การพัฒนาวิธีการอ่านริมฝีปากจากภาพเคลื่อนไหวโดยใช้การเรียนรู้เชิงลึก

โดย
นายเอกภพ จิตตโคติ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
ภาควิชาวิศวกรรมไฟฟ้า
มหาวิทยาลัยศิลปากร
ปีการศึกษา 2566

การพัฒนาวิธีการอ่านริมฝีปากจากภาพเคลื่อนไหวโดยใช้การเรียนรู้เชิงลึก

โดย
นายเอกภพ จิตตโคติ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
ภาควิชาวิศวกรรมไฟฟ้า
มหาวิทยาลัยศิลปากร
ปีการศึกษา 2566

DEVELOPMENT OF LIP READING METHOD FROM VIDEO USING DEEP LEARNING

By

MR. Aekapob JITTAKOTI

A Thesis Submitted in Partial Fulfillment of the Requirements

for Master of Engineering (ELECTRICAL AND COMPUTER ENGINEERING)

Department of ELECTRICAL ENGINEERING

Academic Year 2023

หัวข้อ	การพัฒนาวิธีการอ่านริมฝีปากจากภาพเคลื่อนไหวโดยใช้การเรียนรู้เชิงลึก
โดย	นายเอกภพ จิตตโคติ
สาขาวิชา	วิศวกรรมไฟฟ้าและคอมพิวเตอร์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
อาจารย์ที่ปรึกษาหลัก	อาจารย์ ดร. โสภณ ผู้มีจรรยา

คณะวิศวกรรมศาสตร์และเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยศิลปากร ได้รับพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์และ
(ผู้ช่วยศาสตราจารย์ ดร. อรุณศรี ลีจีรจำเนียร) เทคโนโลยีอุตสาหกรรม

พิจารณาเห็นชอบโดย

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. ระพีพันธ์ แก้วอ่อน)

.....อาจารย์ที่ปรึกษาหลัก
(อาจารย์ ดร. โสภณ ผู้มีจรรยา)

.....ผู้ทรงคุณวุฒิภายใน
(ผู้ช่วยศาสตราจารย์ ดร. ภมร ศิลาพันธ์)

.....ผู้ทรงคุณวุฒิภายนอก
(ผู้ช่วยศาสตราจารย์ ดร. วีรพล จิรจริต)

640920030 : วิศวกรรมไฟฟ้าและคอมพิวเตอร์ แผน ก แบบ ก 2 ระดับปริญญามหาบัณฑิต
คำสำคัญ : การอ่านริมฝีปาก, โครงข่ายประสาทเทียมแบบคอนโวลูชัน, หน่วยความจำสั้นยาว

นาย เอกภพ จิตตโคติ: การพัฒนาวิธีการอ่านริมฝีปากจากภาพเคลื่อนไหวโดยใช้การเรียนรู้เชิงลึก อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก : อาจารย์ ดร. โสภณ ผู้มีจรรยา

วิทยานิพนธ์ฉบับนี้ได้นำเสนอวิธีการพัฒนาประสิทธิภาพของการอ่านริมฝีปากผ่านการวิเคราะห์เฟรมสำคัญโดยใช้ CNN และ LSTM ที่ทำงานร่วมกันซึ่งเป็นการใช้คุณลักษณะของการเรียนรู้แบบรูปภาพร่วมกับคุณลักษณะการเรียนรู้แบบลำดับขั้น หากต้องการเพิ่มประสิทธิของการอ่านริมฝีปากการใช้ชุดข้อมูลดิบทั้งหมดไม่สามารถให้ผลลัพธ์ที่ดีได้ ดังนั้นการเลือกจำนวนเฟรมและเฟรมที่เหมาะสมต่อการเรียนรู้จะส่งผลต่อประสิทธิภาพของแบบจำลองโดยตรง โดยวิธีการเลือกเฟรมได้ถูกนำเสนอผ่านไลบรารี่การตรวจจับใบหน้าของ Mediapipe บนโปรแกรมภาษา Python โดยการศึกษาได้มีการแบ่งการทดลองออกเป็น 3 กลุ่มหลัก นั่นคือ การเลือกจำนวนเฟรมที่ 3 5 และ 10 เฟรม อีกทั้งการเลือกเฟรมดังกล่าวยังแบ่งออกเป็นการเลือกแบบเฟรมเต็มปากและการเลือกแบบเฟรมครึ่งปาก โดยมีที่มาจากสมมติฐานเรื่องของความสมมาตรทางด้านร่างกายซ้ายและขวาของมนุษย์ อีกทั้งยังแสดงถึงการลดขนาดของอินพุตลงครึ่งนึงและเปรียบเทียบประสิทธิภาพของผลลัพธ์ที่ได้ ซึ่งเป็นการนำเสนอวิธีการวิธีการอ่านริมฝีปากที่ไม่มีงานวิจัยใดเคยทำมาก่อน โดยวัตถุประสงค์ของการอ่านริมฝีปากนั้น สามารถช่วยด้านการกู้ข้อมูลคำพูดจากไฟล์วิดีโอที่มีเสียงรบกวนจำนวนมาก รวมถึงการสื่อสารของผู้พิการทางการได้ยินด้วยเช่นกัน ในส่วนของฐานข้อมูลใช้ฐานข้อมูลที่ชื่อ AVDigits ซึ่งเป็นฐานข้อมูลภาษาอังกฤษที่มีการรวบรวมอาสาสมัครที่เป็นเจ้าของภาษาและไม่ใช่เจ้าของภาษากว่า 16 สัญชาติ โดยผลลัพธ์ที่ได้จากการศึกษานี้พบว่า แบบจำลองที่ได้นำเสนอรวมถึงขั้นตอนของการเลือกเฟรมสำคัญทำให้ประสิทธิภาพของการอ่านริมฝีปากทั้งแบบเต็มปากและครึ่งปากให้ผลลัพธ์อยู่ในระดับที่สูงและมีความใกล้เคียงกัน

640920030 : Major (ELECTRICAL AND COMPUTER ENGINEERING)
Keyword : Lip Reading, Convolutional Neural Network, Long Short-Term Memory

MR. Aekapob JITTAKOTI : Development of Lip Reading Method From Video Using Deep Learning Thesis advisor : SOPON PHUMEECHANYA, Ph.D.

This thesis presents a method for improving the efficiency of lip reading through the analysis of keyframes using CNN and LSTM working together, which combines the characteristics of image-based learning with sequential learning features. When attempting to enhance lip reading performance using the entire raw dataset, satisfactory results cannot be achieved. Thus, the selection of an appropriate number of frames and frame selection for learning directly affects the model's efficiency. The frame selection method is proposed through the Mediapipe face detection library in Python. The study divides experiments into three main groups: selecting 3, 5, and 10 frames. Additionally, the frame selection includes full-Lip image frames and half-Lip image frames options, based on the hypothesis of the symmetry of human body parts, both left and right. Furthermore, it demonstrates the reduction of input size by half and compares the performance of the obtained results. This proposes a lip reading method that has not been conducted before. The purpose of lip reading is to aid in speech retrieval from heavily corrupted audio-video files and also to facilitate communication for hearing-impaired individuals. In the database part, the AVDigits database, an English language database consisting of participants who are native and non-native speakers of English from 16 nationalities, is used. The results of this study show that the proposed models, including the crucial frame selection process, significantly improve lip reading performance for both full-Lip image and half-Lip image, achieving high and comparable results.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ได้รับปรึกษาจากอาจารย์ ดร.โสภณ ผู้มีจรรยา อาจารย์ที่ปรึกษาวิทยานิพนธ์ และคณะอาจารย์กรรมการสอบวิทยานิพนธ์ทุกท่าน ได้แก่ผู้ช่วยศาสตราจารย์ ดร.ระพีพันธ์ แก้วอ่อน ประธานสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ภมร ศิลาพันธ์ กรรมการผู้ทรงคุณวุฒิภายใน และผู้ช่วยศาสตราจารย์ ดร.วีรพล จิรจริต กรรมการผู้ทรงคุณวุฒิภายนอก ที่ให้คำปรึกษา ปรับปรุง และแนวทางการแก้ไขให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์ รวมถึงการนำเสนอผลงาน และกระบวนการวิจัย

ขอขอบคุณภาควิชาวิศวกรรมไฟฟ้า และคณะวิศวกรรมศาสตร์และเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยศิลปากร คณาจารย์ประจำภาควิชาวิศวกรรมไฟฟ้า และเจ้าหน้าที่และบุคลากรภาควิชาวิศวกรรมไฟฟ้าทุกท่าน ที่ให้ความช่วยเหลือทั้งด้านการให้คำแนะนำ และให้ความช่วยเหลืออำนวยความสะดวกต่างๆต่อการวิทยานิพนธ์ฉบับนี้

เอกภพ จิตตโคติ

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

งานวิจัยที่เกี่ยวข้องกับการอ่านริมฝีปาก (Lip Reading) เป็นที่รู้จักกันอย่างแพร่หลายในชื่อของการรู้จำคำพูด (Speech Recognition) โดยเนื้องานมีความใกล้เคียงกันและสามารถใช้แทนกันได้ ในบางครั้งงานวิจัยที่ใช้คำว่าการรู้จำคำพูดอาจมีการใช้เสียงเข้ามาในกระบวนการสร้างแบบจำลอง แตกต่างจากงานที่เป็นการอ่านริมฝีปากที่เน้นไปที่รูปภาพเป็นส่วนใหญ่ ซึ่งเป็นการเก็บรวบรวมข้อมูลบริเวณรอบๆริมฝีปากของผู้พูดเพื่อใช้จำแนกออกมาเป็นคำพูดต่างๆ ในขั้นตอนเริ่มต้นของการหาบริเวณรอบๆริมฝีปาก จะต้องใช้การถ่ายภาพเคลื่อนไหวในการพูดประโยค คำ หรือพยางค์ของสิ่งที่ต้องการจำแนก เพื่อคัดกรองข้อมูลคุณลักษณะต่างๆ (Features) ที่จะใช้ในการรู้จำประโยค คำ หรือพยางค์ (Target)

การอ่านริมฝีปากถูกนำมาใช้ในงานประเภทต่างๆมากมาย เช่น การช่วยเหลือผู้บกพร่องทางการได้ยินโดยอุปกรณ์ที่ใช้ในการช่วยแปลการเคลื่อนไหวของริมฝีปากออกมาเป็นคำพูด [1] หรือการแปลงคำพูดออกมาเป็นข้อความในบริเวณที่มีเสียงรบกวนเป็นจำนวนมากโดยการสื่อสารเป็นคำพูดก่อให้เกิดความผิดพลาดหรือความยากลำบาก แต่เนื่องจากในช่วงแรกเริ่มของงานวิจัยการอ่านริมฝีปากนั้นมีประสิทธิภาพที่ค่อนข้างไม่เป็นที่น่าพอใจ โดยเป็นการใช้วิธีการของ non-deep learning ร่วมกับเทคนิคการสร้างด้วยมือ (Hand-Crafted) ซึ่งเป็นยุคแรกเริ่มของงานทางด้านการอ่านริมฝีปาก ตัวอย่างเช่น วิธีของ Hidden Markov Model (HMMs) [2] หรือ Support Vector Machine (SVM) [3] และการคัดกรองข้อมูลคุณลักษณะต่างๆ (Feature Extraction) ที่ร่วมกันในช่วงนั้น ประกอบด้วย Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) [4], Discrete Cosine Transform (DCT) [5], Logicality Discrete Graph (LDG) [6] และ Active Appearance Models (AAMs) เป็นต้น ในช่วงระยะเวลาประมาณ 10 ปีที่ผ่านมางานวิจัยทางด้านการอ่านริมฝีปากก็เริ่มเป็นที่สนใจของนักวิจัยมากขึ้นอย่างต่อเนื่อง เนื่องจากความก้าวหน้าเพิ่มขึ้นอย่างมากในการพัฒนาของการเรียนรู้เชิงลึก (Deep Learning) ซึ่งเป็นส่วนหนึ่งของการเรียนรู้ของเครื่อง (Machine Learning) ที่มีพื้นฐานมาจากโครงข่ายประสาทเทียม (Artificial neural network)[7] ตัวคัดกรองคุณลักษณะข้อมูลแบบดั้งเดิมถูกแทนที่ด้วย Neural Network, Feed-Forward Network, และ Convolutional Neural Network (CNNs) ในส่วนของตัวจำแนก

(Classification) มีการออกแบบ CNN ในรูปแบบต่างๆ [8] และการนำเสนอที่เปรียบเทียบ Pre-trained models [9] อีกทั้งยังมีนำเสนอการใช้ Dilated CNN [10] การรู้จำคำหรือข้อความส่วนใหญ่แล้วจะใช้โครงข่ายการประมวลผลแบบเป็นลำดับ เช่น Recurrent Neural Networks (RNNs) ในรูปแบบของ Long-Short Term Memory networks (LSTMs) [11] และงานวิจัยอีกมากมายที่ต่างให้ความสนใจและนำเสนอวิธีที่ใช้ในการอ่านริมฝีปากโดยการใช้ CNN ทำงานร่วมกับ LSTM [12], [13], [14], [15], [16], [17], [18], เห็นได้ชัดว่าการเรียนรู้เชิงลึกมีบทบาทที่สำคัญในการพัฒนาประสิทธิภาพของการอ่านริมฝีปากอย่างมาก สังเกตจากงานวิจัยช่วงประมาณ 10 ปีที่ผ่านมานั้น นักวิจัยต่างนำเสนอวิธีที่จำเพิ่มประสิทธิภาพของการอ่านริมฝีปากจากการใช้การเรียนรู้เชิงลึกกันทั้งสิ้น ซึ่งการนำเสนอนั้นเป็นการนำเสนอเทคนิคของการเลือกข้อมูลของคุณลักษณะที่จะใช้ในการฝึกสอนที่เน้นไปในด้านของการลดขนาดของอินพุต [19], [20] และการเลือกใช้ตัวจำแนกที่เหมาะสมกับแบบจำลองที่นำเสนอ และเปรียบเทียบผลลัพธ์ที่ได้กับงานวิจัยอื่นๆที่ใกล้เคียง

ดังนั้นงานวิจัยนี้มีจุดมุ่งหมายเพื่อสร้างแบบจำลองของการอ่านริมฝีปากจากการเรียนรู้เชิงลึกประกอบไปด้วย การใช้งาน Convolutional Neural Network ร่วมกับ การใช้งาน Long – Shot Term Memory ในการรู้จำคำพูดจากการอ่านริมฝีปาก ซึ่งการใช้งานร่วมกันเป็นการใช้ข้อดีของ CNN ในการคัดกรองคุณลักษณะข้อมูลที่มีความแข็งแรงสูง และ LSTM ที่มีประสิทธิภาพในการจำแนกที่ดี จะทำให้ประสิทธิภาพผลลัพธ์ของการรู้จำคำพูดสูงขึ้น และการศึกษาสถาปัตยกรรมอื่นๆ ที่เกี่ยวข้อง อีกทั้งยังนำเสนอสถาปัตยกรรมจากแบบจำลองทั้งสองที่ทำงานร่วมกันแล้วยังสามารถลดขนาดของอินพุตลงครึ่งหนึ่งแล้วยังคงไว้ซึ่งประสิทธิภาพที่เทียบเคียงกันได้ เนื่องจากสมมติฐานทางด้านซ้ายและขวาที่นำเสนอโดยผู้วิจัย จากการนำเสนอการลดขนาดของอินพุตโดยผู้วิจัยทำให้เป็นการพูดถึงการใช้ข้อมูลริมฝีปากที่สามารถใช้แค่เพียงครึ่งเฟรมของภาพริมฝีปากที่ไม่มีงานวิจัยใดเคยกล่าวถึงมาก่อน และยังคงมีการเปรียบเทียบผลลัพธ์ของการรู้จำกับฐานข้อมูลที่เกี่ยวข้องและมีความนิยม [21] ซึ่งเป็นฐานข้อมูลที่ถูกนำมาใช้ นอกจากนี้ยังมีฐานข้อมูลที่เกี่ยวข้องกับงานทางด้านการอ่านริมฝีปากโดยนักวิจัยส่วนใหญ่ให้ความสนใจ [22], [23], [24]

1.2 วัตถุประสงค์ของการวิจัย

1.2.1. เพื่อศึกษาสถาปัตยกรรมการเรียนรู้เชิงลึกที่ใช้สำหรับการอ่านริมฝีปาก

1.2.2. เพื่อศึกษาวิธีการอ่านริมฝีปากโดยใช้ CNN ร่วมกับ LSTM

1.2.3. เพื่อศึกษาและวิเคราะห์คีย์เฟรมสำคัญที่ใช้ในการอ่านริมฝีปาก

1.2.5. เพื่อออกแบบและพัฒนาวิธีการอ่านริมฝีปากจากภาพโดยใช้การเรียนรู้เชิงลึกให้มีประสิทธิภาพมากขึ้น

1.3 ขอบเขตของการวิจัย

1.3.1. ออกแบบและพัฒนาวิธีการอ่านริมฝีปากจากภาพ

1.3.2. ออกแบบวิธีการเตรียมและคัดเลือกข้อมูลสำหรับการฝึกสอนแบบจำลอง

1.3.3. ทำการทดลองโดยการวิเคราะห์การเพิ่มขึ้นหรือลดลงของคีย์เฟรมสำคัญรวมถึงการลดขนาดข้อมูลที่ใช้ลงครึ่งหนึ่ง

1.3.4. ใช้ฐานข้อมูลสาธารณะชื่อว่า AV Digits โดยเลือกใช้ที่โหมดการพูดแบบหน้าตรง เสียงปกติชุดข้อมูลตัวเลขภาษาอังกฤษตั้งแต่เลข 0 - 9

1.3.5 ออกแบบและพัฒนาเฟรมเวิร์คที่ใช้สำหรับการอ่านริมฝีปากจากภาพ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

แบบจำลองที่ได้รับการออกแบบใหม่รวมถึงขั้นตอนการเตรียมการข้อมูลที่นำเสนอสามารถเพิ่มประสิทธิภาพในการอ่านริมฝีปากได้ดีขึ้น อีกทั้งการลดขนาดของข้อมูลจะยังคงไว้ซึ่งประสิทธิภาพที่ใกล้เคียงกับข้อมูลแบบดั้งเดิม

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในหัวข้อของทฤษฎีและงานวิจัยที่เกี่ยวข้องได้นำเสนอพื้นฐานที่เกี่ยวข้องกับการพัฒนาการอ่านริมฝีปากจากภาพเคลื่อนไหว ประกอบด้วย การเรียนรู้เชิงลึก โครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neural Network) โครงข่ายประสาทเทียมแบบเกิดซ้ำ (Recurrent Neural Network) ซึ่งเป็นความรู้พื้นฐานของการต่อยอดเป็นหน่วยความจำสั้นยาว (Long – Short Tearm Memory) การตรวจจับใบหน้า (Face Detection) และ การกำหนดจุดบริเวณใบหน้า (Face Localization) และงานวิจัยที่เกี่ยวข้องกับการพัฒนาการอ่านริมฝีอ่าน

2.1 โครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neural Network)

โครงข่ายประสาทเทียมแบบคอนโวลูชันเป็นหนึ่งในรูปแบบของการเรียนรู้เชิงลึกที่มีจำนวนชั้นซ่อนอยู่จำนวนมากจากพื้นฐานของโครงข่ายประสาทเทียมแบบพื้นฐาน และเพิ่มชั้นของคอนโวลูชันเข้าไปเพื่อเพิ่มผลลัพธ์ของการจำแนกให้ดีมากยิ่งขึ้น จากการเพิ่มชั้นคอนโวลูชันส่งผลให้สามารถทำงานได้ดีกับชุดข้อมูลที่เป็นรูปภาพโดยความสามารถในการสกัดคุณลักษณะเด่นออกจากรูปภาพ เช่น เส้นขอบ และ สี การทำงานของโครงข่ายประสาทเทียมแบบคอนโวลูชัน ประกอบไปด้วยเลเยอร์ทั้งหมด 3 ประเภท

2.2.1 ชั้นคอนโวลูชัน (Convolution Layer)

ชั้นคอนโวลูชันสามารถเรียกอีกอย่างได้ว่าเป็น Filter หรือ Kernel เป็นการทำงานทางคณิตศาสตร์รูปแบบหนึ่งที่ทำหน้าที่ในการสกัดเอาคุณลักษณะเด่นออกจากภาพที่รับเข้ามา เพื่อนำคุณลักษณะเด่นเหล่านี้เป็นข้อมูลอินพุตให้กับโครงข่ายประสาทเทียม โดยกระบวนการของคอนโวลูชันเลเยอร์จะทำให้ขนาดของเล็ก จึงการทำ padding เพื่อเพิ่มความสำคัญให้กับข้อมูลบริเวณขอบของภาพ

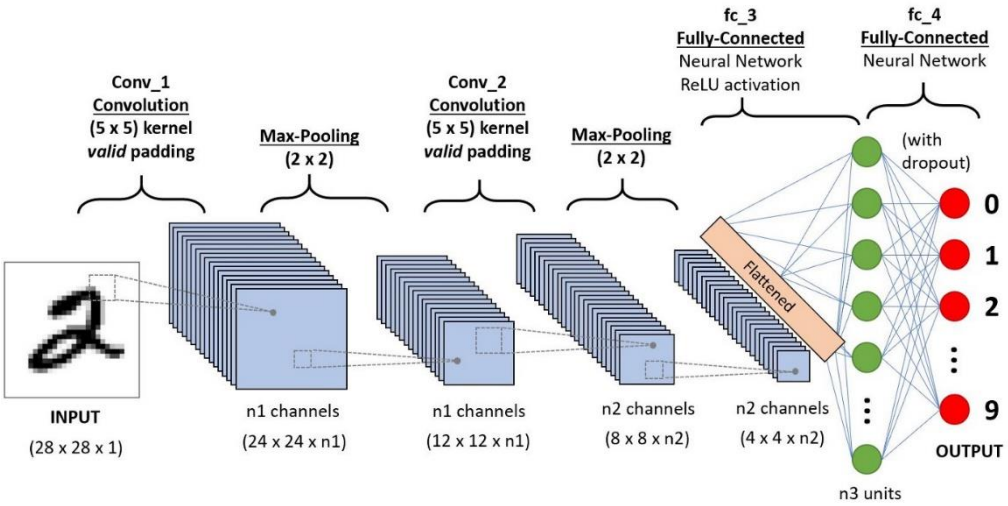
2.2.2 ชั้นพลูลิ่ง (Pooling Layer)

พลูลิ่งเลเยอร์โดยปกติจะอยู่ถัดจากชั้นของคอนโวลูชัน เป็นชั้นของการย่อขนาดของภาพให้เล็กลงทำให้การประมวลผลทำได้เร็วมากขึ้น แทนที่การใช้ข้อมูลของภาพทั้งหมดในการคำนวณ เช่น เดียยับวการที่ตามนุษย์สามารถแยกรูปภาพเดียวกันที่มีขนาดเล็กหรือใหญ่ต่างกันได้นั่นเอง หมายความว่าละเอียดที่บ่งบอกถึงรูปภาพเดิมไม่ถูกตัดออกไปหรือส่งผลต่อการทำงานของชั้นคอนโวลูชันในลำดับ

ถัดไป มี 2 ประเภทหลักที่นิยมใช้งานคือ Max pooling เป็นหาค่ามากสุด และ Average Pooling เป็นการหาค่าเฉลี่ย

2.2.3 ชั้นเชื่อมโยงสมบูรณ์ (Fully Connected Layer)

เป็นชั้นที่รับเอาต์พุตของชั้นก่อนหน้าโดยปกติแล้วคือชั้นพลูลิ่ง นำมาทำ Pooling Feature map หรือ Flattening เพื่อเข้าสู่กระบวนการของการเรียนรู้เชิงลึก ที่เทียบได้กับโครงข่ายประสาทเทียมแบบพื้นฐานที่มี Multilayer perceptron (MLP) ในการปรับค่าของ weight โดย Backpropagation เพื่อการจำแนกผลลัพธ์ต่างๆตามที่ต้องการ



รูปที่ 2.1 โครงข่ายประสาทเทียมแบบคอนโวลูชัน

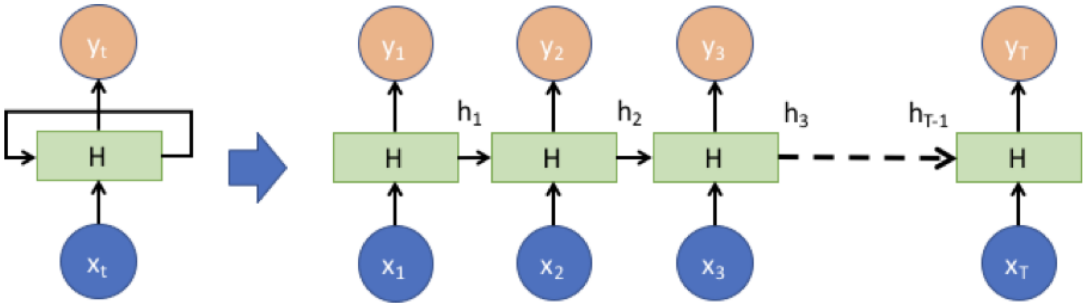
ที่มา https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

2.2 โครงข่ายประสาทเทียมแบบเกิดซ้ำ (Recurrent Neural Network) และ หน่วยความจำสั้นยาว (Long – Short Tearm Memory)

2.2.1 โครงข่ายประสาทเทียมแบบเกิดซ้ำ (Recurrent Neural Network)

แนวคิดที่สำคัญของโครงข่ายประสาทเทียมแบบเกิดซ้ำคือการนำเอาเอาต์พุตที่ได้จากชั้นซ่อน (Hidden State) ชั้นก่อนหน้า มารวมกับอินพุตที่เข้ามา ณ ปัจจุบันเพื่อที่จะใช้งานกับข้อมูลที่มีลักษณะเป็นลำดับ (Sequence) เช่น วิดีโอ (Sequen of image) เป็นการลำดับเหตุการณ์ขึ้นมาจาก

รูปภาพที่นำมาต่อกันหลายๆรูป และ ข้อความ (Sequence of word) เป็นลำดับของการรับคำแต่ละคำเข้ามารวมกันเป็นข้อความหรือประโยค หรือการที่จะประมวลผลข้อความหรือวิดีโอเหล่านั้นว่าเกี่ยวข้องกับอะไร จำเป็นที่จะต้องเอาคำ (กรณีเป็นข้อความ) หรือ เฟรมภาพ (กรณีของวิดีโอ) ก่อนหน้าเข้ามาร่วมในการประมวลผลกับคำหรือเฟรม ณ ปัจจุบันด้วย ซึ่ง RNN ได้ใช้หลักการดังกล่าวในการพัฒนาต่อจากโครงข่ายประสาทเทียมแบบพื้นฐาน เพื่อเอาชั้นซ่อนชั้นก่อนหน้า หรือเรียกว่าความรู้ก่อนหน้า มาคำนวณร่วมกับอินพุตใหม่ที่รับเข้ามา



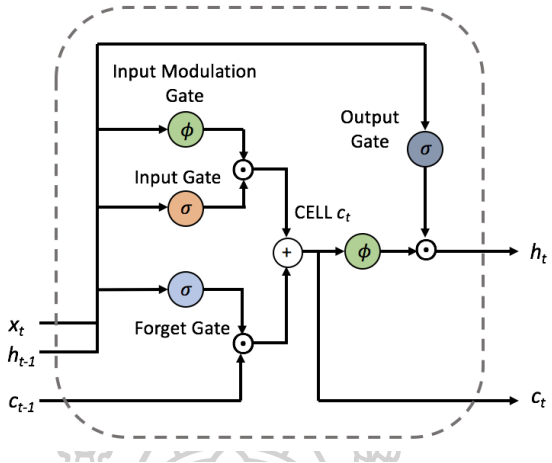
รูปที่ 2.2 Recurrent Neural Network

ที่มา https://medium.com/@sinart.t/long-short-term-memory-lstm-e6cb23b494c6

ปัญหาหลักของ RNN คือการคำนวณค่า weight โดยการคำนวณค่า Gradient ของ Loss function มีความยุ่งยากมากขึ้นตามลำดับที่รับเข้ามา เนื่องค่าเอาต์พุตที่ได้นั้นไม่ได้มาจากช่วงเวลาที่ t=t เพียงอย่างเดียว แต่รวมถึงช่วงเวลาก่อนหน้านั้นด้วย ซึ่งค่าของ Gradient ที่ใช้ว่า weight เกิดจากการคูณกันของค่า derivative จำนวนมากทำให้เกิดความล่าช้าและถ้าหากค่าหากค่า Gradient มีค่าน้อยกว่า 1 ผลของการคูณ Gradient จะมีขนาดน้อยลงไปเรื่อย ๆ เนื่องจากลำดับที่รับเข้ามามีจำนวนมากเกินไป ทำให้ LSTM เข้ามาแก้ปัญหาที่กล่าวไป

2.2.2 หน่วยความจำสั้นยาว (Long – Short Tearm Memory)

หน่วยความจำสั้นยาวเกิดขึ้นมาเพื่อแก้ปัญหาของ RNN เมื่อข้อมูลที่รับเข้ามามีจำนวนที่ยาวเกินไปและเป็นการพัฒนาต่อยอดโดยคงแนวคิดเดิมไว้ แสดงรูปการณ์ทำงานของ LSTM ตามรูปด้านล่าง



รูปที่ 2.3 การทำงานของ LSTM

ที่มา https://medium.com/@sinart.t/long-short-term-memory-lstm-e6cb23b494c6

การทำงานของ LSTM จะมีสิ่งที่เรียกว่า cell state เป็นตัวเก็บ state ของ memory cell โดยตัวควบคุมการทำงานต่างๆเรียกว่า Gate โดยสิ่งที่จะมาควบคุม cell state ผ่าน gate ต่าง ๆ มีดังนี้

1. Forget ทำหน้าที่เป็นเหมือนตัวล้าง cell state จัดการพื้นที่เพื่อเตรียมพร้อมสำหรับรับข้อมูลใหม่ ผ่าน Forger Gate การสร้าง Forget gate จะดูข้อมูลอินพุตร่วมกับชั้นซ่อนก่อนหน้าตามแบบของ RNN และการตัดสินใจจะใช้ sigmoid function

2. Write เมื่อมีอินพุตใหม่เข้ามา อินพุตดังกล่าวจะทำการอัพเดทเข้าไปใน cell state หรือไม่ ถ้าอัพเดทจะต้องอัพเดทด้วยค่าอะไร การที่จะประมวลผลว่าจะทำการอัพเดทหรือไม่จะควบคุมผ่าน Input Gate การตัดสินใจจะใช้ Sigmoid Function หากต้องการอัพเดทข้อมูล การเลือกค่าอัพเดทจะประมวลผลจาก Input Modulation Gate โดยใช้การตัดสินใจแบบ Tanh Function

3. update cell การอัพเดท cell จะเอาต์พุตที่ได้จาก Forget Gate ,Input Gate และ Input Modulation Gate มารวมเข้าด้วยกันแล้วจึงตัดสินใจที่จะอัพเดท cell หรือไม่

4 Read จากการคนวณในลำดับถัดไปของ RNN จะเห็นว่า จะมีการนำเอาเอาต์พุตจากชั้นซ่อนก่อนหน้ามาคำนวณด้วย การ Read ในที่นี้หมายถึงการที่ข้อมูลลำดับถัดไปคือ t+1 สามารถที่จะนำมูล ณ เวลา t ไปคำนวณได้หรือไม่ ตัดสินใจผ่าน Output Gate

2.3 การตรวจจับใบหน้า (Face Detection) และ การกำหนดจุดบริเวณใบหน้า (Face Localization)

2.3.1 การตรวจจับใบหน้า (Face Detection)

คำว่า Face Detection ในบางครั้งสามารถใช้คำว่า Facial Detection ก็ได้ เป็นปัญญาประดิษฐ์รูปแบบหนึ่ง ใช้สำหรับค้นหาใบหน้าของมนุษย์จากข้อมูลรูปแบบในระบบดิจิตอล พัฒนามาจากวิสัยทัศน์คอมพิวเตอร์ (Computer Vision) ในปัจจุบันมีบทบาทสำคัญกับงานหลายหลากประเภทไม่ว่าจะเป็น การตามรอยใบหน้า (Face Tracking) การรู้จำใบหน้า (Face Recognition) การวิเคราะห์ใบหน้า (Face Analysis)

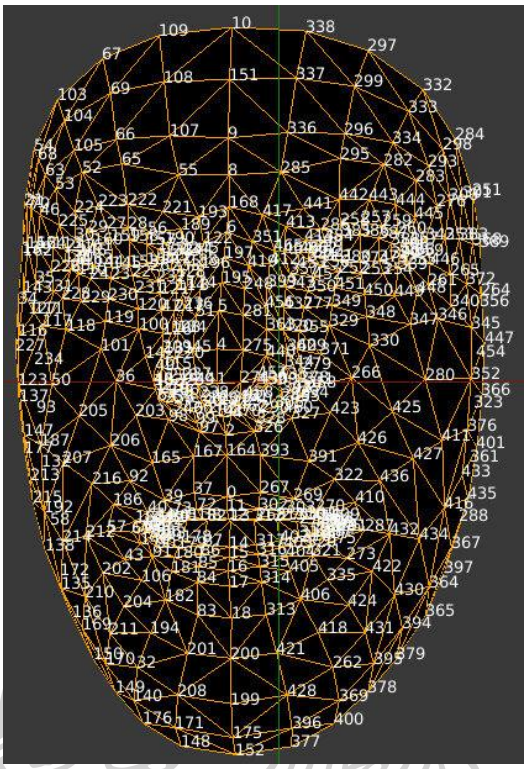
การตรวจจับใบหน้าใช้อัลกอริทึมทางการเรียนรู้ของเครื่อง (Machine Learning) เพื่อค้นหาใบหน้ามนุษย์ภายในภาพ ซึ่งมักจะรวมเอาวัตถุอื่นๆ ที่ไม่ใช่ใบหน้า เช่น วิวทิวทัศน์ อาคาร และส่วนอื่นๆ ของร่างกายมนุษย์ เช่น เท้าหรือมือ โดยปกติแล้ว อัลกอริทึมการตรวจจับใบหน้าจะเริ่มต้นด้วยการค้นหาดวงตาของมนุษย์ ซึ่งเป็นหนึ่งในคุณสมบัติที่ง่ายที่สุดในการตรวจจับ อัลกอริทึมอาจพยายามตรวจหาคิ้ว ปาก จมูก รูจมูก และม่านตา เมื่ออัลกอริทึมสรุปว่าพบบริเวณใบหน้าแล้ว จะใช้การทดสอบเพิ่มเติมเพื่อยืนยันว่าได้ตรวจพบใบหน้าแล้ว ในการยืนยันความถูกต้อง อัลกอริทึมจำเป็นต้องได้รับการฝึกสอนเกี่ยวกับชุดข้อมูลขนาดใหญ่ที่รวมรูปภาพหลายแสนภาพ การฝึกสอนจะช่วยปรับปรุงความสามารถของอัลกอริทึมในการระบุว่ามีใบหน้าอยู่ในรูปภาพหรือไม่และอยู่ที่ไหน

รูปที่ 2.4 Face Detection

ที่มา https://sightcorp.com/knowledge-base/face-detection/

2.3.2 การกำหนดจุดบริเวณใบหน้า (Face Localization)

การกำหนดจุดบริเวณใบหน้าเป็นการนำระบบการตรวจจับใบหน้ามาต่อยอดในการรู้จำตำแหน่งต่างๆบริเวณใบหน้าหรือเรียกอีกอย่างได้ว่า Face Landmark เป็นการกำหนดตำแหน่งสำคัญบนใบหน้าไม่ว่ารูปหน้าที่ปรากฏอยู่ในภาพจะเต็มใบหน้าหรือไม่ก็ตาม อัลกอริทึมจะสามารถจำลองใบหน้าในส่วนที่สามารถมองเห็นได้และไม่สามารถมองเห็นซึ่งในวิทยานิพนธ์ฉบับนี้ได้นำเสนอการใช้ Library ที่ชื่อว่า Mediapipe [25] บนโปรแกรมภาษา Python ในการอ่านริมฝีปากจะใช้เพื่อการตัดภาพบริเวณที่ต้องการนั่นคือริมฝีปากเท่านั้น แบบจำลองที่ใช้ในกากำหนดจุดบนใบหน้าที่นิยมใช้ คือ 68 จุด ภายในวิทยานิพนธ์ฉบับนี้ใช้แบบจำลองการกำหนดจุดอยู่ที่ 468 จุด เพื่อความละเอียดในการตัดรูปภาพบริเวณริมฝีปาก



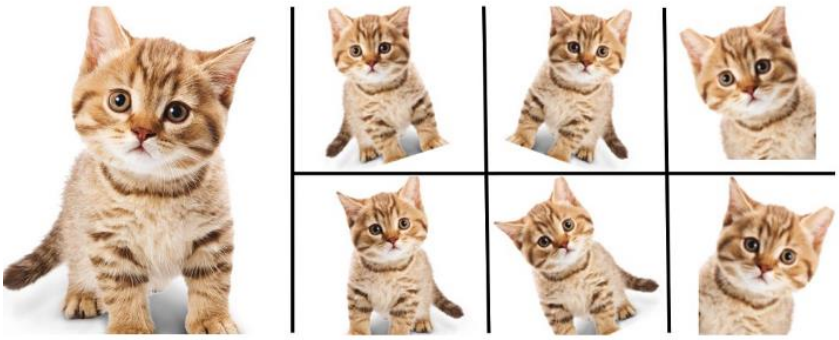
รูปที่ 2.5 Face Localization

ที่มา https://www.analyticsvidhya.com/blog/2021/07/facial-landmark-detection-simplified-with-opencv/

2.4 Data Augmentation

ในการฝึกสอนของแบบจำลองการเรียนรู้เชิงลึกสิ่งที่จำเป็นอย่างหนึ่งคือชุดข้อมูล ในสภาวะที่ไม่สามารถหาชุดข้อมูลเพิ่มได้ หรือการหาชุดมูลทำได้ล่าช้าแต่ต้องการแบบจำลองที่มีความครอบคลุมเพื่อเพิ่มประสิทธิภาพให้ได้มากที่สุด วิธีที่สามารถช่วยได้คือการทำ Data Augmentation เป็นการนำรูปภาพหรือชุดข้อมูลที่มีอยู่มาเพิ่มขนาด ไม่ว่าจะเป็นการ ย่อ ขยาย การตัดมุมขอบของภาพ การใส่ noise การทำให้ภาพเบลอ การทำให้ภาพเอียง ทั้งหมดนี้ทำเพื่อให้ครอบคลุมกับการตรวจจับรูปภาพให้มีความถูกต้องกับสภาวะแวดล้อมที่เกิดขึ้นจริง เช่นการเบลอภาพ เพื่อรองรับในกรณีของการรับภาพมาโดยที่กล้องไม่ได้โฟกัสวัตถุใดๆในรูปภาพทำให้เกิดการเบลอของภาพ หรือ การเพิ่ม noise รองรับในสภาวะของกล้องที่มีคุณภาพต่ำมีความแตกของพิกเซล เป็นต้น ในกรณีเหล่านี้สามารถ

เกิดขึ้นได้จริง และไม่จำเป็นในการหาชุดข้อมูลใหม่ ถ้าหากชุดมูลนั้นมีจำกัด และสามารถเพิ่มประสิทธิภาพของแบบจำลองได้ดีวิธีหนึ่ง



รูปที่ 2.6 การทำ Data Augmentation ในรูปภาพแมว

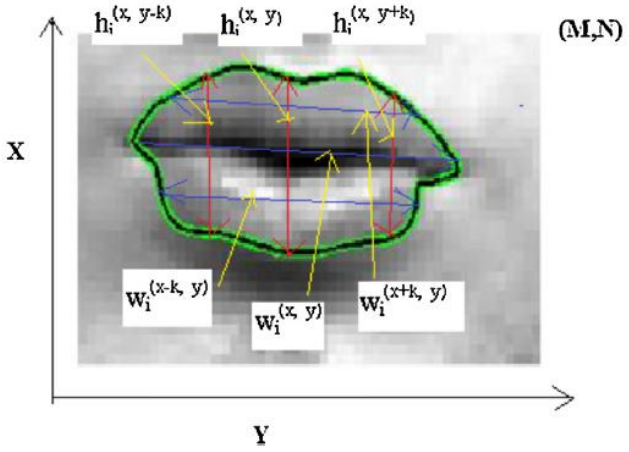
ที่มา https://nanonets.com/blog/data-augmentation-how-to-use-deep-learning-when-you-have-limited-data-part-2/

2.5 งานวิจัยที่เกี่ยวข้อง

2.5.1 A novel lip reading algorithm by using localized ACM and HMM:Tested for digit recognition (Sunil S. Moradea , Suprava Patnaik(2014)) [2]

งานวิจัยนี้นำเสนอการใช้วิธีการอ่านริมฝีปากแบบ non-deep learning โดยการใช้แบบจำลองแอคทีฟคอนทัวร์สำหรับการระบุริมฝีปากและนำเสนอการใช้การแยกคุณลักษณะทางเรขาคณิตในการอ่านริมฝีปาก ผลของคุณสมบัติแต่ละอย่างจะถูกเปรียบเทียบกัน ประกอบด้วยความกว้าง ความยาว และความสูงของริมฝีปาก และคุณลักษณะร่วมที่ใช้พารามิเตอร์ความกว้าง ความยาว และ ความสูงรวมกัน ในการนำเข้าเป็นอินพุตให้กับแบบจำลอง แบบจำลอง Ergodic Hidden markov (HMM) ถูกใช้เป็นตัว แยกประเภทตัวเลข โดยที่แบบจำลองของ Markov นำมาใช้ในการทดลองแบบ 3 ขั้น และ 5 ขั้น โดยจะมีวิดีโอตัว เลขภาษาอังกฤษ ตั้งแต่ เลข 0 ถึงเลข 9 ในภาษาอังกฤษที่ได้ถูกบันทึกไว้ สำหรับการทดสอบประสิทธิภาพการรู้จำ โดยจะทดสอบทั้งฐานข้อมูล

ของ Cuave ซึ่งเป็นฐานข้อมูลแบบสากล กับฐานข้อมูลภายในและเปรียบเทียบผลลัพธ์ที่ได้ ในระหว่างการดำเนินการจะนำไปคำนวณจากคุณลักษณะของเฟรมที่มีความสำคัญที่ได้คัดเลือกมา เพื่อลดการคำนวณที่ มีความซับซ้อน ผลการทดลองของการออกเสียงตัวเลขแสดงให้เห็นค่ามากที่สุดที่รู้จำของตัวเลขแต่ละตัว



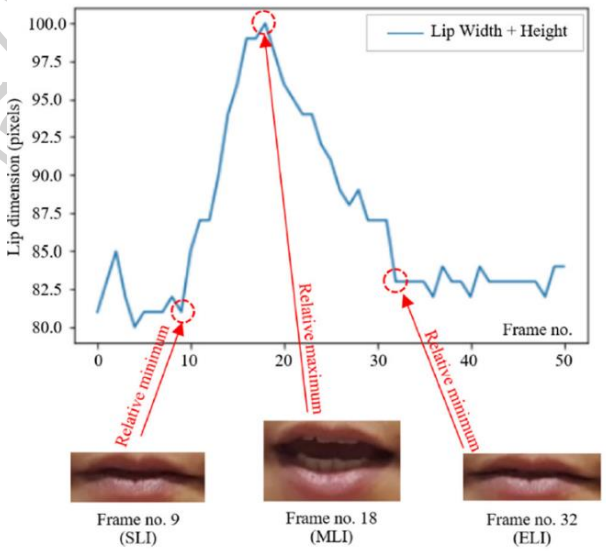
รูปที่ 2.7 การตามรอยริมฝีปากจากอัลกอริทึม ACM

Database	Feature set	3 state HMM R.R. (%)	5 state HMM R.R. (%)
Cuave	*H*	44	58
	W	33	37
	A	40	45
	W+*H*+*A*	66.3	78.33
In-house	*H*	35	49
	W	26	33
	A	33	43
	W+*H*+*A*	64.7	76.6

