การรู้จำอารมณ์จากเสียงพูดภาษาไทยโดยใช้โครงข่ายประสาทเทียม
Thai Speech Emotion Recognition Using Artificial Neural Networks

วัชระ โสธิฤทธิ์ (Watchara Sothirit)*, วรัญญา ปุณณวัฒน์ (Waranya Poonnawat)*,
และณัฐพร เห็นเจริญเลิศ (Nuttaporn Hencharoenlert)*

Received: July 24, 2023
Revised: October 25, 2023
Accepted: November 1, 2023

* ผู้นิพนธ์ประสานงาน: วัชระ โสธิฤทธิ์ (Watchara Sothirit) อีเมล: sothirit.w@gmail.com

DOI:10.14416/j.it.2024.v1.009

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์ (1) เพื่อพัฒนาและประเมินโมเดลการรู้จำอารมณ์จากเสียงพูดภาษาไทยโดยใช้โครงข่ายประสาทเทียม (2) เพื่อให้สามารถจำแนกอารมณ์ของมนุษย์ได้อย่างถูกต้อง และ (3) เพื่อลดช่องว่างในการสื่อสารระหว่างคอมพิวเตอร์และผู้ใช้ โดยใช้ชุดข้อมูลจาก AIResearch.in.th ประกอบด้วยประโยคภาษาไทยจำนวน 27,854 ประโยค แบ่งเป็น อารมณ์ โกรธ เศร้า สุข หงุดหงิด และปกติ ใช้ค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล (MFCC) สำหรับการแยกคุณลักษณะเสียงพูด เตรียมข้อมูลก่อนการประมวลผลโดยใช้เทคนิคการเสริมข้อมูล การยืดเวลาของเสียง การเปลี่ยนระดับเสียง และการฉีดเสียงรบกวน นำข้อมูลที่ได้ไปฝึกในโมเดลในกลุ่มโครงข่ายประสาทเทียม ได้แก่ โครงข่ายประสาทเทียมแบบคอนโวลูชัน 1 มิติ (1D CNN) หน่วยความจำระยะยาว-ระยะสั้น หรือแอลเอสทีเอ็ม (LSTM) และแบบผสมผสาน (1D CNN และ LSTM) ผลลัพธ์แสดงให้เห็นว่าโมเดลแบบผสมผสานมีค่าความถูกต้องสูงสุดที่ 80.36% ตามมาด้วยโมเดล 1D CNN 77.52% และโมเดล LSTM 67.86%

คำสำคัญ: การรู้จำอารมณ์ เสียงพูดภาษาไทย โครงข่ายประสาทเทียม สัมประสิทธิ์เซปสตรัมบนสเกลเมล โมเดลแบบผสม (1D CNN และ LSTM)

Abstract

The research aimed (1) to develop and evaluate an emotion recognition model for Thai speech using artificial neural networks, (2) to enable accurate classification of human emotions, and (3) to bridge communication gaps between computers and users. A dataset from AIResearch.in.th consisting of 27,854 Thai-language sentences categorized into angry, sad, happy, frustrated, and neutral emotions. The Mel Frequency Cepstral Coefficients (MFCC) employed for the speech feature extraction. Data was pre-processed by augmentation techniques, including time stretching, pitch shifting, and noise injection. The pre-processed data trained for artificial neural network models, including a 1-dimensional Convolutional Neural Network (1D CNN), Long Short-Term Memory (LSTM), and a hybrid model (1D CNN and LSTM). Results showed that the hybrid model (1D CNN & LSTM) achieved the highest accuracy of 80.36%, followed by the 1D CNN model (77.52%) and the LSTM model (67.86%).

Keywords: Emotion Recognition, Thai Speech, Artificial Neural Networks, Mel Frequency Cepstral Coefficients (MFCC), Hybrid Model (1D CNN & LSTM).

1. บทนำ

การรู้จำอารมณ์จากเสียงพูดเป็นประเด็นที่มีความสำคัญสำหรับการวิจัยด้านการประมวลผลสัญญาณเสียง โดยสามารถทำให้รับรู้ถึงความรู้สึกและเข้าใจถึงสภาพอารมณ์ของผู้คนได้ อีกทั้งยังมีความสำคัญในการพัฒนาระบบสนสารเทศ อาทิเช่น ระบบคอลเซ็นเตอร์ที่สามารถประเมินอารมณ์ลูกค้า

* แขนงวิชาเทคโนโลยีสารสนเทศและการสื่อสาร สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช
* Information and Communication Technology Program, School of Science and Technology, Sukhothai Thammathirat Open University.

ที่โทรศัพท์เข้ามารับบริการว่ามีอารมณ์เป็นเช่นไร เพื่อบันทึกความรู้สึกจากเสียงพูดตลอดการพูดคุยไว้เป็นสถิติความพึงพอใจในการให้บริการลูกค้า หรือระบบปัญญาประดิษฐ์ที่สามารถแสดงอารมณ์ขณะการสื่อสารกับผู้ใช้ได้อย่างเป็นธรรมชาติมากขึ้น แทนที่จะพูดด้วยน้ำเสียงเรียบ ๆ แบบโมโนโทนแบบที่เราคุ้นเคย [1]

อย่างไรก็ตาม ในรอบหลายปีที่ผ่านมานั้น ยังมีงานวิจัยที่เกี่ยวกับการรู้จำอารมณ์จากเสียงพูดภาษาไทยน้อยมาก และมีความแม่นยำค่อนข้างน้อย นอกจากนั้น คำ ๆ เดียวในภาษาไทย หากใช้ระดับเสียงที่ต่างกัน ความหมายที่ได้ก็แตกต่างออกไป [2] ซึ่งถือว่ามีความท้าทายและน่าสนใจอย่างมากสำหรับการรู้จำอารมณ์ ดังนั้น ผู้วิจัยจึงสนใจสร้างโมเดลการรู้จำอารมณ์จากเสียงพูดภาษาไทยด้วยโครงข่ายประสาทเทียม โดยการเปรียบเทียบ 3 อัลกอริทึม ได้แก่ 1D CNN, LSTM และแบบผสมผสาน (1D CNN&LSTM) เพื่อปรับปรุงประสิทธิภาพของโมเดลการรู้จำอารมณ์จากเสียงพูดภาษาไทย

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการทำวิจัยเรื่องการรู้จำอารมณ์จากเสียงพูดภาษาไทยโดยใช้โครงข่ายประสาทเทียมครั้งนี้ได้ศึกษาและค้นคว้าทฤษฎีและงานวิจัยที่เกี่ยวข้องดังนี้

