computation, 29(9), 2352–2449.

[11] Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L. and Schmauch, B. (2018). Cnn+ lstm architecture for speech emotion recognition with data augmentation. In Proceeding Workshop on Speech, Music and Mind (SMM 2018), 21-25.

[12] Zhao, J., Mao, X. and Chen, L. (2019) Speech emotion recognition using deep 1D \& 2D CNN LSTM networks. Biomedical signal processing and control, 47, 312–323.