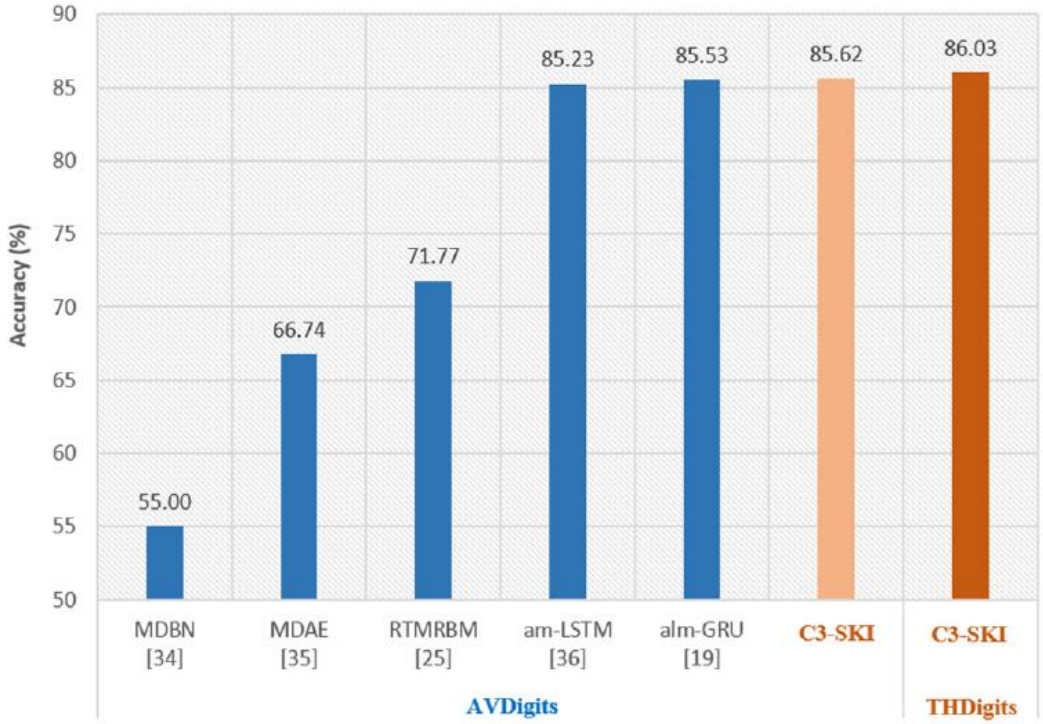
รูปที่ 2.8 ตารางการเปรียบเทียบผลลัพธ์ที่ได้กับฐานข้อมูล Cuave กับฐานข้อมูลที่สร้างขึ้น

2.5.2 Improving the Recognition Performance of Lip Reading Using the Concatenated Three Sequence Keyframe Image Technique (Lap Poomhiran , Phayung Meesad, Sumitra Nuanmeesri (2021)) [20]

ในงานวิจัยนี้ได้นำเสนอวิธีการที่ใช้ในการอ่านริมฝีปากจากการใช้การเรียนรู้เชิงลึกบนพื้นฐานของโครงข่ายประสาทเทียมมาพัฒนาต่อยอด โดยการกำหนดอินพุตที่ได้มาจากการนำเฟรมที่มีการเปลี่ยนแปลงของรูปริมฝีปากทั้งสามเฟรมมาเรียงต่อกันประกอบไปด้วย 1) ภาพเริ่มต้นของการเปิดริมฝีปาก 2) ภาพริมฝีปากที่มีการขยับมากที่สุด 3) ภาพการปิดริมฝีปากที่เป็นการออกเสียงช่วงท้าย โดยตั้งชื่อเทคนิคที่ใช้นี้ว่า C3SKI (Concatenated Three Sequence Keyframe Image) ทุกๆภาพที่นำมาเรียงต่อกันกำหนดมาจากค่าสูงสุดสัมพัทธ์และค่าต่ำสุดสัมพัทธ์ มีการลดขนาดของอินพุตในรูปริมฝีปากแต่ละรูปเหลือ 32x32 พิกเซล และเมื่อนำมาต่อกัน 3 ภาพ ทำให้ภาพมีขนาด 96x32 พิกเซล เป็นการลดขนาดของอินพุตที่ยังคงมีประสิทธิภาพที่ดีในการรู้จำคำพูดจากการอ่านริมฝีปากได้ดี วิธีดังกล่าวได้นำไปทดสอบกับฐานข้อมูล AVDigits ซึ่งเป็นฐานข้อมูลตัวเลขสากลประกอบด้วยเลข 0 – 9 ในภาษาอังกฤษและได้สร้างฐานข้อมูลของตัวเอง ให้ชื่อว่า THDigits ที่ประกอบด้วยตัวเลข 0 – 9 ในภาษาไทย และทำการแสดงผลลัพธ์และเปรียบเทียบกับวิธีอื่นๆที่ใช้อ่านริมฝีปาก



รูปที่ 2.9 เทคนิค C3-SKI ที่ใช้ในการหาค่าอินพุต



รูปที่ 2.10 กราฟแสดงประสิทธิภาพของวิธีที่นำเสนอกับวิธีการก่อนหน้า

2.5.3 End-To-End Low-Resource Lip-Reading With Maxout CNN And LSTM (Ivan Fung, Brian Mak (2018)) [13]

บทความนี้นำเสนอสถาปัตยกรรมที่ใช้ในการอ่านริมฝีปากโดยใช้ CNN ร่วมกับ LSTM ผ่าน Activation unit ที่ชื่อ MAX-OUT ภายใต้คลังข้อรูปที่มีปริมาณน้อย ในส่วนของ MAX-OUT เป็น Activation ที่มีความเรียบง่ายและมีประสิทธิภาพ ทำงานได้ดีเมื่อใช้ร่วมกับ Dropout สามารถที่จะใช้แทน Activation ReLu ได้ และฐานข้อมูลที่นำมาพิจารณามีชื่อว่า OULUVS2 ภายในประกอบไปด้วยผู้พูด 52 คน ชาย 39 หญิง 13 แบ่งออกเป็น 3 ส่วน ส่วนที่ 1 ตัวเลข 0 – 9 ส่วนที่ 2 วลีที่ใช้บ่อย 10 คำ ส่วนที่ 3 ประโยคที่ใช้เอามาจากฐานข้อมูล TIMIT ถือเป็นมีปริมาณทีน้อยเมื่อเทียบกับฐานข้อมูลอย่าง LRW ในส่วนของสถาปัตยกรรมเป็นการใช้งานของ CNN กับ LSTM ประกอบด้วย ConV 8 layer และ BLSTM 1 layer ไม่มีชั้นของ pooling layer ผลการทดลองเป็นดังนี้ แบบจำลอง MAX-OUT ทั้งส่วนของ CNN และ LSTM ให้ประสิทธิภาพสูงกว่าแบบจำลองอื่นอยู่ที่ 87.6 % แต่ก็เพิ่มมาด้วยระยะเวลาการฝึกสอนสูงที่สุดที่ 7.8 ชั่วโมงซึ่งมากกว่าไม่ใช้ถึง 3 เท่า และ

การเพิ่มจำนวนชั้นของ MAX-OUT นั้นสามารถเพิ่มประสิทธิภาพได้ แต่จำนวนชั้นที่มากเกินไปไม่สามารถจะเพิ่มประสิทธิภาพได้สูงขึ้นและกินเวลานาน ในการทดลองการเพิ่มจำนวนชั้น ประสิทธิภาพที่ดีที่สุดอยู่ที่ 4 ชั้น 87.6% การเพิ่มเป็น 5 ชั้น ไม่สามารถเพิ่มประสิทธิภาพได้มาขึ้น 86.3% และใช้เวลาในการฝึกสอนมากถึง 10 ชั่วโมง

Method ($k = 4$ for maxout)	Accuracy (%)
Auto-encoder with tanh-BLSTM [3]	84.5
ReLU-CNN with tanh-BLSTM	84.6
ReLU-CNN with maxout-BLSTM	84.4
maxout-CNN with tanh-BLSTM	85.6
maxout-CNN-BLSTM	87.6

รูปที่ 2.11 ตารางการเปรียบเทียบความแม่นยำของตัวจำแนกในแต่ละแบบจำลอง

Method ($k = 4$ for maxout)	Time (hr)
ReLU-CNN with tanh-BLSTM	2.4
ReLU-CNN with maxout-BLSTM	2.5
maxout-CNN with tanh-BLSTM	7.8
maxout-CNN-BLSTM	7.8

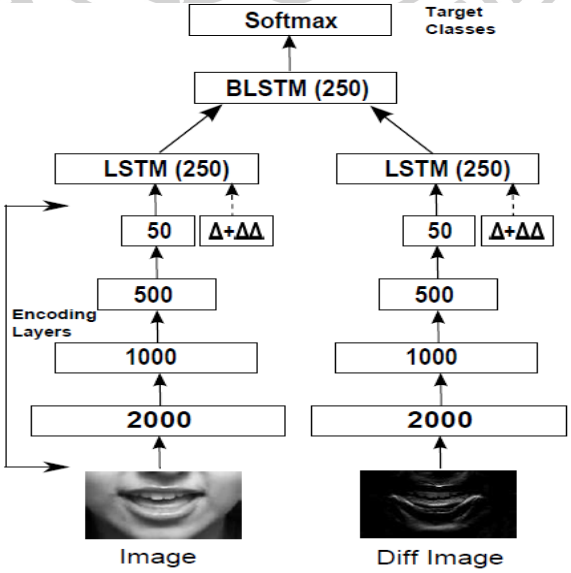
รูปที่ 2.12 ตารางการเปรียบเทียบระยะเวลาที่ใช้ฝึกสอนในแต่ละแบบจำลอง

maxout-CNN-BLSTM	Accuracy (%)	Time (hr)
$k = 2$	85.6	4.2
$k = 3$	86.1	6.2
$k = 4$	87.6	7.8
$k = 5$	86.3	10.0

รูปที่ 2.13 ตารางการเปรียบเทียบการเพิ่มจำนวน Feature maps (Maxout)

2.5.4 End-To-End Visual Speech Recognition With LSTMS (Stavros Petridis, Zuwei Li, Maja Pantic (2017)) [11]

บทความนี้นำเสนอสถาปัตยกรรมที่ใช้ในการอ่านริมฝีปากโดยใช้ Deep Learning ในส่วนของ LSTM เนื่องจากในช่วงเวลานั้นงานวิจัยในเรื่องของการอ่านริมฝีปากมีการใช้งาน LSTM ค่อนข้างน้อย แบบจำลองนี้เป็นแบบจำลองแรกที่สามารถเรียนรู้ที่จะสกัดคุณลักษณะออกมาจากพิกเซลได้โดยตรงและสามารถทำการจำแนกได้พร้อมๆกัน ซึ่งเป็นการทำงานที่มีความก้าวหน้าในการอ่านริมฝีปากในช่วงเวลานั้น โดยแบบจำลองจะถูกแบ่งออกเป็นสองส่วนที่จะทำงานไปพร้อมๆกันในการสกัดคุณลักษณะจากรูปภาพของปากและรูปภาพที่ผ่านการ Diff โดยทั้งสองส่วนนี้สร้างจาก LSTM เรียกว่า Bidireactional LSTM (BLSTM) ภาพรวมของอัลกอริทึมแสดงได้ ดังรูป ในส่วนด้านซ้ายจะเป็นการสกัดคุณลักษณะจากภาพบริเวณริมฝีปากโดยตรงเป็นข้อมูลแบบ static ในส่วนด้านขวาจะเป็นการสกัดคุณลักษณะจากภาพที่ผ่านการ Diff เพื่อเก็บคุณลักษณะการเคลื่อนไหวในขณะนั้นเป็นข้อมูลแบบ Dynamic โดยฐานข้อมูลที่ใช้จะมีสองฐานข้อมูล 1) OuluVS2 2) CUAVE ฐานข้อมูล OulUVS2 ประกอบด้วยผู้พูด 52 พูด 10 คำ 3 ครั้งทั้งหมด 156 ครั้งในการพูด 1 คำ ต่อ 1 คน เป็นคำภาษาอังกฤษ และฐานข้อมูล CUAVE มีคนพูด 36 คน พูดเลข 0 – 9 จำนวน 5 ครั้ง ในแต่ละคน ทั้งหมดจะมี 180 ครั้งในการพูด 1 เลข ต่อ 1 คน รูปภาพริมฝีปากที่ถูกตัดออกมาจะถูกลดขนาดเหลือ 30x50 พิกเซล ในการทดลองจะทดลองทั้งสองฐานข้อมูลเปรียบเทียบผลลัพธ์ในแต่ละฐานข้อมูล



รูปที่ 2.14 สถาปัตยกรรมที่นำเสนอ

Method	Classification Accuracy
End-to-End (Raw Image)	78.0
End-to-End (Diff Image)	75.8
End-to-End (Raw + Diff Images, Fig. 1)	84.5
DCT + HMM [22] †	74.8
Latent Variable Models [22] †	73.0

รูปที่ 2.16 ตารางการเปรียบประสิทธิภาพความแม่นยำของการจำแนกในแต่ละแบบจำลองของฐานข้อมูลOuluVS2

Method	Classification Accuracy
End-to-End (Raw Image)	71.4
End-to-End (Diff Image)	65.9
End-to-End (Raw + Diff Images, Fig. 1)	78.6
Deep Autoencoder + SVM [4]	68.7
Deep Boltzmann Machines + SVM [23]	69.0
AAM +HMM [24] †	75.7
Patch-based Features + HMM [25] ∗	77.1
Visemic AAM + HMM [26] † ‡	83.0

รูปที่ 2.15 ตารางการเปรียบเทียบประสิทธิภาพความแม่นยำของการจำแนกในแต่ละแบบจำลองของฐานข้อมูล CUAVE

2.5.5 Lip Reading Word Classification (Abiel Gutierrez, Zoe-Alanah Robert (2016)) [12]

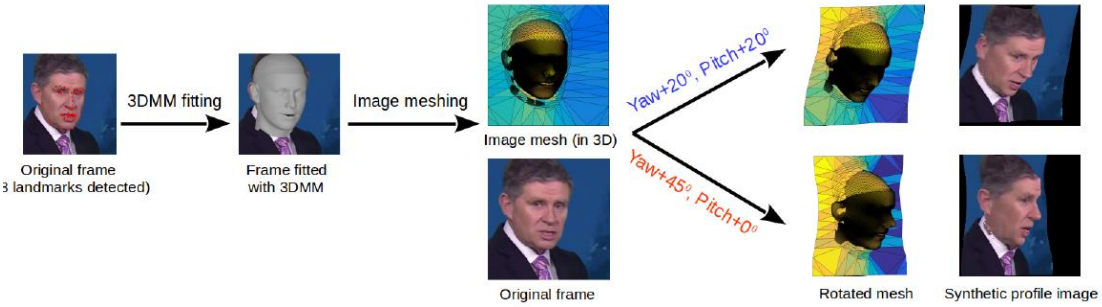
บทความนี้นำเสนอการอ่านริมฝีปากจากการศึกษาการทำงานร่วมกันของสถาปัตยกรรม CNN และ LSTM ในรูปแบบต่างๆ และนำมาเปรียบเทียบผลลัพธ์ของประสิทธิภาพของการจำแนกในแต่ละแบบจำลองที่สร้างขึ้น ประกอบด้วย แบบจำลองที่สร้างจาก 1) CNN+LSTM แบบพื้นฐาน 2) Deep layered CNN + LSTM 3) ImageNet Pretrained VGG-16 Features + LSTM 4) Fine-tuned VGG-16 + LSTM ชุดข้อมูลที่ใช้มีชื่อว่า MIRACL-CV1 โดยภายในประกอบไปด้วย ภาพสีที่มีความตื้นลึกของผู้พูดทั้งหมด 15 คน โดยที่แต่คนละจะทำการออกเสียงคำพูด 10 คำ และ 10 วลี 10 ครั้งในแต่ละรอบ ความละเอียด 640x480 พิกเซล วิดีโอมีเฟรมเรทที่ 15 เฟรมต่อวินาที มีการใช้เฟรมตั้งแต่ 4 – 27 เฟรม ตัวอย่าง คำ และ วลีที่ใช้มีดังนี้ I am sorry, start, begin, Nice to meet you เป็นต้น ภายในชุดข้อมูลมีการเพิ่มจำนวนอินพุตโดยการทำ Data Augmentation ประกอบไปด้วยการทำ การเพิ่มขนาดของรูปภาพเป็น 3 เท่าในแนวตั้ง มีการกลับด้านรูปภาพ และสุ่มการเพิ่ม noise ในแต่ละรูป การพูดในแต่ละคำจะนำในแต่ละเฟรมเป็นอินพุตแบบอนุกรมโดยที่ไม่มีการเลือกเฟรมและนำเข้าแบบจำลองเพื่อทำการฝึกสอนและได้ผลลัพธ์ดังนี้

Model	**Training**	**Validation**	**Test**
Baseline	85%	64%	39%
Deep CNN + LSTM	52%	39%	25%
Frozen VGG + LSTM	100%	76%	55%
Fine-tuned VGG + LSTM	100%	79%	59%

รูปที่ 2.17 ตารางการเปรียบความประสิทธิภาพความแม่นยำการจำแนกในรูปแบบของการฝึกสอน การตรวจสอบและการทดสอบของแบบจำลองในแต่ละแบบจำลอง

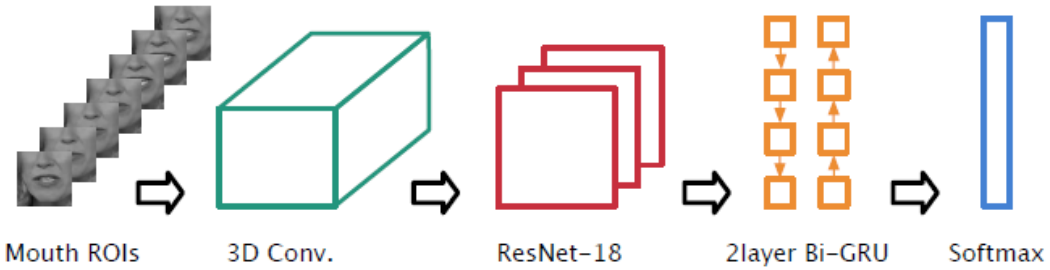
2.5.6 Towards Pose-Invariant Lip-Reading (Shiyang Cheng, Pingchuan Ma, Georgios Tzimiropoulos, Stavros Petridis, Adrian Bulat, Jie Shen, Maja Pantic (2020)) [15]

บทความนำเสนอแบบจำลองการอ่านริมฝีปากจากการฝึกสอนของท่าทางในรูปแบบต่างๆ จากข้อมูลที่สังเคราะห์ขึ้นมาเองแทนการเก็บรวบรวม แบบจำลองที่นำเสนอมีประสิทธิภาพเพิ่มขึ้นอย่างมากจากงานที่นำเสนอมาก่อนหน้าในด้านของมุมมองหน้าที่ไม่หันตรง และยังคงมีประสิทธิภาพที่เหนือกว่ากับงานที่ใช้ในมุมมองที่หันตรงหรือมุมมองที่ใช้บริเวณรอบๆริมฝีปาก โดยการใช้แบบจำลองชื่อว่า 3D Morphable Model (3DMM) กับฐานข้อมูล LRW และฐานข้อมูล LRS2



รูปที่ 2.18 การทำ Data Augmentation จาก 3DMM

ในส่วนของการทำ Pose Augmentation เป็นการสร้างแบบจำลอง 3D เพื่อขยับโครงหน้าแบบสามมิติจากฐานข้อมูล LRW และการทำ Augmentation ในแบบ 2D คือการ 1) การเพิ่มหรือลดขนาดจาก 0.8x - 1.2x 2) การทำ downsampling ของภาพบริเวณปาก 0.4-0.8 ของขนาดในแต่ละภาพ จากนั้นก็ upsampling กลับมาในขนาดของรูปดังเดิม 3) สุ่มการเพิ่ม noise บริเวณรอบปาก



รูปที่ 2.19 สถาปัตยกรรมที่นำเสนอ

ฐานข้อมูลที่ใช้ในการเปรียบเทียบการทดลองประกอบไปด้วย 2 ฐานข้อมูล

1) LRW เป็นฐานข้อมูลที่มีข้อมูลทั้งภาพและเสียงที่มี 500 คำที่แตกต่างกัน ภายในประกอบด้วยผู้พูด 1000 คน ในแต่ละการออกเสียงจะมี 29 เฟรม ฐานข้อมูลถูกแบ่งออกเป็น การฝึกสอน 800 คำในแต่ละคลาส และการตรวจสอบและการทดสอบแบ่งออกเป็น 50 คำในแต่ละคลาส

2) LRS2 เป็นฐานข้อมูลแบบภาพและเสียงและข้อความจากใบหน้าผู้พูดเก็บรวมรวบจาก BBC TV มีความแปรปรวนของความยาวในการพูดมาก และท่าทางตำแหน่งศีรษะของผู้พูด มีการแบ่งออกมาเป็นคำที่ใช้ในการฝึกสอน

ผลการทดลองเป็นไปดังนี้

Models	Accuracy (%) on different test sets			
	LRW	LP	LRS2	LRS2-Ba
M[LRW]	82.78	69.86	57.05	54.39
M[LP]	81.67	79.08	57.25	54.43
M[LRW+LP]	83.08	79.38	58.86	56.02
M[LRW]+Aug2D	**83.20**	72.14	58.84	56.07
M[LRW+LP]+Aug2D	83.08	**79.53**	**59.60**	56.78
M[LRW+LRS2-Ba]	82.73	69.62	-	**59.59**

รูปที่ 2.20 ตารางการเปรียบประสิทธิภาพความแม่นยำจากชุดทดสอบในแต่ละชุดในแต่ละแบบจำลอง

Models	Accuracy (%) on different poses				
	0°-15°	15°-30°	30°-45°	45°-60°	60°-90°
M[LRW]	58.11	55.18	49.62	41.73	23.07
M[LRW]+Aug2D	60.26	55.64	50.55	44.64	29.41
M[LRW+LP]	59.19	55.77	50.67	48.16	38.99
M[LRW+LP]+Aug2D	59.69	56.24	52.86	**49.54**	**39.68**
M[LRW+LRS2-Ba]	**63.42**	**59.24**	**54.29**	49.34	35.76

รูปที่ 2.21 ตารางการเปรียบเทียบประสิทธิภาพจากคลิปข้อมูลทดสอบ LRS2-Ba ในแต่ละมุมของการหัน

Models	Accuracy (%) on different poses				
	0°-15°	15°-30°	30°-45°	45°-60°	60°-90°
M[LRW]	55.77	44.86	28.22	9.52	7.89
M[LRW]+Aug2D	57.35	47.2	32.78	9.52	10.53
M[LRW+LP]	57.08	48.28	**38.59**	**30.16**	**23.68**
M[LRW+LP]+Aug2D	57.77	49.95	37.34	26.98	21.05
M[LRW+LRS2-Ba]	**60.79**	**51.6**	34.44	22.22	10.53

รูปที่ 2.22 ตารางการเปรียบเทียบประสิทธิภาพจากคลิปข้อมูลทดสอบ LRS2-Ba ในแต่ละมุมของการก้มหน้า

2.5.7 การเปรียบเทียบข้อดีและข้อจำกัดของงานวิจัยข้างต้น

ตารางที่ 2.1 การเปรียบเทียบข้อดีและข้อจำกัดของงานวิจัยข้างต้น

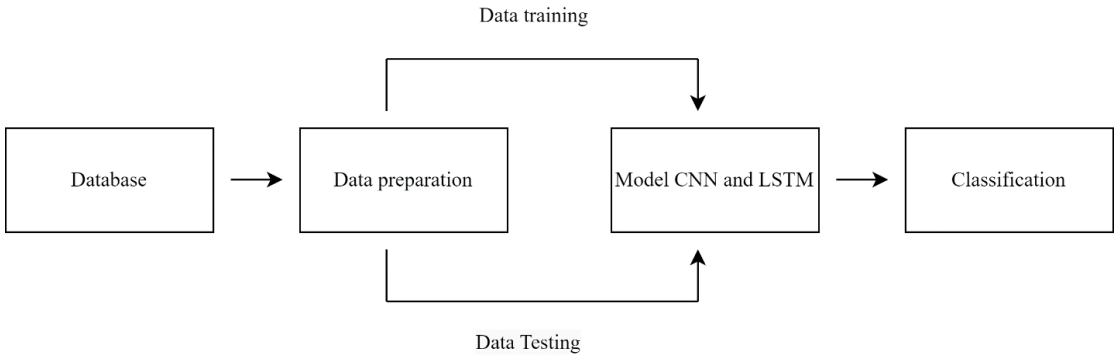
ชื่อบทความวิจัย	ข้อดี	ข้อจำกัด
A novel lip reading algorithm by using localized ACM and HMM:Tested for digit recognition (2014)	เป็นวิธีการใหม่ในช่วงเวลานั้น โดยการใช้ HMM model	เป็นวิธีการเก่า โดยในปัจจุบันใช้ Deep Learning ให้ประสิทธิภาพดีกว่า
Improving the Recognition Performance of Lip Reading Using the Concatenated Three Sequence Keyframe Image (2021)	ใช้ CNN และนำเสนอเทคนิคการเลือกเฟรมสำคัญเพื่อลดจำนวนรูปภาพ	-
End-To-End Low-Resource Lip-Reading With Maxout CNN And LSTM (2018)	Activation Unit Max-out เพิ่มประสิทธิภาพได้สูง	ใช้เวลาในการฝึกสอนนาน
End-To-End Visual Speech Recognition With LSTMS (2017)	เพิ่มประสิทธิภาพโดยใช้ LSTM ผ่าน Raw and diff image	-
Lip Reading Word Classification (2016)	ประสิทธิภาพของการใช้ pre-train model ค่อนข้างสูง	เกิด overfiting กับรูปภาพใหม่ที่ไม่เคยเห็น ทำให้ประสิทธิภาพที่ได้จากการทดสอบน้อย
Towards Pose-Invariant Lip-Reading (2020)	เปลี่ยนฐานมูลที่เป็นแบบ frontal ให้เอียงในลักษณะต่างๆ และนำไปทดสอบกับข้อมูลแบบเอียงหน้า	-

บทที่ 3
วิธีดำเนินการวิจัย

ในส่วนของวิธีการดำเนินการวิจัยในวิทยานิพนธ์เล่มนี้ได้ออกแบบและพัฒนาแบบจำลองการอ่านริมฝีปาก โดยได้ศึกษาค้นคว้าข้อมูลพบว่า ข้อมูลส่วนใหญ่ที่ใช้เป็นฐานข้อมูลภาษาอังกฤษที่ได้รับความนิยมและมีความเป็นสากล โดยข้อมูลดังกล่าวอยู่ในรูปแบบของคลิปวิดีโอ หรือลำดับของรูปภาพที่เรียงต่อกันในเชิงเวลา (Temporal Keyframe Sequence) และมีความหลากหลายตามความต้องการของผู้วิจัยที่ต้องการจะกำหนดขอบเขตของงาน เช่น ชุดข้อมูลที่เป็นตัวเลข ชุดข้อมูลที่เป็นคำหรือวลี ชุดข้อมูลที่เป็นตัวอักษร ตลอดจนชุดข้อมูลที่เป็นประโยค และในงานวิจัยนี้ได้กำหนดขอบเขตของการวิจัยโดยใช้ชุดข้อมูลที่เป็นตัวเลขในภาษาอังกฤษตั้งแต่เลข 0 ถึงเลข 9 จากฐานข้อมูลภาษาอังกฤษที่มีความเป็นสากลชื่อว่า AV Digits โดยในปัจจุบันมีการนำเสนองานวิจัยจำนวนมากที่เกี่ยวข้องกับการศึกษาการอ่านริมฝีปากและจากการค้นคว้าพบว่า มีความน่าสนใจเกี่ยวกับการใช้งานสถาปัตยกรรมแบบ CNN ที่ใช้งานร่วมกับ LSTM โดยสถาปัตยกรรมทั้งสองมีจุดเด่นของการสร้างแบบจำลองที่นำมาใช้งานร่วมกันได้คือ การเรียนรู้ของภาพและการเรียนรู้ของความสัมพันธ์ของข้อมูลที่เปลี่ยนตามเวลา ในงานวิจัยนี้จึงเลือกใช้ CNN ที่ทำงานร่วมกับ LSTM เป็นส่วนสำคัญของการสร้างแบบจำลองการอ่านริมฝีปาก

เนื่องจากการทดลองพบว่าข้อมูลจากฐานข้อมูลนั้นมีความหลากหลายของผู้พูดจำนวนมากที่รวมถึงวิธีการพูดของแต่ละผู้พูดที่มีความแตกต่างกัน บางผู้พูดมีการเปิดปากค้างเอาไว้แทบจะตลอดเวลา บางผู้พูดสามารถเปล่งเสียงออกมาได้โดยแทบจะไม่เห็นการขยับของริมฝีปาก ด้วยเหตุผลนี้ ทำให้แบบจำลองไม่สามารถที่จะเรียนรู้ข้อมูลที่มีจำนวนมากเกินไปและส่งผลกระทบต่อการเรียนรู้ทำให้ไม่สามารถสร้างแบบจำลองได้ โดยในบทนี้จะกล่าวถึงขั้นตอนของการดำเนินการวิจัยและอธิบายโครงสร้างของแบบจำลองการอ่านริมฝีปาก โดยเริ่มจากภาพรวมของการสร้างแบบจำลองการอ่านริมฝีปาก วิธีการที่ได้มาซึ่งข้อมูลริมฝีปากที่พร้อมสำหรับการสร้างแบบจำลอง รวมถึงแบบจำลองที่สร้างเพื่อใช้ในการอ่านริมฝีปากของงานวิจัยเล่มนี้

3.1 ภาพรวมของวิธีการสร้างแบบจำลองการอ่านริมฝีปาก



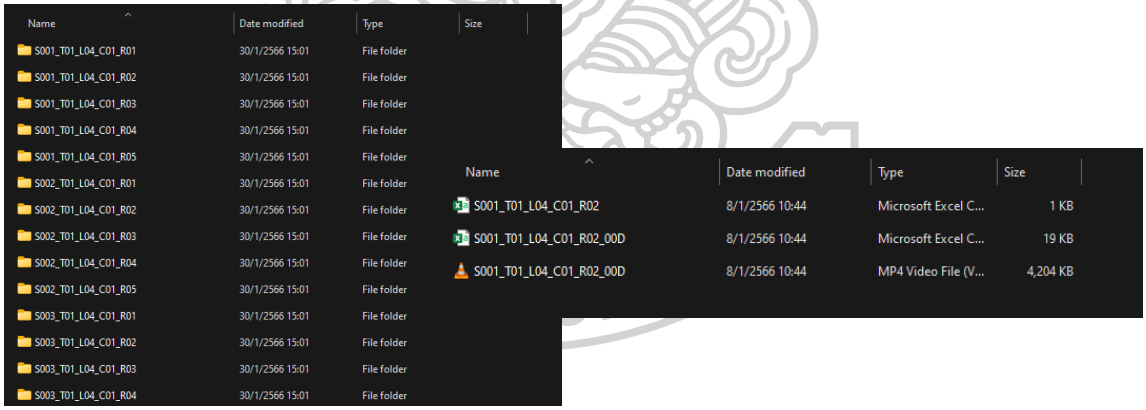
รูปที่ 3.1 ภาพรวมของขั้นตอนการสร้างแบบจำลองการอ่านริมฝีปาก

ดังที่ได้กล่าวไปแล้วในงานวิจัยนี้ได้นำฐานข้อมูลที่ชื่อว่า AV Digits มาใช้ โดยได้กำหนดขอบเขตของงานคือชุดข้อมูลที่เป็นตัวเลข 0 ถึง 9 ในภาษาโดยมีตัวเลือกเพิ่มเติมคือใช้ข้อมูลจากผู้พูดแบบหน้าตรงภายในฐานข้อมูลดังกล่าวเท่านั้น โดยข้อมูลดังกล่าวมาจากผู้เข้าร่วมการทดลองทั้งหมด 53 คน 16 สัญชาติที่มีทั้งเจ้าของภาษาและไม่ใช่เจ้าของภาษา ซึ่งถือว่าเป็นข้อมูลที่มีความหลากหลาย เมื่อเราได้ฐานข้อมูลพร้อมทั้งขอบเขตของการทดลองของเราแล้ว ข้อมูลทั้งหมดจะต้องผ่านการคัดเลือกจำนวนเฟรมในขั้นตอนของการเตรียมการของข้อมูลใน block ที่ชื่อว่า Data preparation หลังจากนั้นข้อมูลจะแบ่งออกเป็น 2 กลุ่ม คือ กลุ่มของ Training เพื่อสร้างแบบจำลอง และกลุ่มของการ Testing เพื่อประเมินความสามารถที่แบบจำลองเราเรียนรู้ได้ เมื่อได้ 2 ส่วนนี้แล้วข้อมูลจะส่งเข้าแบบจำลองใน block ที่ชื่อว่า Model CNN and LSTM โดยข้อมูลที่จะส่งเข้าแบบจำลองจะถูกเรียงลำดับเฟรมอย่างถูกต้องเอาไว้แล้วตามจำนวนเฟรมที่มีการทดลอง แบบจำลองจะทำการเรียนรู้ข้อมูลที่เป็นรูปภาพไปพร้อมๆกับเรียนรู้ความสัมพันธ์ของการเปลี่ยนแปลงในแต่ละรูปภาพที่ส่งเข้ามา และผลลัพธ์สุดท้ายอยู่ที่ block ที่ชื่อว่า Classification ที่จะบอกค่าการเรียนรู้ต่างๆ ของแบบจำลอง ดังนี้ Accuracy Val Accuracy Loss Val Loss โดยภาพรวมของการสร้างแบบจำลองการอ่านริมฝีปากแสดงได้ดังรูป 3.1

3.2 การเตรียมการข้อมูล

3.2.1 ฐานข้อมูล Av Digits

ภายในฐานข้อมูลที่ชื่อว่า AV Digits นั้นเป็นฐานข้อมูลที่สร้างขึ้นเพื่อการศึกษาเกี่ยวกับการอ่านริมฝีปากโดยเฉพาะโดยภายในประกอบไปด้วยข้อมูลที่สามารถนำมาใช้เพื่อการอ่านริมฝีปากจำนวนโดยได้รวบรวมข้อมูลเอาไว้ 2 ชุด คือ ชุดข้อมูลที่เป็นตัวเลขในภาษาอังกฤษ 0 – 9 และชุดข้อมูลที่เป็นวลีสั้นๆ 10 วลีในภาษาอังกฤษอีกทั้งยังมีโหมดการพูดให้เลือกอีก 3 โหมดนั่นคือ โหมดปกติ โหมดเสียงกระซิบ และโหมดไม่มีเสียง รวมไปถึงมุมกล้องของการถ่ายทำทั้งหมด 3 มุม คือ หน้าตรง เอียง 45 องศา และด้านข้าง โดยมีผู้เข้าร่วมการทดลองสร้างฐานข้อมูลนี้ทั้งหมด 53 คน จาก 16 สัญญาติ ผู้เข้าร่วมจะทำการบันทึกการพูดของตัวเองที่ได้รับมอบหมายทั้งหมด 5 ครั้ง โดยในแต่ละครั้งจะไม่มีครั้งไหนที่ลำดับการพูดจะซ้ำกัน และใช้ปุ่ม space bar เพื่อแยกแต่ละคำพูดบอกเพื่อบอกเวลาและบันทึกลงไฟล์ excel ดังนั้น เมื่อเราต้องการใช้ข้อมูลจากทั้ง 53 ผู้เข้าร่วม และใช้หน้าตรงจากชุดข้อมูลที่เป็นตัวเลข 0 – 9 ในภาษาอังกฤษ เราจะมีข้อมูลทั้งหมด 2,650 ตัวอย่างที่จะนำมาใช้ในการเตรียมข้อมูลสำหรับสร้างแบบจำลองนี้



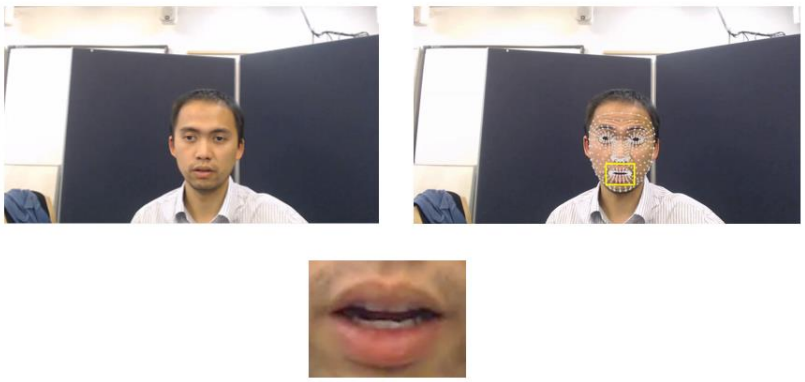
รูปที่ 3.2 โฟลเดอร์ที่ได้หลังจากดาวน์โหลดข้อมูลจากฐานข้อมูล AV Digits

เมื่อดาวน์โหลดข้อมูลที่ต้องการมาแล้ว เราจะได้ข้อมูลดังรูปที่ 3.2 โดยชื่อของโฟลเดอร์จะระบุลักษณะเฉพาะของชุดข้อมูลนั้น โดยที่ S ย่อมาจาก Subject ซึ่งหมายถึงลำดับที่ของผู้พูด T ย่อมาจาก Task หมายถึง งานที่ได้รับมอบหมายเช่น 01 หมายถึงชุดข้อมูลตัวเลขเป็นต้น L ย่อมาจาก Language หมายถึงภาษา ฐานข้อมูลนี้มีภาษาเดียวซึ่งก็คือภาษาอังกฤษโดยจะแทนด้วย L04 C ย่อ

มาจาก Condition หมายถึงโหมดในการพูดชุดข้อมูลนั้น C01 หมายถึงโหมดปกติ และ R ย่อมาจาก Repetition หมายถึงจำนวนรอบที่ทำการพูดมีตั้งแต่ R01 ถึง R05 ในแต่ละผู้พูดของแต่ละโหมดการพูด ในแต่ละโฟลเดอร์สิ่งที่ให้มาคือไฟล์ .CSV ที่บอกถึงช่วงเวลาของตัวเลขที่พูดและลำดับของเลขที่ผู้พูดนั้นๆพูดหรือเรียกว่า timestamp file โดยชื่อต่างๆ จำเป็นต้องรู้เพื่อทำการเข้าถึงที่อยู่อ้างอิงของโฟลเดอร์ต้นฉบับกับโฟลเดอร์ปลายทางเพื่อความถูกต้องของข้อมูลตัวอย่างของไฟล์ .CSV ใช้เพื่อแสดงความสัมพันธ์ของช่วงเวลากับลำดับการพูดของแต่ละเลขภายในคลิปวิดีโอแสดงได้ดังรูปที่ 3.3

Relative_Start_Time	Relative_Stop_Time	Absolute_S	Absolute_St	Utterance
0	18310000	1.5E+16	1.5E+16	9
18310000	35870000	1.5E+16	1.5E+16	5
35870000	54510000	1.5E+16	1.5E+16	0
54510000	71000000	1.5E+16	1.5E+16	6
71000000	87900000	1.5E+16	1.5E+16	7
87900000	108200000	1.5E+16	1.5E+16	2
108200000	123450000	1.5E+16	1.5E+16	4
123450000	142500000	1.5E+16	1.5E+16	3
142500000	161120000	1.5E+16	1.5E+16	1
161120000	180160000	1.5E+16	1.5E+16	8

รูปที่ 3.3 Time stamp file สำหรับแสดงความสัมพันธ์ของเวลากับลำดับการพูดของการพูดตัวเลขภายในคลิปวิดีโอ



รูปที่ 3.4 ภาพรวมของขั้นตอนการตัดริมฝีปาก

3.3.2 Face Detection and Face Localization บน Mediapipe

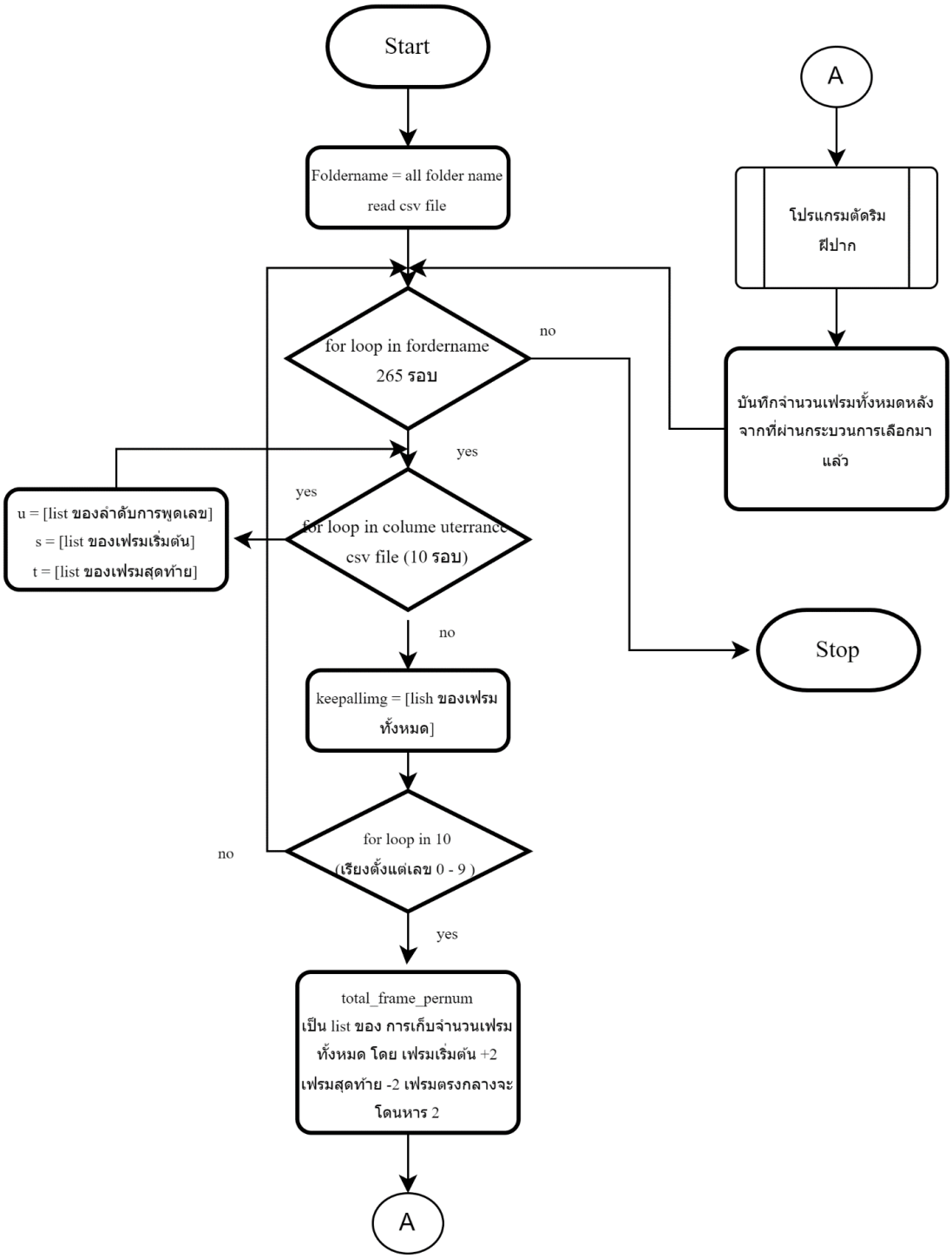
ในแต่ละโฟลเดอร์ของชุดข้อมูลจะประกอบไปด้วยคลิปวิดีโอสำหรับการอ่านริมฝีปากพร้อมกับไฟล์ time stamp เพื่อระบุช่วงเวลาดังกล่าว ในขั้นตอนการสกัดเอารูปริมฝีปาก โฟลเดอร์ทุกโฟลเดอร์จะถูกสร้างขึ้นโดยมีจำนวนเท่ากับโฟลเดอร์ที่ดาวน์โหลดมาจากฐานข้อมูล หลังจากนั้นจะเป็นขั้นตอนของการเขียนโปรแกรมเพื่อเชื่อมระหว่างไฟล์ time stamp กับ คลิปดิโอ โดยมีไลบรารี่ที่มีจำเป็นนั่นก็คือ Mediapipe ซึ่งเป็นไลบรารี่ที่มีทั้ง Face Detection และ Face Localization ที่ใช้เพื่อตรวจจับใบหน้าพร้อมๆ กับการกำหนดจุดบริเวณใบหน้า โดยสามารถระบุความแม่นยำของจุดได้ทั่วทั้งใบหน้า 468 จุด ภายในงานวิจัยนี้ได้คัดเลือกจุดที่เหมาะสมสำหรับนำมาใช้เพื่อการสกัดรูปภาพริมฝีปากมีทั้งหมด 4 จุด ดังนี้ 57 164 200 และ 287 โดยทั้ง 4 จุดที่เลือกมานั่นจะทำการวาดกรอบสี่เหลี่ยมเป็นบริเวณที่มีความเหมาะสมสามารถมองเห็นการเคลื่อนไหวของริมฝีปากได้ชัดเจนในระหว่างที่ผู้พูดทำการพูด รูปภาพของริมฝีปากที่ได้มานั้นจะไม่กว้างจนลดรายละเอียดของริมฝีปากมากเกินไปและไม่แคบจนมองไม่เห็นการเคลื่อนไหวส่วนอื่นๆ บริเวณรอบๆริมฝีปาก ตัวอย่างการใช้ไลบรารี่ Mediapipe ในขั้นตอนของการสกัดรูปภาพริมฝีปากแสดงได้ดังรูปที่ 3.4