2.1 การรู้จำอารมณ์จากเสียงพูด

การรู้จำอารมณ์จากเสียงพูด (Speech Emotion Recognition) เป็นกระบวนการให้ระบบคอมพิวเตอร์สามารถระบุและรู้ถึงอารมณ์หรือความรู้สึกที่แสดงผ่านเสียงพูดได้ โดยวิเคราะห์ลักษณะเสียงต่าง ๆ ในเสียงพูด เช่น ความสูงและความต่ำของเสียง (pitch), ความเร็วในการพูด (speech rate), ความเปลี่ยนแปลงความถี่ (frequency variation), และคุณลักษณะเสียงอื่น ๆ เพื่อระบุอารมณ์หรือความรู้สึกที่เกี่ยวข้องกับเสียงพูด การรู้จำอารมณ์จากเสียงพูดมีความสำคัญในหลายแนวทางการประยุกต์ใช้ เช่น ในงานวิจัยทางการแพทย์นำไปใช้ในการระบุอารมณ์อ่อนไหวของผู้ป่วยในการวินิจฉัยโรคหรือติดตามความก้าวหน้าของการรักษา เป็นต้น [3]

2.2 การสกัดคุณลักษณะ

การสกัดคุณลักษณะสามารถใช้ไลบรารี Librosa ของภาษา Python สกัดคุณลักษณะ Mel Frequency Cepstral Coefficient (MFCC) ของแต่ละเสียงออกมาได้ [4] MFCC เป็นค่าพลังงานสเปกตรัมของสัญญาณเสียงในช่วงสั้น ๆ ถูกนำมาใช้ในงานรู้จำเสียงพูดมนุษย์อย่างกว้างขวาง มีกระบวนการทำงานสอดรับกับระบบการได้ยินของมนุษย์ รองรับสัญญาณเสียงตั้งแต่ 1 กิโลเฮิร์ตซ์ขึ้นไป ซึ่งเป็นช่วงความถี่ที่มนุษย์จะสามารถได้ยิน ทำให้ MFCC นั้นเหมาะที่จะนำมาใช้ในงานรู้จำเสียงพูด [2]

2.3 การเสริมข้อมูล

การเสริมข้อมูล (Data Augmentation) คือกระบวนการเพิ่มข้อมูลในชุดข้อมูลเดิมโดยการใช้เทคนิคต่าง ๆ เพื่อเพิ่มความหลากหลายและความเป็นตัวอย่างให้กับโมเดล และยังช่วยลดความเสี่ยงของโมเดลที่จะเกิดการเรียนรู้ที่เกินจำเป็น (overfitting) ได้ด้วย การเสริมข้อมูลสัญญาณเสียง MFCC สามารถใช้เทคนิคต่าง ๆ ได้ เช่น การยืดเวลา การเปลี่ยนระดับเสียง และการฉีดเสียงรบกวน [5]

2.4 Convolutional Neural Network (CNN)

โมเดล CNN มีความสามารถในการแยกแยะลักษณะเด่นของข้อมูล เช่น เส้นโค้ง รูปร่าง หรือลักษณะอื่น ๆ โดยมีการใช้ตัวกรองเคอร์เนล (kernel) ที่สามารถตรวจจับลักษณะเหล่านี้ได้ โมเดล CNN จะคืนค่าข้อมูลนำออกที่เป็นแผนที่ฟีเจอร์ (feature map) ที่ถูกสกัดมาจากข้อมูลนำเข้า [6]

2.5 Long Short-Term Memory (LSTM)

โมเดล LSTM สามารถจำได้ว่าข้อมูลที่ผ่านมามีลักษณะอย่างไร และสามารถเรียนรู้จากข้อมูลเหล่านี้เพื่อทำนายต่อไปได้อย่างมีประสิทธิภาพ โดยโมเดล LSTM จะเป็นโมเดลที่เหมาะสำหรับการประมวลผลข้อมูลที่มีลำดับการเกิดเหตุการณ์ เช่น ข้อมูลสัญญาณเสียงพูด เนื่องจากสัญญาณเสียงพูดเป็นสัญญาณอนุกรมเวลา โมเดล LSTM จะสามารถรับรู้ลักษณะและลำดับของสัญญาณเสียงพูดได้ [6]

2.6 1D CNN&LSTM Hybrid Algorithms

1D CNN-LSTM Hybrid Algorithms คือการใช้โมเดล CNN และ LSTM ร่วมกันในการประมวลผลข้อมูลที่มีลักษณะเป็นชุดข้อมูลชุดเดียว (one-dimensional data) เช่น สัญญาณเสียง ซึ่งมีความยาวแต่ไม่มีความกว้างหรือความลึกเช่นเดียวกับข้อมูลภาพ [6]

2.7 การประเมินประสิทธิภาพ

2.7.1 การตรวจสอบโมเดลด้วยการแบ่งข้อมูลเป็นส่วน ๆ (K fold Cross-validation) เป็นวิธีที่นิยมในการทำงานวิจัยในปัจจุบัน เพื่อใช้ในการทดสอบประสิทธิภาพของโมเดล เนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัดประสิทธิภาพด้วยวิธี Cross-validation จะทำการแบ่งข้อมูลออกเป็นหลายส่วนซึ่งมักจะแสดงด้วยค่า k

เช่น 5 k-fold cross-validation คือ ทำการแบ่งข้อมูลออกเป็น 5 ส่วน (k) โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน [7]

2.7.2 การหยุดฝึกล่วงหน้า (Early Stopping) เป็นเทคนิคที่นิยมใช้ในการฝึกโมเดลการเรียนรู้เชิงลึก (deep learning) เพื่อป้องกันการเรียนรู้เกินจำเป็น (overfitting) ซึ่งเป็นลักษณะที่โมเดลมีประสิทธิภาพสูงในการจัดกลุ่มข้อมูลสำหรับการฝึก แต่ไม่มีประสิทธิภาพในการจัดกลุ่มข้อมูลทดสอบ ดังนั้น การใช้ Early Stopping จะช่วยลดการเรียนรู้เกินจำเป็น และช่วยให้โมเดลมีประสิทธิภาพในการจัดกลุ่มข้อมูลทดสอบได้อย่างมีประสิทธิภาพมากขึ้น [8]

2.7.3 เมตริกซ์ความสับสน (Confusion Matrix) เป็นเครื่องมือในการประเมินผลลัพธ์ของการทำนาย (Prediction) ที่ทำนายผลจากโมเดลที่สร้างขึ้น โดยมีแนวคิดจากการวัดว่าสิ่งที่คิด (การทำนายโมเดล) กับ สิ่งที่เกิดขึ้นจริง มีสัดส่วนเป็นอย่างไร [9] ดังภาพที่ 1