3.2.3 ขั้นตอนที่ได้มาซึ่งข้อมูลริมฝีปาก 10 เฟรมสุดท้าย

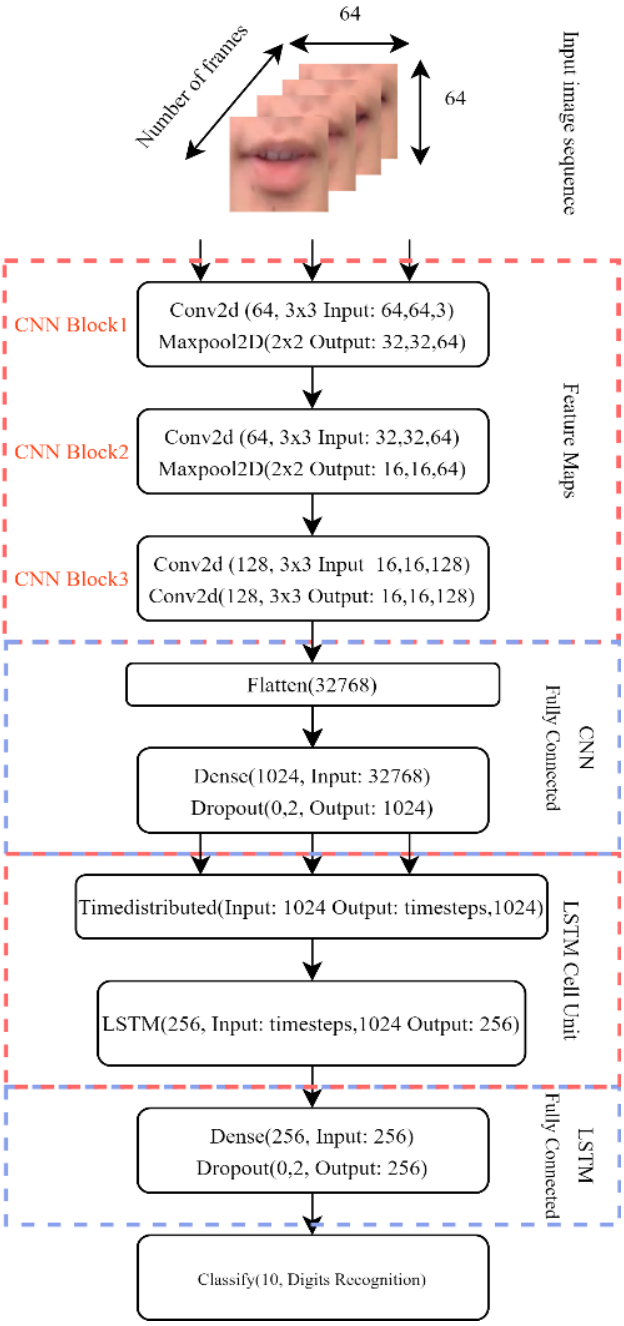
หลังจากที่กระบวนการของ Mediapipe ถูกนำมาใช้งานกับฐานข้อมูล AV Digits โดยรายละเอียดของการเขียนโปรแกรมเพื่อสร้างชุดข้อมูลที่ผ่านการเตรียมแล้วมีขั้นตอนดังนี้

เราจะทำการอ่านชื่อของโฟลเดอร์จากฐานข้อมูล AV Digits ที่เราดาวน์โหลดมา เพื่อทำการวนลูปการเข้าถึงข้อมูลภายในโดยข้อมูลภายในประกอบไปด้วย 2 ไฟล์ที่มีความสำคัญนั่นคือ ไฟล์วิดีโอและไฟล์ time tamp เราจะอ่านข้อมูลทุกอย่างมาจากไฟล์ time stamp และรู้ช่วงเวลาที่ผู้พูดพูดตัวเลขต่างๆ จากรูปที่ 3.3 เราจะได้ column ที่ชื่อ Relative_Start_time และ Relative_Stop_Time ค่าทั้งสองนี้จะถูกแปลงจากข้อมูลที่เกี่ยวข้องกับเวลาเป็นการระบุช่วงระยะของเฟรมคือการนำค่าดังกล่าวไปหารด้วย 107 และคูณด้วย 30 ที่เป็นค่าของเฟรมเรทของคลิปวิดีโอและเป็นคอลัมที่แสดงถึงลำดับเฟรมที่มีการเริ่มต้นการพูดตัวเลขนั้นและลำดับเฟรมสุดท้ายที่ตัวเลขนั้นถูกพูด และอีก 1 column ชื่อ Utterance เป็นคอลัมบอกถึงตัวเลขที่กำลังพูดในขณะนั้นโดยใน 1 ตัวเลขที่พูดมีจำนวนเฟรมมากมายตั้งแต่หลักสิบจนถึงหลักร้อยเฟรม เมื่อได้ข้อมูลตัวเลขทั้ง 10 จากไฟล์ time stamp และ เราจะอ่านข้อมูลจากคลิปวิดีโอ และทำการวนลูปไปตามชุดข้อมูลของไฟล์ time stamp ที่อ่านมานั่นคือ 10 รอบ ในแต่ละรอบจะทำการเชื่อมข้อมูล Utterance กับลำดับเฟรมแรกของเลขนั้นและลำดับเฟรมสุดท้ายของแรกนั้น หลังจากนั้นเราจะทำการเขียนเฟรมแรกด้วยเฟรม

ที่อยู่ถัดจากเฟรมแรกจริงๆ ไป 2 เฟรม และเฟรมสุดท้ายด้วยเฟรมที่อยู่ก่อนเฟรมสุดท้ายจริงๆ 2 เฟรม หลังจากนั้นเราจะเลือกข้อมูลที่อยู่ระหว่างกลางด้วยการเลือกเฟรมที่มีการหาร 2 ลงตัว และทำการบันทึกเฟรมทั้งลงโฟลเดอร์ที่ถูกจัดเรียงและตรวจสอบความถูกต้องแล้ว หลังจากนั้นก็จะตรวจสอบข้อมูลทั้งหมดพร้อมทั้งเลือกออกมา 10 เฟรมสุดท้ายจากทั้งหมด และทำการเปลี่ยนชื่อไฟล์ให้มีการเรียงลำดับที่ถูกต้องเหมือนเดิม ภาพรวมของขั้นตอนการเลือกเฟรมแสดงได้ดัง flowchart ในรูปที่ 3.5



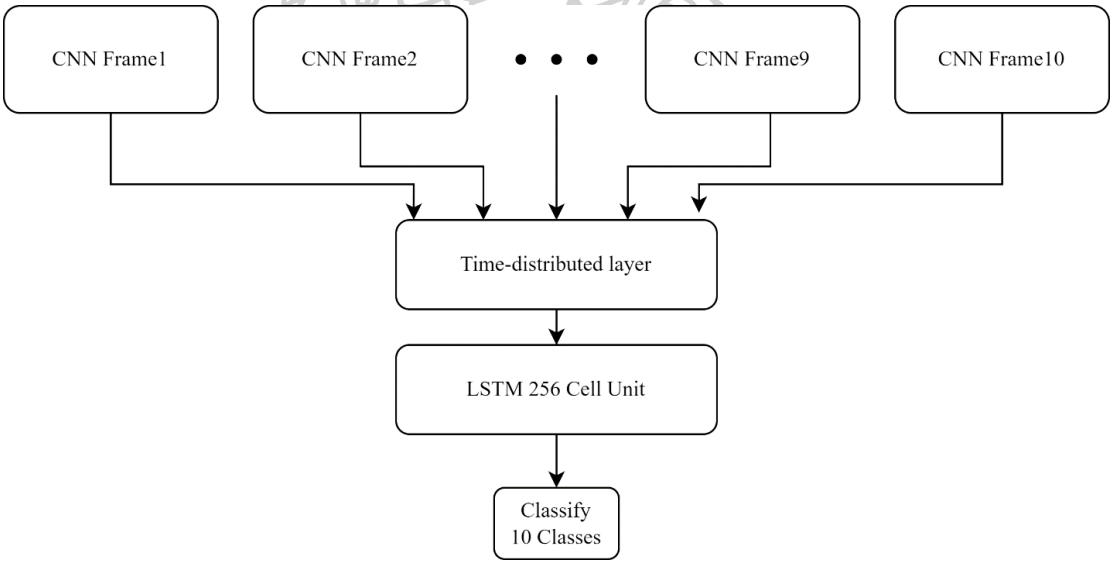
รูปที่ 3.5 flowchart แสดงการทำงานของขั้นตอนการเขียนโปรแกรมของการเลือกเฟรม



รูปที่ 3.6 สถาปัตยกรรมของแบบจำลอง CNN และ LSTM ที่นำเสนอเพื่อใช้ในการอ่านริมฝีปาก

3.3 แบบจำลอง CNN และ LSTM

แบบจำลอง CNN และ LSTM ทำงานร่วมกันผ่านชั้นที่ชื่อว่า time-distributed โดยในงานวิจัยเล่มนี้ต้องการที่จะนำเสนอการวิเคราะห์ของข้อมูลการอ่านริมฝีปากที่เรียกว่า keyframe โดยมีความหมายถึงเฟรมที่มีความสำคัญต่อการเรียนรู้ของแบบจำลอง โดยการวิเคราะห์เจาะจงไปที่คีย์เฟรมดังกล่าวที่มีความเกี่ยวข้องกับการเปลี่ยนแปลงตามเวลาที่เรียกว่า Temporal ซึ่งนำความหมายของทั้งสองคำมารวมกันจะหมายถึง คีย์เฟรมที่มีความสำคัญของการเรียนรู้ของแบบจำลองโดยจะเปลี่ยนแปลงไปตามช่วงเวลา ด้วยเหตุนี้จึงมีการเลือกจำนวนเฟรมและขนาดของคีย์เฟรมในการทดลอง ดังนั้นในขั้นตอนของการนำข้อมูลเข้าแบบจำลอง ต้องมีการระบุถึงขนาดของคีย์เฟรมและจำนวนของคีย์เฟรม โดยมีดังนี้ ขนาดของคีย์เฟรมมีทั้งแบบเต็มปากคือ 64×64 พิกเซลและครึ่งปากมีขนาดเป็น 64×32 พิกเซลและจำนวนเฟรมที่เลือกคือ 3 เฟรม 5 เฟรม และ 10 เฟรม โดยข้อมูลทั้งสองอย่างจะมีการเปลี่ยนแปลงตัวเลขภายในของแบบจำลอง ผ่านชั้นของ time-distributed ตัวอย่างเช่น การรับข้อมูลเข้ามาแบบ 10 เฟรมและเต็มปากจะแสดงได้ดังรูปที่ 3.6



รูปที่ 3.7 การทำงานของชั้น time-distributed

จากรูปที่ 3.7 จะแสดงการทำงานของชั้น time-distributed ซึ่งเป็นหัวใจสำคัญของการทำงานร่วมกันระหว่าง CNN และ LSTM เนื่องจากชั้นดังกล่าวจะจัดการข้อมูลที่มีความเป็นลำดับเกี่ยวเนื่องกับเวลาขึ้นอยู่กับ time step ที่เลือกในกรณีนี้คือ 10 เฟรม รูปภาพของริมฝีปากตั้งแต่เฟรมที่ 1 ถึง 10 จะผ่านการประมวลผลของแบบจำลองของ CNN แบบอิสระจากกัน ซึ่งเป็นการประมวลผลแบบ CNN ทั่วไปคือรูปภาพจะถูกนำผ่านสิ่งที่เรียกว่า kernel เพื่อสกัดเอาคุณลักษณะต่างๆ เช่น ภาพขอบ เส้นในแนวตั้งหรือแนวดิ่ง เส้นในแนวนอนหรือแนวขวาง เป็นต้น สิ่งเหล่านั้นจะถูกฝึกสอนและนำมาปรับค่า weight ในส่วนของการฝึกสอนของ CNN หลังจากนั้นจะมาร่วมกันในชั้นของ time-distributed เนื่องจากพารามิเตอร์สุดท้ายของแบบจำลอง CNN คือ 1024 เมื่อเข้าสู่ชั้น time-distributed จะมีเอาต์พุตอยู่ที่ 10,1024 โดยเอาต์พุตตั้งแต่ลำดับที่ 1 จะเข้าไปอัพเดทค่าใน LSTM ทั้ง 256 Hidden State วนไปจนครบ 10 รอบ โดยการเรียนรู้ของชั้น LSTM เป็นการเรียนรู้ผ่านชุดข้อมูลที่ได้จากการสกัดเอาคุณลักษณะออกมาแล้วของรูปภาพและค้นหาสิ่งที่มีความสัมพันธ์และเปลี่ยนแปลงไปในแต่ละเฟรมรูปภาพที่เข้ามาประมวลผล เพื่อสร้างความสัมพันธ์ของข้อมูลในแต่ละลำดับและทำการจำแนกเป็น 10 ประเภทของชุดข้อมูลตัวเลขในภาษาอังกฤษต่อไป

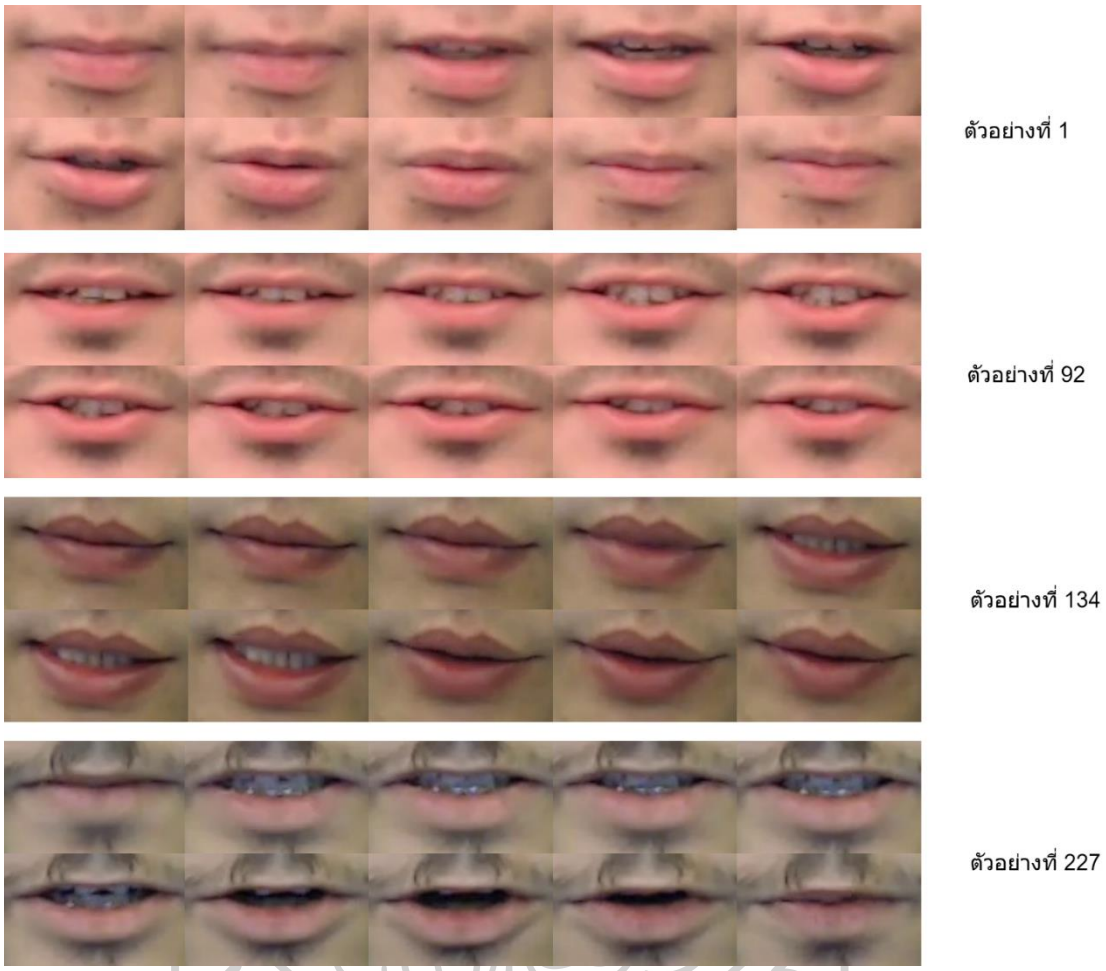
บทที่ 4
ผลการทดลองของงานวิจัย

การทดลองครั้งนี้ได้เขียนในรูปแบบของโปรแกรมภาษา Python ซึ่งมีความสามารถที่โดดเด่นของด้านของการประมวล AI และได้ใช้บริการ Cloud Computing จาก Google ที่ชื่อ Google Collaboratory โดยมีการสมัครการใช้บริการ Collab Pro ซึ่งสามารถเข้าถึงการใช้ประสิทธิภาพของ GPU A100 ที่มีประสิทธิภาพสูงกว่าคอมพิวเตอร์ที่ใช้ในเชิงพาณิชย์ในปัจจุบัน

4.1 ชุดข้อมูลที่ใช้ในการทดลอง

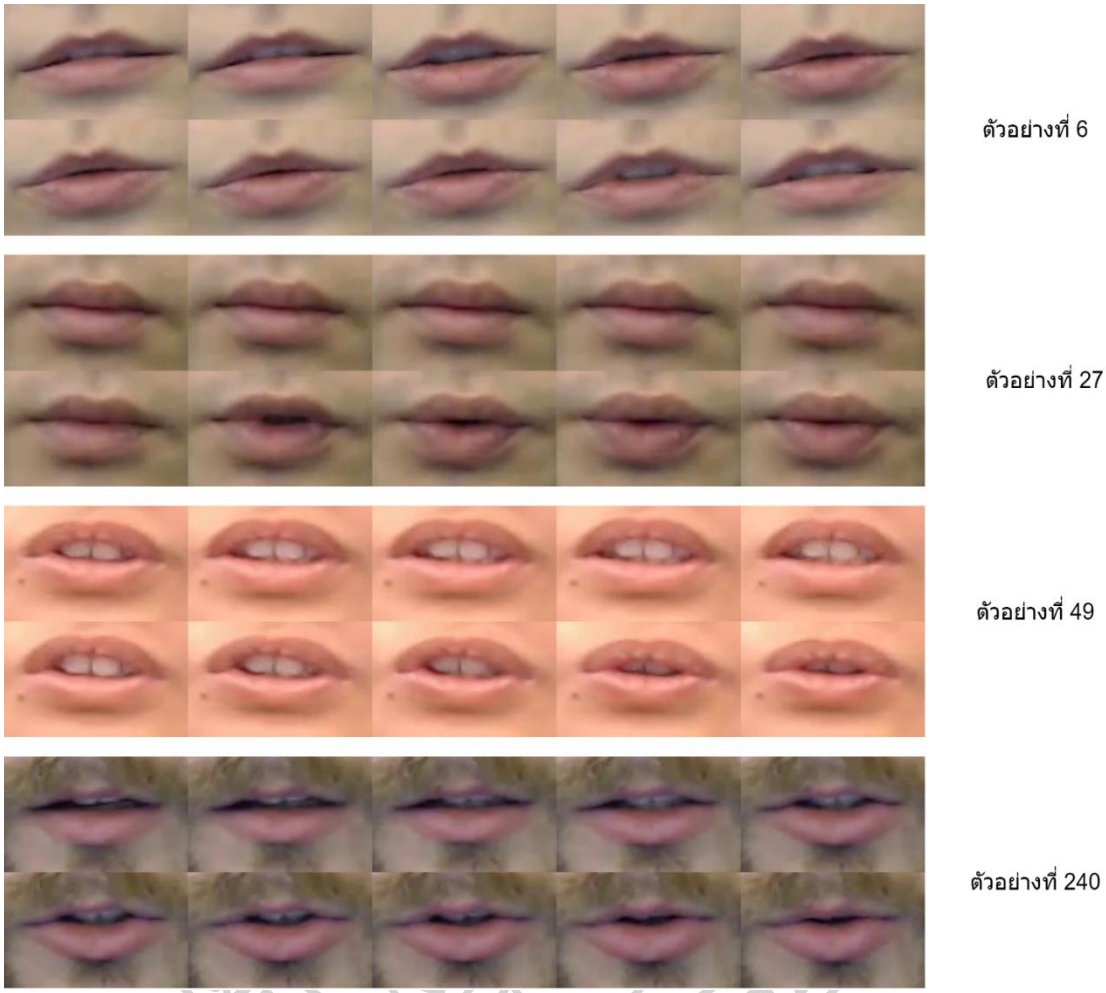
4.1.1 ชุดข้อมูลจากรูปภาพตามแบบต้นฉบับที่ตัดครอบมา

ในการทดลองสร้างแบบจำลองในการอ่านริมฝีปากในงานวิจัยนี้ใช้ฐานข้อมูลที่ชื่อว่า AVDIgit มีผู้เข้าร่วมการทดลองทั้งหมด 53 คน โดยแบ่งการทดลองออกมาเป็น 2 กลุ่มหลักนั่นคือ การทดลองกับรูปภาพแบบเต็มปากมีขนาดของรูปภาพอยู่ที่ 64x64 พิกเซลและการทดลองกับรูปภาพแบบครึ่งปากมีขนาดีรูปภาพอยู่ที่ 64x32 พิกเซล ชุดข้อมูลทั้ง 2 กลุ่มนี้เป็นชุดข้อมูลเดียวกันแตกต่างกันที่ขนาดของรูปภาพซึ่งส่งผลโดยตรงต่อการเรียนรู้ของแบบจำลองที่ขนาดของรูปภาพอินพุตหายไปครึ่งหนึ่งผลลัพธ์ที่ได้จากแบบจำลองมีความแตกต่างกันมากน้อยเพียงใด แต่ลำดับการเรียงของเฟรมเหมือนเดิมทุกประการ โดยชุดข้อมูลประกอบด้วย 10 คลาสที่เป็นตัวเลขภาษาอังกฤษ 0 – 9 จำนวนข้อมูลที่ใช้ในการฝึกสอนทั้งหมดอยู่ที่ 2120 ตัวอย่างแบ่งเป็นคลาสละ 212 ตัวอย่าง และข้อมูลสำหรับการทดสอบอยู่ที่ 530 ตัวอย่างแบ่งเป็นคลาสละ 53 ตัวอย่าง ตัวอย่างของชุดข้อมูลในแต่ละคลาสหลังจากที่ผ่านขั้นตอนการเตรียมข้อมูลแล้วสามารถแสดงได้ดังรูปที่ 4.1 ถึง 4.10 จากชุดข้อมูลตัวอย่างดังกล่าวจะสังเกตเห็นได้ว่า ต่อให้ข้อมูลนั้นมาจากผู้พูดคนเดียวกัน แต่ความกว้างของรูปปากหรือลักษณะของริมฝีปากในการพูดเลขนั้นไม่มีความเหมือนกันแต่อย่างใด และรวมไปถึงการพูดของผู้เข้าร่วมการทดลองมีอุปนิสัยของการพูดที่ไม่ค่อยขยับริมฝีปากหรือไม่การขยับที่น้อยก็ตาม สิ่งเหล่านี้คือความท้าทายในการสร้างแบบจำลองที่ได้กล่าวไปข้างต้น

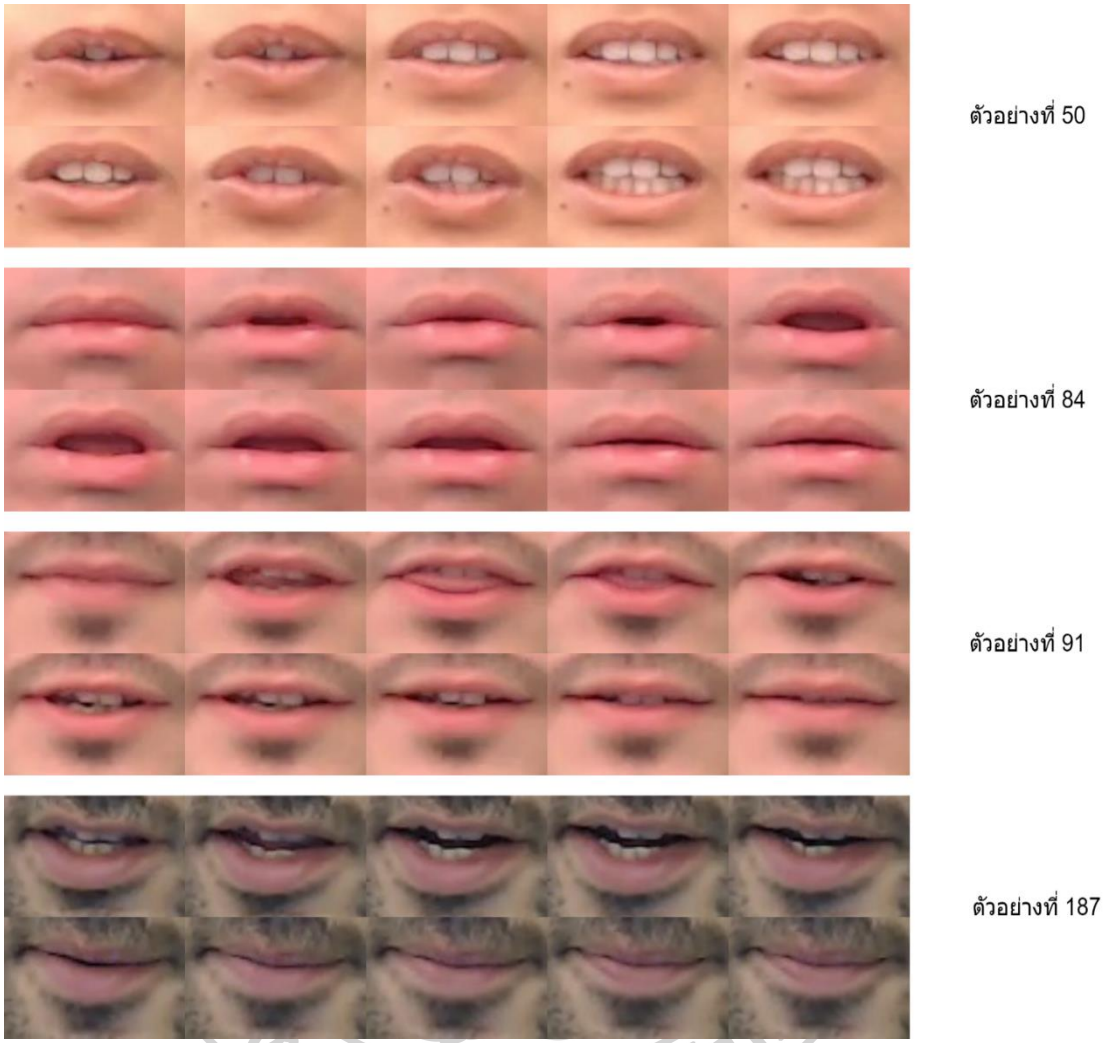


รูปที่ 4.1 ตัวอย่างชุดข้อมูลของการพูดเลขศูนย์ทั้งสิบเฟรมแถวบนซ้ายไปขวาเรียงลำดับเฟรมที่ 1 – 5 แถวล่างซ้ายไปขวาเรียงลำดับเฟรมที่ 6 – 10

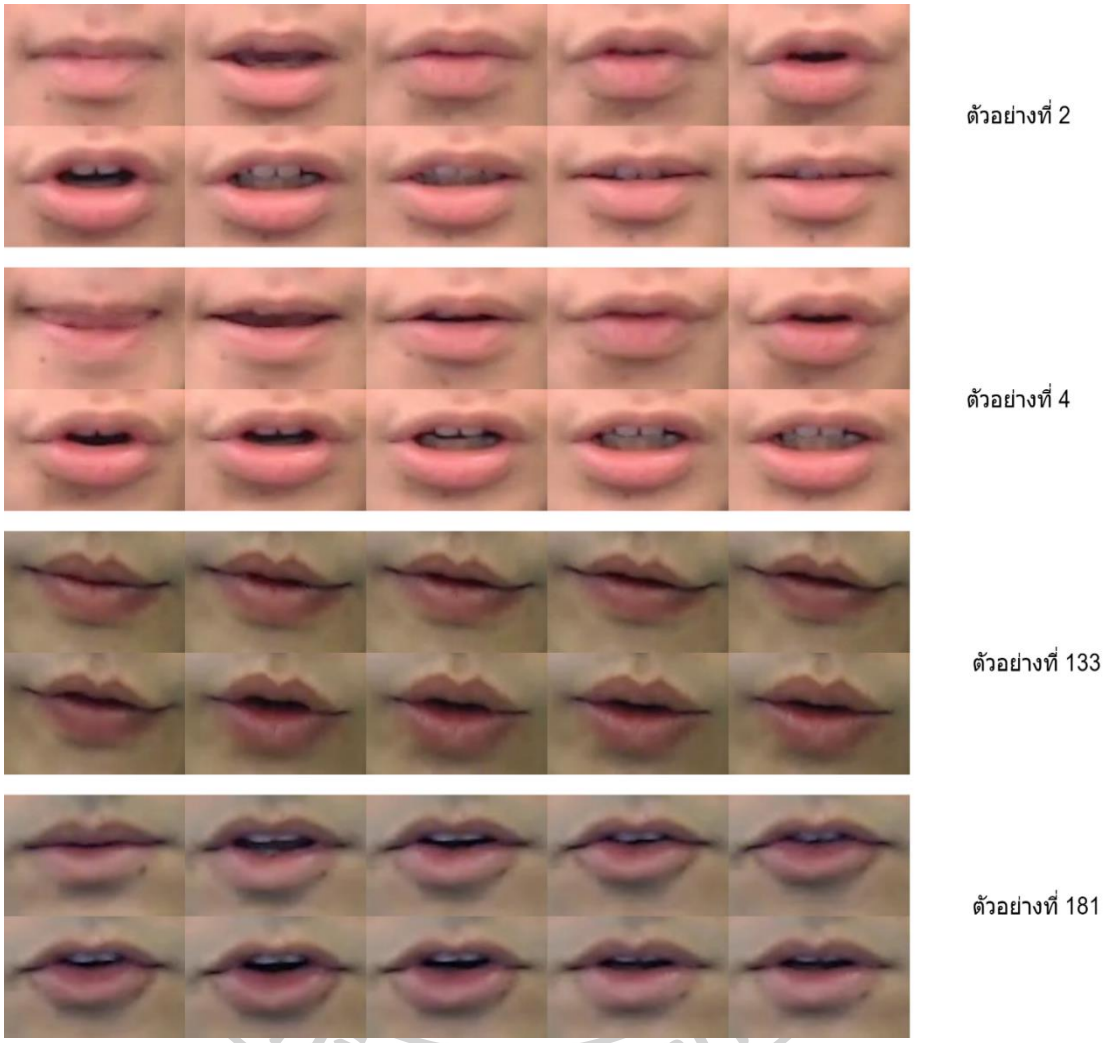
รูปที่ 4.2 ตัวอย่างชุดข้อมูลของการพูดเลขหนึ่งทั้งสิบเฟรมแถวบนซ้ายไปขวาเรียงลำดับเฟรมที่ 1 – 5 แถวล่างซ้ายไปขวาเรียงลำดับเฟรมที่ 6 – 10



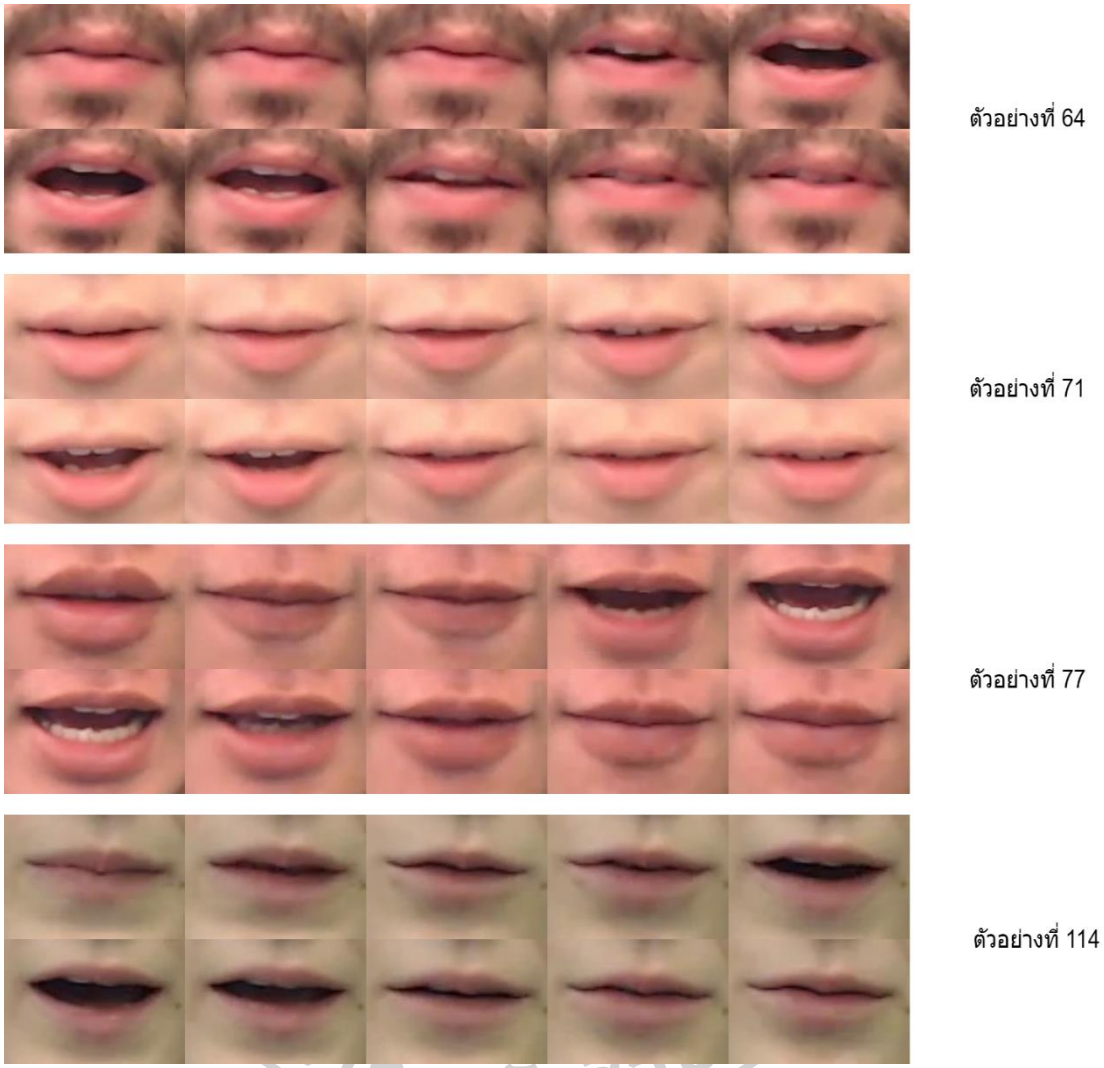
รูปที่ 4.3 ตัวอย่างชุดข้อมูลของการพูดเลขสองทั้งสิบเฟรมแถวบนซ้ายไปขวาเรียงลำดับเฟรมที่ 1 – 5 แถวล่างซ้ายไปขวาเรียงลำดับเฟรมที่ 6 – 10



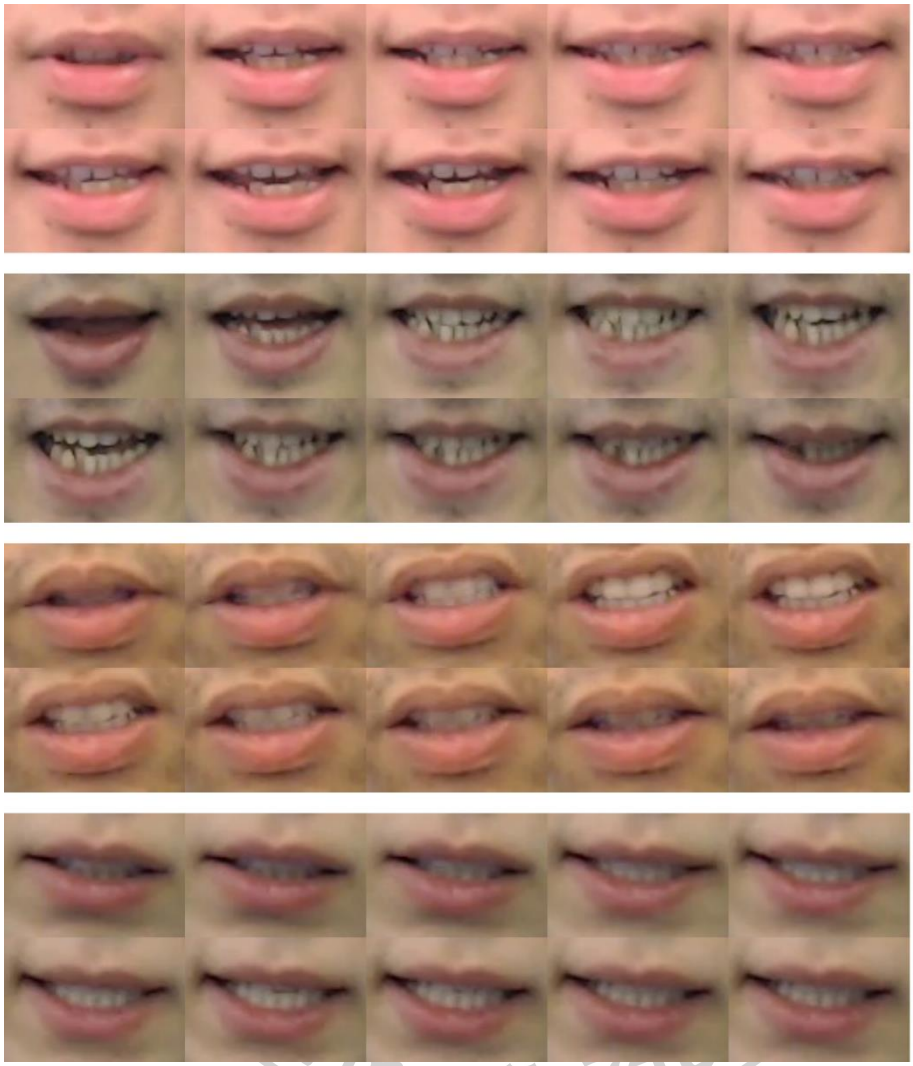
รูปที่ 4.4 ตัวอย่างชุดข้อมูลของการพูดเลขสามทั้งสิบเฟรมแถวบนซ้ายไปขวาเรียงลำดับเฟรมที่ 1 – 5 แถวล่างซ้ายไปขวาเรียงลำดับเฟรมที่ 6 – 10



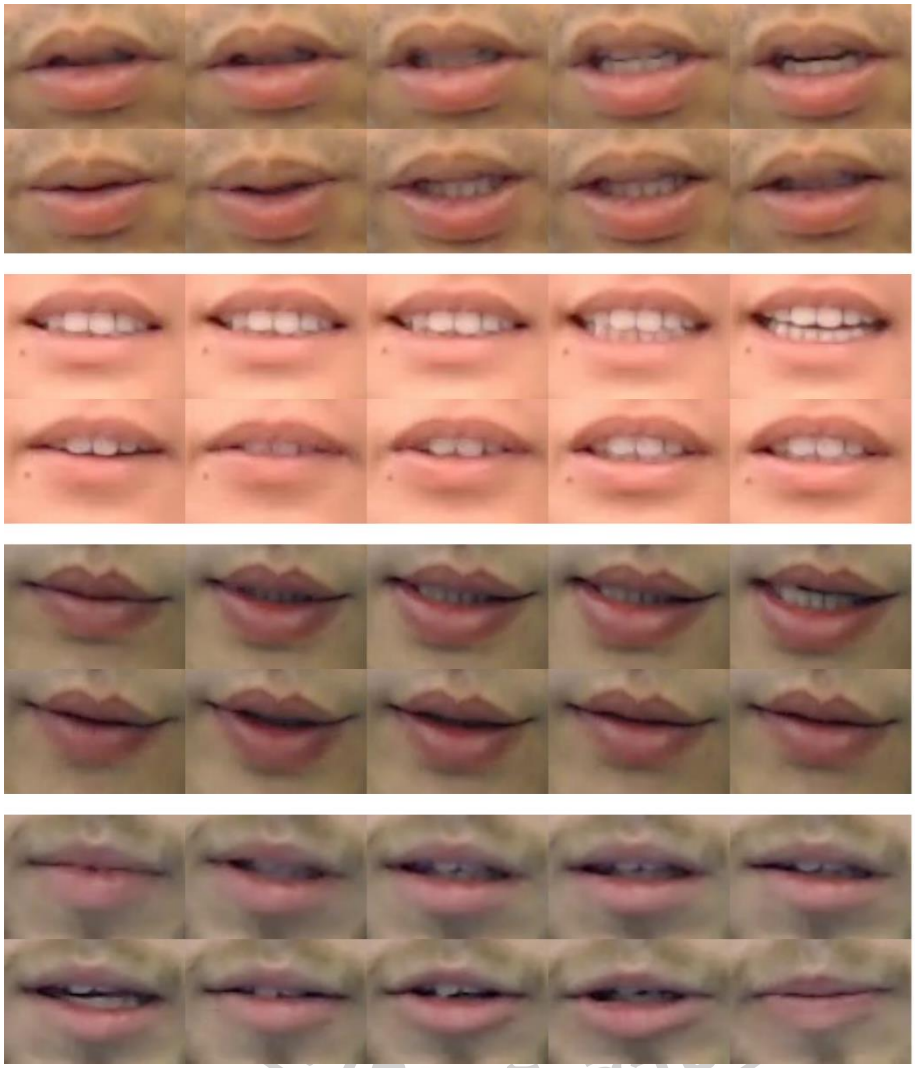
รูปที่ 4.5 ตัวอย่างชุดข้อมูลของการพูดเลขสี่ทั้งสิบเฟรมแถวบนซ้ายไปขวาเรียงลำดับเฟรมที่ 1 – 5 แถวล่างซ้ายไปขวาเรียงลำดับเฟรมที่ 6 – 10



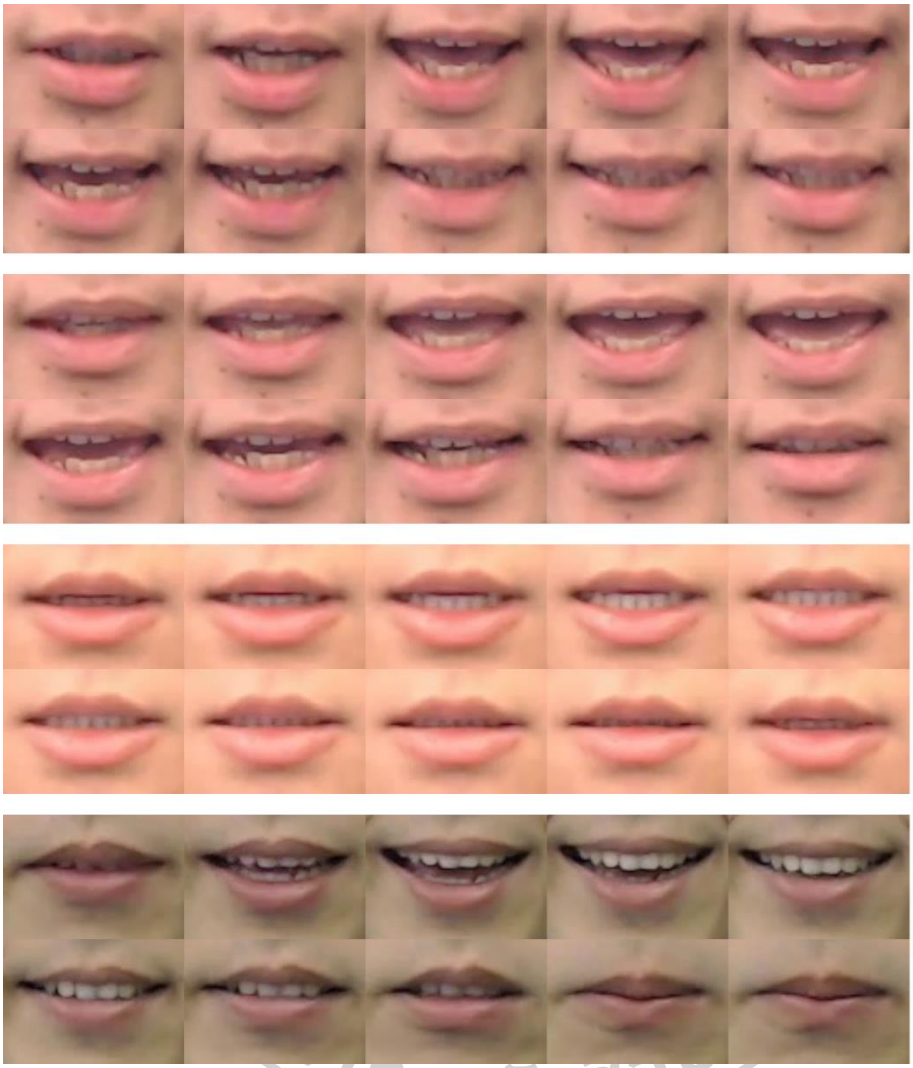
รูปที่ 4.6 ตัวอย่างชุดข้อมูลของการพูดเลขห้าทั้งสิบเฟรมแถวบนซ้ายไปขวาเรียงลำดับเฟรมที่ 1 – 5 แถวล่างซ้ายไปขวาเรียงลำดับเฟรมที่ 6 – 10



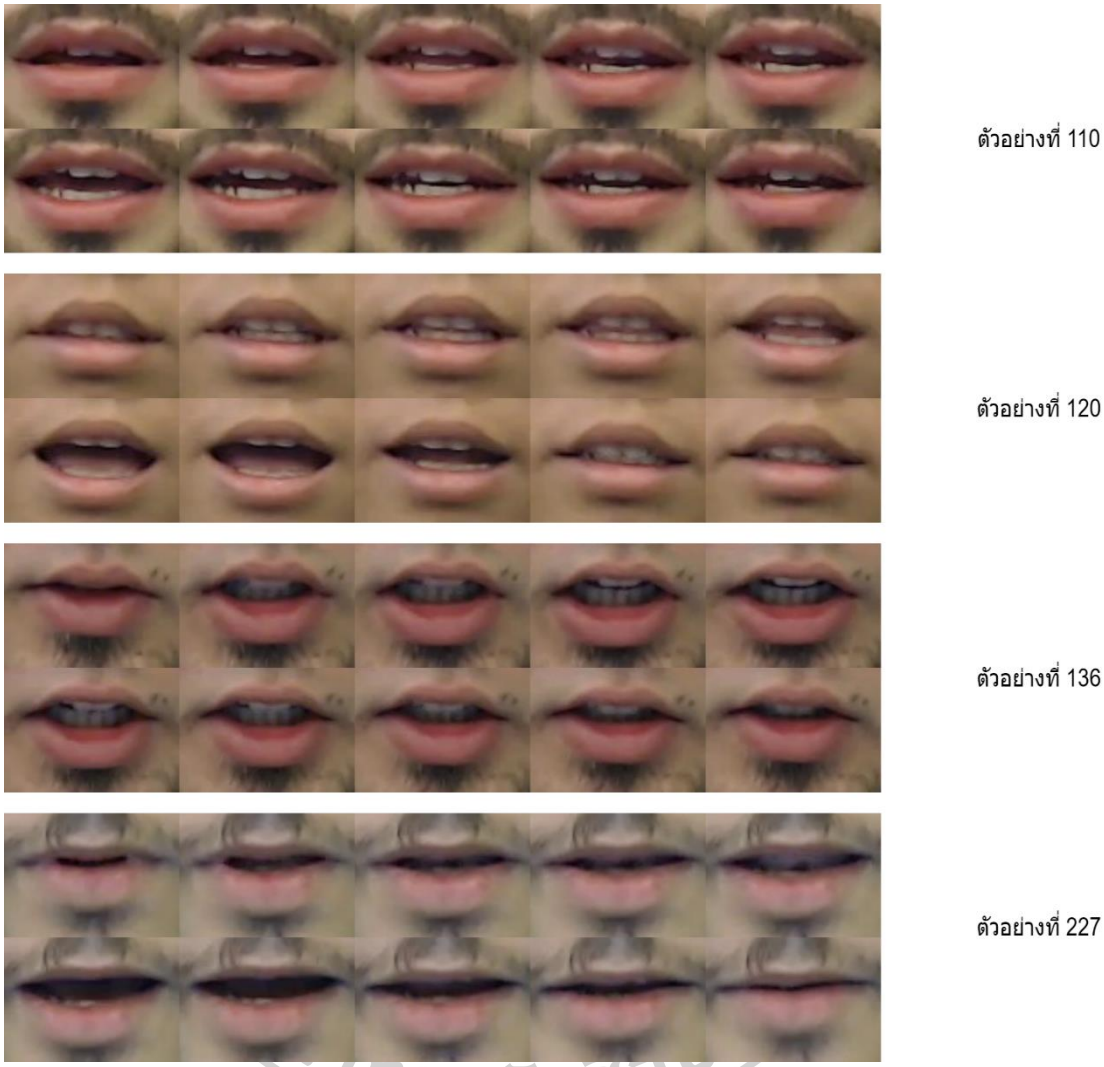
รูปที่ 4.7 ตัวอย่างชุดข้อมูลของการพูดเลขหกทั้งสิบเฟรมแถวบนซ้ายไปขวาเรียงลำดับเฟรมที่ 1 – 5 แถวล่างซ้ายไปขวาเรียงลำดับเฟรมที่ 6 – 10



รูปที่ 4.8 ตัวอย่างชุดข้อมูลของการพูดเลขเจ็ดทั้งสิบเฟรมแถวบนซ้ายไปขวาเรียงลำดับเฟรมที่ 1 – 5 แถวล่างซ้ายไปขวาเรียงลำดับเฟรมที่ 6 – 10



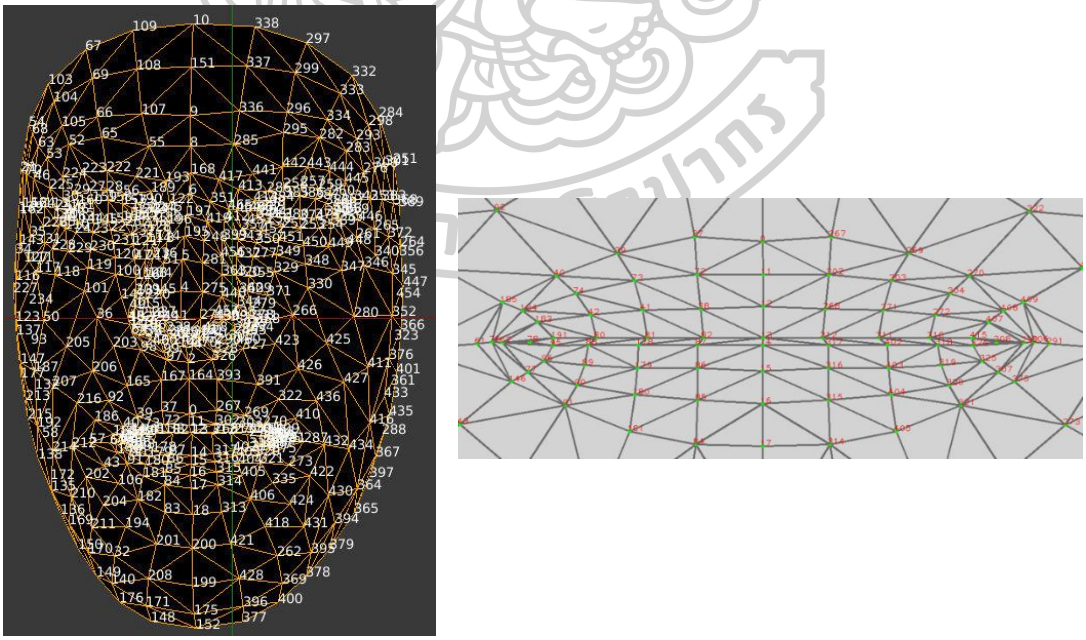
รูปที่ 4.9 ตัวอย่างชุดข้อมูลของการพูดเลขแปดทั้งสิบเฟรมแถวบนซ้ายไปขวาเรียงลำดับเฟรมที่ 1 – 5 แถวล่างซ้ายไปขวาเรียงลำดับเฟรมที่ 6 – 10



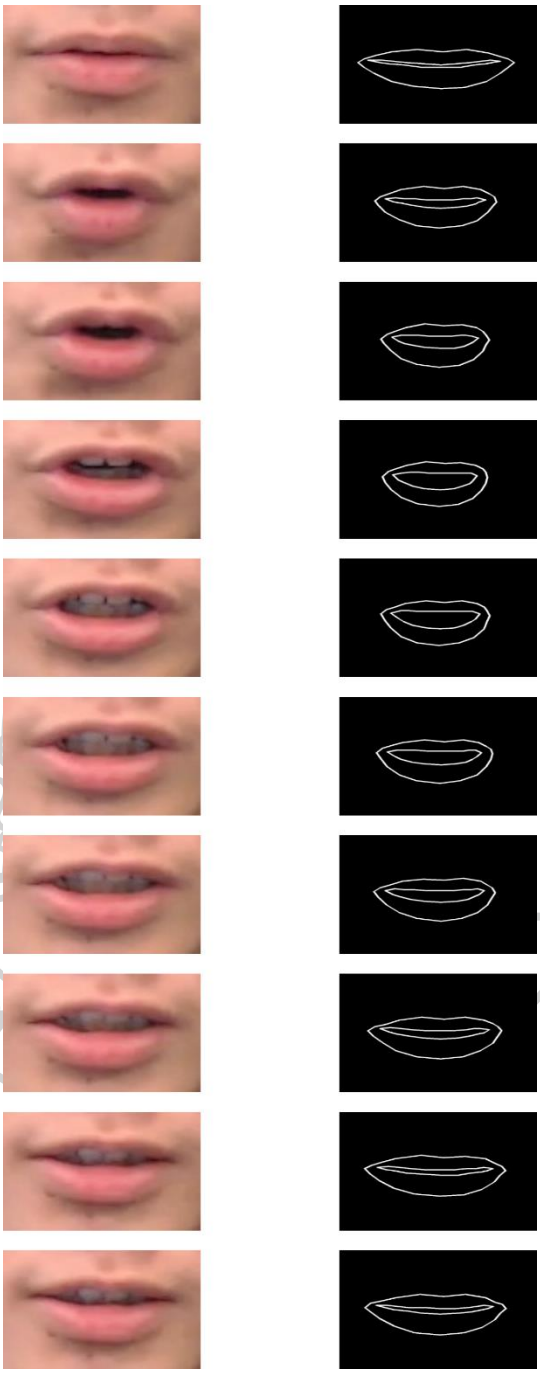
รูปที่ 4.10 ตัวอย่างชุดข้อมูลของการพูดเลขเก้าทั้งสิบเฟรมแถวบนซ้ายไปขวาเรียงลำดับเฟรมที่ 1 – 5 แถวล่างซ้ายไปขวาเรียงลำดับเฟรมที่ 6 – 10

4.1.2 ชุดข้อมูลจากรูปภาพแบบเส้นรอบริมฝีปาก

ชุดข้อมูลจากรูปภาพแบบเส้นรอบริมฝีปากเป็นชุดข้อมูลที่นำมาดัดแปลงลักษณะที่ปรากฏภายในรูปภาพโดยการใช้จุดจากไลบรารี่ Mediapipe ก่อนหน้านี้โดยการเพิ่มจำนวนจุดที่ตัดบริเวณโดยรอบริมฝีปากเข้าไปทั้งหมด 40 จุด แบ่งออกเป็นริมฝีปากด้านนอกที่บอกถึงระยะขอบของริมฝีปาก 20 จุดและริมฝีปากด้านในที่บอกถึงความกว้างของการขยับเพื่อพูดของริมฝีปากอีก 20 จุด รายละเอียดของตำแหน่งของจุดที่เลือกทั้งหมดมีดังนี้ [185, 40, 39, 37, 0, 267, 269, 270, 409, 291, 375, 321, 405, 314, 17, 84 181, 91, 146, 61] สำหรับริมฝีปากด้านนอกและ [78, 191, 80, 81 , 82, 13, 312, 311, 310, 415, 308, 324 , 318, 402, 317, 14, 87, 178, 88, 95] สำหรับริมฝีปากด้านใน จุดและตำแหน่งที่ใช้ตัดบริเวณริมฝีปากแสดงได้ดังรูปที่ 4.11 ทั้งนี้เพื่อการศึกษาการทำงานของการเรียนรู้แบบจำลอง CNN และ LSTM ผ่านรูปภาพที่ดัดแปลงให้เหลือแค่คุณลักษณะของเส้นรอบๆ ริมฝีปากว่า มีประสิทธิภาพมากพอต่อการเรียนรู้ของแบบจำลองหรือไม่ แต่เนื่องจากรูปภาพที่ได้เป็นการนำจุดต่างๆ มาวางในลักษณะเดียวกันแล้วลากเส้นเชื่อมใหม่ ซึ่งหมายถึงเป็นการวาดขึ้นใหม่ที่ทำให้รูปที่ได้มีขนาดที่เปลี่ยนแปลงขึ้นอยู่กับการขยับของริมฝีปากโดยที่จุดศูนย์กลางของรูปไม่เปลี่ยนแปลงไป เพื่อเน้นไปที่การศึกษาเรื่องของการขยับและการเปลี่ยนแปลงไปของริมฝีปากในแต่ละเฟรมที่มีการออกเสียง โดยรูปภาพดังกล่าวสามารถแสดงได้ดังรูปที่ 4.12



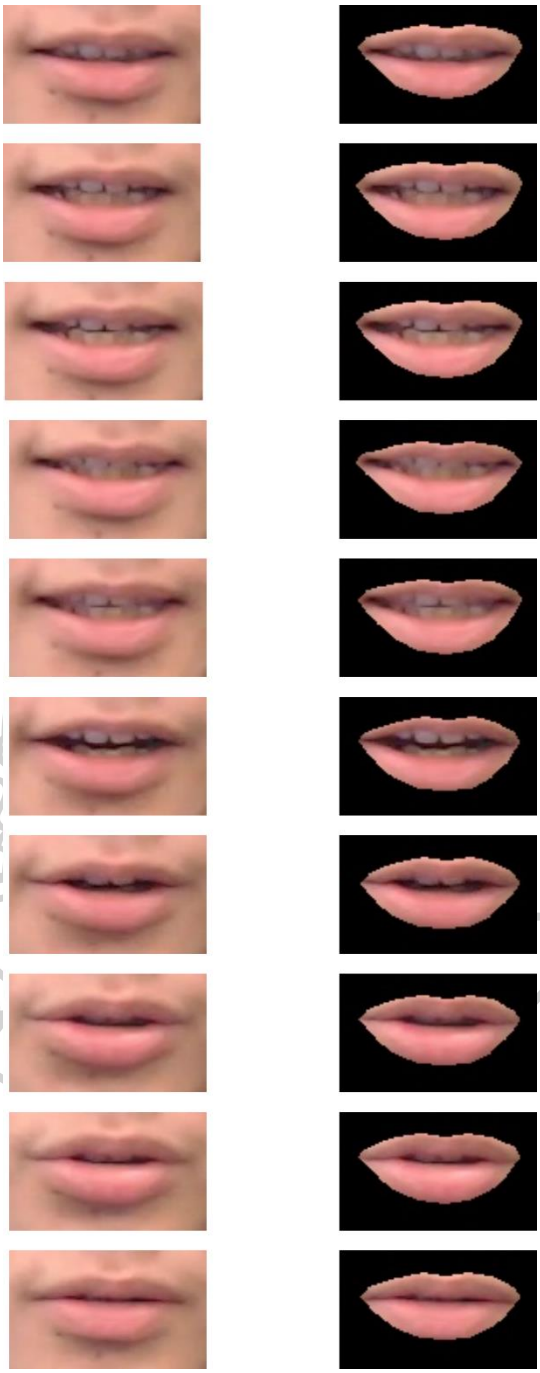
รูปที่ 4.11รูปภาพแสดงจุดและตำแหน่งที่ใช้ในการตัดบริเวณรอบริมฝีปาก รูปด้านซ้ายแสดงภาพรวมของจุดต่างๆ รูปด้านขวาแสดงภาพรวมของจุดที่ใช้ตัดเฉพาะบริเวณริมฝีปาก



รูปที่ 4.12 รูปภาพแสดงการเปรียบเทียบเฟรมต่อเฟรมของการพูดเลขสี่ระหว่างชุดข้อมูลริมฝีปากแบบต้นฉบับกับชุดข้อมูลรูปภาพริมฝีปากแบบเส้นรอบริมฝีปาก

4.1.3 ชุดข้อมูลจากรูปภาพแบบมีสีเฉพาะบริเวณที่เป็นริมฝีปาก

ชุดข้อมูลรูปภาพนี้ได้มาจากชุดข้อมูลต้นฉบับที่นำมาดัดแปลงใส่สีเฉพาะบริเวณที่มีริมฝีปากเท่านั้น วิธีการที่ได้มาแตกต่างจากชุดข้อมูลรูปภาพแบบเส้นรอบริมฝีปาก เนื่องจากชุดข้อมูลนี้เป็นการใช้วิธีที่เลือกว่า Mask ซึ่งเป็นหน้ากากสำหรับคัดกรองสี โดยเมื่อเปรียบเทียบกับรูปแบบต้นฉบับแล้วพบว่าระยะความกว้างความยาวของรูปภาพ ไม่มีความเปลี่ยนแปลง การใส่สีเฉพาะบริเวณที่เป็นริมฝีปากเท่านั้นเป็นการเน้นการศึกษาไปที่การให้ความสำคัญกับบริเวณที่เป็นริมฝีปากจริงๆ โดยองค์ประกอบรอบๆ นอกที่ดวงตาจริงๆ จะสามารถมองเห็นจะถูกตัดออก เพื่อเน้นยำอีกครั้งว่าเฉพาะบริเวณที่เป็นริมฝีปากจริงๆ ส่งผลอย่างไรกับการเรียนรู้ของแบบจำลอง โดยชุดข้อมูลรูปภาพแบบมีสีเฉพาะบริเวณที่เป็นริมฝีปากแสดงได้ดังรูป 4.13



รูปที่ 4.13 รูปภาพแสดงการเปรียบเทียบเฟรมต่อเฟรมของการพูดเลขศูนย์ระหว่างชุดข้อมูลริมฝีปากแบบต้นฉบับกับชุดข้อมูลรูปภาพริมฝีปากแบบมีสีเฉพาะบริเวณที่เป็นริมฝีปาก

4.2 การทดลองหาแบบจำลองที่ดีที่สุดจากการเลือก 3 เฟรม

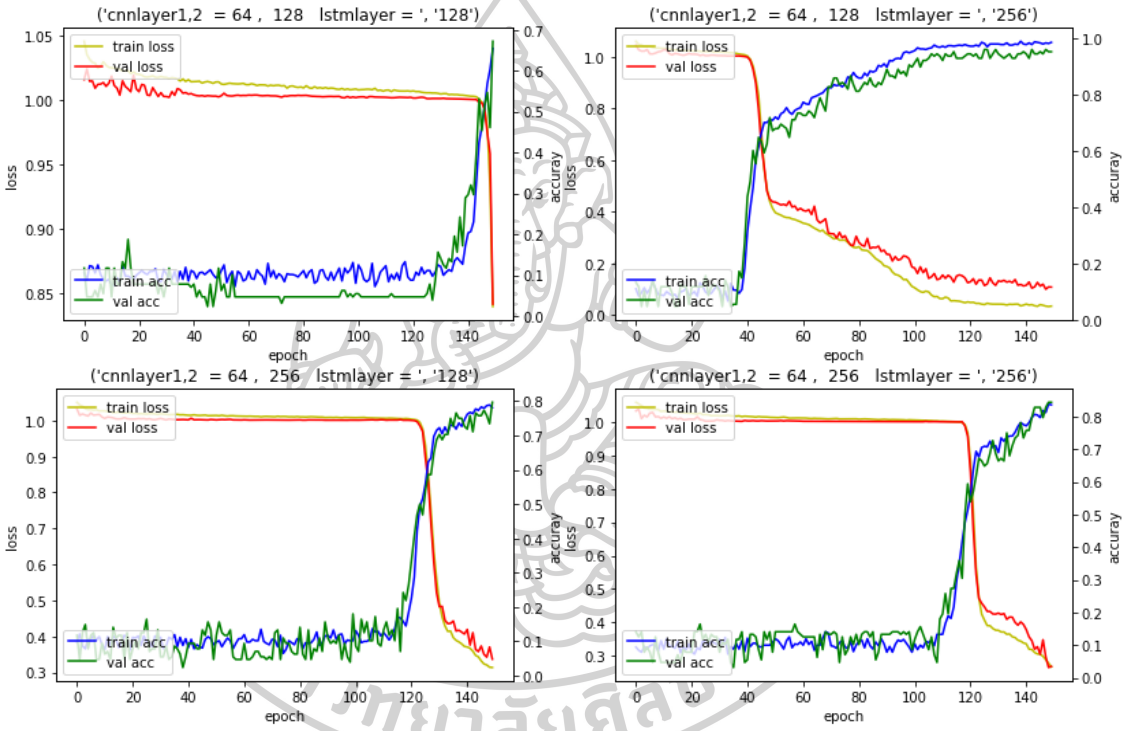
ในการสร้างแบบจำลองของการเรียนรู้เชิงลึก สิ่งที่สำคัญและจำเป็นที่จะต้องกำหนดเพื่อใช้ในการคำนวณการปรับค่าน้ำหนักของพารามิเตอร์เพื่อให้โมเดลเกิดการเรียนรู้นั่นคือการเลือกฟังก์ชันของ optimizer และ loss ตัวเลือกที่ดีที่สุดจะช่วยให้แบบจำลองแสดงประสิทธิภาพออกมาได้มากที่สุดเช่นกัน โดยในแต่ละงานไม่มีความตายตัวว่าฟังก์ชันไหนเหมาะกับงานประเภทไหน ดังนั้นแล้วการเลือก optimizer และ loss ขึ้นอยู่กับงานและในขณะเดียวกันก็ขึ้นอยู่กับแบบจำลองที่ใช้ ในบางครั้งจำเป็นต้องทดลองเลือกตัวเลือกหลายๆ ตัว เพื่อหาค่าที่ดีที่สุดสำหรับงานนั้นๆ ในการทดลองของงานวิจัยเล่มนี้ได้กำหนด optimizer สำหรับการหาค่าที่ดีที่สุดดังนี้ Adadelta Adagrad Adam Adamax FTRL Nadam RMSprop และ SGD พร้อมกันนี้ได้มีการกำหนด loss ในการหาค่าที่ดีที่สุดดังนี้ BinaryCrossentropy CategoricalCrossentropy CategoricalHinge MeanAbsoluteError MeanSquaredError และ SparseCategoricalCrossentropy ในการทดลองนี้ใช้ข้อมูลจากชุดข้อมูลต้นฉบับโดยมีการใช้ข้อมูลทั้งหมดสำหรับการ train โดยมีการแบ่ง 80% สำหรับ train process และ 20% สำหรับ validation process ซึ่งกำหนดจำนวนเฟรมทั้งหมดที่ใช้หาค่าดังกล่าวคือ 3 เฟรมแบบเต็มปากเพื่อความเร็วของการทดลอง เฟรมที่ใช้คือ 2 5 และ 8 โดยมีการปรับค่าในจำนวนชั้นของ CNN และ LSTM ร่วมด้วยโดยที่โปรแกรมสำหรับใช้สร้างสถาปัตยกรรมที่ใช้ในการสร้างแบบจำลองในการทดสอบแสดงได้ดังรูปที่ 4.14 จากรูปค่าในแต่ละชั้นจะเปลี่ยนตามตัวแปร cnnlayer1 cnnlayer2 cnnlayer3 และ lstmlayer ผลลัพธ์ที่จะนำมาแสดงต่อไปนี้ เป็นผลลัพธ์ที่เกิดเป็นรูปร่างและมีประสิทธิภาพที่ดีจากการเลือกตัวเลือกนั้นๆ ผลลัพธ์ที่ไม่ทำให้แบบจำลองเกิดการเรียนรู้จะไม่นำมาแสดง

```
model = Sequential()

model.add(Conv2D(cnnlayer1, (3,3), activation='relu', input_shape=(64,64,3)))
model.add(MaxPool2D((2, 2)))
model.add(Conv2D(cnnlayer2, (3,3), activation='relu'))
model.add(MaxPool2D((2, 2)))
model.add(Conv2D(cnnlayer3, (3,3), activation='relu'))
model.add(MaxPool2D((2, 2)))
model.add(Flatten())
model.add(Dense(128, activation='relu'))
model.add(Dropout(0.2))
model2 = Sequential()
model2.add(TimeDistributed(model,input_shape = (timesteps,64,64,3)))
model2.add(LSTM(lstmlayer))
model2.add(Dense(256, activation='relu'))

model2.add(Dropout(0.2))
model2.add(Dense(n_labels, activation="softmax"))
```

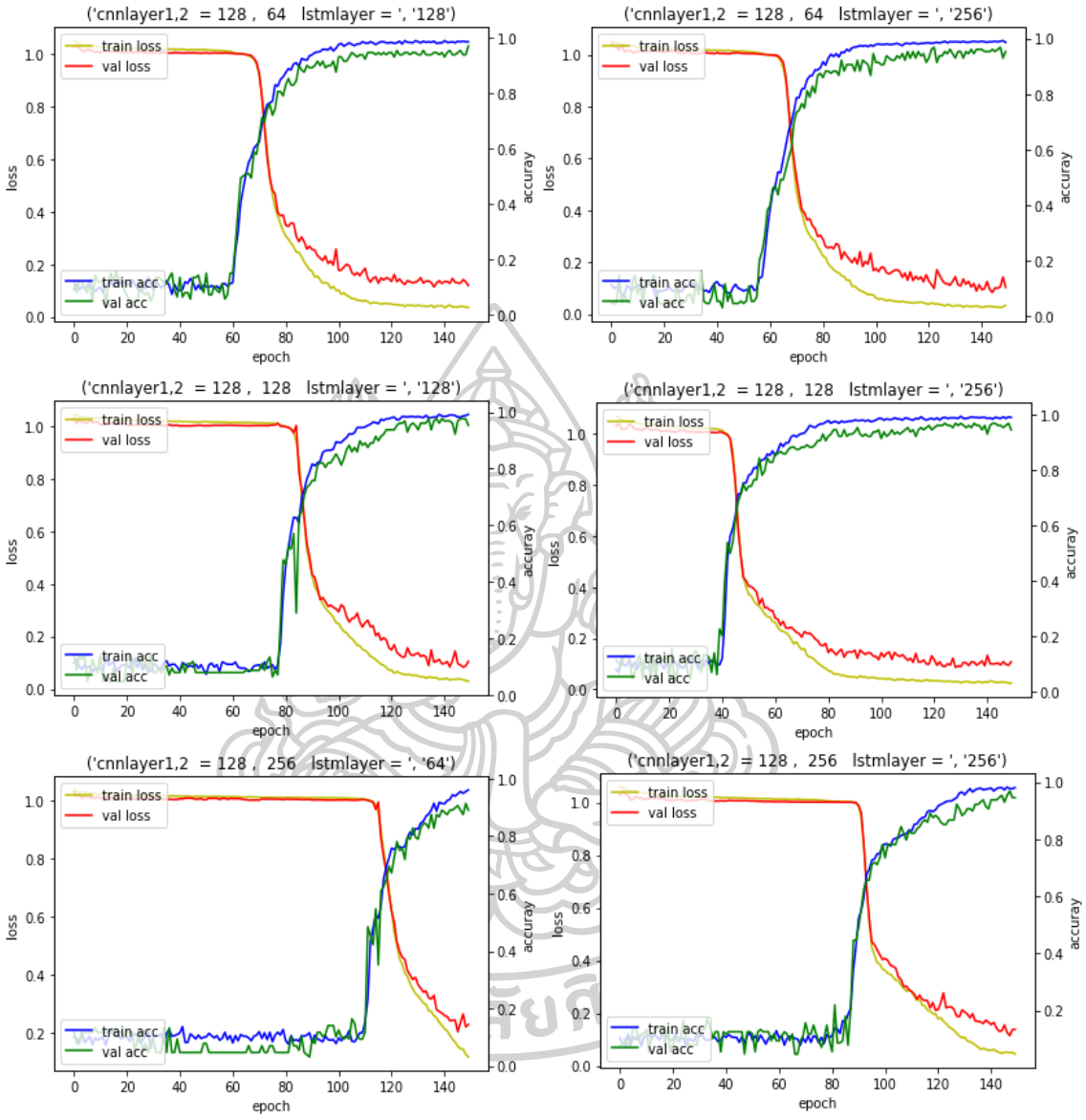
รูปที่ 4.14 แสดงตัวอย่างโปรแกรมสร้างสถาปัตยกรรมที่ใช้ในการทดลอง

4.2.1 Optimizer Adagrad และ Loss Categorical Hinge 2 เลเยอร์

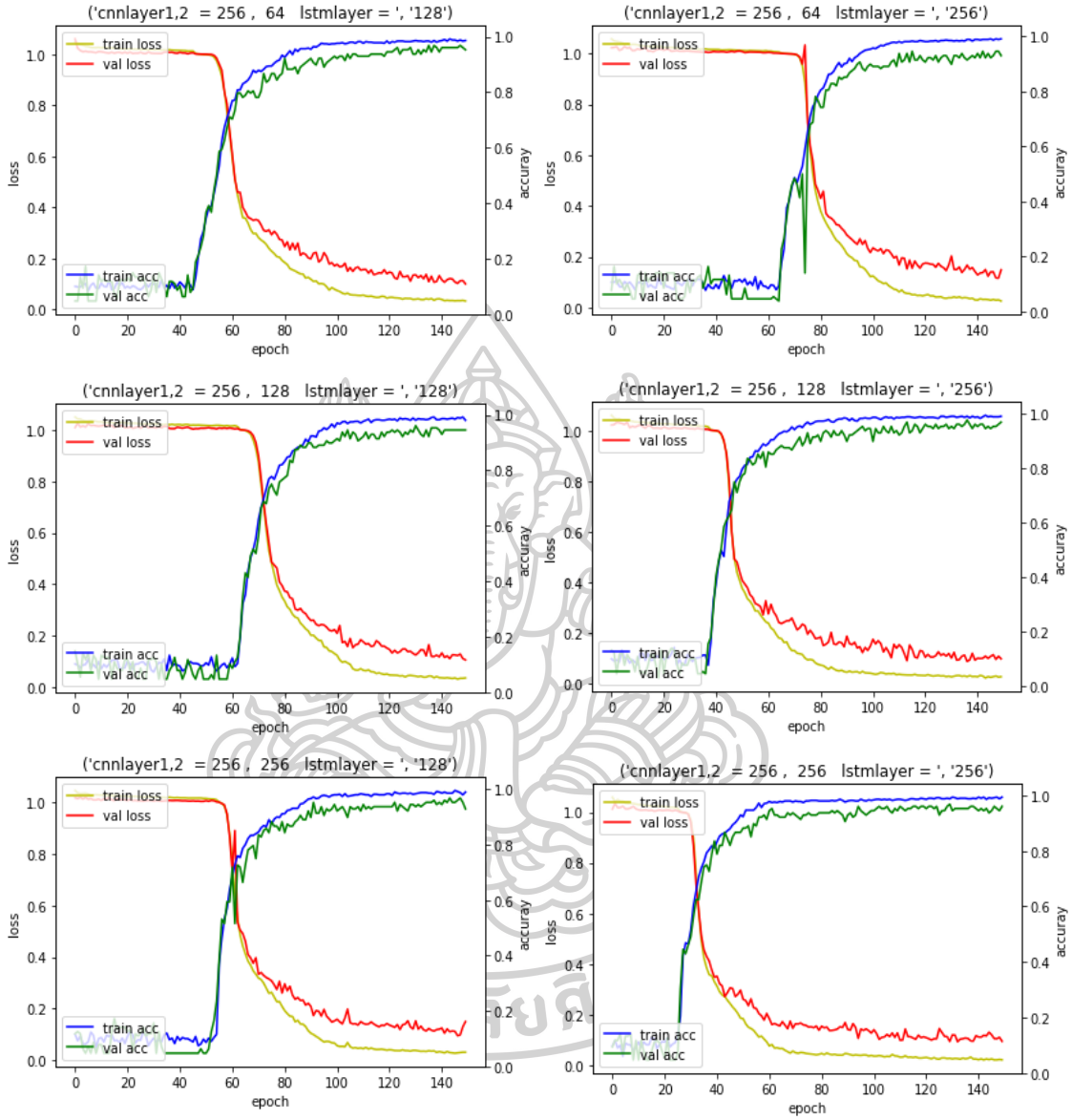
หลังจากการทดลองดังกล่าวผลลัพธ์ตัวหนึ่งที่สามารถทำให้แบบจำลองเกิดการเรียนรู้ได้คือการเลือก optimizer เป็น Adagrad จับคู่กับ loss เป็น CategoricalHinge โดยในช่วงแรกกำหนดให้มีการเรียกการประมวลผลที่ชั้นของ CNN 2 ชั้น และมีการกำหนดเลือก kernel ตั้งแต่ 64 128 และ 256 รวมถึงการปรับตั้งค่าของ LSTM ที่มี Units ทั้ง 64 128 และ 256 ด้วยเช่นกัน การจัดกลุ่มของรูปภาพที่นำเสนอจะเรียงจากการกำหนดที่ชั้น CNN ชั้นแรก โดยกราฟของการเรียนรู้ของแบบจำลองแสดงได้ดังรูป 4.15 – 4.18



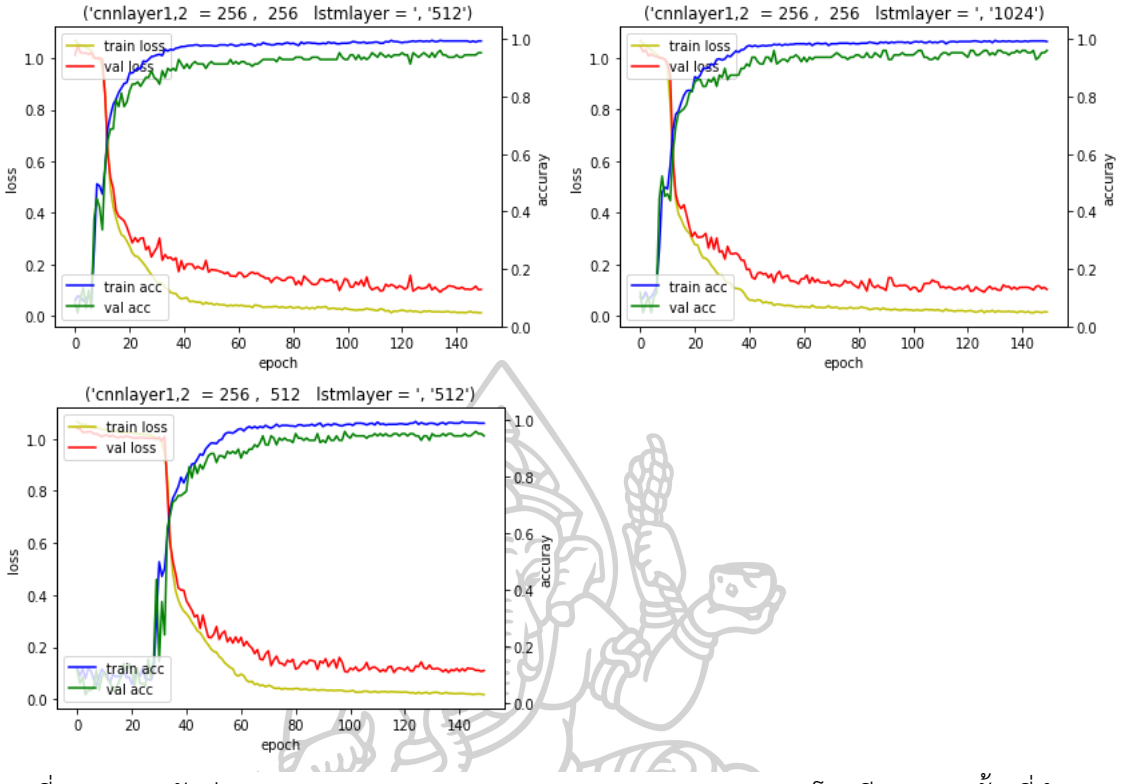
รูปที่ 4.15 การจับคู่ optimizer Adagrad และ Loss Categorical Hinge โดยมี CNN 2 ชั้น ที่กำหนดชั้นแรกเป็น 64



รูปที่ 4.16 การจับคู่ optimizer Adagrad และ Loss Categorical Hinge โดยมี CNN 2 ชั้น ที่กำหนดชั้นแรกเป็น 128



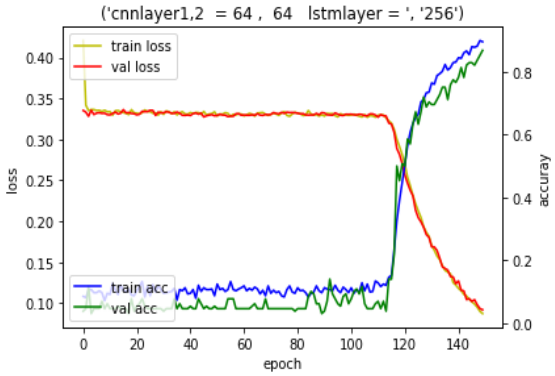
รูปที่ 4.17 การจับคู่ optimizer Adagrad และ Loss Categorical Hinge โดยมี CNN 2 ชั้น ที่กำหนดชั้นแรกเป็น 256 ชุดที่ 1



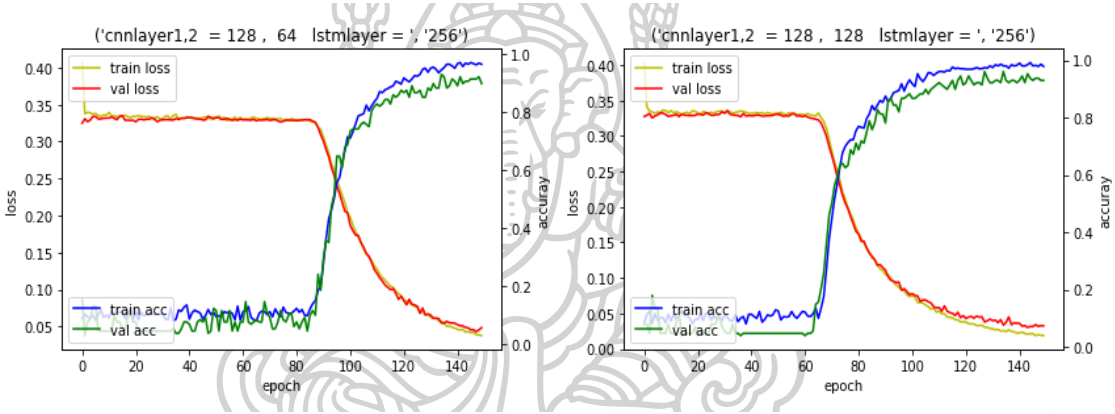
รูปที่ 4.18 การจับคู่ optimizer Adagrad และ Loss CategoricalHinge โดยมี CNN 2 ชั้น ที่กำหนดชั้นแรกเป็น 256 ชุดที่ 2

4.2.2 Optimizer Adagrad และ Loss Binary Cross-Entropy 2 เลเยอร์

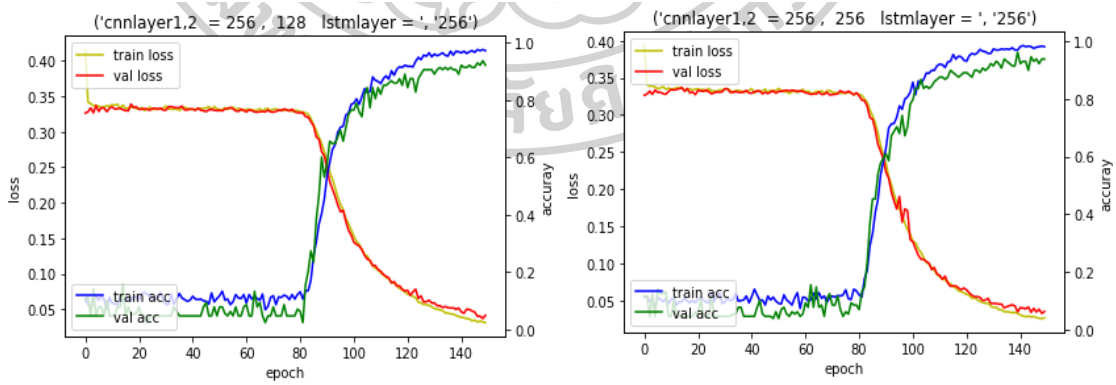
หลังจากการทดลองดังกล่าวผลลัพธ์อีกตัวหนึ่งที่สามารถทำให้แบบจำลองเกิดการเรียนรู้ได้ คือ การเลือก optimizer เป็น Adagrad จับคู่กับ loss เป็น Binary Cross-Entropy เช่นเดียวกันในการตั้งค่าตัวแปรต่างๆ โดยมีการเรียกการประมวลผลที่ชั้นของ CNN 2 ชั้น และมีการกำหนดเลือก kernel ตั้งแต่ 64 128 และ 256 รวมถึงการปรับตั้งค่าของ LSTM ที่มี Units ทั้ง 64 128 และ 256 ด้วยเช่นกัน และเมื่อเปรียบเทียบผลลัพธ์ที่มองได้ด้วยตาเปล่าจะเห็นว่า กราฟการเรียนรู้ของแบบจำลองที่นำเสนอด้วยการเลือก CNN 2 ชั้น การเลือก Loss Binary Cross-Entropy สามารถให้ผลลัพธ์ได้ดีกว่า แม้ว่าจะมีกราฟที่แสดงการเรียนรู้ได้น้อยกว่า โดยกราฟของการเรียนรู้ของแบบจำลองแสดงได้ดังรูป 4.19 – 4.21



รูปที่ 4.19 การจับคู่ optimizer Adagrad และ Loss Binary Cross-Entropy โดยมี CNN 2 ชั้น ที่กำหนดชั้นแรกเป็น 64