Confusion Matrix	Predicted Positive	Predicted Negative	Sum
Actual Positive	TP = 40	FN = 10	50
Actual Negative	FP = 20	TN = 60	80
Sum	60	70	

ภาพที่ 1 *ตัวอย่างค่าของเมตริกซ์ความสับสน*

จากภาพที่ 1 สามารถอธิบายได้ดังนี้

TP คือจำนวนข้อมูลที่ทำนายถูกต้อง ตรงกับสิ่งที่เกิดขึ้น ในกรณีทำนายว่าจริง และสิ่งที่เกิดขึ้น ก็คือ จริง

TN คือจำนวนข้อมูลที่ทำนายถูกต้อง ตรงกับสิ่งที่เกิดขึ้น ในกรณีทำนายว่าไม่จริง และสิ่งที่เกิดขึ้น ก็คือ ไม่จริง

FP คือจำนวนข้อมูลที่ทำนายผิด ไม่ตรงกับสิ่งที่เกิดขึ้น คือทำนายว่า จริง แต่สิ่งที่เกิดขึ้น คือ ไม่จริง

FN คือจำนวนข้อมูลที่ทำนายผิด ไม่ตรงกับสิ่งที่เกิดขึ้น คือทำนายว่าไม่จริง แต่สิ่งที่เกิดขึ้น คือ จริง

จากเมตริกซ์ความสับสน (Confusion Matrix) สามารถนำมาคำนวณหาค่าประสิทธิภาพของโมเดลในรูปแบบค่าต่าง ๆ ได้ดังนี้

1) ค่าความถูกต้อง (Accuracy) คือผลรวมของตัวเลขบนเส้นทแยงมุมในตาราง Confusion Matrix หารด้วยจำนวนทั้งหมด โดยใช้สมการดังนี้

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

เช่น $$Accuracy = \frac{40+60}{40+20+10+60}$$

2) ค่าความสามารถในการตรวจจับคลาส (Recall) คือความถูกต้องของการทำนายว่าจะเป็น "จริง" เทียบกับจำนวนครั้งของเหตุการณ์ ทั้งทำนายและ เกิดขึ้น ว่า "เป็นจริง" โดยใช้สมการดังนี้

$$Recall = \frac{TP}{TP+FN}$$

เช่น $$Recall = \frac{40}{40+10}$$

3) ค่าความแม่นยำ (Precision) คือการเปรียบเทียบการทำนายที่ถูกต้องว่า จริง และก็เกิดขึ้นจริง (TP) กับการทำนายว่า จริง แต่สิ่งที่เกิดขึ้น คือ ไม่จริง (FP) โดยใช้สมการดังนี้

$$Precision = \frac{TP}{TP+FP}$$

เช่น $$Precision = \frac{40}{40+20}$$

4) ค่า F1 (F1-Score) คือค่าที่รวมความสามารถของ Precision และ Recall เข้าด้วยกัน เพื่อให้ได้ค่าเฉลี่ยที่ดีทั้งในการทำนาย Positive และการตรวจจับ Positive โดยใช้สมการดังนี้

$$F1\text{-}Score = \frac{2*(Precision*Recall)}{Precision+Recall}$$

2.8 งานวิจัยที่เกี่ยวข้อง

การศึกษาเกี่ยวกับงานวิจัยด้านการรู้จำอารมณ์จากเสียงพูด ได้ศึกษางานวิจัยดังต่อไปนี้

ศรัญญา กาญจนวัฒนา และคณะ [10] นำเสนอการจำแนกอารมณ์จากการรู้จำเสียงพูดด้วยการเรียนรู้แบบเชิงลึก โดยเปรียบเทียบอัลกอริทึมระหว่าง CNN และ LSTM โดยใช้ชุดข้อมูลที่สร้างขึ้นเอง 500 ชุดประกอบด้วย 5 อารมณ์ และใช้ข้อมูลของ AIResearch.in.th 27,854 ชุด นำข้อมูลมาสกัดคุณลักษณะ (MFCC) แยกประเภทอารมณ์ด้วยอัลกอริทึม CNN และ LSTM และประเมินโมเดลโดยใช้ค่าความสูญเสีย เพื่อหาความเหมาะสมของโมเดล ผลการวิจัยพบว่าอัลกอริทึม LSTM มีความเหมาะสมกับการจำแนกอารมณ์จากเสียงพูดมากกว่า CNN เนื่องจาก LSTM มีค่าความสูญเสีย

น้อยกว่าและมีความแม่นยำสูงกว่า CNN อยู่ที่ 67.38%

ศิรสิทธิ์ กิจทวีสินพูน และเอกรัฐ รัฐกาญจน์ [4] นำเสนอการวิเคราะห์อารมณ์จากเสียงพูดภาษาไทย โดยนำชุดข้อมูลเสียงพูดภาษาไทยประกอบด้วย 5 อารมณ์ มาสกัดเอาคุณลักษณะออกมาเป็นเวกเตอร์ขนาด 193 คุณลักษณะ และทำการปรับเวกเตอร์ที่ได้ให้อยู่ในรูปแบบที่มีความยาวเท่ากัน (Vector Normalization) จากนั้นนำไปสร้างโมเดลด้วยอัลกอริทึม SVM, Random Forest, Neural Network และ 1D CNN วัดประสิทธิภาพของโมเดลโดยแบ่งข้อมูลเป็นสำหรับการฝึก 80% และสำหรับการทดสอบ 20% ผลการทดสอบค่าความแม่นยำพบว่า Neural Network มากที่สุด 70% และรองลงมาได้แก่ 1D CNN 66%, SVM 61% และ Random Forest 58%

Pratama และ Sihwi [11] ใช้อัลกอริทึม SVM ในการจำแนกอารมณ์จากเสียงพูด สกัดคุณลักษณะด้วย MFCC และใช้เคอร์เนลฟังก์ชันในการจำแนกอารมณ์ความเหมาะสมของเสียงโดยใช้ข้อมูลจาก 2 แหล่ง คือ RAVDESS และ TESS นำมารวมกันและเลือกเฉพาะอารมณ์ที่ตรงกัน 6 อารมณ์ในการสร้างโมเดล การทดสอบแบ่งออกเป็น 3 สถานการณ์ ได้แก่ 1) ใช้ชุดข้อมูล RAVDESS 2) ใช้ชุดข้อมูล TESS และ 3) ใช้ชุดข้อมูลผสมผสาน และในทุกสถานการณ์จะใช้เคอร์เนลฟังก์ชัน (ประกอบด้วย Linear, Polynomial และ Radial basis) ในการทดสอบและกำหนดพารามิเตอร์เคอร์เนล 3 ค่า คือ 10, 100 และ 1000 ในการทดสอบเพื่อหาค่าความถูกต้องที่เหมาะสมมากที่สุด ผลการทดสอบพบว่า เคอร์เนลฟังก์ชัน แบบ radial basis ให้ค่าความถูกต้องสูงสุดคือ 70% ในการทดสอบแต่ละครั้ง