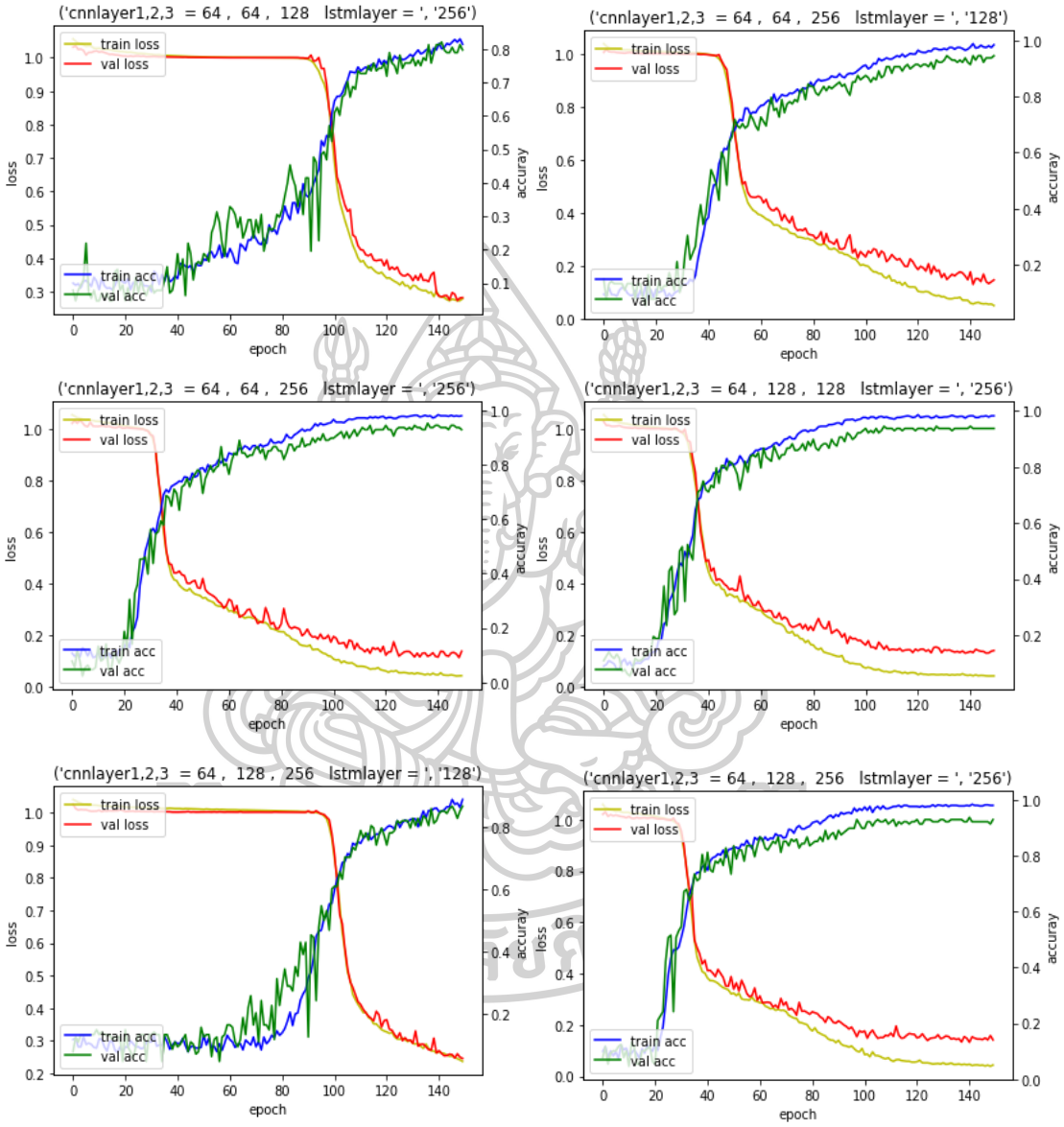
รูปที่ 4.20 การจับคู่ optimizer Adagrad และ Loss Binary Cross-Entropy โดยมี CNN 2 ชั้น ที่กำหนดชั้นแรกเป็น 128



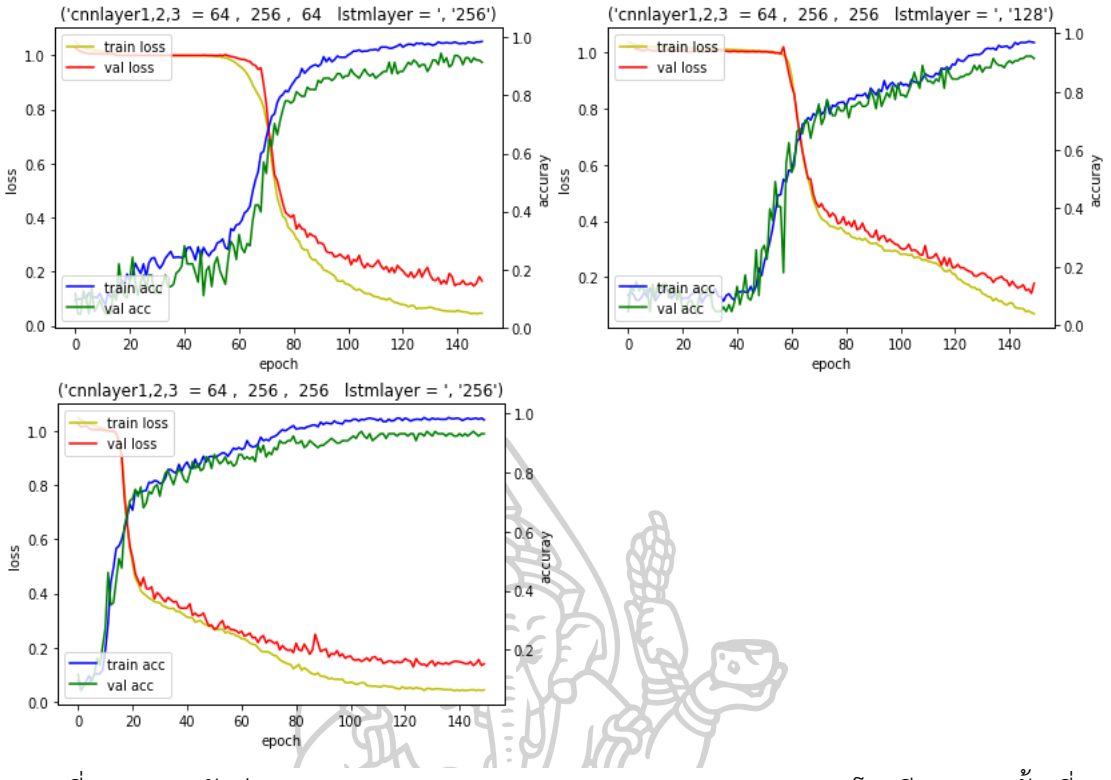
รูปที่ 4.21 การจับคู่ optimizer Adagrad และ Loss Binary Cross-Entropy โดยมี CNN 2 ชั้น ที่กำหนดชั้นแรกเป็น 256

4.2.3 Optimizer Adagrad และ Loss Categorical Hinge 3 เลเยอร์

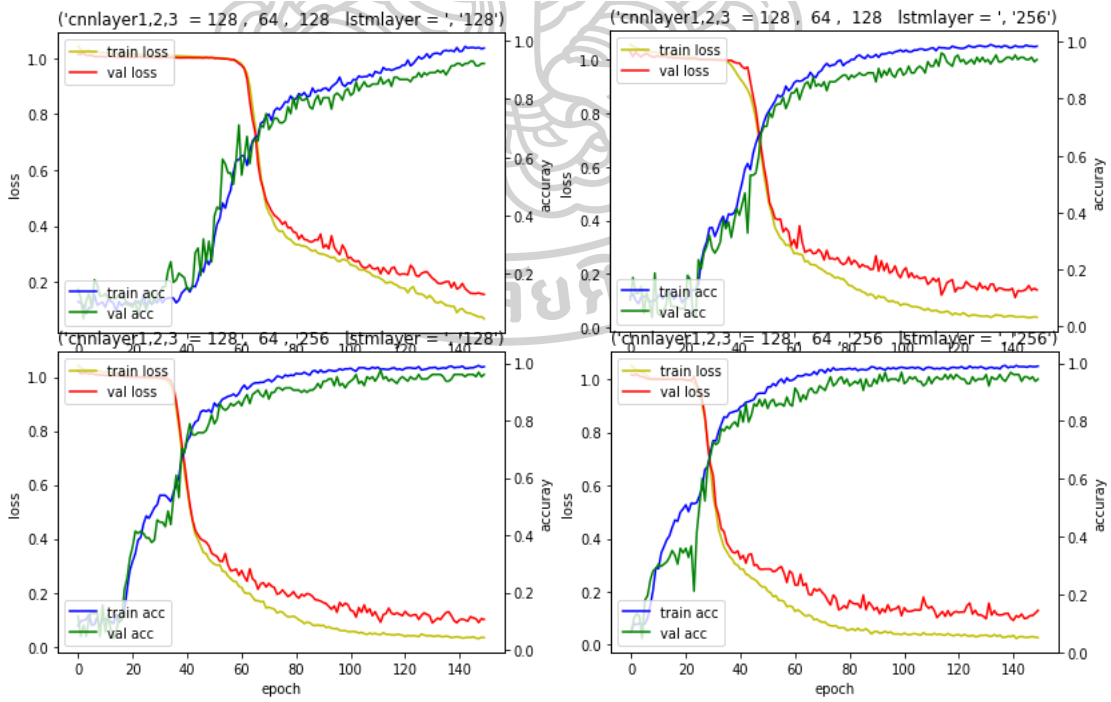
หลังจากปรับจำนวนชั้นของ CNN ที่ใช้ในการทดลองเป็น 2 ชั้นแล้ว ต่อไปเป็นปรับการตั้งค่าของจำนวนชั้นของ CNN เป็น 3 ชั้น เพื่อสังเกตการณ์เปลี่ยนแปลงกราฟของการเรียนรู้ของแบบจำลอง ในทางทฤษฎีแล้วการเพิ่มขึ้นของชั้น CNN จะส่งผลให้แบบจำลองมีความซับซ้อนมากขึ้น แต่จะมีความเร็วที่น้อยลง ในงานที่ต้องการความซับซ้อนมากขึ้น การเพิ่มขึ้นของชั้น CNN อาจส่งผลดีต่อแบบจำลองและให้ประสิทธิภาพการรู้จำที่ดีขึ้น ดังนั้นแล้วการตั้งค่าจะเป็นไปตามการทดลองก่อนหน้า เพียงแต่เพิ่มชั้นส่วนการประมวลผลของ CNN อีก 1 ชั้น ร่วมกับการปรับเปลี่ยนค่าของ LSTM ตามลำดับ ซึ่งการจับคู่ Optimizer เป็น Adagrad และ Loss เป็น Categorical Hinge ยังคงส่งผลให้แบบจำลองเกิดการเรียนรู้ได้ โดยกราฟการเรียนรู้ของแบบจำลองแสดงได้ดังรูปที่ 4.22 – 4.27



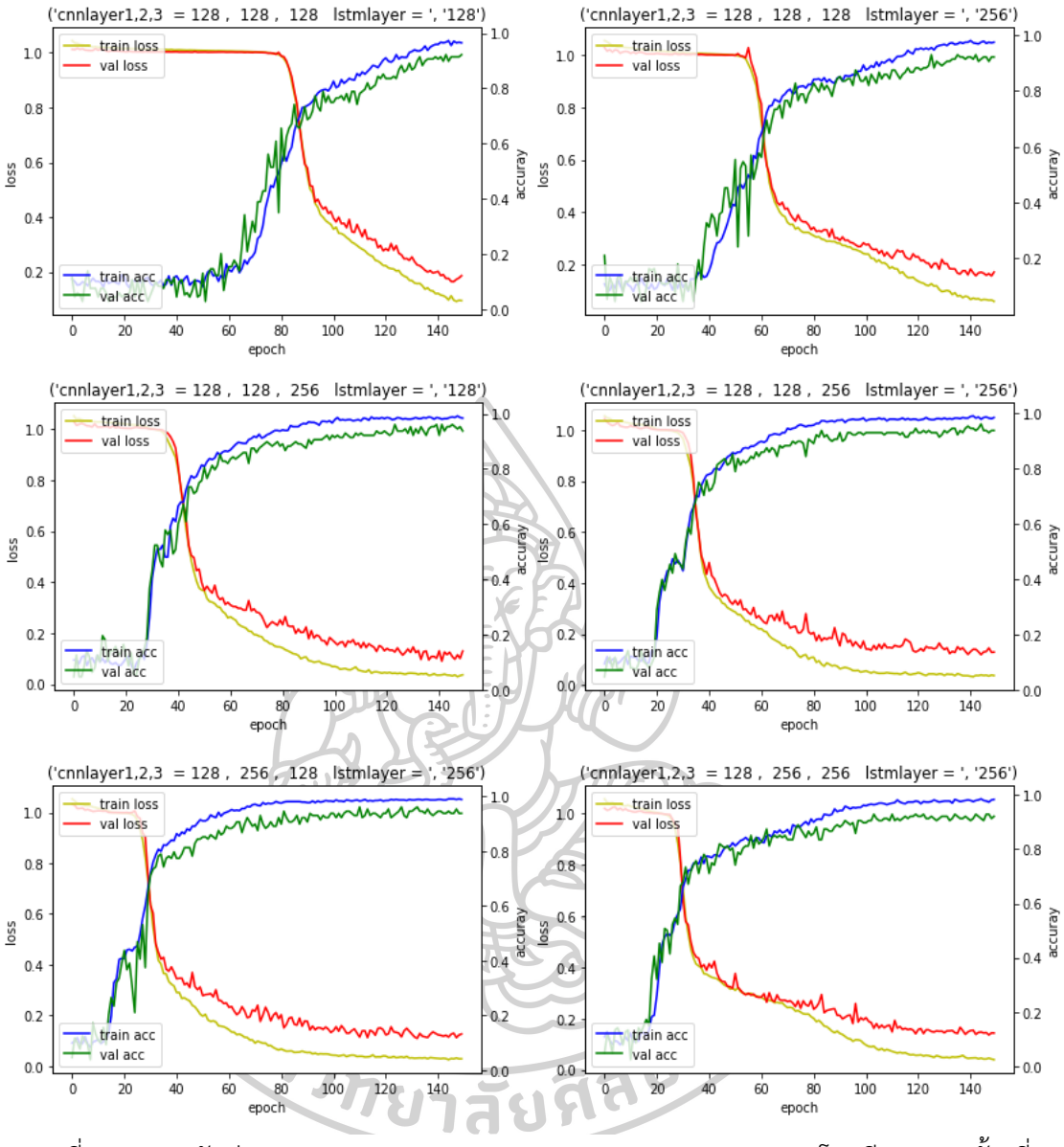
รูปที่ 4.22 การจับคู่ optimizer Adagrad และ Loss Categorical Hinge โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 64 ชุดที่ 1



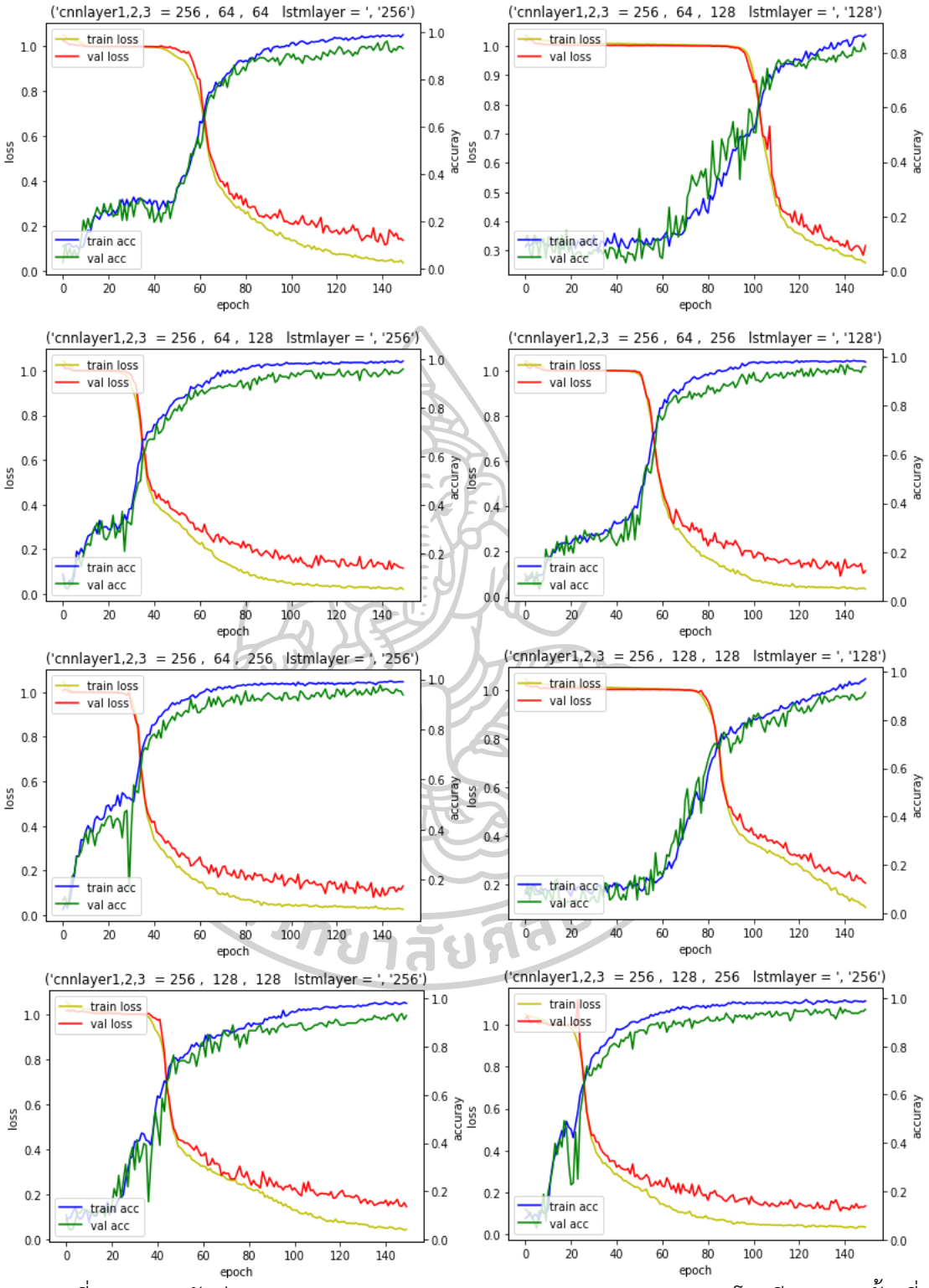
รูปที่ 4.23 การจับคู่ optimizer Adagrad และ Loss Categorical Hinge โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 64 ชุดที่ 2



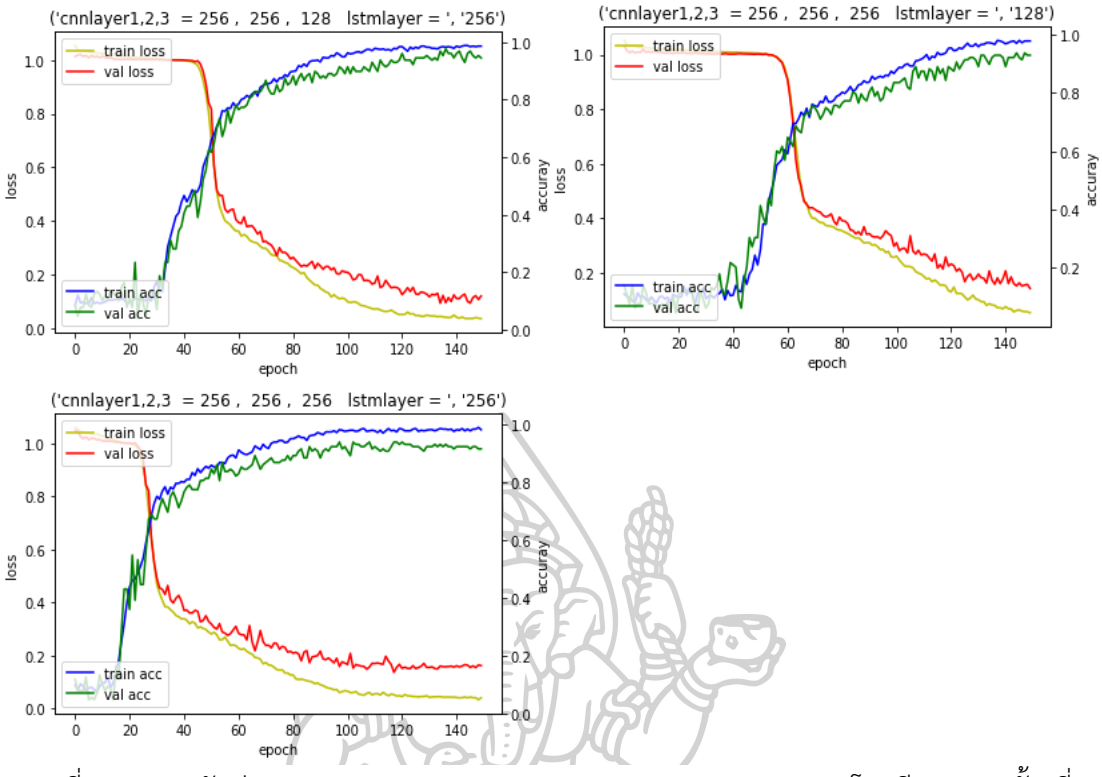
รูปที่ 4.24 การจับคู่ optimizer Adagrad และ Loss Categorical Hinge โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 128 ชุดที่ 1



รูปที่ 4.25 การจับคู่ optimizer Adagrad และ Loss Categorical Hinge โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 128 ชุดที่ 2



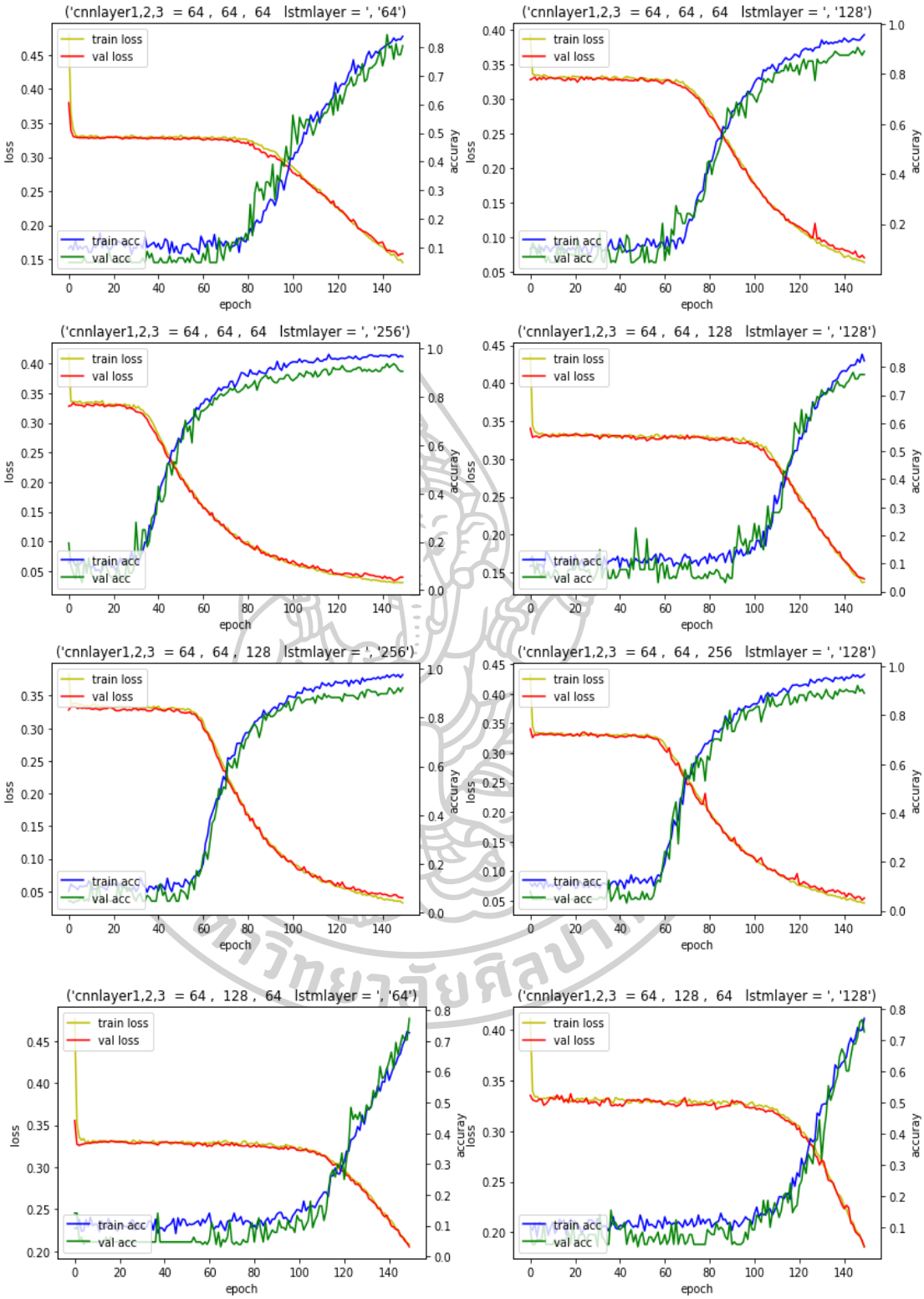
รูปที่ 4.26 การจับคู่ optimizer Adagrad และ Loss Categorical Hinge โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 256 ชุดที่ 1



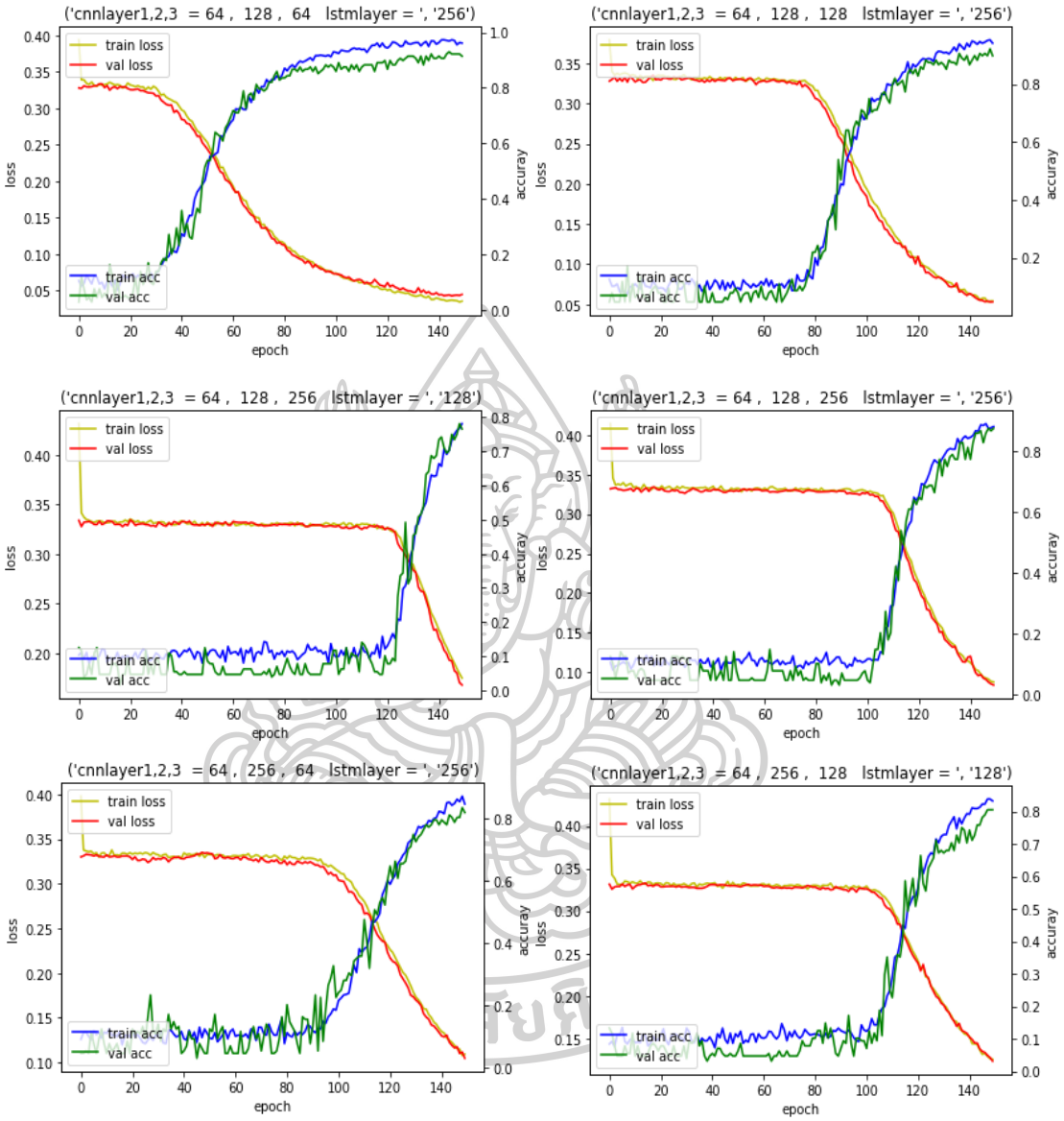
รูปที่ 4.27 การจับคู่ optimizer Adagrad และ Loss Categorical Hinge โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 256 ชุดที่ 2

4.2.3 Optimizer Adagrad และ Loss Binary Cross-Entropy 3 เลเยอร์

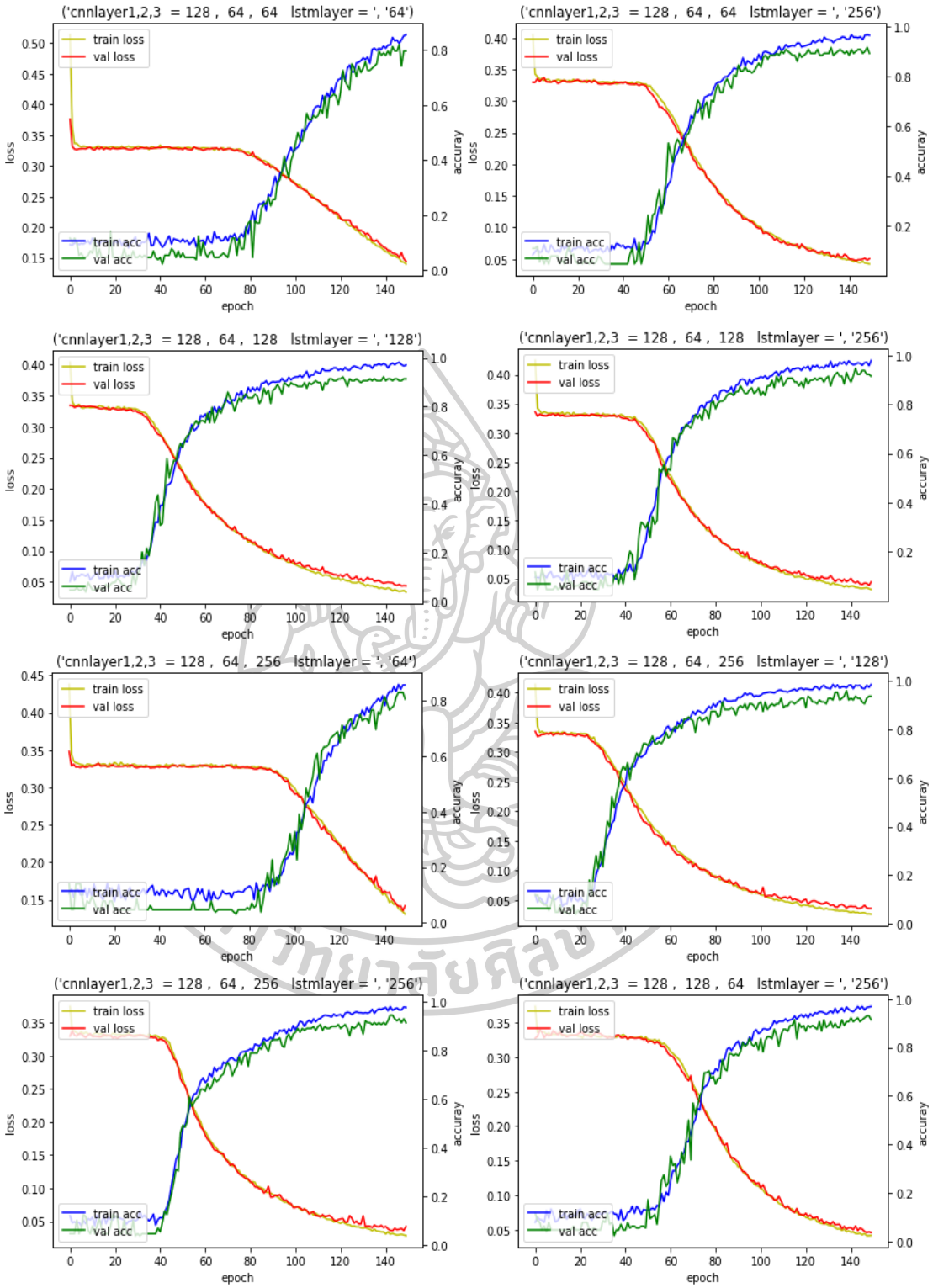
เนื่องจากการเลือก Optimizer Adagrad และ Loss Binary Cross-Entropy ในหัวข้อก่อนหน้าที่มีการกำหนดชั้นของ CNN ที่ 2 ชั้น นั้นให้ผลลัพธ์ของกราฟแสดงผลเส้นของการเรียนรู้ของแบบจำลองที่ดีเมื่อเปรียบเทียบกับการเลือก Loss เป็น Categorical Hinge จะสังเกตเห็นว่าเส้นโค้งของกราฟการเรียนรู้ที่กำหนดใช้ Loss Binary Cross-Entropy ดังนั้นแล้วจึงมีการทดลองเลือก Loss ดังกล่าวที่มีการกำหนดชั้นของ CNN ที่ 3 ชั้นเพื่อเปรียบเทียบผลลัพธ์ และการทดลองจับคู่ที่ Optimizer Adagrad และ Binary Cross-Entropy ยังคงส่งผลให้แบบจำลองเกิดการเรียนรู้ได้เช่นเคย โดยกราฟการเรียนรู้ของแบบจำลองแสดงได้ดังรูปที่ 4.28 – 4.34



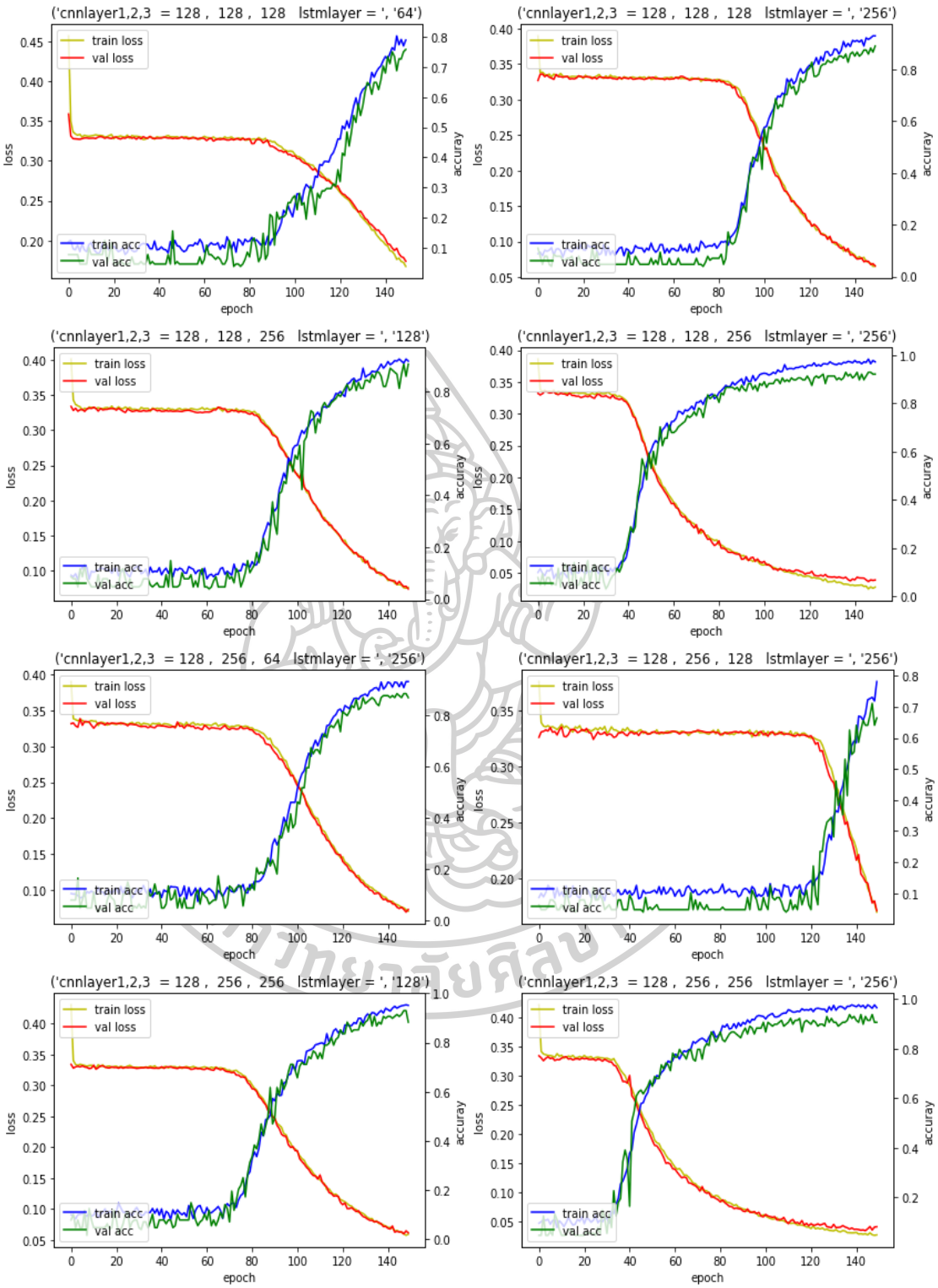
รูปที่ 4.28 การจับคู่ optimizer Adagrad และ Loss Binary Cross-Entropy โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 64 ชุดที่ 1



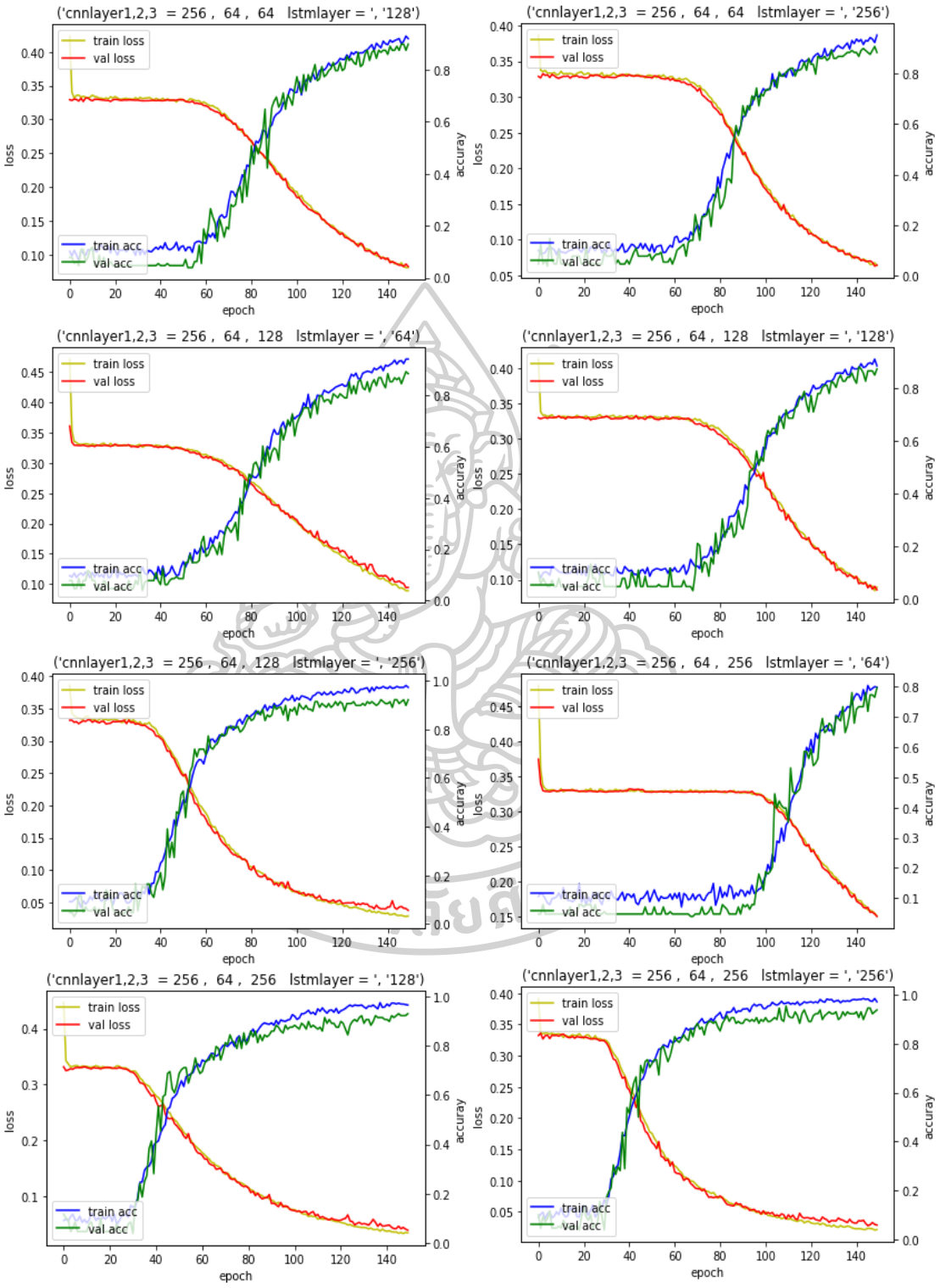
รูปที่ 4.29 การจับคู่ optimizer Adagrad และ Loss Binary Cross-Entropy โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 64 ชุดที่ 2



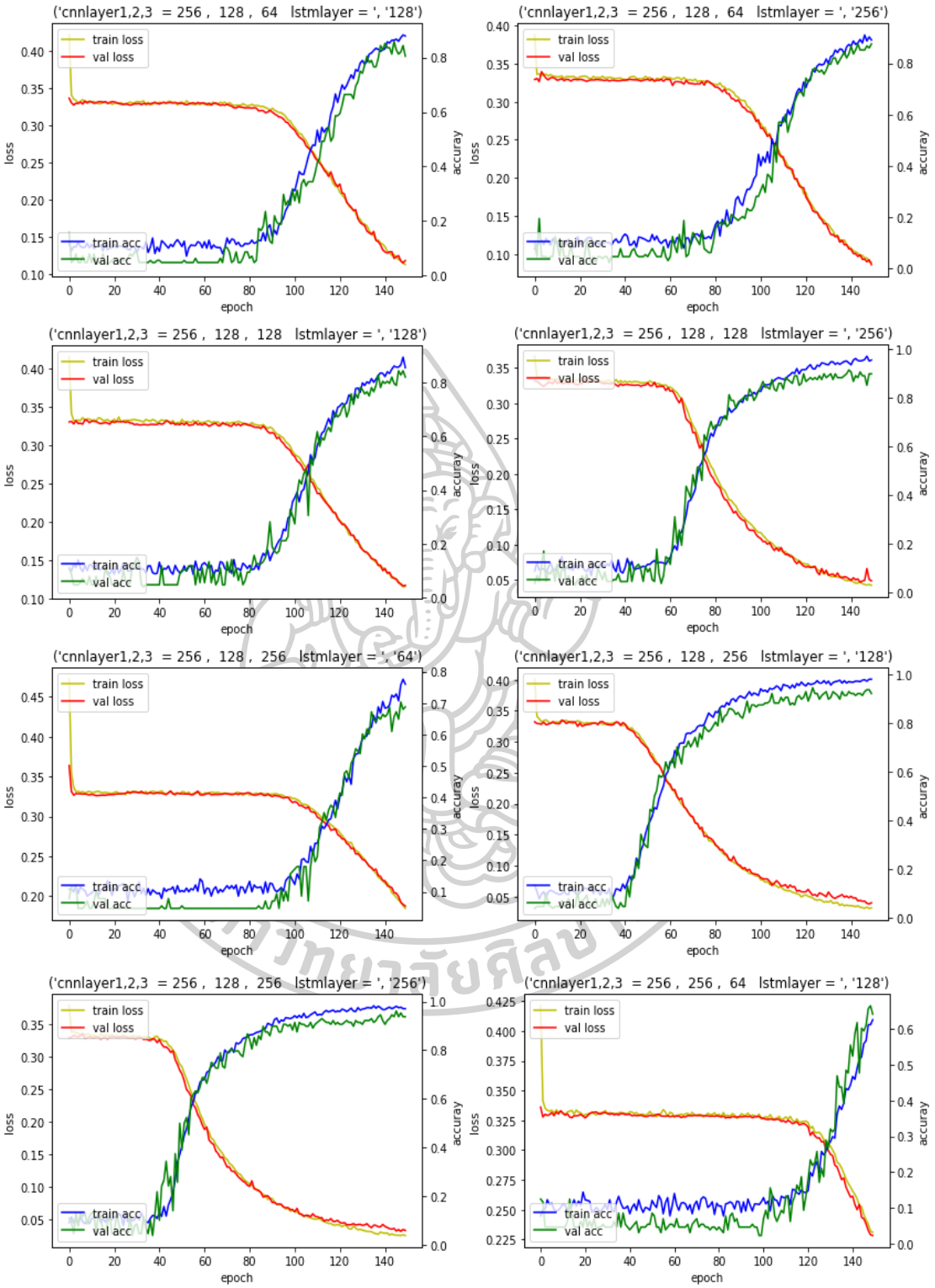
รูปที่ 4.30 การจับคู่ optimizer Adagrad และ Loss Binary Cross-Entropy โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 128 ชุดที่ 1



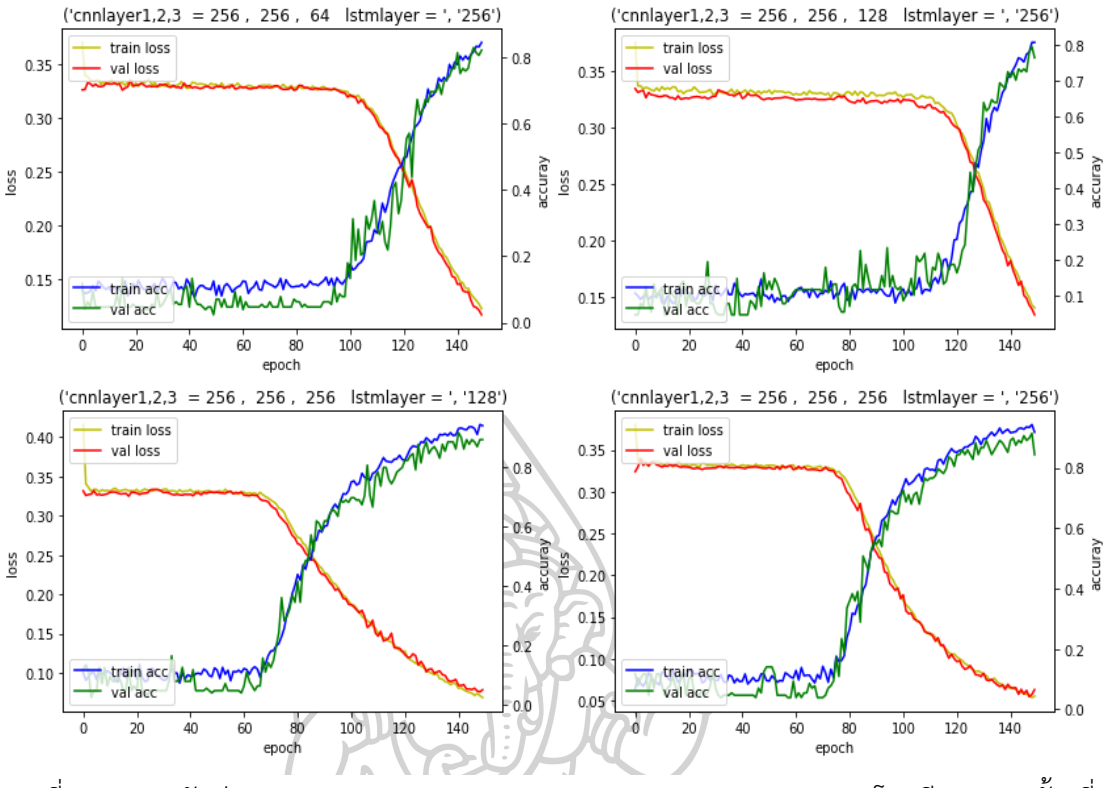
รูปที่ 4.31 การจับคู่ optimizer Adagrad และ Loss Binary Cross-Entropy โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 128 ชุดที่ 2



รูปที่ 4.32 การจับคู่ optimizer Adagrad และ Loss Binary Cross-Entropy โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 256 ชุดที่ 1



รูปที่ 4.33 การจับคู่ optimizer Adagrad และ Loss Binary Cross-Entropy โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 256 ชุดที่ 2

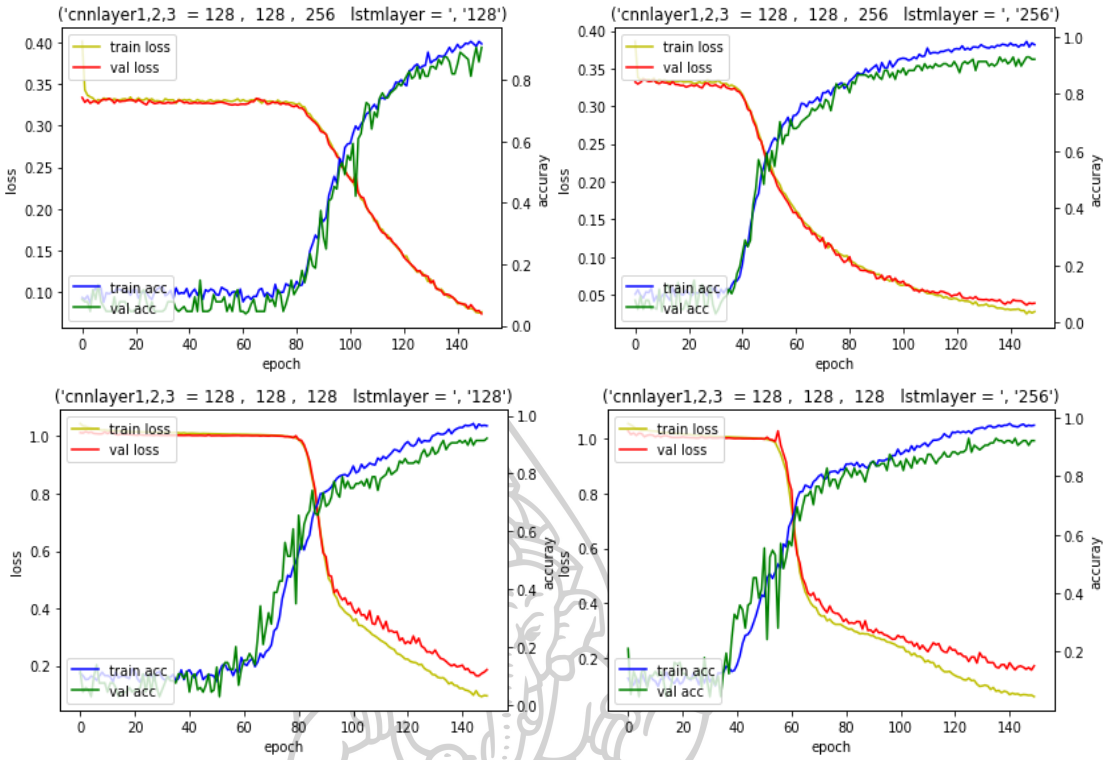


รูปที่ 4.34 การจับคู่ optimizer Adagrad และ Loss Binary Cross-Entropy โดยมี CNN 3 ชั้น ที่กำหนดชั้นแรกเป็น 256 ชุดที่ 3

4.2.4 เปรียบเทียบผลลัพธ์ของการทดลองและคัดเลือกแบบจำลองที่ความเหมาะสมที่สุด

เนื่องจากการทดลองมีกราฟของเส้นการเรียนรู้ของแบบจำลองจำนวนมาก แสดงให้เห็นว่าแบบจำลองสามารถเรียนรู้จากชุดข้อมูลที่ใช้ในการศึกษาครั้งนี้ได้จากการกำหนดในแต่ละชั้นของ CNN และจำนวน Units ของ LSTM ตั้งแต่ 64 128 256

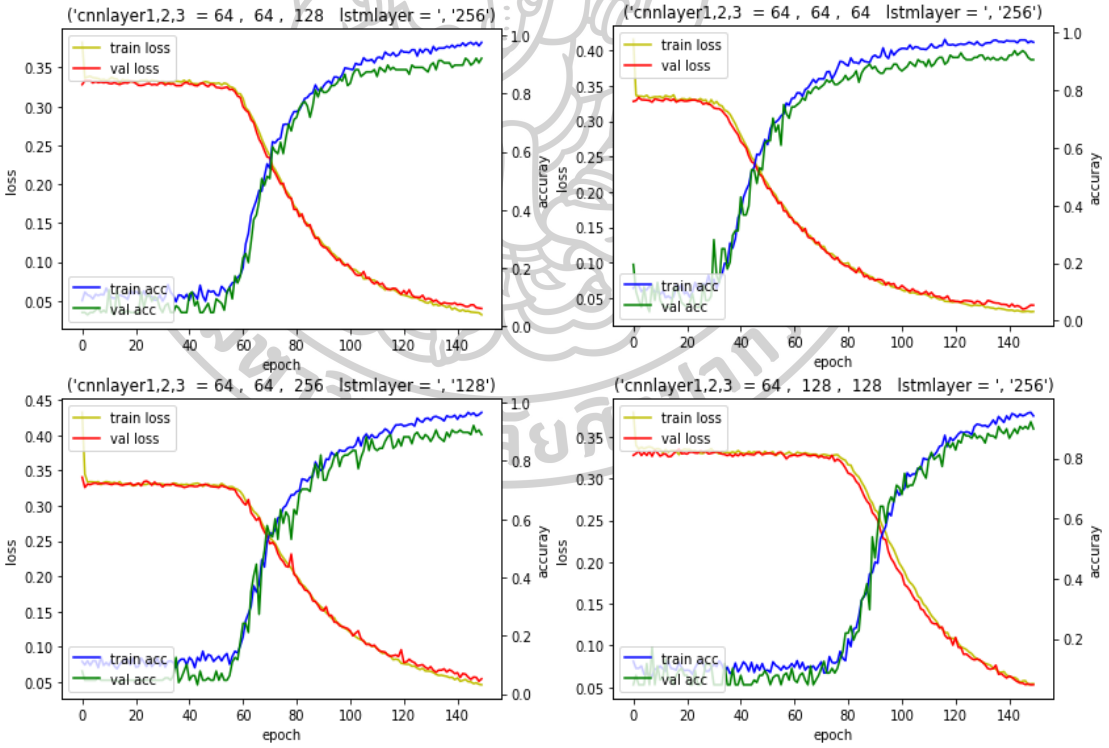
เมื่อเปรียบเทียบการเปลี่ยนแปลงในแต่ละจำนวน Units ของ LSTM พบว่าถ้าหากมีการเพิ่มขึ้นในแต่ละชั้นของ LSTM ในขณะที่มีการกำหนดชั้นของ CNN ที่เท่ากันส่งผลให้แบบจำลองสามารถที่จะเรียนรู้การรู้จำได้ดียิ่งขึ้น ไม่ว่าจะเป็นการเลือก Loss แบบ Categorical Hinge และ Binary Cross-Entropy หรือ การเลือกจำนวนชั้นของ CNN เป็น 2 ชั้น และ 3 ชั้น โดยกราฟตัวอย่างที่แสดงให้เห็นถึงการเพิ่มขึ้นของจำนวน Units ของ LSTM แสดงได้ดังรูป 4.35 จากข้อมูลรูปดังกล่าวจะสังเกตเห็นว่า กราฟแสดงผลลัพธ์ของการเรียนรู้ได้ดีขึ้นมากเมื่อเปลี่ยน Units ที่ 128 เป็น 256 ดังนั้นการกำหนดเลือกจำนวน Units ในชั้นของ LSTM ที่ 256 มีความเหมาะสมที่สุด



รูปที่ 4.35 การเปรียบเทียบการเปลี่ยนแปลงจำนวน Units ของ LSTM โดย 2 รูปด้านบนคือ Loss แบบ Binary Cross-Entropy และ 2 รูปด้านล่างแบบ Categorical Hinge

เนื่องจากการทดลองดังกล่าวแสดงให้เห็นว่าการเลือกจำนวน kernel แบบใดส่งผลให้แบบจำลองสามารถแสดงกราฟการเรียนรู้ออกมาได้ดี โดยการทดลองดังกล่าวกำหนดให้ใช้จำนวนเฟรมที่น้อยที่สุดนั่นคือ 3 เฟรม และใช้จำนวนของชุดข้อมูลทั้งหมดที่มีในกระบวนการของการฝึกสอน ซึ่งในการนำเสนอนั้น ชุดข้อมูลจะถูกแบ่งออกเป็นชุดข้อมูลสำหรับการทดสอบ ทำให้ชุดข้อมูลในการฝึกสอนลดน้อยลงและส่งผลต่อการเรียบรู้ของแบบจำลองในด้านลบด้วย อีกทั้งยังต้องทำการทดลองกับชุดข้อมูลทั้ง 5 เฟรม และ 10 เฟรมอีกด้วย และอีกหนึ่งเหตุผลของการเลือกแบบจำลองในแต่ละชั้นของ CNN จะสังเกตว่า ในแต่ละชั้นที่มีการทดลองแบบ 2 ชั้น ก็สามารถให้ผลลัพธ์ของการเรียนรู้ได้ดี แต่ในการทดลองของการเลือกจำนวนเฟรมที่ 5 เฟรม หรือ 10 เฟรม ที่มีความซับซ้อนมากกว่าจำนวนชั้นเพียง 2 ชั้น ไม่เพียงพอต่อการเรียนรู้ของแบบจำลองเนื่องจากความซับซ้อนของข้อมูลที่มากขึ้น และการเลือกจำนวน Kernel ในการนำเสนอของงานวิจัย Deep learning โดยใช้ CNN ไม่นิยมที่จะเลือกให้ในแต่ละชั้นของ CNN ให้มีจำนวน Kernel ที่มากหรือน้อยสลับกันไป แต่การเลือกนิยมเลือกแบบน้อยไปมากซึ่งหมายถึงการเลือกชั้นที่ CNN ชั้นแรกที่น้อยที่สุดและเพิ่ม kernel ขึ้นใน

ชั้นถัดไปหรือเท่าเดิม เพื่อการสกัดที่มากขึ้นในชั้นถัดๆ ไป จากการทดลองการเลือก Loss แบบ Categorical Hinge ให้ผลลัพธ์ของการฝึกสอนได้ดีแต่เมื่อพิจารณาที่รูปแบบของกราฟจากการเลือก Loss แบบ Binary Cross-Entropy ทำได้ดีกว่า แม้ว่าในตอนสุดท้ายจะแสดงค่าของการฝึกสอนที่ใกล้เคียงกันก็ตาม โดยกราฟที่จะนำมาพิจารณาในขั้นตอนสุดท้ายแสดงได้ดังรูป 4.36 เมื่อพิจารณาจากกราฟ ทุกกราฟแสดงผลลัพธ์ของการเรียนรู้ของแบบจำลองได้ดีและการกำหนดชั้นของ CNN ที่ 64 เท่ากันทุกชั้นดูจะเป็นกราฟที่ดีที่สุดเนื่องจากความซับซ้อนของข้อมูลที่น้อยกับความเรียบง่ายของแบบจำลองส่งผลให้ประสิทธิภาพและความเหมาะสมของการทดลองนี้ออกมาดีและมีความน่าสนใจ แต่ด้วยเหตุผลที่กล่าวไปข้างต้นจึงขอเลือกการตั้งค่าการเลือกจำนวน Kernel แต่ละชั้นของ CNN และจำนวน Units ของ LSTM ที่ CNN = 64, 64, 128 และ LSTM = 256 เพื่อรองรับการสร้างแบบจำลองจากข้อมูลที่มีความซับซ้อนมากขึ้นอีกระดับหนึ่ง เป็นแบบจำลองที่มีความเหมาะสมที่สุดในการทดลองครั้งนี้

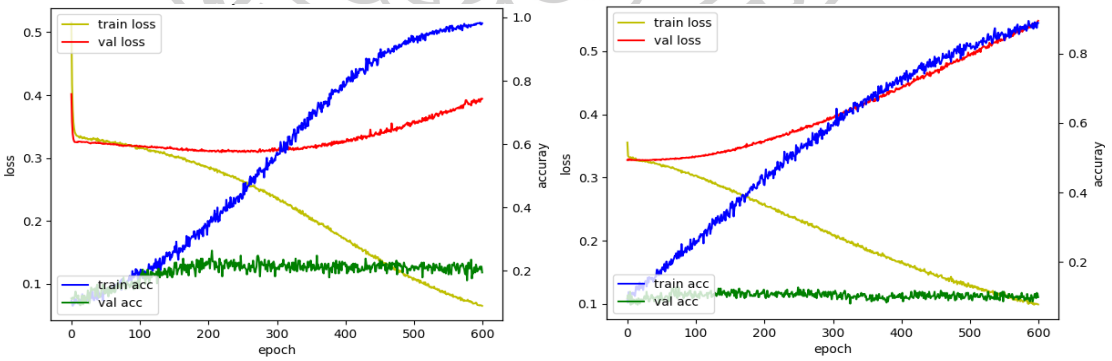


รูปที่ 4.36 การเปรียบเทียบการเปลี่ยนแปลงจำนวน Kernel แต่ละชั้นของ CNN

4.3 การทดลองใช้แบบจำลองกับชุดข้อมูลรูปภาพที่นำเสนอ

4.3.1 การทดลองกับชุดข้อมูลรูปภาพแบบเส้นรอบริมฝีปาก

หลังจากการคัดเลือกแบบจำลองในหัวข้อก่อนหน้าแล้ว การทดลองครั้งนี้เป็นการทดสอบกับแบบจำลองที่ผ่านการคัดเลือกกับชุดข้อมูลรูปภาพแบบเส้นรอบริมฝีปากเพื่อวัดประสิทธิภาพของชุดข้อมูลที่ได้นำเสนอไปก่อนหน้านี้ในเรื่องของการขยับและการเปลี่ยนแปลงไปของริมฝีปากในแต่ละเฟรมที่มีการออกเสียง โดยที่เส้นดังกล่าวมาจากรูปภาพตามแบบต้นฉบับทุกประการซึ่งเส้นที่ได้เป็นการสร้างขึ้นใหม่โดยการสร้างใช้การอ้างอิงจากจุดในตำแหน่งเดียวกันและลากเส้นเชื่อมตามที่ได้กล่าวไปในหัวข้อของชุดข้อมูลรูปภาพแบบเส้นรอบริมฝีปาก ในการตั้งค่าการทดลองเราจะใช้จำนวนชุดมูล 1590 ตัวอย่างใน Training Process ที่จะแบ่งออกเป็น 80% สำหรับการ train และ 20% สำหรับการ Validation และอีก 1060 สำหรับการทดสอบ รวมข้อมูลทั้งหมดเป็น 2650 ตัวอย่างจากฐานข้อมูลที่ได้นำมาใช้ ในการทดลองครั้งนี้แตกต่างจากการทดลองในการหาแบบจำลองที่มีความเหมาะสมที่สุดเนื่องจากข้อมูลในการ train น้อยลงมากอีกทั้งยังใช้จำนวนเฟรมที่ 10 เฟรมซึ่งแบ่งออกเป็นเฟรมรูปภาพแบบเต็มริมฝีปากและเฟรมรูปภาพแบบครึ่งริมฝีปากอีกด้วย ทั้งนี้เพื่อการสังเกตการเรียนรู้ของแบบจำลองในกรณีที่มีความซับซ้อนมากที่สุด และจำนวนรอบที่ใช้ในการ train คือ 600 รอบ และแบบจำลองจะถูกบันทึกค่าพารามิเตอร์ต่างๆที่รอบสุดท้าย กราฟการ train แสดงได้ดังรูปที่ 4.37



รูปที่ 4.37 กราฟแสดงการเรียนรู้ของแบบจำลองที่พัฒนาในระยะที่ 1 กับชุดข้อมูลรูปภาพแบบเส้นรอบริมฝีปากโดยกราฟด้านซ้ายเป็นแบบเฟรมเต็มริมฝีปากและกราฟด้านขวาเป็นแบบครึ่งเฟรมริมฝีปาก

เมื่อพิจารณาจากกราฟของการของการ train แล้วพบว่าแบบจำลองที่ใช้ชุดข้อมูลรูปภาพแบบเส้นรอบริมฝีปากไม่สามารถที่จะเรียนรู้จากคุณลักษณะดังกล่าวได้ในกระบวนการ train นั้นให้ผลลัพธ์ที่ดีแต่กระบวนการ Validation แบบจำลองไม่สามารถรู้จำชุดข้อมูลได้ ผลลัพธ์ของการทดสอบกับชุดข้อมูลสำหรับการทดสอบเป็นดังตารางที่ 4.1 เมื่อพิจารณาจากตารางผลลัพธ์ที่ได้ถือว่าต่ำมากเทียบกับชุดข้อมูลที่ใช้ในการทดสอบ 1060 ที่แต่ละคลาสมี 106 ซึ่งคลาสของเลขหนึ่งหรือว่า Zero ให้ผลลัพธ์ที่ดีที่สุดที่ 35 ครั้งในรูปแบบของเต็มเฟรมริมฝีปาก

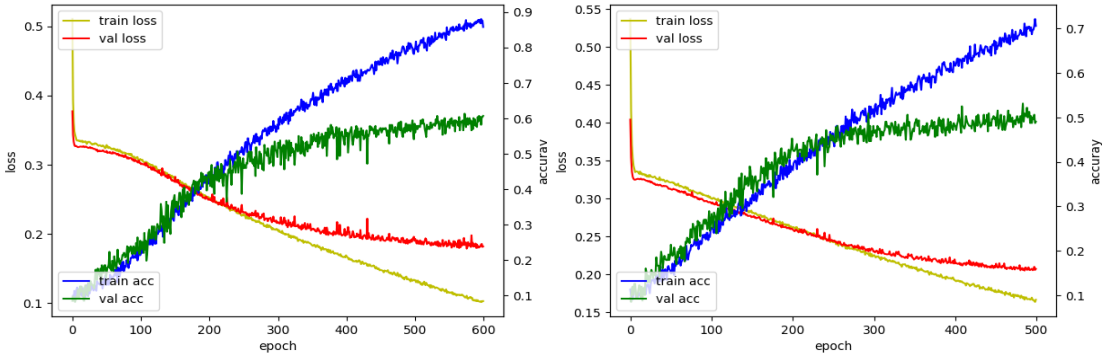
ตารางที่ 4.1 ผลลัพธ์ของการรู้จำจากชุดข้อมูลรูปภาพสำหรับทดสอบแบบเส้นรอบริมฝีปาก

	อัตราการรู้จำ	จำนวนที่ตอบถูกในแต่ละคลาส									
		0	1	2	3	4	5	6	7	8	9
เต็มเฟรม	0.142	14	35	9	15	18	15	5	8	22	10
ครึ่งเฟรม	0.130	17	24	8	17	12	16	7	15	17	5

4.3.2 การทดลองกับชุดข้อมูลรูปภาพแบบมีสีเฉพาะบริเวณที่เป็นริมฝีปาก

การทดลองครั้งนี้เป็นการทดสอบกับแบบจำลองที่ผ่านการคัดเลือกกับชุดข้อมูลรูปภาพแบบมีสีเฉพาะบริเวณที่เป็นริมฝีปากเพื่อวัดประสิทธิภาพของชุดข้อมูลที่ได้นำเสนอไปก่อนหน้านี้ในเรื่องของการให้ความสำคัญกับบริเวณที่เป็นริมฝีปากเท่านั้นในการศึกษาว่าข้อมูลคุณลักษณะที่จะผ่านการสกัดจากแบบจำลองออกเป็นพารามิเตอร์ต่างๆ ที่ใช้ในการเรียนรู้การรู้จำของแบบจำลองว่า ชุดข้อมูลดังกล่าวมีความเหมาะสมหรือไม่ เช่นเดียวกับการตั้งค่าการทดลองก่อนหน้านั่นคือข้อมูล 1590 ตัวอย่างสำหรับการ Training Process ที่จะแบ่งออกเป็น 80% สำหรับการ train และ 20% สำหรับการ Validation และอีก 1060 สำหรับการทดสอบ รวมข้อมูลทั้งหมดเป็น 2650 และการทดลองที่ 10 เฟรมซึ่งแบ่งออกเป็นเฟรมรูปภาพแบบเต็มริมฝีปากและเฟรมรูปภาพแบบครึ่งริมฝีปาก จำนวนรอบคือ 600 รอบ และแบบจำลองจะถูกบันทึกค่าพารามิเตอร์ต่างๆที่รอบสุดท้าย กราฟการ train แสดงได้ดังรูปที่ 4.38 จากรูปจะสังเกตเห็นว่าแบบจำลองแสดงผลลัพธ์ของการเรียนรู้ได้ดีทั้งแบบเต็มเฟรมริมฝีปากและแบบครึ่งเฟรมริมฝีปาก ในชุดข้อมูลการทดลองนี้ จะเห็นผลลัพธ์ได้ชัดเจนกว่าการทดลองก่อนหน้าโดยยังมีความแตกต่างที่มองเห็นได้จากการทดลองในรูปแบบเต็มเฟรมริมฝีปากที่มีค่า

Accuracy ของการ Train และ Validation อยู่ที่ 0.857 และ 0.606 ตามลำดับ ในส่วนของครึ่งเฟรมริมฝีปากจะอยู่ที่ 0.706 และ 0.490 ตามลำดับ



รูปที่ 4.38 กราฟแสดงการเรียนรู้ของแบบจำลองที่พัฒนาในระยะที่ 1 กับชุดข้อมูลรูปภาพแบบมีสีเฉพาะบริเวณที่เป็นริมฝีปากโดยกราฟด้านซ้ายเป็นแบบเฟรมเต็มริมฝีปากและกราฟด้านขวาเป็นแบบครึ่งเฟรมริมฝีปาก

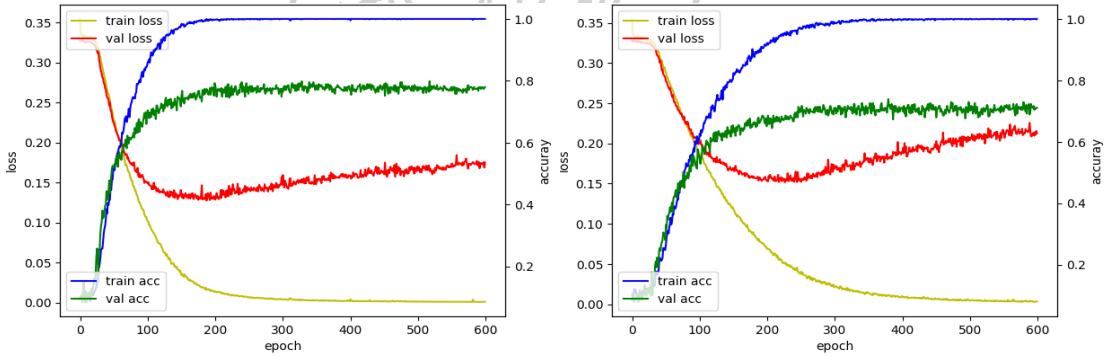
ผลลัพธ์ของการทดสอบกับชุดข้อมูลสำหรับการทดสอบเป็นไปดังตารางที่ 4.2 เมื่อพิจารณาผลลัพธ์ที่ได้จากตารางพบว่า แบบจำลองสามารถให้ประสิทธิภาพของการรู้จำที่ดีขึ้นมากเทียบกับการทดลองก่อนหน้าในการทดสอบโดยชุดข้อมูลที่ไม่ผ่านการ Train และข้อมูลแบบเต็มเฟรมริมฝีปากและแบบครึ่งเฟรมริมฝีปากยังให้ผลลัพธ์ที่ใกล้เคียงกันได้แม้ว่าขนาดของข้อมูลอินพุตจะหายไปครึ่งหนึ่งก็ตาม

ตารางที่ 4.2 ผลลัพธ์ของการรู้จำจากชุดข้อมูลรูปภาพสำหรับทดสอบแบบเติมสีเฉพาะริมฝีปาก

	อัตราการรู้จำ	จำนวนที่ตอบถูกในแต่ละคลาส									
		0	1	2	3	4	5	6	7	8	9
เต็มเฟรม	0.687	79	94	81	53	88	88	79	45	62	59
ครึ่งเฟรม	0.609	70	72	84	38	69	82	73	42	66	50

4.3.3 การทดลองกับชุดข้อมูลรูปภาพตามแบบต้นฉบับที่ตัดครอบมา

การทดลองโดยใช้ข้อมูลรูปตามแบบต้นฉบับเป็นการทดลองโดยใช้ข้อมูลรูปภาพริมฝีปากจากวิดีโอภายในฐานข้อมูลที่ไม่ได้ผ่านการดัดแปลงหรือปรับเปลี่ยนสิ่งอื่นใดจากตัวข้อมูลรูปภาพ โดยข้อมูลรูปภาพจะเห็นบริเวณที่เป็นริมฝีปากและบริเวณรอบนอกที่ติดกับบริเวณที่เป็นริมฝีปากระยะหนึ่ง ซึ่งการตั้งค่าการทดลองเป็นไปในทิศทางเดียวกันคือ ข้อมูล 1590 ตัวอย่างสำหรับการ Training Process ที่จะแบ่งออกเป็น 80% สำหรับการ train และ 20% สำหรับการ Validation และอีก 1060 สำหรับการทดสอบ รวมข้อมูลทั้งหมดเป็น 2650 และการทดลองที่ 10 เฟรมซึ่งแบ่งออกเป็นเฟรมรูปภาพแบบเต็มริมฝีปากและเฟรมรูปภาพแบบครึ่งริมฝีปาก จำนวนรอบคือ 600 รอบ และแบบจำลองจะถูกบันทึกค่าพารามิเตอร์ต่างๆที่รอบสุดท้าย รูปภาพกราฟการ Train แสดงได้ดังรูปที่ 4.39 จากรูปกราฟของการ Train แสดงผลลัพธ์ได้ดีที่สุดโดยที่ค่า Accuracy ของการ Train และ Validation สำหรับเต็มเฟรมริมฝีปากอยู่ที่ 1.000 และ 0.780 ตามลำดับ และ 1.000 และ 0.711 ตามลำดับ สำหรับครึ่งเฟรมริมฝีปาก



รูปที่ 4.39 กราฟแสดงการเรียนรู้ของแบบจำลองที่พัฒนาในระยะที่ 1 กับชุดข้อมูลรูปภาพแบบมีสีเฉพาะบริเวณที่เป็นริมฝีปากโดยกราฟด้านซ้ายเป็นแบบเฟรมเต็มริมฝีปากและกราฟด้านขวาเป็นแบบครึ่งเฟรมริมฝีปาก

ผลลัพธ์ของการทดสอบกับชุดข้อมูลสำหรับการทดสอบเป็นไปดังตารางที่ 4.3 เมื่อพิจารณาผลลัพธ์ที่ได้จากตารางพบว่า แบบจำลองให้ผลลัพธ์ของการเรียนรู้ที่ดีมากที่สุดเนื่องจากชุดข้อมูลในการทดสอบ 1 คลาสมี 106 ตัวอย่าง โดยคลาสที่ 5 หรือ Five สามารถตอบถูกได้มากถึง 103 ตัวอย่างซึ่งผิดเพียงแค่ 3 ตัวอย่างเท่านั้น นอกจากนี้ผลลัพธ์อื่นได้ก็ให้ประสิทธิภาพของการรู้จำที่ดีมากเช่นเดียวกัน ประสิทธิภาพของคลาสเลข 0 หรือ Zero และ 9 หรือ Nine ให้ผลลัพธ์ในรูปแบบ

ของเต็มเฟรมริมฝีปากและครึ่งเฟรมริมฝีปากที่ค่อนข้างแตกต่างกันมาก โดยที่คลาส Zero มีจำนวนที่ตอบถูกอยู่ที่ 84 ในแบบเต็มเฟรมในขณะที่ครึ่งเฟรมตอบถูกอยู่ที่ 69 และคลาส Nine จำนวนที่ตอบถูกในรูปแบบเต็มเฟรมมีเพียงแค่ 67 ครั้งน้อยกว่าแบบครึ่งเฟรมที่ตอบถูกอยู่ที่ 81 ครั้ง ทั้งนี้คลาสอื่นๆ มีการตอบถูกมากหรือน้อยแต่ยังมีความเกาะกลุ่มกันของตัวเลขความถูกต้อง

ตารางที่ 4.3 ผลลัพธ์ของการรู้จำจากชุดข้อมูลรูปภาพสำหรับทดสอบตามแบบต้นฉบับ

	อัตราการรู้จำ	จำนวนที่ตอบถูกในแต่ละคลาส									
		0	1	2	3	4	5	6	7	8	9
เต็มเฟรม	0.820	84	101	94	75	94	103	86	84	82	67
ครึ่งเฟรม	0.775	69	95	81	76	79	95	87	75	83	81

4.3.4 เปรียบเทียบผลลัพธ์ที่ได้จากการทดลอง

เมื่อเปรียบเทียบในแต่ละผลลัพธ์ของทดสอบจะสังเกตเห็นว่า ผลลัพธ์ของการทดลองมีค่าการรู้จำที่เพิ่มขึ้นโดยเรียงลำดับหัวข้อของการทดลองเป็นไปดังนี้ การทดลองโดยใช้ชุดข้อรูปภาพแบบเส้นรอบริมฝีปาก(แบบจำลองไม่เกิดการเรียนรู้จากชุดมูลที่กำหนดให้) การทดลองโดยใช้ชุดข้อมูลรูปภาพแบบเติมสีรอบริมฝีปาก(แบบจำลองมีการเรียนรู้) และ การทดลองโดยใช้ชุดข้อมูลรูปภาพตามแบบต้นฉบับ(แบบจำลองเรียนรู้ได้ดีที่สุด) และเมื่อสังเกตที่ความเข้าใกล้กันของประสิทธิภาพในรูปแบบเต็มเฟรมริมฝีปากและครึ่งเฟรมริมฝีปากพบว่า การใช้รูปภาพตามแบบต้นฉบับมีความเข้าใกล้กันมากที่สุดเมื่อใช้แบบจำลองที่ผ่านการคัดเลือก ซึ่งการเปรียบเทียบการทดลองครั้งนี้ทำให้ทราบว่าการใช้ข้อมูลตามแบบต้นฉบับให้ผลลัพธ์การเรียนรู้ของแบบจำลองที่มีประสิทธิภาพมากที่สุด

ตารางที่ 4.4 การเปรียบเทียบผลลัพธ์ในแต่ละชุดข้อมูลที่นำเสนอ

การทดลอง	อัตราการรู้จำ
ชุดข้อมูลรูปภาพแบบเส้นรอบริมฝีปากเต็มเฟรม	0.142
ชุดข้อมูลรูปภาพแบบเส้นรอบริมฝีปากครึ่งเฟรม	0.130
ชุดข้อมูลรูปภาพแบบเติมสีรอบริมฝีปากเต็มเฟรม	0.687
ชุดข้อมูลรูปภาพแบบเติมสีรอบริมฝีปากครึ่งเฟรม	0.609
ชุดข้อมูลรูปภาพตามแบบต้นฉบับเต็มเฟรม	0.820
ชุดข้อมูลรูปภาพตามแบบต้นฉบับครึ่งเฟรม	0.775

4.4 การพัฒนาปรับปรุงแบบจำลองเป็นระยะที่ 2

ในการทดลองก่อนหน้านี้เป็นการทดลองเพื่อทดสอบชุดข้อมูลต่างๆ จึงมีการกำหนดจำนวนเฟรมที่มากที่สุด และจำนวนรอบที่มากที่สุด แต่เมื่อพิจารณาจากกราฟของการ Train แบบจำลองแล้วจะสังเกตเห็นว่า กรณีที่ใช้ชุดข้อมูลตามแบบต้นฉบับที่แสดงดังรูปที่ 4.39 นั้น เส้นการเรียนรู้ของค่า Validation Loss นั้นเริ่มที่จะเพิ่มสูงขึ้นหลังจากผ่านรอบการเรียนรู้ที่ 300 รอบ และเมื่อแบบจำลองเสร็จสิ้นในขั้นตอนของการ Train ก็จะบันทึกพารามิเตอร์ที่รอบสุดท้าย ส่งผลให้แบบจำลองบันทึกที่รอบที่มีค่า Validation Loss ที่สูงขึ้น เพื่อแก้ปัญหานี้จึงลดจำนวนรอบของการ Train ลงมาเหลือ 300 รอบ และเรียกใช้ฟังก์ชัน ModelCheckpoint ซึ่งเป็นไลบรารี่ใน Tensorflow keras ฟังก์ชันดังกล่าวจะมีการกำหนดคำสั่งที่ชื่อว่า save_best_only = True และ monitor = ‘val_accuracy’ เมื่อเรียกใช้คำสั่งเหล่านี้ในกระบวนของการ Train แบบจำลองจะบันทึกพารามิเตอร์ทุกๆครั้งที่ แบบจำลองสามารถที่จะเรียนรู้และให้ค่า Validation Accuracy ใหม่ที่สูงขึ้นกว่าที่เคยบันทึกก่อนหน้านี้เท่านั้น ซึ่งในหัวข้อนี้จะแบ่งการทดลองออกเป็น การเลือกใช้จำนวนเฟรมที่ 3 5 และ 10 เฟรมร่วมด้วยและยังคงใช้การทดลองแบบเต็มเฟรมริมฝีปากและครึ่งเฟรมริมฝีปากอยู่ด้วย เพื่อแสดงถึงว่าแบบจำลองที่จะนำมาปรับเปลี่ยนใหม่นั้นให้ประสิทธิภาพที่ดีกว่าแบบจำลองเดิมในกรณีใดบ้าง

4.4.1 การพัฒนาแบบจำลอง

เนื่องจากแบบจำลองได้มาก่อนหน้าเป็นแบบจำลองที่สร้างขึ้นอย่างง่ายเพื่อการคัดเลือกจำนวน Kernel ในแต่ละชั้นของ CNN และจำนวน Units ที่ใช้ใน LSTM ซึ่งในการพัฒนานี้จะยังคงรูปแบบของแบบจำลองเดิมเอาไว้โดยมีการปรับเปลี่ยนดังนี้

1) การเพิ่มอัลกอริทึม padding = ‘same’

การใช้เทคนิคนี้จะช่วยให้อินพุตและเอาต์พุตนั้นมีค่าที่เท่ากันช่วยรักษา feature maps ในกระบวนการของ CNN โดยการเพิ่มพิกเซลบริเวณขอบของรูปภาพอินพุต เพื่อป้องกันของสูญเสียข้อมูลบริเวณขอบของรูปภาพอินพุตเนื่องจากรูปภาพบริเวณขอบจะถูกคอนโวลูชันน้อยกว่าบริเวณที่เป็นตรงกลาง การใช้ padding = ‘same’ จะช่วยทำให้ในแต่ละพิกเซลได้รับการประมวลผลที่เท่าเทียมกัน ซึ่งการเพิ่มฟังก์ชันนี้เข้ามา เนื่องจากการทดลองที่ใช้ข้อมูลรูปภาพแบบมีสีเฉพาะบริเวณที่เป็นริมฝีปากนั้นให้ประสิทธิภาพที่น้อยกว่าการใช้ข้อมูลรูปภาพตามแบบต้นฉบับที่มองเห็นข้อมูล

รอบๆ บริเวณริมฝีปาก นั่นหมายความว่าข้อมูลรูปภาพบริเวณขอบนั้นมีความสำคัญต่อการสร้างแบบจำลอง

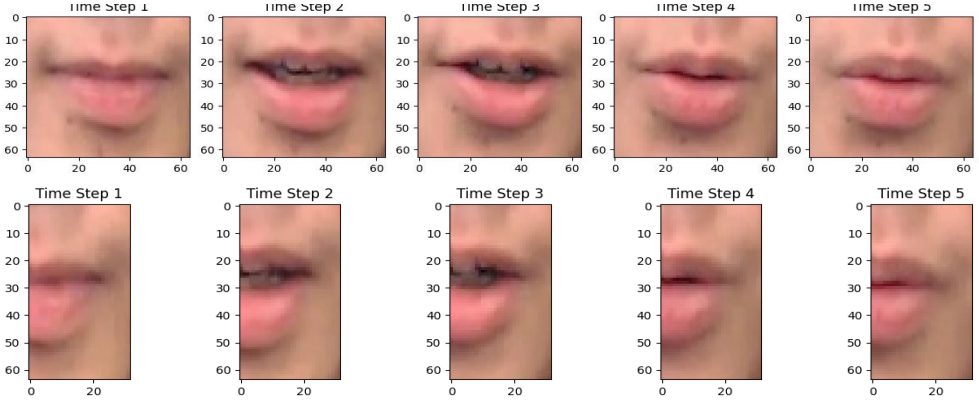
2) การเพิ่มชั้นคอนโวลูชันแทนที่ชั้นของพลูลิ่ง

การใช้ padding = ‘same’ เป็นการคงขนาดของรูปภาพเอาไว้ และการลดชั้นพลูลิ่งและเพิ่มชั้นการประมวลผลคอนโวลูชันก็เป็นการกระทำในทำนองเดียว โดยแบบจำลองเดิมที่มีชั้น Max pooling ถึง 3 ชั้น ในแบบจำลองจะนำเสนอนี้จะลดชั้น Max pooling ลงเหลือ 2 ชั้น โดยที่ชั้นสุดท้ายจะถูกแทนที่ด้วยชั้นคอนโวลูชั้น 128 เป็นการเพิ่มการสกัดเอาคุณลักษณะออกมาแทนที่จะลดขนาดของรูปภาพลง ซึ่งทำให้เกิดการเพิ่มชั้น Dense จาก 128 เป็น 1024 เพื่อรองรับการเรียนรู้ของพารามิเตอร์ที่มีจำนวนมากขึ้น

ซึ่งโปรแกรมที่สร้างแบบจำลองที่พัฒนาในระยะที่ 2 แสดงได้ดังรูปที่ 4.40 และตัวอย่างของชุดข้อมูลที่ใช้ในการทดลองทั้งแบบเต็มเฟรมและครึ่งเฟรมแสดงได้ดังรูปที่ 4.41

```
model = Sequential()

model.add(Conv2D(64, (3,3), activation='relu',padding='same', input_shape=(64,64,3)))
model.add(MaxPool2D((2, 2)))

model.add(Conv2D(64, (3,3), activation='relu',padding='same'))
model.add(MaxPool2D((2, 2)))

model.add(Conv2D(128, (3,3), activation='relu',padding='same'))
model.add(Conv2D(128, (3,3), activation='relu',padding='same'))
#model.add(MaxPool2D((2, 2)))
model.add(Flatten())
model.add(Dense(1024, activation='relu'))
model.add(Dropout(0.2))
model2 = Sequential()
model2.add(TimeDistributed(model,input_shape = (10,64,64,3)))
model2.add(LSTM(256))
model2.add(Dense(256, activation='relu'))
model2.add(Dropout(0.2))
model2.add(Dense(n_labels, activation="softmax"))
```