จากการทบทวนงานวิจัยพบว่า งานวิจัยที่เป็นการรู้จำอารมณ์ที่เป็นภาษาอื่นหรือภาษาไทยยังมีจำนวนน้อย และมีเปอร์เซ็นต์ความถูกต้องที่ค่อนข้างน้อย เช่น งานวิจัยของศรัญญา กาญจนวัฒนา และคณะ [10] ที่มีค่าความถูกต้องอยู่ที่ 67.38% ดังนั้น ผู้วิจัยจึงสนใจประเด็นการพัฒนาโมเดลการรู้จำอารมณ์จากเสียงพูดภาษาไทย เพื่อเพิ่มความถูกต้องให้สูงขึ้น โดยดำเนินการดังนี้ การสกัดคุณลักษณะเสียงใช้ MFCC การสร้างโมเดลการรู้จำอารมณ์จากเสียงพูดภาษาไทยใช้ 3 อัลกอริทึม ได้แก่ 1D CNN, LSTM และแบบผสมผสาน (1D CNN&LSTM) เพื่อเปรียบเทียบกัน และการประเมินประสิทธิภาพใช้ k-fold cross validation เพื่อหาโมเดลการรู้จำอารมณ์จากเสียงพูดภาษาไทยที่มีค่าความถูกต้องที่สูงขึ้น

3. วิธีดำเนินการวิจัย

3.1 ประชากรและกลุ่มตัวอย่าง

ชุดข้อมูล Thai-SER dataset จาก AIResearch.in.th [12] ซึ่งเป็นชุดข้อมูลอารมณ์จากเสียงพูดภาษาไทยประกอบด้วยเสียงพูด จำนวน 27,854 ประโยค มีทั้งหมด 5 อารมณ์ ได้แก่ โกรธ เศร้า สุข หงุดหงิด และปกติ จากนักแสดง 200 คน (ชาย 87 คน และหญิง 113 คน) รวมระยะเวลา 41 ชั่วโมง 36 นาที

3.2 เครื่องมือที่ใช้ในการวิจัย

ในการวิจัยครั้งได้ใช้เครื่องมือในการวิจัย ได้แก่ เครื่องคอมพิวเตอร์พกพา (MacOS, i7, Ram 16 GB), โปรแกรม PyCharm CE, โปรแกรม MongoDB Compass และภาษา Python

3.3 การเตรียมข้อมูล

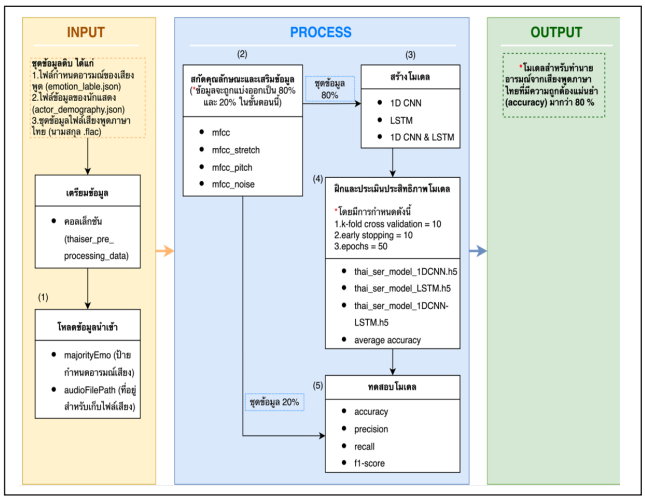
ในการเตรียมข้อมูลสำหรับการวิเคราะห์ข้อมูลมีขั้นตอนแบ่งออกเป็น 2 ขั้นตอนดังนี้

3.3.1 การดาวน์โหลดชุดข้อมูลเสียง ดาวน์โหลดข้อมูลจาก AIResearch.in.th โดยข้อมูลประกอบด้วยไฟล์ emotion_label.json เป็นไฟล์ข้อมูลที่กำหนดอารมณ์ของเสียง (Emotion Label) actor_demography.json เป็นไฟล์ข้อมูลของนักแสดง และไฟล์เสียงพูดจะในรูปแบบ zip ไฟล์

3.3.2 การแปลงข้อมูล ในขั้นตอนนี้จะเป็นการรวมข้อมูลจาก 3 ส่วนให้กลายเป็นไฟล์ json ไฟล์เดียว อยู่ในรูปแบบที่พร้อมใช้งาน และบันทึกข้อมูลที่พร้อมใช้งานลงในฐานข้อมูล MongoDB

3.4 การสร้างโมเดลสำหรับวิเคราะห์ข้อมูล

ขั้นตอนการสร้างโมเดลสำหรับการวิเคราะห์ข้อมูลแบ่งออกเป็น 5 ขั้นตอน แสดงดังภาพที่ 2



ภาพที่ 2 ขั้นตอนการสร้างโมเดล

จากภาพที่ 2 ขั้นตอนการสร้างโมเดลสามารถอธิบายแต่ละขั้นตอนได้ดังนี้

3.4.1 โหลดชุดข้อมูลนำเข้า ทำการโหลดชุดข้อมูลนำเข้าจากฐานข้อมูล MongoDB ที่เก็บไว้ในคอลเล็กชันที่ชื่อว่า thaiser_pre_processing_data ตั้งแต่ในขั้นตอนการเตรียมข้อมูลเพื่อที่จะนำไปใช้ในขั้นตอนการสกัดคุณลักษณะ

3.4.2 การสกัดคุณลักษณะและการเสริมข้อมูล

1) การสกัดคุณลักษณะ ทำการสกัดคุณลักษณะของเสียงพูดโดยใช้ชุดข้อมูลเสียงที่ได้จากขั้นตอนโหลดชุดข้อมูลนำเข้า นำมาสกัดคุณลักษณะ MFCC ของเสียงต้นฉบับ และนำเสียงต้นฉบับมาปรับแต่งเพิ่มเติมโดยใช้เทคนิคการยืดเวลาของเสียง (Time Stretching) การเปลี่ยนระดับเสียง (Pitch Shifting) และการฉีดเสียงรบกวน (Noise Injection) หลังจากนั้นนำเสียงที่ปรับแต่งแล้วมาสกัดคุณลักษณะ MFCC อีกครั้งเพื่อที่จะนำคุณลักษณะเหล่านี้ไปใช้ในขั้นตอนการเสริมข้อมูล