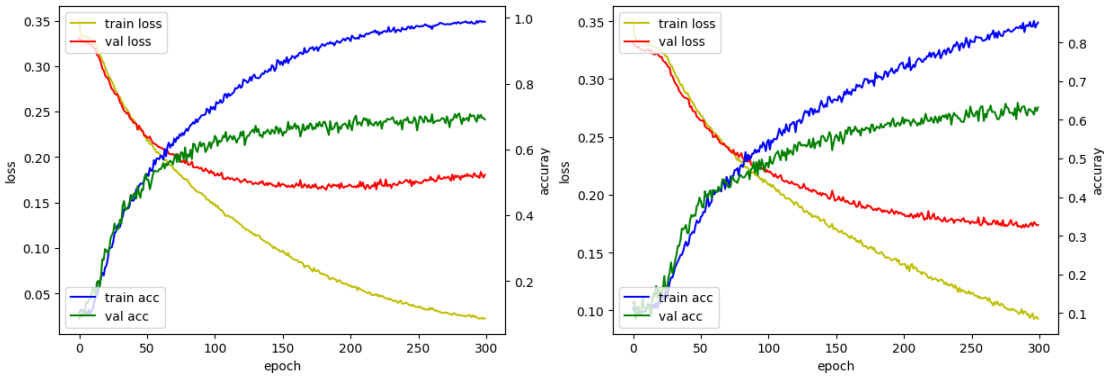
รูปที่ 4.40 โปรแกรมที่ใช้สร้างแบบจำลองที่พัฒนาในระยะที่ 2



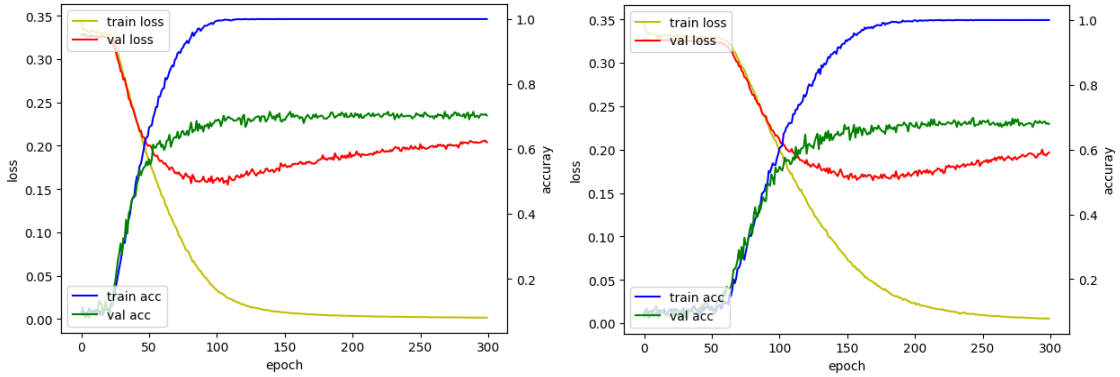
รูปที่ 4.41 ตัวอย่างแสดงการเปรียบเทียบชุดข้อมูลแบบเต็มเฟรมและครึ่งเฟรม

4.4.2 การเปรียบเทียบประสิทธิภาพของแบบจำลองกรณี 3 เฟรม

การเปรียบเทียบประสิทธิภาพของแบบจำลองที่พัฒนาในระยะที่ 1 และที่พัฒนาในการทดลองใช้ข้อมูลจำนวนเฟรมทั้งหมด 3 เฟรม ครั้งนี้มีการตั้งค่าข้อมูลต่างๆ ดังนี้ ข้อมูลที่ใช้ใน Training Process มี 2120 ตัวอย่าง และแบ่งออกเป็น 80% สำหรับการ train และ 20% สำหรับการ Validation และอีก 530 สำหรับการทดสอบ ทั้งนี้จำนวนที่ใช้ในการทดสอบ 1060 นั้นเยอะเกินไปคิดเป็น 40% ของข้อมูลทั้งหมดที่มีดังนั้นจึงลดลงมาที่ 530 คิดเป็น 20% ของข้อมูลทั้งที่มี รวมข้อมูลทั้งหมด 2650 กราฟของการ Train 3 เฟรมแบบเต็มเฟรมและครึ่งเฟรมของแบบจำลองที่พัฒนาในระยะที่ 1 และแบบจำลองที่พัฒนาในระยะที่ 2 แสดงได้ดังรูปที่ 4.42 และ 4.43 และตารางแสดงมาตรวัดต่างๆ แสดงดังตารางที่ 4.5 ในส่วนผลของการทดสอบแสดงได้ดังรูปที่ 4.44 และ 4.45 รวมถึง ตารางที่ 4.6



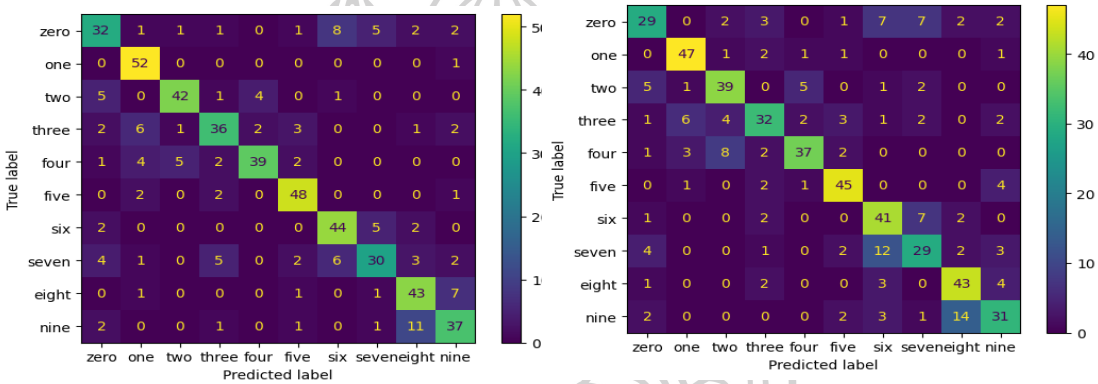
รูปที่ 4.42 กราฟการ Train ของแบบจาลองที่พัฒนาในระยะที่ 1 3 เฟรม โดยที่ด้านซ้ายเป็นแบบ เต็มเฟรมและด้านขวาเป็นแบบครึ่งเฟรม



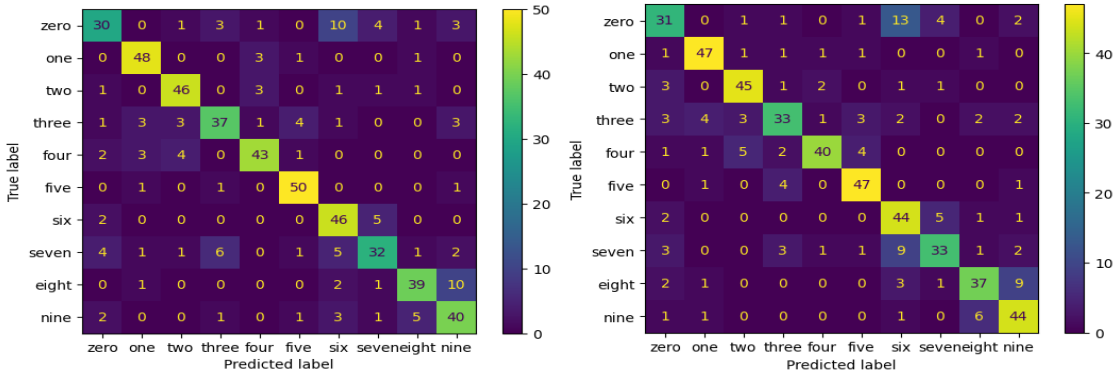
รูปที่ 4.43 กราฟการ Train ของแบบจาลองที่พัฒนาในระยะที่ 2 3 เฟรม โดยที่ด้านซ้ายเป็นแบบ เต็มเฟรมและด้านขวาเป็นแบบครึ่งเฟรม

ตารางที่ 4.5 ตารางการเปรียบเทียบผลลัพธ์ที่ได้ของการ Train ของแบบจำลองที่พัฒนาในระยะที่ 1 และแบบจำลองที่พัฒนาในระยะที่ 2 3 เฟรม

	มาตรวัด		
	Acc/Loss	Val Acc/Loss	Checkpoint
แบบจำลองที่พัฒนาในระยะที่ 1 เต็มเฟรม	0.985/0.026	0.710/0.179	281
แบบจำลองที่พัฒนาในระยะที่ 2 เต็มเฟรม	1,000/0.010	0.715/0.170	140
แบบจำลองที่พัฒนาในระยะที่ 1 ครึ่งเฟรม	0.838/0.098	0,644/0.172	285
แบบจำลองที่พัฒนาในระยะที่ 2 ครึ่งเฟรม	0.998/0.011	0.698/0.182	244



รูปที่ 4.44 Confusion Matrix ของแบบจำลองที่พัฒนาในระยะที่ 1 3 เฟรม โดยที่ด้านซ้ายเป็นแบบเต็มเฟรมและด้านขวาเป็นแบบครึ่งเฟรม



รูปที่ 4.45 Confusion Matrix ของแบบจำลองที่พัฒนาในระยะที่ 2 3 เฟรม โดยที่ด้านซ้ายเป็นแบบเต็มเฟรมและด้านขวาเป็นแบบครึ่งเฟรม

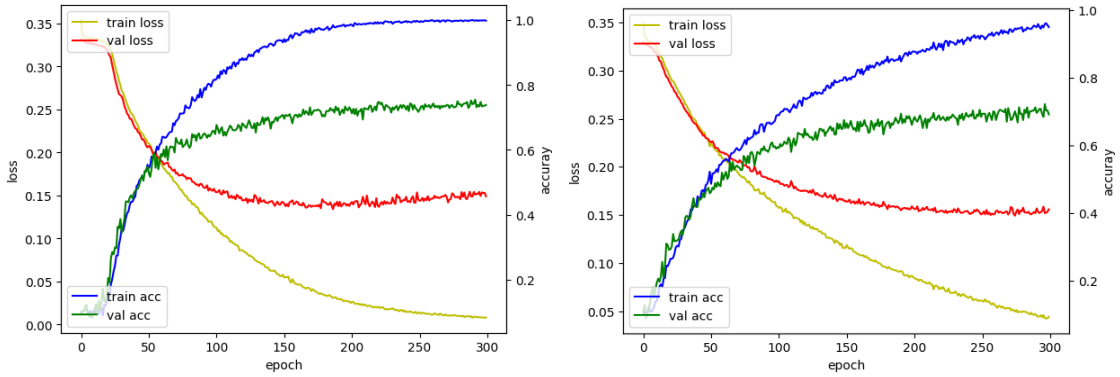
ตารางที่ 4.6 ตารางการเปรียบเทียบประสิทธิภาพของแบบจำลองที่พัฒนาในระยะที่ 1 และแบบจำลองที่พัฒนาในระยะที่ 2 3 เฟรม

	อัตราการรู้จำ
แบบจำลองที่พัฒนาในระยะที่ 1 เต็มเฟรม	0.760
แบบจำลองที่พัฒนาในระยะที่ 2 เต็มเฟรม	0.775
แบบจำลองที่พัฒนาในระยะที่ 1 ครึ่งเฟรม	0.703
แบบจำลองที่พัฒนาในระยะที่ 2 ครึ่งเฟรม	0.757

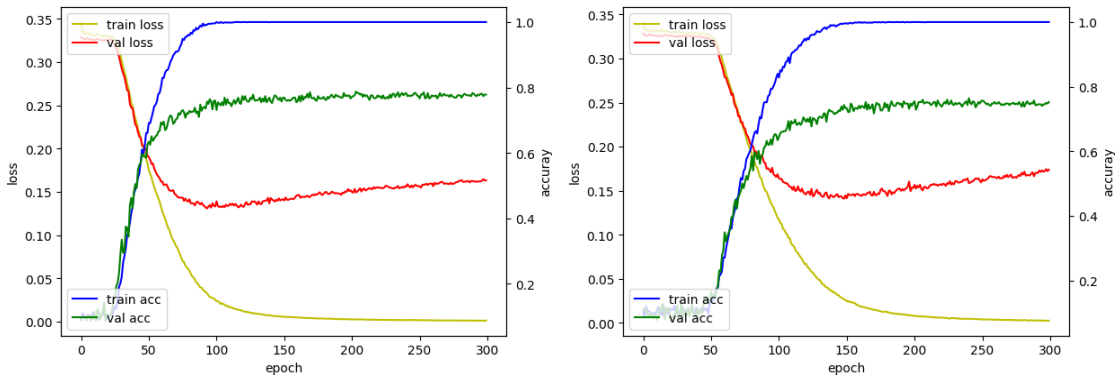
จากรูปที่ 4.41 และ 4.42 เมื่อพิจารณากราฟการเรียนรู้ของแบบจำลองพบว่า มีความแตกต่างอย่างเห็นได้ชัดว่าแบบจำลองมีความพัฒนาขึ้นอย่างมากจากการปรับเปลี่ยนโครงสร้างของสถาปัตยกรรม แต่ค่าของการ Train ในผลลัพธ์สุดท้ายของทั้ง 2 แบบจำลองให้ผลลัพธ์ที่ใกล้เคียงกัน และเมื่อเทียบกับประสิทธิภาพในการทดสอบของแบบจำลองในชุดมูลที่ไม่ได้ผ่านการ Train มาก่อน พบว่า แบบจำลองมีความพัฒนาขึ้นมากในกรณีเป็นครึ่งเฟรมจาก 0.703 เป็น 0.757 และเมื่อเปรียบความแตกต่างของรูปแบบเต็มเฟรมริมฝีปากและครึ่งเฟรมริมฝีปากพบว่ามีความแตกต่างของประสิทธิภาพอัตราการรู้จำที่น้อยลง

4.4.3 การเปรียบเทียบประสิทธิภาพของแบบจำลองกรณี 5 เฟรม

การเปรียบเทียบประสิทธิภาพของแบบจำลองที่พัฒนาในระยะที่ 1 และที่พัฒนาในการทดลองใช้ข้อมูลจำนวนเฟรมทั้งหมด 5 เฟรม ครั้งนี้มีการตั้งค่าข้อมูลต่างๆ ดังนี้ ข้อมูลที่ใช้ใน Training Process มี 2120 ตัวอย่าง และแบ่งออกเป็น 80% สำหรับการ train และ 20% สำหรับการ Validation และอีก 530 สำหรับการทดสอบ ทั้งนี้จำนวนที่ใช้ในการทดสอบ 1060 นั้นเยอะเกินไปคิดเป็น 40% ของข้อมูลทั้งหมดที่มีดังนั้นจึงลดลงมาที่ 530 คิดเป็น 20% ของข้อมูลทั้งที่มี รวมข้อมูลทั้งหมด 2650 กราฟของกรา Train 5 เฟรมแบบเต็มเฟรมและครึ่งเฟรมของแบบจำลองที่พัฒนาในระยะที่ 1 และแบบจำลองที่พัฒนาในระยะที่ 2 แสดงได้ดังรูปที่ 4.46 และ 4.47 และตารางแสดงมาตรวัดต่างๆ แสดงดังตารางที่ 4.7 ในส่วนผลของการทดสอบแสดงได้ดังรูปที่ 4.48 และ 4.49 รวมถึง ตารางที่ 4.8



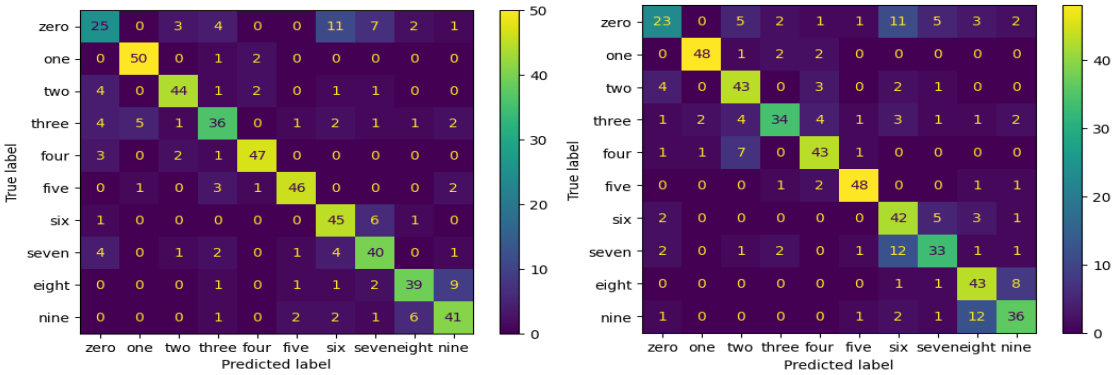
รูปที่ 4.46 กราฟการ Train ของแบบจาลองที่พัฒนาในระยะที่ 1 5 เฟรม โดยที่ด้านซ้ายเป็นแบบ เต็มเฟรมและด้านขวาเป็นแบบครึ่งเฟรม



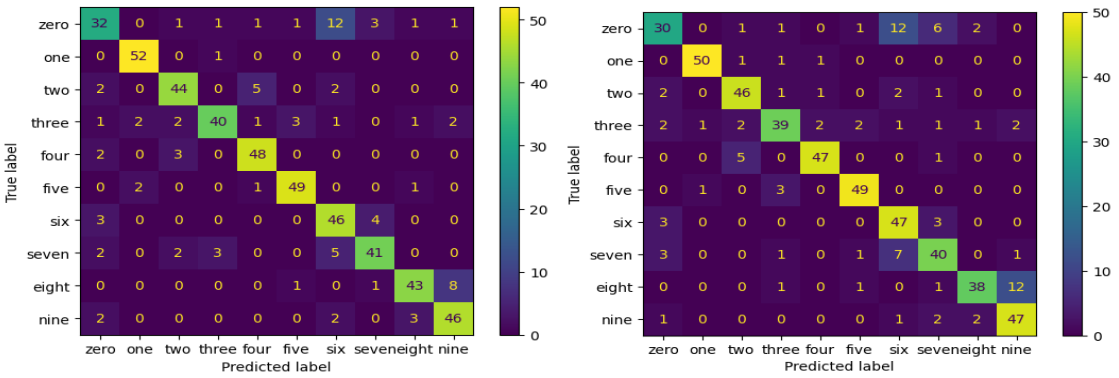
รูปที่ 4.47 กราฟการ Train ของแบบจำลองที่พัฒนาในระยะที่ 2 5 เฟรม โดยที่ด้านซ้ายเป็นแบบเต็มเฟรมและด้านขวาเป็นแบบครึ่งเฟรม

ตารางที่ 4.7 ตารางการเปรียบเทียบผลลัพธ์ที่ได้ของการ Train ของแบบจำลองที่พัฒนาในระยะที่ 1 และแบบจำลองที่พัฒนาในระยะที่ 2 5 เฟรม

	มาตรวัด		
	Acc/Loss	Val Acc/Loss	Checkpoint
แบบจำลองที่พัฒนาในระยะที่ 1 เต็มเฟรม	0.997/0.009	0.755/0.151	292
แบบจำลองที่พัฒนาในระยะที่ 2 เต็มเฟรม	1.000/0.003	0.788/0.149	204
แบบจำลองที่พัฒนาในระยะที่ 1 ครึ่งเฟรม	0.959/0.043	0.722/0.152	297
แบบจำลองที่พัฒนาในระยะที่ 2 ครึ่งเฟรม	1.000/0.005	0.764/0.159	241



รูปที่ 4.48 Confusion Matrix ของแบบจาลองที่พัฒนาในระยะที่ 1 5 เฟรม โดยที่ด้านซ้ายเป็น แบบเต็มเฟรมและด้านขวาเป็นแบบครึ่งเฟรม



รูปที่ 4.49 Confusion Matrix ของแบบจำลองที่พัฒนาในระยะที่ 2 5 เฟรม โดยที่ด้านซ้ายเป็นแบบเต็มเฟรมและด้านขวาเป็นแบบครึ่งเฟรม

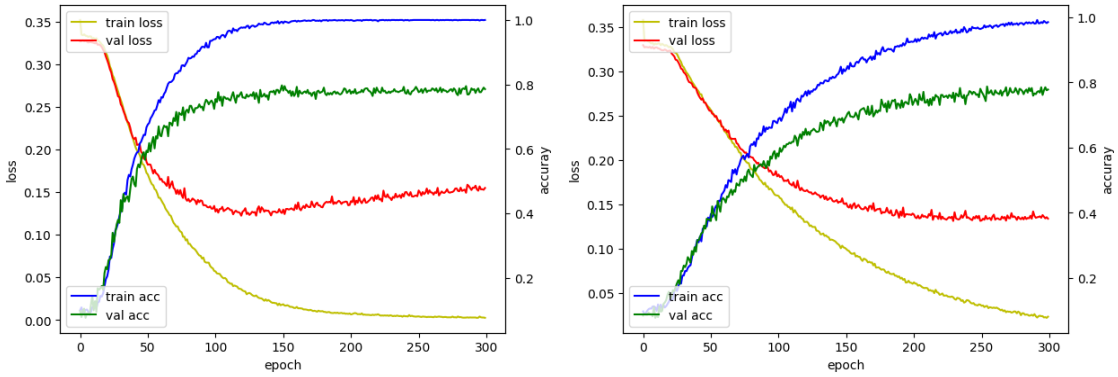
ตารางที่ 4.8 ตารางการเปรียบเทียบประสิทธิภาพของแบบจำลองที่พัฒนาในระยะที่ 1 และแบบจำลองที่พัฒนาในระยะที่ 2 5 เฟรม

	อัตราการรู้จำ
แบบจำลองที่พัฒนาในระยะที่ 1 เต็มเฟรม	0.779
แบบจำลองที่พัฒนาในระยะที่ 2 เต็มเฟรม	0.832
แบบจำลองที่พัฒนาในระยะที่ 1 ครึ่งเฟรม	0.742
แบบจำลองที่พัฒนาในระยะที่ 2 ครึ่งเฟรม	0.817

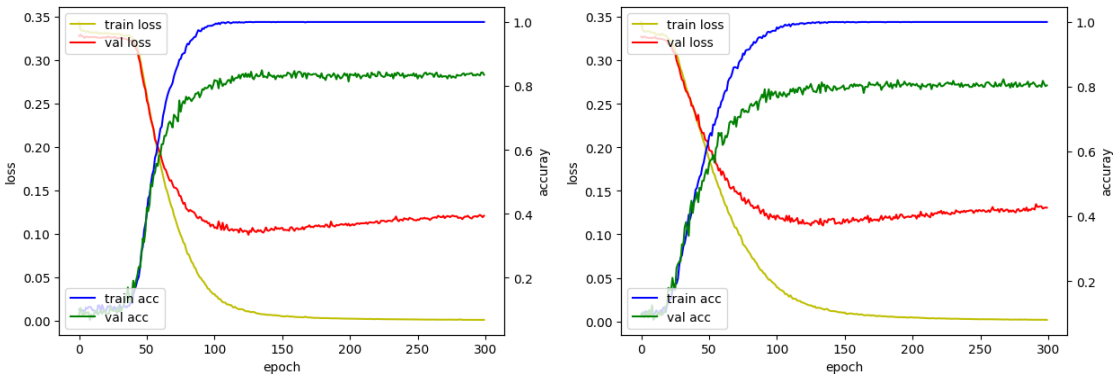
จากรูปที่ 4.46 และ 4.47 กราฟการ Train ของแบบจำลองในช่วงแรกพบว่าแบบจำลองสามารถเรียนรู้ได้เร็วกว่าอย่างเห็นได้ชัดจากการใช้แบบจำลองที่พัฒนาในระยะที่ 2 รวมถึงในส่วนท้ายของการเรียนรู้ของแบบจำลองพบว่าแบบจำลองที่พัฒนาในระยะที่ 2 นั้นสามารถให้ผลลัพธ์ได้ดีกว่าในทุกกรณีอีกทั้ง Checkpoint สุดท้ายของแบบจำลองยังให้ผลลัพธ์ที่น้อยลงอีกด้วยซึ่งผลลัพธ์ของการ Train แบบจำลองทั้งแบบเต็มเฟรมและครึ่งก็เฟรมก็มีความเข้าใกล้กันมากขึ้น และเมื่อเทียบกับประสิทธิภาพในการทดสอบของแบบจำลองในชุดมูลที่ไม่ได้ผ่านการ Train มาก่อนพบว่าแบบจำลองที่พัฒนาในระยะที่ 2 โดยใช้ชุดข้อมูลแบบครึ่งเฟรมยังสามารถมีประสิทธิภาพเหนือว่า (0.817) แบบจำลองที่พัฒนาในระยะที่ 1 โดยใช้ชุดข้อมูลแบบเต็ม (0.779) โดยที่แบบจำลองที่พัฒนาในระยะที่ 2 แบบครึ่งเฟรมให้ประสิทธิภาพการรู้จำที่ 0.817 และแบบเต็มเฟรมอยู่ที่ 0.832 ซึ่ง เป็นประสิทธิภาพที่สามารถเทียบเคียงกันได้เช่นเดียว แต่ให้ผลลัพธ์ที่สูงกว่าการทดลองเปรียบเทียบแบบ 3 เฟรม

4.4.4 การเปรียบเทียบประสิทธิภาพของแบบจำลองกรณี 10 เฟรม

การเปรียบเทียบประสิทธิภาพของแบบจำลองที่พัฒนาในระยะที่ 1 และที่พัฒนาในการทดลองใช้ข้อมูลจำนวนเฟรมทั้งหมด 10 เฟรม ครั้งนี้มีการตั้งค่าข้อมูลต่างๆ ดังนี้ ข้อมูลที่ใช้ใน Training Process มี 2120 ตัวอย่าง และแบ่งออกเป็น 80% สำหรับการ train และ 20% สำหรับการ Validation และอีก 530 สำหรับการทดสอบ ทั้งนี้จำนวนที่ใช้ในการทดสอบ 1060 นั้นเยอะเกินไปคิดเป็น 40% ของข้อมูลทั้งหมดที่มีดังนั้นจึงลดลงมาที่ 530 คิดเป็น 20% ของข้อมูลทั้งที่มี รวมข้อมูลทั้งหมด 2650 กราฟของกรา Train 10 เฟรมแบบเต็มเฟรมและครึ่งเฟรมของแบบจำลองที่พัฒนาในระยะที่ 1 และแบบจำลองที่พัฒนาในระยะที่ 2 แสดงได้ดังรูปที่ 4.50 และ 4.51 และตารางแสดงมาตรวัดต่างๆ แสดงดังตารางที่ 4.9 ในส่วนผลของการทดสอบแสดงได้ดังรูปที่ 4.51 และ 4.52 รวมถึง ตารางที่ 4.10



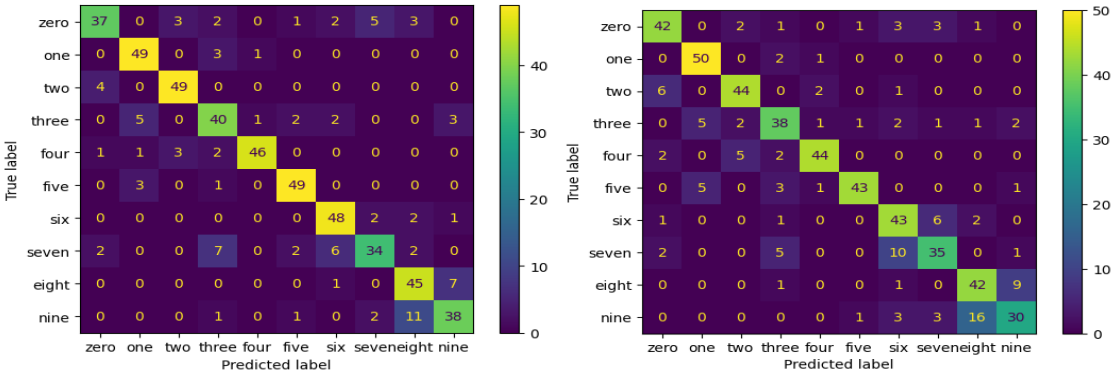
รูปที่ 4.50 กราฟการ Train ของแบบจาลองที่พัฒนาในระยะที่ 1 10 เฟรม โดยที่ด้านซ้ายเป็นแบบเต็มเฟรมและด้านขวาเป็นแบบครึ่งเฟรม



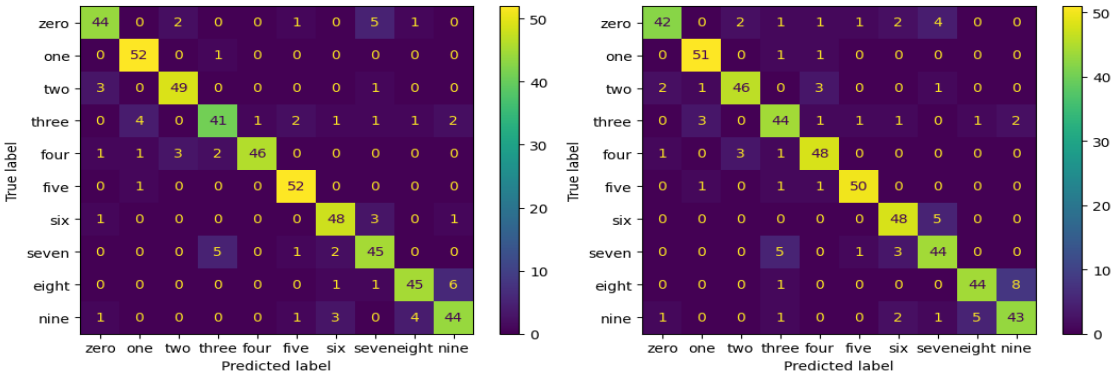
รูปที่ 4.51กราฟการ Train ของแบบจำลองที่พัฒนาในระยะที่ 2 10 เฟรม โดยที่ด้านซ้ายเป็นแบบเต็มเฟรมและด้านขวาเป็นแบบครึ่งเฟรม

ตารางที่ 4.9 ตารางการเปรียบเทียบผลลัพธ์ที่ได้ของการ Train ของแบบจำลองที่พัฒนาในระยะที่ 1 และแบบจำลองที่พัฒนาในระยะที่ 2 10 เฟรม

	มาตรวัด		
	Acc/Loss	Val Acc/Loss	Checkpoint
แบบจำลองที่พัฒนาในระยะที่ 1 เต็มเฟรม	0.995/0.018	0.797/0.125	150
แบบจำลองที่พัฒนาในระยะที่ 2 เต็มเฟรม	1.000/0.007	0.849/0.102	136
แบบจำลองที่พัฒนาในระยะที่ 1 ครึ่งเฟรม	0.972/0.035	0.787/0.133	255
แบบจำลองที่พัฒนาในระยะที่ 2 ครึ่งเฟรม	1.000/0.002	0.823/0.126	268



รูปที่ 4.52 Confusion Matrix ของแบบจาลองที่พัฒนาในระยะที่ 1 10 เฟรม โดยที่ด้านซ้ายเป็น แบบเต็มเฟรมและด้านขวาเป็นแบบครึ่งเฟรม



รูปที่ 4.53 Confusion Matrix ของแบบจำลองที่พัฒนาในระยะที่ 2 10 เฟรม โดยที่ด้านซ้ายเป็นแบบเต็มเฟรมและด้านขวาเป็นแบบครึ่งเฟรม

ตารางที่ 4.10 ตารางการเปรียบเทียบประสิทธิภาพของแบบจำลองที่พัฒนาในระยะที่ 1 และแบบจำลองที่พัฒนาในระยะที่ 2 10 เฟรม

	อัตราการรู้จำ
แบบจำลองที่พัฒนาในระยะที่ 1 เต็มเฟรม	0.821
แบบจำลองที่พัฒนาในระยะที่ 2 เต็มเฟรม	0.879
แบบจำลองที่พัฒนาในระยะที่ 1 ครึ่งเฟรม	0.775
แบบจำลองที่พัฒนาในระยะที่ 2 ครึ่งเฟรม	0.868

จากการทดลอง กราฟในรูปที่ 4.50 และ 4.51 แสดงถึงการเรียนรู้ของแบบจำลองที่เร็วอย่างมาก พิจารณาจากกราฟ ที่ค่า Validation Accuracy หลังจากที่ผ่านไป 150 รอบทั้งแบบจำลองที่พัฒนาในระยะที่ 1 และแบบจำลองที่พัฒนาในระยะที่ 2 ในรูปแบบเต็มเฟรมริมฝีปากดูมีแนวโน้มที่จะเพิ่มขึ้นเล็กน้อย และเมื่อเทียบกับประสิทธิภาพในการทดสอบของแบบจำลองในชุดมูลที่ไม่ได้ผ่านการ Train มาก่อนพบว่าแบบจำลองให้ค่าประสิทธิภาพของการรู้จำที่สูงขึ้นอย่างมากเมื่อเทียบในรูปแบบเต็มเฟรมเพิ่มขึ้น 0.821 เป็น 0.879 โดยที่แบบจำลองที่พัฒนาในระยะที่ 2 ในรูปแบบครึ่งเฟรมก็ยังคงให้ประสิทธิภาพที่สูงเทียบเคียงกับแบบจำลองที่พัฒนาในระยะที่ 2 ในรูปแบบเต็มเฟรมที่ 0.868 ซึ่งเป็นการเทียบเคียงที่มีค่าใกล้กันมากที่สุดในการทดลองทั้ง 3 กรณี

4.4.5 สรุปผลการเปรียบเทียบของแบบจำลองที่พัฒนาในระยะที่ 1 และแบบจำลองที่พัฒนาในระยะที่ 2

เมื่อเปรียบเทียบประสิทธิภาพของการทดลองทั้ง 3 กรณีในรูปแบบของกราฟการ Train ของแบบจำลองและประสิทธิภาพเมื่อทดสอบโดยใช้ชุดข้อมูลที่ไม่ผ่านการ Train พบว่า

1) กราฟของการ Train แบบจำลองทั้ง 3 กรณีมีแนวโน้มที่จะเรียนรู้พารามิเตอร์จากชุดข้อมูลที่กำหนดให้เร็วขึ้นเมื่อใช้แบบจำลองที่พัฒนาในระยะที่ 2 เมื่อเทียบกันในขณะที่ชุดข้อมูลในการฝึกสอนมีขนาดและจำนวนที่เท่ากัน แม้ในกรณีของการทดลอง 3 เฟรม จะพบว่าในส่วนท้ายของการทดลอง แบบจำลองให้ผลลัพธ์ที่ใกล้เคียงกันที่ค่าของ Val Accuracy ระหว่างแบบจำลองที่พัฒนาในระยะที่ 1 และแบบจำลองที่พัฒนาในระยะที่ 2 ก็ตาม แต่ในการทดลองที่ 5 เฟรมค่าของ Val Accuracy ที่ได้จากการฝึกสอนแม้จะมีค่าที่ใกล้เคียงกันแต่เมื่อนำแบบจำลองมาทดสอบกับชุดข้อมูลสำหรับการทดสอบพบว่า แบบจำลองที่พัฒนาในระยะที่ 2 ให้ประสิทธิภาพที่เหนือกว่าแบบจำลองที่พัฒนาในระยะที่ 1 อย่างเห็นได้ชัด หรือกระทั่งการทดลองในกรณีของ 10 เฟรมนั้น แบบจำลองสามารถที่จะให้ผลลัพธ์การเรียนรู้ที่ดีขึ้นได้ จากการทดลองพบว่าเนื่องจากแบบจำลองที่พัฒนาในระยะที่ 2 นั้นมีพารามิเตอร์ที่มากขึ้น เมื่อใส่ชุดข้อมูลที่มากขึ้นเข้าไป จึงเห็นความแตกต่างของการเรียนรู้ของแบบจำลองได้ดียิ่งขึ้น ส่งผลไปยังประสิทธิภาพของการทดสอบกับชุดข้อมูลสำหรับทดสอบที่ดีขึ้นด้วย

2) ประสิทธิที่ดีขึ้นจากชุดข้อมูลของการทดสอบที่เหมือนกันๆ เนื่องจากการทดลองใช้ข้อมูลของการทดสอบที่เหมือนกันทั้ง 3 กรณี เมื่อเปรียบเทียบการทดลองในกรณีเดียวกันแล้ว แม้ว่ากรณี

ของ 3 เฟรมจะให้ความแตกต่างที่น้อยกว่ากรณีอื่น แต่ก็ยังคงให้ผลลัพธ์ของประสิทธิภาพที่ดีกว่าแบบจำลองที่พัฒนาในระยะที่ 1 ในกรณีของการทดลองที่ครึ่งเฟรมริมฝีปากและเมื่อเปรียบเทียบในทุกกรณี ประสิทธิภาพของแบบจำลองที่มีการใช้จำนวนเฟรมที่เพิ่มขึ้นมีแนวโน้มที่จะสามารถให้ประสิทธิการรู้จำของแบบจำลองที่สูงขึ้นเมื่อทดสอบโดยชุดข้อมูลทดสอบ อีกทั้งแบบจำลองที่พัฒนาในระยะที่ 2 ยังสามารถทำให้การเปรียบเทียบของชุดข้อมูลรูปแบบเต็มเฟรมและครึ่งเฟรมนั้นมีความใกล้เคียงกันมากยิ่งขึ้นแม้ว่าขนาดของอินพุตจะหายไปครึ่งหนึ่งก็ตาม

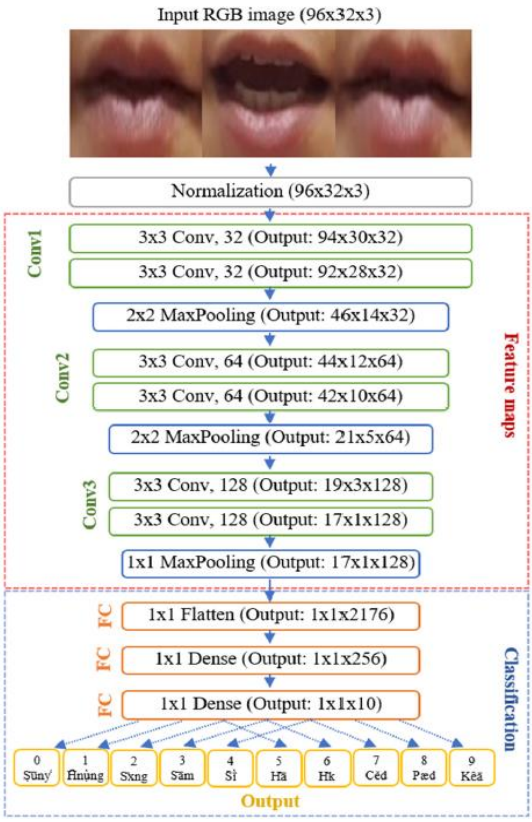
3) การนำแบบจำลองมาทดสอบผ่านชุดข้อมูลสำหรับการทดลอบจะแสดงค่าที่ทำนาย (Predicted Label) และค่าจริง (True Label) และถ้าหากค่าทั้งสองนี้เหมือนกันหมายความว่าการรู้จำของแบบจำลองที่มีต่อข้อมูลตัวอย่างที่นำมาทดสอบนั้นถือว่าตอบถูก ซึ่งค่าการสอบถูกในแต่ละคลาสจะแสดงได้โดย Confusion Matrix ที่แสดงไปก่อนหน้านี้ โดยที่สีของการตอบถูกแสดงเป็นโทนสีม่วงมีความหมายว่าแบบจำลองนั้นมีอัตราการรู้จำต่อคลาสนั้นๆ น้อย แต่ถ้าหากสีของการตอบถูกเป็นสีโทนเหลืองสว่างมีความหมายว่าแบบจำลองนั้นมีอัตราการรู้จำต่อคลาสนั้นๆ สูง เมื่อพิจารณาในแต่ละคลาสพบว่า คลาส 1 (One) มีอัตราของการตอบถูกที่ค่อนข้างสูงในทุกๆ แบบจำลอง ตามมาด้วยคลาส 5 (Five) ขึ้นอยู่กับกรณีของจำนวนเฟรมที่ทำการทดลอง ในด้านของการตอบผิด คลาสที่มีการตอบผิดมากที่สุดคือ 0 (Zero) ซึ่งการทำนายของแบบจำลองในตอบในคลาสนี้จะทำนายว่าเป็นคลาส 6 (Six) เป็นส่วนใหญ่ ซึ่งเหตุการณ์นี้เกิดขึ้นกับคลาสของ 7 (Seven) ด้วย อีกหนึ่งคลาสที่มีการตอบผิดมากนั่นคือ 9 (Nine) ซึ่งจะมีการทำนายเป็นคลาส 8 (Eight) เสียส่วนใหญ่ ซึ่งอัตราการตอบผิดของแบบจำลองจะค่อยๆ ลดน้อยลงเมื่อมีการเพิ่มขึ้นของจำนวนเฟรมและการปรับเปลี่ยนการเรียนรู้ของพารามิเตอร์ที่มากขึ้นเช่นการพัฒนาแบบจำลองในระยะที่ 2 ตามที่ได้นำเสนอ

4) ประสิทธิภาพของการเรียนรู้ของแบบจำลองที่ทำการฝึกสอนโดยชุดข้อมูลครึ่งเฟรมริมฝีปาก จากการทดลองพัฒนาประสิทธิภาพของการรู้จำของแบบจำลองโดยใช้ชุดข้อมูลรูปภาพแบบครึ่งเฟรมริมฝีปากที่ให้ประสิทธิภาพที่ใกล้เคียงกับการรู้จำของแบบจำลองโดยใช้ชุดข้อมูลรูปภาพแบบเต็มเฟรมริมฝีปากซึ่งเป็นประสิทธิภาพของการรู้จำที่ค่อนข้างสูงโดยลดลงมาเล็กน้อย แต่ก็ไม่สามารถที่จะทำให้ประสิทธิภาพนั้นเพิ่มขึ้นจนเท่ากันได้ แต่ผลลัพธ์ของเวลาในการใช้ต่อชุดข้อมูลทดสอบ 1 ตัวอย่างเร็วกว่าประมาณ 0.2 วินาที ขึ้นอยู่กับการใช้งานจริงว่าพอใจในรูปแบบใด ต้องการความเร็วที่เพิ่มขึ้นเล็กน้อยโดยที่ประสิทธิภาพก็ไม่ได้ต่างกันมากสามารถเลือกใช้แบบครึ่งเฟรมริมฝีปากได้ แต่ถ้าหากเน้นประสิทธิภาพที่สูงเห็นสมควรให้ใช้แบบเต็มเฟรมริมฝีปาก

4.5 การวัดประสิทธิของแบบจำลอง

4.5.1 แบบจำลองที่จะนำมาเปรียบเทียบ

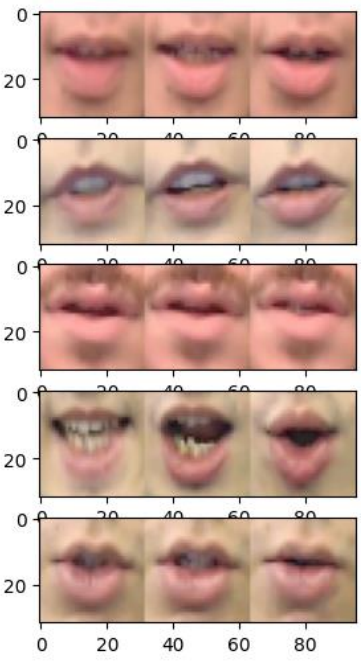
การวัดประสิทธิภาพของแบบจำลองในหัวข้อนี้ได้เปรียบเทียบกับงานวิจัยที่ได้นำเสนอเนื้อหาของการวิจัยที่มีความคล้ายคลึงกับวิทยานิพนธ์ฉบับนี้ โดยชื่อของงานวิจัยดังกล่าวมีชื่อว่า Improving the Recognition Performance of Lip Reading Using the Concatenated Three Sequence Keyframe Image Technique หรือจะใช้ชื่อย่อว่าเทคนิค C3-SKI โดยภายในงานวิจัยดังกล่าวได้พูดถึงวิธีการของเลือกเฟรมที่มีความสำคัญเพื่อจะนำมาสร้างแบบจำลองแล้วสามารถทำให้แบบจำลองเกิดการเรียนรู้ได้ดียิ่งขึ้น การทดลองดังกล่าวกำหนดการเลือกจำนวนเฟรมแบบ 3 เฟรม กำหนดจากช่วงของการเริ่มต้นปิดปาก หรือ Start-Lip Image (SLI) ช่วงกลางระหว่างการพูด หรือ Middle-Lip Image (MLI) ช่วงท้ายของการพูด End-Lip Image (ELI) โดยที่ขนาดของอินพุตที่ใช้คือ 96×32 พิกเซล เป็นการนำเสนอโดยการนำรูปภาพที่เลือกมาต่อกันแบบ Concatenated Image และมีแบบจำลองที่นำเสนอแสดงดังรูปที่ 4.53