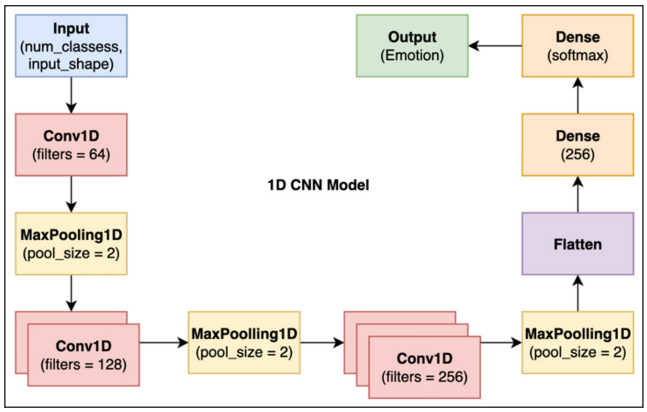
2) การเสริมข้อมูล ทำการเสริมข้อมูลด้วยการวนลูปเพิ่มข้อมูลจากข้อมูลเสียงที่ได้ปรับแต่งแล้ว และวนลูปเพิ่มข้อมูลในกลุ่มอารมณ์ที่มีข้อมูลจำนวนน้อย ๆ เพื่อให้ข้อมูลในแต่ละอารมณ์มีความสมดุลกันมากขึ้น จะได้ข้อมูลที่ใช้ในการวิเคราะห์เพิ่มขึ้นเป็น 70,294 ชุดข้อมูล จากนั้นนำชุดข้อมูลทั้งหมดที่ได้มาแบ่งออกเป็น 80% สำหรับใช้ในการฝึกโมเดล และ 20% สำหรับใช้ในการทดสอบโมเดล เพื่อประเมินประสิทธิภาพของโมเดล

3.4.3 สร้างโมเดลด้วยอัลกอริทึมในกลุ่มโครงข่ายประสาทเทียม ได้แก่ 1D CNN, LSTM และแบบผสมผสาน (1D CNN&LSTM)

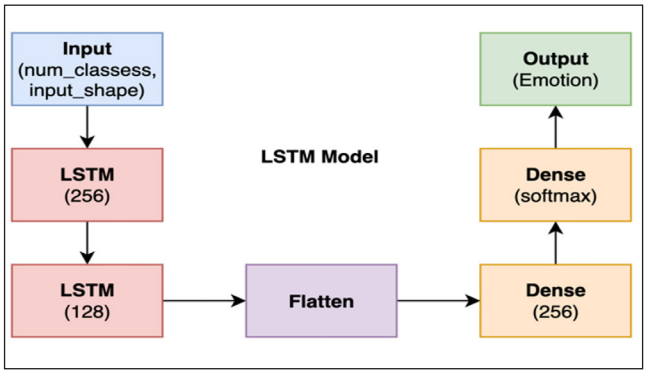
1) การสร้างโมเดล 1D CNN โดยสามารถแสดงโครงสร้างของโมเดลได้ดังภาพที่ 3

จากภาพที่ 3 เป็นการแสดงโครงสร้างของโมเดล 1D CNN เริ่มจากส่วนของ input รับพารามิเตอร์เป็น num_classess, input_shape (ข้อมูลที่จะใช้ในการฝึกโมเดล) จากนั้นเป็นการทำงานของชั้นคอนโวลูชันแบบหนึ่งมิติ (Conv1D) มีอยู่สามชั้น และกำหนด filters เท่ากับ 64, 128, 256 ตามลำดับ ในแต่ละชั้นของคอนโวลูชันจะต่อด้วยการทำพูลลิ่งเสมอ โดยใช้ Max pooling กำหนด pool_size เท่ากับ 2 แล้วต่อด้วยชั้น Flatten สำหรับแปลง Output จากชั้นก่อนหน้าให้กลายเป็นเวกเตอร์เดียวกัน และสุดท้ายต่อด้วยชั้นของ Dense อีกสองชั้นเพื่อเชื่อมต่อโหนดทุกโหนดในชั้นก่อนหน้าเข้าด้วยกัน โดย Dense ชั้นแรกกำหนดจำนวนเท่ากับ 256 โหนด และชั้นที่สอง softmax ใช้สำหรับจำแนกหมวดหมู่ของ Output หรือ Class

2) การสร้างโมเดล LSTM แสดงโครงสร้างโมเดลดังภาพที่ 4



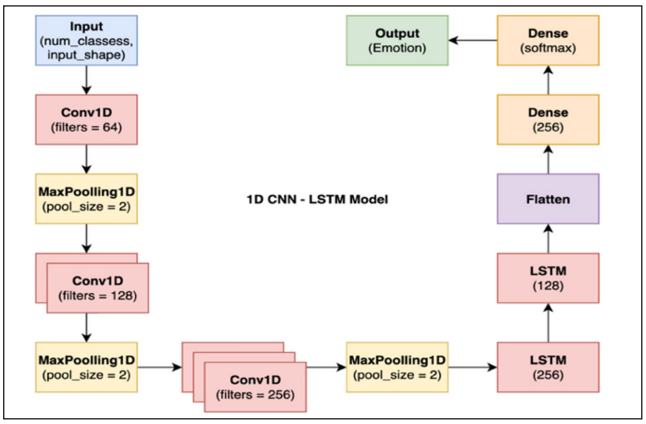
ภาพที่ 3 *โครงสร้างโมเดล 1D CNN*



ภาพที่ 4 *โครงสร้างโมเดล LSTM*

จากภาพที่ 4 เป็นการแสดงโครงสร้างของโมเดล LSTM เริ่มจากส่วนของ input รับพารามิเตอร์เป็น num_classess, input_shape (ข้อมูลที่จะใช้ในการฝึกโมเดล) จากนั้นเป็นการทำงานของชั้นหน่วยความจำระยะยาว-ระยะสั้นหรือแอลเอสทีเอ็ม (LSTM) มีอยู่สองชั้น ชั้นแรกกำหนดหน่วยความจำเซลล์ที่ 256 เซลล์ และชั้นที่สองกำหนดที่ 128 เซลล์ แล้วต่อด้วยชั้น Flatten สำหรับแปลง Output จากชั้นก่อนหน้าให้กลายเป็นเวกเตอร์เดียวกันและสุดท้ายต่อด้วยชั้นของ Dense อีกสองชั้นเพื่อเชื่อมต่อโหนดทุกโหนดในชั้นก่อนหน้าเข้าด้วยกัน โดย Dense ชั้นแรกกำหนดจำนวนเท่ากับ 256 โหนด และชั้นที่สอง softmax ใช้สำหรับจำแนกหมวดหมู่ของ Output หรือ Class

3) การสร้างโมเดลแบบผสมผสาน (1D CNN& LSTM) โดยสามารถแสดงโครงสร้างของโมเดลได้ดังภาพที่ 5



ภาพที่ 5 *โครงสร้างโมเดล 1D CNN&LSTM*

จากภาพที่ 5 เป็นการแสดงโครงสร้างของโมเดลแบบผสมผสาน (1D CNN & LSTM) เริ่มจากส่วนของ input จะรับพารามิเตอร์เป็น num_classess, input_shape (ข้อมูลที่จะใช้ในการฝึกโมเดล) จากนั้นเป็นการทำงานของชั้นคอนโวลูชันแบบหนึ่งมิติ (Conv1D) มีอยู่สามชั้น และกำหนด filters เท่ากับ 64, 128, 256 ตามลำดับ ในแต่ละชั้นของคอนโวลูชันจะต่อด้วยการทำพูลลิ่งเสมอ โดยใช้ Max pooling กำหนด pool_size เท่ากับ 2 ต่อด้วยการทำงานของ LSTM มีอยู่สองชั้น ชั้นแรกกำหนดหน่วยความจำเซลล์ที่ 256 เซลล์ และชั้นที่สองกำหนดที่ 128 เซลล์ แล้วจากนั้นต่อด้วยชั้น Flatten สำหรับแปลง Output จากชั้นก่อนหน้าให้กลายเป็นเวกเตอร์เดียวกัน และสุดท้ายต่อด้วยชั้นของ Dense อีกสองชั้นเพื่อเชื่อมต่อโหนดทุกโหนดในชั้นก่อนหน้าเข้าด้วยกัน โดย Dense ชั้นแรกจะกำหนดจำนวนเท่ากับ 256 โหนด และชั้นที่สอง softmax ใช้สำหรับจำแนกหมวดหมู่ของ Output หรือ Class

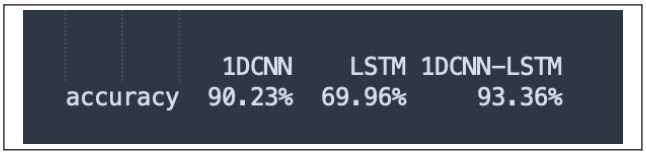
3.4.4 การฝึกโมเดลและประเมินประสิทธิภาพของโมเดล นำชุดข้อมูล 80% ที่ได้แบ่งไว้มาทำการฝึกโมเดล เพื่อเปรียบเทียบและประเมินประสิทธิภาพของโมเดลทั้ง 3 โมเดลที่สร้างขึ้น โดยโมเดลทั้งสามถูกฝึกด้วยขั้นตอนที่เหมือนกันคือ ใช้การตรวจสอบโมเดลด้วยการแบ่งข้อมูลเป็นส่วน ๆ (K Fold Cross-validation) โดยกำหนดที่ 10 K Fold และในแต่ละรอบของ Fold จะกำหนดจำนวนครั้งในการฝึกที่ 50 epochs โดยการกำหนดจำนวนของ epochs อาจจะฝึกไม่ถึง 50 ครั้ง เพราะจะมีการใช้การหยุดฝึกล่วงหน้า (Early Stopping) เพื่อป้องกันการเกิดการเรียนรู้เกินจำเป็น (Overfitting)

3.4.5 การทดสอบโมเดล ใช้ชุดข้อมูลใหม่ 20% ที่ได้แบ่งไว้แล้วนำมาทดสอบกับทั้ง 3 โมเดลที่ได้ผ่านการฝึกมาแล้วในขั้นตอนการฝึกโมเดลและประเมินประสิทธิภาพของโมเดล

4. ผลการดำเนินงาน

4.1 ผลการฝึกโมเดล

ในการฝึกโมเดลทั้ง 3 โมเดล ได้แก่ 1D CNN, LSTM และแบบผสมผสาน (1D CNN & LSTM) ฝึกในขั้นตอนที่เหมือนกัน และกำหนดพารามิเตอร์ที่เหมือนกัน ผลลัพธ์ค่าเฉลี่ยความถูกต้อง (accuracy) ในการฝึกของทั้ง 3 โมเดล ได้แก่ 1D CNN, LSTM และแบบผสมผสาน (1D CNN & LSTM) ค่าความถูกต้อง (accuracy) อยู่ที่ 90.23%, 69.96% และ 93.36% ตามลำดับ ดังภาพที่ 6



ภาพที่ 6 *ผลลัพธ์ค่าเฉลี่ยความถูกต้องในการฝึก*

4.2 ผลการทดสอบโมเดล

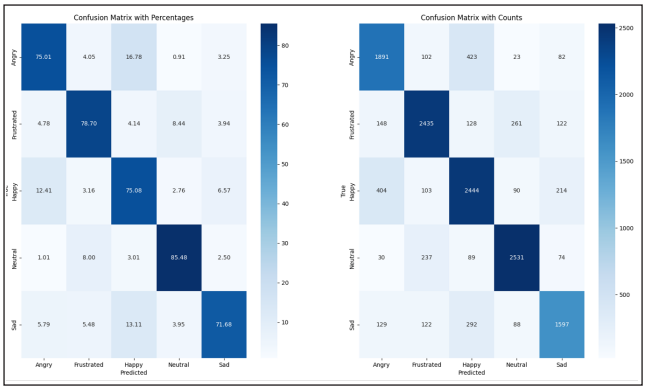
การทดสอบโมเดลของทั้ง 3 โมเดล จากข้อมูลชุดใหม่ 20% ผลการทดสอบของทั้ง 3 โมเดล มีดังนี้

4.2.1 ผลการทดสอบโมเดล 1D CNN มีค่าความถูกต้องอยู่ที่ 77.52% โดยสามารถสรุปเป็นรายงานการจำแนกประเภท (Classification report) ได้ดังตารางที่ 1

ตารางที่ 1 *รายงานการจำแนกประเภท สำหรับโมเดล 1D CNN*

	precision	recall	f1-score
Angry	72.67%	75.01%	73.82%
Frustrated	81.19%	78.70%	79.93%
Happy	72.39%	75.08%	73.71%
Neutral	84.56%	85.48%	85.02%
Sad	76.45%	71.68%	73.99%
accuracy			77.52%
macro avg	77.45%	77.19%	77.29%
weighted avg	77.59%	77.52%	77.53%

จากตารางที่ 1 สามารถแสดงผลอยู่ในรูปแบบเมตริกซ์ความสับสน ของโมเดล 1D CNN โดยใช้ค่า recall มาแสดงเป็นเปอร์เซ็นต์ และจำนวนความถูกต้องของแต่ละอารมณ์ได้ดังภาพที่ 7



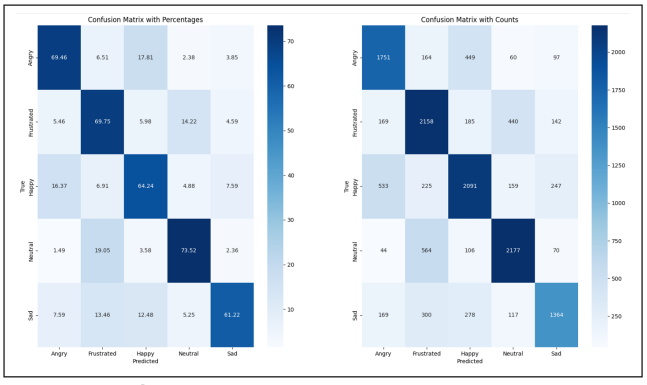
ภาพที่ 7 *เมตริกซ์ความสับสนของโมเดล 1D CNN*

4.2.2 ผลการทดสอบโมเดล LSTM มีค่าความถูกต้องอยู่ที่ 67.86% โดยสามารถสรุปเป็นรายงานการจำแนกประเภท (Classification report) ได้ดังตารางที่ 2

ตารางที่ 2 *รายงานการจำแนกประเภทสำหรับโมเดล LSTM*

	precision	recall	f1-score
Angry	65.68%	69.46%	67.51%
Frustrated	63.27%	69.75%	66.35%
Happy	67.26%	64.24%	65.71%
Neutral	73.72%	73.52%	73.62%
Sad	71.04%	61.22%	65.77%
accuracy			67.86%
macro avg	68.19%	67.64%	67.79%
weighted avg	68.06%	67.86%	67.85%

จากตารางที่ 2 สามารถแสดงผลอยู่ในรูปแบบเมตริกซ์ความสับสน ของโมเดล LSTM โดยใช้ค่า recall มาแสดงเป็นเปอร์เซ็นต์ และจำนวนความถูกต้องของแต่ละอารมณ์ได้ดังภาพที่ 8



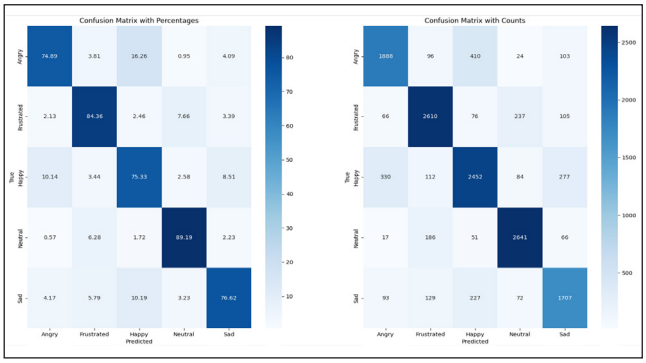
ภาพที่ 8 *เมตริกซ์ความสับสนของโมเดล LSTM*

4.2.3 ผลการทดสอบโมเดล 1D CNN&LSTM มีค่าความถูกต้องอยู่ที่ 80.36% โดยสามารถสรุปเป็นรายงานการจำแนกประเภท (Classification report) ได้ดังตารางที่ 3

ตารางที่ 3 *รายงานการจำแนกประเภท สำหรับโมเดล 1D CNN&LSTM*

	precision	recall	f1-score
Angry	78.86%	74.89%	76.83%
Frustrated	83.31%	84.36%	83.83%
Happy	76.24%	75.33%	75.78%
Neutral	86.36%	89.19%	87.76%
Sad	75.60%	76.62%	76.10%
accuracy			80.36%
macro avg	80.08%	80.08%	80.06%
weighted avg	80.30%	80.36%	80.31%

จากตารางที่ 3 สามารถแสดงผลอยู่ในรูปแบบเมตริกซ์ความสับสน ของโมเดล 1D CNN&LSTM โดยใช้ค่า recall มาแสดงเป็นเปอร์เซ็นต์ และจำนวนความถูกต้องของแต่ละอารมณ์ได้ดังภาพที่ 9



ภาพที่ 9 *เมตริกซ์ความสับสนของโมเดล 1D CNN&LSTM*

4.2.4 เปรียบเทียบผลการทดสอบโมเดลของทั้ง 3 โมเดล โดยสามารถสรุปเป็นตารางการเปรียบเทียบค่าความถูกต้อง (Accuracy) ได้ดังตารางที่ 4

ตารางที่ 4 *เปรียบเทียบค่าความถูกต้องของทั้ง 3 โมเดล*

	Accuracy
1D CNN	77.52%
LSTM	67.86%
แบบผสมผสาน (1D CNN & LSTM)	80.36%

5. สรุปและอภิปรายผล

ในการวิจัยครั้งนี้มีวัตถุประสงค์หลักคือสร้างโมเดลในการรู้จำอารมณ์จากเสียงพูดภาษาไทยและประเมินประสิทธิภาพของโมเดลที่สร้างขึ้นโดยใช้ชุดข้อมูลเสียงพูดจาก AIResearch.in.th ประกอบด้วยเสียงพูดภาษาไทยจำนวน 27,854 ประโยค

แบ่งเป็น 5 ประเภทอารมณ์ ได้แก่ โกรธ เศร้า สุข หงุดหงิด และปกติ โดยใช้การสกัดคุณลักษณะ MFCC ของเสียง จากนั้นเสริมข้อมูลด้วยวิธีการยืดเวลาของเสียง การเปลี่ยนระดับเสียง และการฉีดเสียงรบกวน ก่อนนำไปสร้างเป็นโมเดลด้วยอัลกอริทึมในกลุ่มโครงข่ายประสาทเทียม ได้แก่ โครงข่ายประสาทเทียมแบบคอนโวลูชัน 1 มิติ (1D CNN), หน่วยความจำระยะยาว-ระยะสั้นหรือแอลเอสทีเอ็ม (LSTM) และแบบผสมผสาน (1D CNN & LSTM) ซึ่งผลการทดสอบในการจำแนกประเภทอารมณ์ของแต่ละโมเดลผลปรากฏว่าโมเดลแบบผสมผสาน (1D CNN & LSTM) มีค่าความถูกต้อง (Accuracy) มากที่สุด คือ 80.36% โมเดล 1D CNN 77.52% และ โมเดล LSTM ถูกต้องน้อยที่สุดคือ 67.86%

สำหรับการวิจัยในครั้งนี้เมื่อได้ทำการเปรียบเทียบกับงานวิจัยในอดีตที่มีการดำเนินการที่คล้ายกัน เช่น งานวิจัยของศรัญญา กาญจนวัฒนา และคณะ [10] ได้นำเสนอการจำแนกอารมณ์จากการรู้จำเสียงพูดด้วยการเรียนรู้แบบเชิงลึก และใช้แหล่งข้อมูลจากที่เดียวกัน ผลการวิจัยโมเดล LSTM มีค่าความถูกต้องสูงสุดอยู่ที่ 67.38% แต่การวิจัยครั้งนี้ได้ปรับโมเดลเป็นแบบผสมผสาน (1D CNN & LSTM) ซึ่งเป็นการผสมระหว่างการทำงานของโมเดล CNN ที่มีความสามารถในการตรวจจับลำดับของข้อมูล และ LSTM ที่มีความสามารถในการจัดการกับลำดับข้อมูลที่มีความยาวได้ดี โดยมีการเสริมข้อมูลด้วยเทคนิคการยืดเวลาของเสียง การเปลี่ยนระดับเสียง และการฉีดเสียงรบกวน เพื่อเพิ่มจำนวนของชุดข้อมูลให้โมเดลมีการเรียนรู้จากข้อมูลที่หลากหลายมากขึ้น และมีการปรับใช้พารามิเตอร์ในการฝึกโมเดลที่ต่างกัน โดยการฝึกโมเดลแบบ K Fold Cross Validation ปรับใช้ร่วมกับ Early Stopping ในการหยุดฝึกล่วงหน้า จึงทำให้โมเดลแบบผสมผสาน (1D CNN & LSTM) ได้ค่าความถูกต้องที่สูงขึ้น (80.36%)

6. ข้อเสนอแนะ

6.1 การนำโมเดลไปประยุกต์ใช้งาน

นำโมเดลไปประยุกต์ใช้กับชุดข้อมูลจริงเพื่อทดสอบและเปรียบเทียบต่อไป เช่น ภาษาพูดในภาษาไทยหลายสำเนียง เช่น สำเนียงใต้ เหนือ กลาง อีสาน จะทำให้สื่ออารมณ์แตกต่างกันหรือไม่

6.2 ข้อเสนอแนะในการวิจัยครั้งต่อไป

เพื่อเพิ่มประสิทธิภาพของการรู้จำอารมณ์จากเสียงพูดภาษาไทยในอนาคต อาจมีการสำรวจชุดข้อมูลที่ใช้ในการฝึกโมเดลเพิ่มเติม เพื่อให้มีข้อมูลที่หลากหลายมากยิ่งขึ้น นอกจากนี้อาจพิจารณาใช้โมเดลที่มีประสิทธิภาพสูงสุดในที่นี้คือโมเดลแบบผสมผสาน (1D CNN & LSTM) นำไปปรับพารามิเตอร์เพิ่มเติมอาจจะใช้วิธีค้นหาแบบกริด (Grid Search) เพื่อค้นหาพารามิเตอร์ที่ดีที่สุดเพื่อที่จะได้ผลลัพธ์ที่ดีที่สุดในการทดสอบ และอาจจะนำเอาอัลกอริทึม LSTM แบบสองทิศทาง (BI LSTM) เข้ามาปรับใช้ร่วมกับอัลกอริทึม 1D CNN เพื่อเปรียบเทียบประสิทธิภาพก็ได้ และเนื่องจากความยาวของเสียงบางข้อมูลสั้นเกินไปโดยเฉพาะเสียงที่มีประโยค 3 คำ อาจพิจารณาเพิ่มความยาวของเสียงให้ยาวขึ้น และอาจพิจารณาแยกเสียงตามเพศ อายุ ซึ่งอาจจะเพิ่มความถูกต้องของโมเดลได้มากขึ้น

7. เอกสารอ้างอิง

[1] T. Wangvanichapan, *Artificial intelligence can now read human voices "Data set and Emotional Classification Model from Thai Speech" the work of Professor Chulalongkorn University Available for free download today.* Available Online at https://www.chula.ac.th/highlight/47227/, accessed on 26 June 2023.

[2] P. Kulkasem, S. Rasameekhwan, B. Chandrakongkul, S. Rimcharoen, K. Chinsarn, P. Boonthong, and M. Chansuphap. *Emotion recognition of affective speech based on hybrid classifiers,* A Complete Research Report, Faculty of Information Sciences, Burapha University, 2015.

[3] M. El Ayadi, M. S. Kamel, and F. Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." *Pattern Recognition,* Vol. 44, No. 3, pp. 572-587, 2011.

[4] S. Kitthaweesinpoon and E. Rattagan. *Speech Emotion Recognition of Thai Language.* Master's thesis, Data Science Courses, Faculty of Applied Statistics, National Institute of Development Administration, 2021.

[5] J. Salamon and P. J. Bello. "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification." *IEEE Signal Processing Letters,* Vol. 24, No. 3, pp. 279-283, 2017.

[6] R. Kawade, R. Konade, P. Majukar, and S. Patil. "Speech Emotion Recognition Using 1D CNN-LSTM Network on Indo-Aryan Database." *International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT),* Vol. 3, pp. 1288-1293, 2022.

[7] E. Pacharawongsakda, *Dividing the data to test the efficiency of the model.* Available Online at https://www.linkedin.com/in/eakasit-pacharawongsakda-ph-d-475a8452/recent-activity/posts/, accessed on 26 June 2023.

[8] J. Brownlee, *Use Early Stopping to Halt the Training of Neural Networks At the Right Time, Machine Learning Mastery.* Available Online at https://machinelearningmastery.com/how-to-stop-training-deep-neural-networks-at-the-right-time-using-early-stopping/, accessed on 26 June 2023.

[9] P. Gatchalee, *Confusion Matrix is an important tool for evaluating prediction results in machine learning.* Available Online at https://medium.com/@pagongatchalee/, accessed on 17 October 2023.

[10] S. Kanjanawattana, A. Jarat, and P. Praneetpholkrang. "Classification of Human Emotion from Speech Recognition Using Deep Learning." *Science and Technology Journal Sisaket Rajabhat University,* Vol. 2, No. 2, pp. 1-11, July-December, 2022.

[11] A. Pratama and S. W. Sihwi. "Speech Emotion Recognition Model using Support Vector Machine Through MFCC Audio Feature." *International Conference on Information Technology and Electrical Engineering (ICITEE),* Vol. 14, pp. 303-307, 2022.

[12] VISTEC-depa Thailand Artificial Intelligence Research Institute, *Emotion classification dataset from Thai speech.* Available Online at https://airesearch.in.th/releases/speech-emotion-dataset/, accessed on 20 January 2023.