รูปที่ 4.54 แสดงแบบจำลองที่นำเสนอโดยงานวิจัยที่จะนำเปรียบเทียบ

4.5.2 การตั้งค่าข้อมูลที่ใช้ในการทดลอง

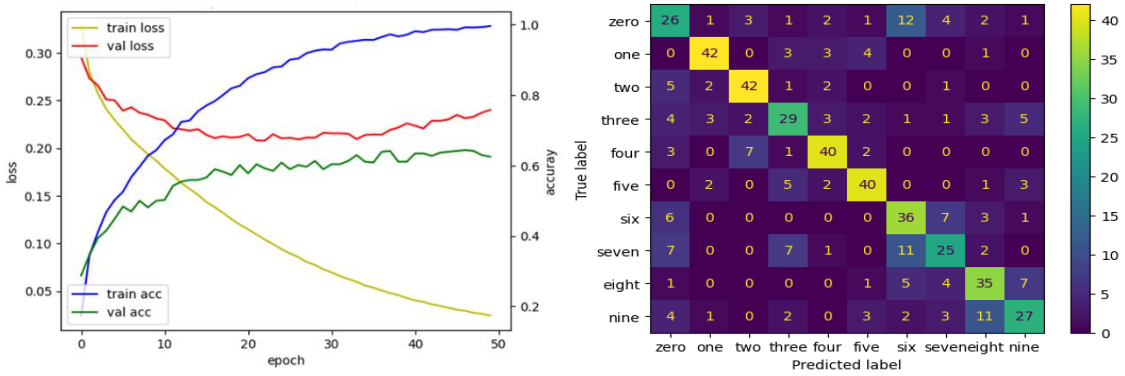
ข้อมูลที่จะนำมาใช้ในการวัดประสิทธิภาพของแบบจำลองที่ได้นำเสนอในวิทยานิพนธ์ฉบับนี้ แบ่งออกเป็นข้อมูลสำหรับการฝึกสอนและข้อมูลสำหรับการทดสอบ โดยที่ชุดข้อมูลทั้งสองอย่างมีปริมาณ 2650 ตัวอย่าง ข้อมูลที่ใช้ใน Training Process มี 2120 ตัวอย่าง และแบ่งออกเป็น 80% สำหรับการ train และ 20% สำหรับการ Validation และอีก 530 สำหรับการทดสอบ ซึ่งเป็นการตั้งค่าการทดสอบในลักษณะเดียวกับหัวข้อก่อนหน้าในการพัฒนาประสิทธิภาพของแบบจำลองที่จะนำเสนอในวิทยานิพนธ์ฉบับนี้ โดยแบบจำลองที่จะนำมาเปรียบเทียบจะใช้การตั้งค่าทุกอย่างเหมือนๆ กัน เปลี่ยนเพียงรูปแบบของการป้อนข้อมูลเข้าแบบจำลองและสถาปัตยกรรมที่นำเสนอ โดยที่ขนาดของอินพุตกรณี 3 เฟรมจะอยู่ที่ 32×96 พิกเซล ซึ่งเป็นขนาดของอินพุตที่ใช้ในการนำเสนอของงานวิจัยดังกล่าว ในการเปรียบครั้งนี้ได้เพิ่มเติมการเปรียบเทียบแบบ 5 เฟรมและ 10 เฟรมเข้าไปด้วย ทำให้จำเป็นต้องเป็นใช้การต่อกันของรูปภาพตามที่ได้นำเสนอในงานวิจัยดังกล่าว ในกรณี 5 เฟรมจะอยู่ที่ 32×160 พิกเซล และในกรณี 10 เฟรมจะอยู่ที่ 32×320 พิกเซล อีกทั้งยังใช้ Optimizer Adagrad และ Loss Binary Cross-Entropy ในการวัดประสิทธิภาพเนื่องจากการ Train แบบจำลองของ CNN แบบจำลองเกิด Overfit ได้ง่ายจึงกำหนดจำนวนรอบที่ 50 รอบ และใช้คำสั่ง Checkpoint เหมือนเดิม เพื่อให้เงื่อนไขต่างๆยังคงเหมือนกัน รูปตัวอย่างการต่อกันของอินพุตที่จะใช้ในแบบจำลองที่จะนำมาเปรียบเทียบแสดงดังรูปที่ 4.54



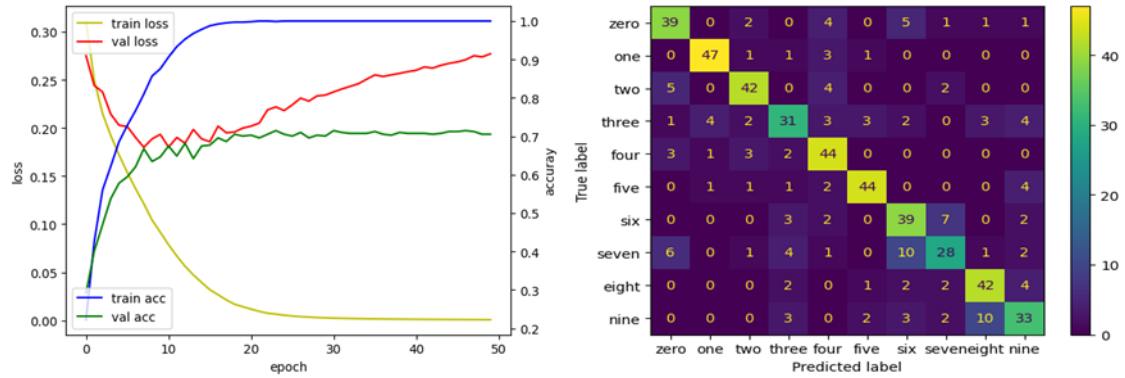
รูปที่ 4.55 ตัวอย่างของอินพุตที่ใช้ในแบบจำลองที่นำมาเปรียบเทียบ

4.5.3 การทดลองของแบบจำลองที่นำมาเปรียบเทียบ

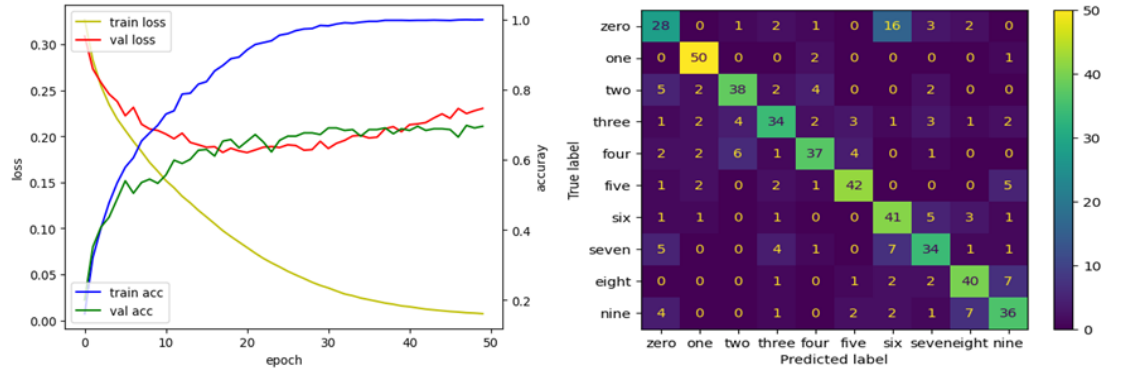
การทดลองจะเป็นการแสดงกราฟของการ Train แบบจำลองและ Confusion Matrix ของผลลัพธ์ที่ได้จากการทดสอบกับชุดข้อมูลการทดสอบโดยเรียงลำดับจาก 3 เฟรม 5 เฟรม และ 10 เฟรม รูปแสดงผลการทดลองดังกล่าวแสดงได้ดังรูปที่ 4.55 – 4.57 รวมถึงตารางที่ 4.11 และผลลัพธ์ของการทดสอบแสดงได้ดังตารางที่ 4.12



รูปที่ 4.56 กราฟการ Train และ Confusion Matrix ของแบบจำลองที่นำมาเปรียบเทียบ 3 เฟรม



รูปที่ 4.57 กราฟการ Train และ Confusion Matrix ของแบบจำลองที่นำมาเปรียบเทียบ 5 เฟรม



รูปที่ 4.58 กราฟการ Train และ Confusion Matrix ของแบบจำลองที่นำมาเปรียบเทียบ 10 เฟรม

ตารางที่ 4.11 ตารางการเปรียบเทียบผลลัพธ์ที่ได้ของการ Train ของแบบจำลองที่นำมาเปรียบเทียบ

	มาตรวัด		
	Acc/Loss	Val Acc/Loss	Checkpoint
3 เฟรม	0.993/0.029	0.644/0.232	47
5เฟรม	1.000/0.008	0.698/0.225	48
10 เฟรม	0.998/0.007	0.715/0.222	24

ตารางที่ 4.12 ตารางการเปรียบเทียบประสิทธิภาพของแบบจำลองที่นำมาเปรียบเทียบ

	อัตราการรู้จำ
3 เฟรม	0.645
5 เฟรม	0.717
10 เฟรม	0.734

4.5.4 เปรียบเทียบประสิทธิภาพของแบบจำลองที่นำเสนอกับแบบจำลองที่นำมาเปรียบเทียบ

ในหัวข้อนี้เป็นการวัดประสิทธิภาพของการรู้จำระหว่างแบบจำลองที่นำเสนอกับแบบจำลองที่นำมาเปรียบเทียบ เพื่อแสดงประสิทธิภาพของวิธีและแบบจำลองที่นำเสนอกับงานวิจัยในรูปแบบสากล ในการเปรียบเทียบของขั้นตอนนี้เป็นการเปรียบเทียบในรูปแบบเต็มเฟรมริมฝีปากเนื่องจากการดทอลงที่นำเสนอในงานวิจัยดังกล่าวนำเสนอในรูปแบบเต็มเฟรมในรูปแบบที่สามารถพบเจอได้เป็นมาตรฐาน แสดงได้ดังตารางที่ 4.13

ตารางที่ 4.13 ตารางการเปรียบเทียบประสิทธิภาพของแบบจำลองที่นำเสนอกับแบบจำลองที่จำมาเปรียบเทียบโดยชุดข้อมูลทดสอบในรูปแบบเต็มริมฝีปาก

แบบจำลอง	อัตราการรู้จำในแต่ละจำนวนเฟรม		
	3 เฟรม	5 เฟรม	10 เฟรม
C3-SKI	0.645	0.717	0.734
แบบจำลองที่นำเสนอ	**0.775**	**0.832**	**0.879**

จากรูปที่ 4.55 – 4.57 พบว่ากราฟของการ Train มีค่า Val Accuracy ที่เพิ่มขึ้นเมื่อการฝึกสอนของแบบจำลองผ่านไปได้ 20 รอบ หมายความว่าแบบจำอลงของ CNN นั้นสามารถที่จะเรียนรู้ข้อมูลรูปภาพได้เร็วและเกิด Overfit ได้ง่าย เพื่อความเท่าเทียมจึงใช้ฟังก์ชัน ModelCheckpoint เพื่อบันทึกเอาช่วงที่ดีที่สุดของการฝึกสอนของแบบจำลองเช่นเดียวกัน และหากเปรียบเทียบที่ค่าของ Accuracy และ Loss ของการ Train จะพบว่าเมื่อข้อมูลที่ใช้ในการฝึกสอนมีจำนวนเฟรมที่เพิ่มขึ้น แบบจำลองจะเรียนรู้ได้เร็วมากขึ้นถึงแม้ในตอนสุดท้ายของการ Train จะให้ค่าที่ใกล้เคียงกัน และจากตารางที่ 4.12 นำมาเปรียบเทียบกับตารางที่ 4.6 4.8 และ 4.10 โดยค่าของแบบจำลองที่พัฒนาในระยะที่ 2 ในรูปแบบเต็มเฟรมที่นำเสนอโดยวิทยานิพนธ์ฉบับนี้จะแสดงได้ดังตารางที่ 4.13 และเมื่อพิจารณาจากตารางดังกล่าวพบว่า แบบจำลองที่นำเสนอสามารถให้ประสิทธิภาพของการรู้จำที่ดีกว่าในทุกกรณีอีกทั้งยังสามารถบอกได้ว่า การเพิ่มขึ้นของจำนวนเฟรมหลังจากที่มีการใช้การคัดเลือกนำเอาชุดข้อมูลที่มีความเหมาะสมต่อการเรียนรู้ของแบบจำลองสามารถเพิ่มประสิทธิภาพของแบบจำลองได้จริงในทุกกรณี อีกทั้งแบบจำลองที่นำเสนอสามารถให้ค่าประสิทธิภาพที่ใกล้เคียงกันของอัตราการรู้จำในรูปแบบของชุดข้อมูลแบบเต็มเฟรมริมฝีปากและครึ่งเฟรมริมฝีปากโดยที่ยังคงมีศักยภาพที่สูง โดยที่แบบจำลองที่นำเสนอให้ประสิทธิภาพการรู้จำสูงสุดที่ 0.879 ซึ่งแบบจำลองที่นำมาเปรียบเทียบให้ประสิทธิภาพอยู่ที่ 0.734 ซึ่งน้อยกว่าแบบจำลองที่นำเสนอในรูปแบบ 3 เฟรม แบบครึ่งเฟรมริมฝีปากจากตารางที่ 4.6 ที่ให้ประสิทธิภาพอยู่ที่ 0.757

บทที่ 5
สรุปและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการพัฒนาประสิทธิภาพของการอ่านริมฝีปากผ่านการวิเคราะห์เฟรมสำคัญโดยใช้ CNN และ LSTM ที่ทำงานร่วมกันซึ่งมีความเหมาะสมกับการประมวลผลข้อมูลที่เป็นลำดับขั้นในแบบรูปภาพเช่นการใช้รูปภาพริมฝีปากที่เปลี่ยนแปลงไปในแต่ละช่วงของการเปล่งเสียงในงานการอ่านริมฝีปากนี้ อีกทั้งยังนำเสนอการลดขนาดของข้อมูลอินพุตของแบบจำลองที่สามารถคงไว้ซึ่งประสิทธิภาพของอัตราการรู้จำที่สูงคือการใช้ครึ่งเฟรมริมฝีปาก

ฐานข้อมูลที่ใช้และนำเสนอในวิทยานิพนธ์ฉบับนี้ใช้ฐานข้อมูลที่มีความเป็นสากลชื่อว่า AV Digits ที่ประกอบไปด้วยอาสาสมัคร 53 คน ในชุดข้อมูลที่เป็นตัวเลขภาษาอังกฤษมีทั้งอาสาสมัครที่เป็นเจ้าของภาษาและไม่ใช่เจ้าของภาษา ซึ่งความท้าทายของงานทางด้านการอ่านริมฝีปากคือรูปแบบที่เป็นคุณลักษณะเฉพาะของผู้พูดในการพูด เช่น บางผู้พูดมีการขยับริมฝีปากที่มากหรือน้อยแตกต่างกัน และในบางคำก็มีรูปแบบของการพูดที่คล้ายคลึงกัน ด้วยเหตุนี้สิ่งที่จำเป็นในการเพิ่มประสิทธิภาพของงานทางด้านการอ่านริมฝีปากคือการเลือกชุดข้อมูลในการฝึกสอนที่มีความเหมาะสมและเห็นถึงความเปลี่ยนแปลงของริมฝีปากได้ดี ดังนั้นแล้วภายในวิทยานิพนธ์ฉบับนี้ได้นำเสนอวิธีที่ใช้เพื่อคัดเลือกเฟรมริมฝีปากที่มีความสำคัญ

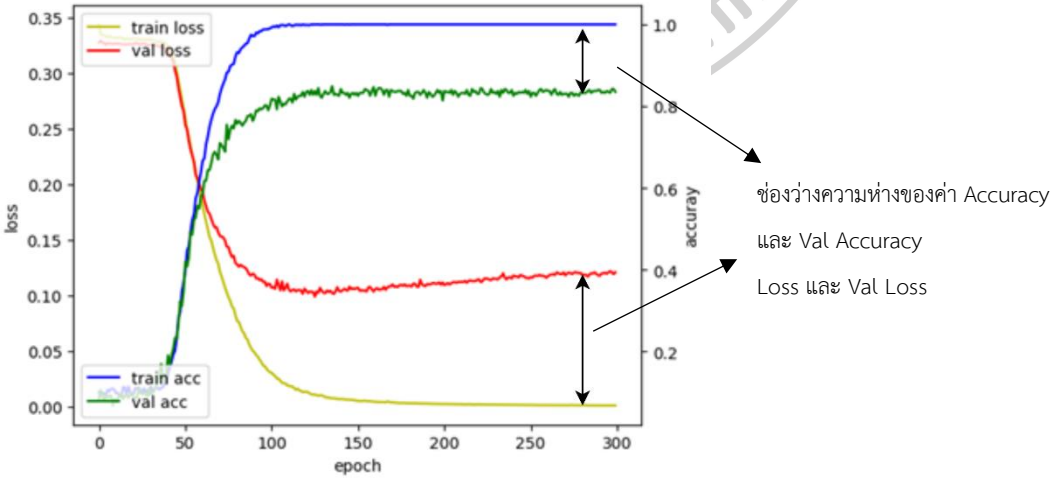
ในส่วนของการทดลองกำหนดตัวแปรในการเลือกใช้แบบจำลองที่จะนำเสนอโดยได้กำหนดการเลือกในแต่ละชั้น หรือ การเลือกใช้ Optimizer และ Loss ที่ดีที่สุด และได้นำเสนอการเปรียบเทียบที่มีความแตกต่างของชุดข้อมูลภายในฐานข้อมูลเดียวกัน เพื่อศึกษาความเหมาะสมหรือคุณลักษณะต่างๆของรูปภาพที่ได้นำเสนอในวิทยานิพนธ์ฉบับนี้ โดยผลลัพธ์ของการทดลองพบว่าการใช้ชุดข้อมูลรูปภาพตามแบบต้นฉบับให้ผลลัพธ์ที่ดีต่อการเรียนรู้ของแบบจำลองดีที่สุด นอกจากนี้เพื่อสร้างแบบจำลองที่ดีที่สุดจึงได้การพัฒนาต่อยอดแบบจำลองให้มีประสิทธิภาพเพิ่มขึ้นและลดช่องว่างระหว่างการใช้ชุดข้อมูลรูปภาพแบบเต็มเฟรมริมฝีปากและครึ่งเฟรมริมฝีปาก และผลลัพธ์ของการพัฒนาประสิทธิภาพแบบจำลองส่งผลให้แบบจำลองให้ประสิทธิภาพของการรู้จำที่สูง อีกทั้งยังลดช่องว่างของประสิทธิภาพระหว่างชุดข้อมูลรูปภาพแบบเต็มเฟรมริมฝีปากและครึ่งเฟรมริมฝีปากได้เป็นอย่างดี โดยที่มีประสิทธิภาพการรู้จำอยู่ที่ 0.879 สำหรับเต็มเฟรมริมฝีปาก และ 0.868 สำหรับครึ่งเฟรมริมฝีปากใน กรณี 10 เฟรมเพิ่มขึ้นจากแบบจำลองก่อนการพัฒนาที่ทำได้อยู่ที่ 0.821

สำหรับเต็มเฟรมริมฝีปาก และ 0.775 สำหรับครึ่งเฟรมริมฝีปาก ซึ่งเป็นประสิทธิภาพที่สูงที่สุดจากการพิจารณาทั้ง 3 กรณีคือ 3 เฟรม 5 เฟรม และ 10 เฟรม โดยการทดลองดังกล่าวแสดงให้เห็นว่าเมื่อชุดข้อมูลในการทดลองมีความเหมาะสมต่อการเรียนรู้ การเพิ่มขึ้นของจำนวนข้อมูลที่ใช้ในการเรียนรู้ส่งผลต่อประสิทธิภาพการเรียนรู้ของแบบจำลองได้เป็นอย่าดีในวิทยานิพนธ์เล่มนี้ได้เสนอเอาไว้ที่ 10 เฟรม และในด้านของการใช้แบบจำลองจริงหากต้องการความเร็วแต่ประสิทธิภาพที่ได้มีความแม่นยำที่สูงพอสมควรสามารถใช้ครึ่งเฟรมริมฝีปากได้ แต่ถ้าหากต้องการเน้นประสิทธิภาพให้สูงที่สุดควรที่จะใช้เต็มเฟรมริมฝีปาก ทั้งนี้ขึ้นกับความพอใจของผู้ใช้งาน

การวัดประสิทธิของแบบจำลองกับงานที่มีความคล้ายกันและได้รับการตีพิมพ์ทางวารสารไปก่อนหน้านี้พบว่า แบบจำลองที่นำเสนอสามารถให้ประสิทธิภาพการรู้จำที่ดีกว่า ภายใต้เงื่อนไขเดียวกันแตกต่างกันด้วยวิธีการที่นำเสนอ

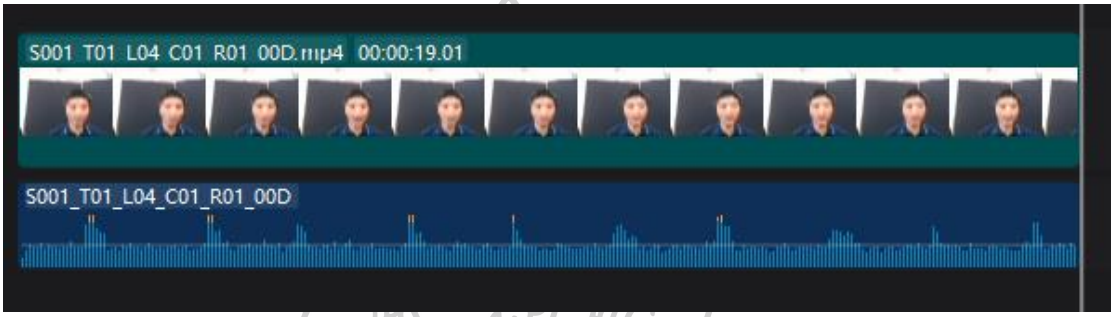
5.2 ปัญหาและข้อเสนอแนะ

1) ที่กราฟของการ Train เมื่อพิจารณาที่กราฟของการทดลองแล้ว แบบจำลองยังเกิด Overfit ที่เยอะพอสมควร ค่า Accuracy และ Validation Accuracy ยังคงมีความห่างกันอยู่ ถ้าหากลดความห่างตรงนี้ได้ ประสิทธิภาพกับชุดข้อมูลที่ใช้ทดสอบก็จะเพิ่มสูงขึ้นได้



รูปที่ 5.1 กราฟการ Train จากการทดลองที่มีความห่างกันเส้นการเรียนรู้

2) เมื่อใช้โปรแกรมตัดต่อจะสามารถมองเห็นช่วงเวลาที่เกิดการเปล่งเสียงได้ และถ้าหากใช้โปรแกรมอื่นใดหรือสามารถการสร้างแบบจำลองเฉพาะสำหรับในขั้นตอนของการเตรียมการข้อมูลจะช่วยลดระยะเวลาการเตรียมการข้อมูลลงได้มากถ้าหากจำนวนข้อมูลมีจำนวนกว่านี้ รูปแสดงกราฟแท่งของการเปล่งเสียงแสดงได้ดังรูป 4.59



รูปที่ 5.2 กราฟแท่งแสดงการเปล่งเสียงในโปรแกรมตัดต่อ

5.3 แนวทางการพัฒนาต่อยอด

1) พัฒนาแบบจำลองให้มีประสิทธิภาพสูงขึ้นจากปรับเปลี่ยนโครงสร้างสถาปัตยกรรมใหม่

2) พัฒนารูปแบบของงานให้เป็นแบบครบวงจร หรือ End–To–End โดยการสร้างแบบจำลองที่สามารถทำได้ตั้งแต่การเตรียมการข้อมูลตลอดจนการอ่านริมฝีปาก

3) พัฒนาให้มีแบบจำลองชุดข้อมูลอื่น เช่น คำพูด หรือ ประโยคสั้นๆ ที่ใช้บ่อย

รายการอ้างอิง

1. Nittaya, W., K. Wetchasit, and K. Silanon. *Thai Lip-Reading CAI for hearing impairment student*. in 2018 *Seventh ICT International Student Project Conference (ICT-ISPC)*. 2018. IEEE.
2. Morade, S.S. and S. Patnaik, *A novel lip reading algorithm by using localized ACM and HMM: Tested for digit recognition.* optik, 2014. 125(18): p. 5181-5186.
3. Shaikh, A.A., et al. *Lip reading using optical flow and support vector machines*. in 2010 3*Rd international congress on image and signal processing*. 2010. IEEE.
4. Potamianos, G., H.P. Graf, and E. Cosatto. *An image transform approach for HMM based automatic lipreading*. in *Proceedings* 1998 *International Conference on Image Processing. ICIP*98 (*Cat. No.* 98*CB*36269). 1998. IEEE.
5. Nefian, A.V., et al., *Dynamic Bayesian networks for audio-visual speech recognition.* EURASIP Journal on Advances in Signal Processing, 2002. 2002: p. 1-15.
6. Fu, Y., et al. *Lipreading by locality discriminant graph*. in 2007 *IEEE International Conference on Image Processing*. 2007. IEEE.
7. Rathee, N. *Investigating back propagation neural network for lip reading*. in 2016 *International Conference on Computing, Communication and Automation (ICCCA)*. 2016. IEEE.
8. NadeemHashmi, S., et al. *A lip reading model using CNN with batch normalization*. in 2018 *eleventh international conference on contemporary computing (IC*3). 2018. IEEE.
9. Ozcan, T. and A. Basturk, *Lip reading using convolutional neural networks with and without pre-trained models.* Balkan journal of electrical and computer engineering, 2019. 7(2): p. 195-201.
10. Akman, N.P., et al. *Lip reading multiclass classification by using dilated CNN with Turkish dataset*. in 2022 *International Conference on Electrical, Computer and Energy Technologies (ICECET)*. 2022. IEEE.
11. Petridis, S., Z. Li, and M. Pantic. *End-to-end visual speech recognition with LSTMs*. in 2017 *IEEE international conference on acoustics, speech and signal*

processing (ICASSP). 2017. IEEE.

12. Gutierrez, A. and Z. Robert, *Lip reading word classification.* Comput Vision-ACCV, 2017.
13. Fung, I. and B. Mak. *End-to-end low-resource lip-reading with maxout CNN and LSTM*. in 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018. IEEE.
14. Lu, Y. and J. Yan, *Automatic lip reading using convolution neural network and bidirectional long short-term memory.* International Journal of Pattern Recognition and Artificial Intelligence, 2020. 34(01): p. 2054003.
15. Cheng, S., et al. *Towards pose-invariant lip-reading*. in *ICASSP* 2020-2020 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020. IEEE.
16. Dweik, W., S. Altorman, and S. Ashour, *Read my lips: Artificial intelligence word-level arabic lipreading system.* Egyptian Informatics Journal, 2022. 23(4): p. 1-12.
17. Amit, A.G., J. Jnoyola, and S. Sameepb, *Lip reading using CNN and LSTM.* Technical report, 2016.
18. He, L., et al., *An optimal 3D convolutional neural network based lipreading method.* IET Image Processing, 2022. 16(1): p. 113-122.
19. Jang, D.-W., et al., *Lip reading using committee networks with two different types of concatenated frame images.* IEEE Access, 2019. 7: p. 90125-90131.
20. Poomhiran, L., P. Meesad, and S. Nuanmeesri, *Improving the recognition performance of lip reading using the concatenated three sequence keyframe image technique.* Engineering, Technology & Applied Science Research, 2021. 11(2): p. 6986-6992.
21. Petridis, S., et al. *Visual-only recognition of normal, whispered and silent speech*. in 2018 *ieee international conference on acoustics, speech and signal processing (icassp)*. 2018. IEEE.
22. Patterson, E.K., et al. *CUAVE: A new audio-visual database for multimodal human-computer interface research*. in 2002 *IEEE International conference on acoustics, speech, and signal processing*. 2002. IEEE.
23. Yang, S., et al. *LRW*-1000: *A naturally-distributed large-scale benchmark for lip*

reading in the wild. in 2019 14*th IEEE international conference on automatic face & gesture recognition (FG* 2019). 2019. IEEE.

24. Rekik, A., A. Ben-Hamadou, and W. Mahdi. *A new visual speech recognition approach for RGB-D cameras*. in *Image Analysis and Recognition:* 11*th International Conference, ICIAR* 2014*, Vilamoura, Portugal, October* 22-24*,* 2014*, Proceedings, Part II* 11. 2014. Springer.

25. Kartynnik, Y., et al., *Real-time facial surface geometry from monocular video on mobile GPUs.* arXiv preprint arXiv:1907.06724, 2019.

ภาคผนวก

ภาคผนวก ก

ผลงานที่ได้รับการตีพิมพ์


EEAAT
Electrical Engineering Academic Association (Thailand)
iEECON 2024
The 2024 International Electrical Engineering Congress
CERTIFICATE OF PRESENTATION
TO
Aekapob Jittakoti and Sopon Phumeechanya
Silpakorn University
for the paper entitled
Temporal Keyframe Technique Based on CNN and LSTM for Enhancing Lip Reading Performance
The 12th International Electrical Engineering Congress 2024 (12th iEECON 2024)
"Smart Factory and Intelligent Technology for Tomorrow"
March 6-8, 2024
The Zign Hotel & Villa, Pattaya, Thailand
(Assoc. Prof. Dr. Athikom Roeksabutr)
President , EEAAT
(Asst. Prof. Dr. Saichol Chudjuarjeen)
General Chair
(Assoc. Prof. Dr. Jessada Konpang)
Technical Conference Chair

Conference Topics

COMMUNICATIONS

Communication Theory, Antennas and Propagation Optical Communications, Microwaves, Wireless Communications, Signal Processing for Communication, Channel Coding, Multimedia Communication, Remote Sensing and Applications, Metamaterials, etc.

ELECTRONICS & CONTROL

Analog Circuits, Filters and Data Conversion, Analog and Mixed Signal Processing, Embedded Computer System, Robotics, VLSI Design, Biomedical Electronics, Industrial Electronics and automation, Adaptive Control, Electric Circuit Technology, Fault Tolerance and Detection, Semiconductor Materials, Magnetic Materials, Themoelectric materials and devices, Sensor, Organic Electronics and Printed Electronic etc.

DIGITAL SIGNAL PROCESSING

Image and Video Processing, Audio and Speech Processing, Pattern Recognition, Biomedical Signal Processing, Computer Vision and Pattern Recognition, Adaptive Signal Processing, Machine Learning for Signal Processing, etc.

POWER & ENERGY

Smart Grid Technology, Planning, Management Operation, and Control; Electric Power Systems : Generation Transmission and Distribution, Electrical Machines, Energy Conversions, Renewable Energy Sources, Power Electronics, Energy Systems, Power Quality, High Voltage Engineering, Insulation and Materials, Piezoelectric ceramic and thin films, Energy storage materials and technology, energy harvesting and energy storage designs, Advanced charecterization and properties of ferroelectric materials etc.

COMPUTER & IT

Computer Networks, Cloud Communication and Networking, Data Mining, Artificial Intelligence, Computational Theory, Information System, High Performance Computing, Computer Security, Software Engineering, Distributed and Parallel Computing, Web Services and Internet Computing, Multi-agent Systems, Human Computer Interaction, Internet of Thing (IoT) etc.

INDUSTRIAL TECHNOLOGY

Applied Sciences, Technology Management, Digital Businesses, Engineering, Innovation, Industrial Education, Asefa Technology etc.

Important Dates :

Paper submission deadline: October 12, 2023
Paper acceptance notification: December 22, 2023
Camera-ready submission deadline: January 12, 2024
Early-bird registration deadline: January 12, 2024
Conference dates: March 6-8, 2024

Contact information :

สถาบันวิจัยและพัฒนา
Research and Development Institute
Research and Development Institute
Rajamangala University of Technology Krungthep
Tel : +(66) 2 287 9600 EXT 3111 Fax : +(66) 2 287 9684
ieecon@mail.rmutk.ac.th

http://www.ieecon.org/ieecon2024/

2024 International Electrical Engineering Congress (iEECON 2024)
March 6-8, 2024, Pattaya Chonburi, THAILAND

Temporal Keyframe Technique Based on CNN and LSTM for Enhancing Lip Reading Performance

Aekapob Jittakoti
Department of Electrical Engineering
Faculty of Engineering and Industrical Techhnology
Silpakorn University
Nakhon Pathom, Thailand
jittakoti_a@su.ac.th

Sopon Phumeechanya
Department of Electrical Engineering
Faculty of Engineering and Industrical Techhnology
Silpakorn University
Nakhon Pathom, Thailand
phumeechanya_s@su.ac.th*
*Corresponding author

***Abstract*—The purpose of this research is to present a method for improving lip reading performance through the analysis of temporal keyframes using CNN and LSTM. The performance of the model is determined by employing the input data from the data preparation process. The lip image data is divided into three groups for development: 3 frames, 5 frames, and 10 frames, which include half-lip image data. Lip reading research has never been conducted before based on the assumption of a typical human body shape with left and right symmetry, as in this study. When the unseen test set was evaluated, 10 frames achieved the best recognition rate, with results of 0.879% accuracy for the full lip image dataset and 0.868% accuracy for the half lip image dataset, exhibiting comparable performance.**

Keywords—lip reading; convolutional neural network; long short-term memory; temporal keyframe analysis

I. Introduction

Reading lips in general is the capacity of humans to understand communication without sound, but it can be seen in the way the lips move. Later, humans developed mathematical equations for modeling the transformation of words spoken into text by lip detection using various methods. In image processing, lip reading is one of the applications used to build models that can learn and recognize words, phrases, and sentences by analyzing the visual information around the lips. Lip reading can be known as speech recognition in some research and may provide both images and audio. The process is referred to as Visual Speech Recognition if only images are used. Similar to this, if sound is added to the information, it is referred to as Audio-Visual Speech Recognition. Lip reading is a difficult task because many spoken words are accompanied by lip and facial movements. For instance, a lip movement may signify a different word depending on the sentence's context. There are many applications that take advantage of lip reading's capabilities, such as recovering lost audio information due to noise disturbances in video by transferring lips movements into words, communicating with the hearing impaired [1] and so on. It is obvious from the aforementioned advantages that lip reading research is constantly improving and advantageous to education and development.

In this paper, we design a novel architecture for lip reading using Convolutional Neural Network (CNN) to extract features from images and Long Short Term Memory (LSTM) to classify targets, which can produce better results. A larger and more recent audiovisual database [2], which contains normal, whispered, and silent speech but only normal and frontal visual information, was used for all these experiments. According to the large number of databases that are accessible, including those that contain words, phrases, and sentences, we would like to restrict the scope of our research by defining all the data utilized as pictures of speaking the digits 0 to 9 in English from the database stated.

II. Related Work

The model building of the lip reading process has three main parts: first, input selection; second, an extraction of features method from the lip image that mostly provides video data; and third, a classification method to classify targets. Due to technological constraints and the low efficiency of image feature extraction, the early research work was difficult and resulted in a low performance recognition rate. Duchnowski et al. [3], used a traditional mathematical linear transformation method named Linear Discriminant Analysis (LDA) and presented it. Shortly after, Principal Components Analysis (PCA) [4] was published with a similar method and improved slightly in efficacy. Thereupon, Discrete Cosine Transform (DCT) [5], Logicality Discriminant Graph (LDG) [6], and Action Contour Model (ACM) are all results presented that had the potential to produce higher recognition rate performance, respectively. Hidden Markov Model (HMM) [7] has been widely used for speech recognition and has been presented by these studies above as a classifier. Another is Support Vector Machine (SVM) [8] also received attention during that time. These are a few examples of how to build model lip reading using different methods early in the research development.

Presently, lip reading research is also becoming increasingly popular as a result of the effective development of deep learning models. Numerous studies have developed different deep learning techniques. At the beginning of the neural network process, N. Rathee [9] approved lip reading using the Back Propagation Neural Network as a classifier and the Local Binary Pattern (LBP) as a feature's extractor. CNN is one of the most popular and is used to perform feature extraction from images, which is a trademark of CNN's. S. NadeemHashmi et al. [10] designed twelve-layer CNN with Batch Normalization investigation. T. Ozcan et al. [11] demonstrate the design architecture of CNN compared with CNN pre-trained models such as AlexNet, GoogleNet and the number of minibatch sizes used. N. P. Akman et al. [12]

2024 12th International Electrical Engineering Congress (iEECON) | 979-8-3503-8359-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/iEECON60677.2024.10537881

2024 International Electrical Engineering Congress (iEECON 2024)
March 6-8, 2024, Pattaya Chonburi, THAILAND

proposed the architecture of a dilated CNN. The underlying concept behind Concatenated Three Sequence Keyframe Image, or C3-SKI [13], which we are going to discuss in the next part, is the technique to select three frames that are crucial for learning the model. Another is LSTM, which has drawn a lot of interest for its capacity to classify time-related data. There has been a combination of CNN and LSTM architecture [14], [15], [16], and [17], which use CNN as a feature extractor and LSTM as a Classifier.

Choosing input data is a crucial component of the system; the model's effectiveness may simultaneously increase or decrease as a result. The data must be accurate for the system's performance to be at its best. The model learns slowly because there is too much redundancy in the data, which doesn't give it any learning value. The majority of that information comes from sizable, unnormalized databases (some databases are in-house) and is stored as video clips on platforms of varying lengths, depending on the length of the words or sentences spoken, silence moments, and mouth movements that influence the next speech. To get the best results, more diverse databases need to be organized more precisely. Many researchers have offered their own data sorting to compare outcomes. The usage of these approaches has been covered in the research that has previously been mentioned, whether it be the conventional CNN method or the combination of CNN and LSTM. P. Sindhura et al. [18] crated a 4x4 matrix of concatenated images, including 16 frames, that can be used as a demonstration of how they fed this type of image information to the system that is most commonly used in the CNN model. There might be more or less than that in some literature.

The database serves as a platform for comparing algorithms that have been published in various research articles. The database contains video clips, picture, or other forms of data for each database. Grid Corpus [19] is a database that contains audio-visual information about sentences spoken by 34 people (18 males and 16 females) and totals more than 34,000 sentences in 2006. AVletters [20], the Avletters database has collected 10 speakers (5 males and 5 females) who pronounced each letter from A to Z three times in 2002. MIRACL-VC1 [21] was published as a database, including both depth and color images. There are ten words and ten phrases both uttered by 15 speakers (5 males and 10 females) 10 times in this database in 2014. In this study, we use a database called AV Digits [3] which has been available since 2018 and has been cited by a few researchers. This database contains normal, whispered, and silent speech. There were 53 participants from three different perspectives (frontal, 45, profile) and uttered five times digits and phrases in three modes. The number of digits is utilized from 0 to 9 which is essentially used in the accustomed dataset. The AV Digits database is arduous and sophisticated enough to investigate and develop lip-reading system algorithms.

III. THE PROPOSED METHOD

This section describes the content of a proposed framework for lip reading that was used to learn the English digits dataset 0 - 9 from the database, which includes three key sections: Datasets, Data Preparation, Model Architecture.

A. Datasets

First, it is critical to show the database's details. This database, known as AV Digits, was released in 2018. Both digits and short phrases are present in the database. There are three independent speech modes—normal, whispered, and silent—as well as three different viewing angles: straight face, 45 degrees, and 180 degrees away from the three camera recorders. We primarily offer the straight face and normal condition form digit datasets that were utilized in this research. All participants were asked to utter the English numbers 0 to 9 five times in the digit section. This indicates that among five videos of one individual speaking in each single mode, different numbers will be mentioned. The spoken numerals are uttered in a random sequence. Participants who were not native speakers were asked to complete the identical objectives as those who were. The average age and standard deviation of the 53 people participating (41 males and 12 females) from 16 different nationalities were 26.7 and 4.3 years, respectively. The recordings were made in a laboratory environment utilizing three cameras with a resolution of 1280×780 at 30 frames per second at three different angles. Audio from the camera's connected microphone was also included in the video clip at a sampling rate of 44.1 kHz. The speaker records each digit that is responsible, and the space bar follows by being used to separate each digit in the video to provide a timestamp for each utterance.

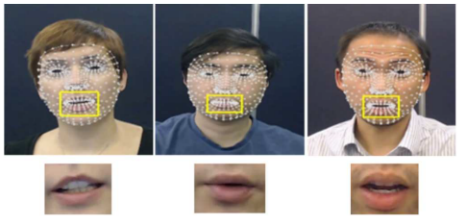


Fig. 1. Example of face landmark 468 for lip extraction on the dataset.

B. Data Preparation

Each time a number is spoken, the length of the duration varies based on the uttered digit or the person speaking it; for instance, number 0 has two syllables and number 1 has one syllable. The number of frames utilized in capturing speech varies depending on the pace of the speaker. Speech recognition is challenging since the same lip motions might result in distinct words, as was mentioned in the preceding section. Therefore, to identify the most useful frames for learning, the visual information allowed by the model lip reading system for analysis must be normalized. Choosing the most appropriate number of frames for speech is the one that fluctuates, identifies lip movements, and enables the model to learn as well as recognize uttered digits more effectively. The steps for data preparation are as follows:

- Face landmark 468 point proposed in [22] was used in this section via Mediapipe [23] on the library Python language program and can set the spot around the lip to be more precise. The following points were taken: 57, 164, 200, and 287. As shown in Fig. 1.

2024 International Electrical Engineering Congress (iEECON 2024)
March 6-8, 2024, Pattaya Chonburi, THAILAND

- A timestamp is an indication of the time that digits are uttered in a video. All speaking numbers of digits have between tens and hundreds of frames. The speech's starting and ending frames are moved slightly toward the center in the next two frames, and the middle frames are preserved as variables by skipping two frames at a time. The sections of the frames that indicated lip movement were then manually selected from the collected frames. In the end, 10 frames are selected randomly to represent a spoken number with complete lip movement.
- Because the data is compiled randomly from the beginning, there are many factors that dictate that the same sequence of frame in each utterance is not the same frames in every utterance. As a result, we configured the frames to be utilized in this experiment as frames 2, 5, and 8 for three frames and 2, 4, 5, 7, and 8 for five frames.

C. Model Architecture

The proposed model, which consists of 3 CNN blocks with a total of 6 layers and utilizes "same" padding, 2 Dense layers, 1 Flatten layer, 2 Dropout layers, Timedistributed 1 layer and LSTM 1 layer, combines both types of CNN and LSTM for development. All layers use the activation function as RELU, except for the classification layer applied by SoftMax. The shape of the input and model architecture are shown in Fig. 2.

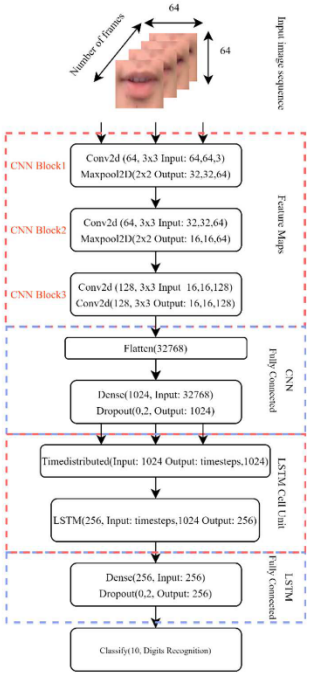
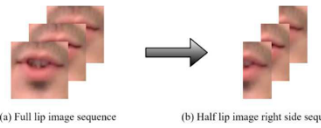


Fig. 2. Input image sequence and Purposed model.



(a) Full lip image sequence (b) Half lip image right side sequence

Fig. 3. Example of input with full lip image sequence and half lip image right side sequence.

IV. EXPERIMENTS

In order to determine the best performance, we separated the experimental into three groups in this section: 3 frames, 5 frames, and 10 frames, with half-lip image datasets included. This dataset is offered to examine the experimental performance under the presumption of a typical human body shape with left and right symmetry, which has never been done previously in lip reading research. Fig. 3 depicts the shape of the input sequence, which includes full and half-lip images.

A. Trainning Setting Information

Because the input labels contain a one-hot encoder, we use an Adagrad optimizer with a learning rate of 0.001 and a loss function of binary cross-entropy to train the model with one batch size. In order to save the training's best value, the function ModelCheckpoint is called for an epoch of 300. The total speech data is 2650, and 2120 were utilized for the training process, with 80% being used for the training set, 20% for the validation set, and 530 for testing. The dimensions of 64x64 for the full lip image sequence and 64x32 for the half lip image right side sequence are presented.

B. Results of Each Frame Sequences on Our Famework

The outcomes from the model we proposed across different frame sequences will be investigated using full frame and half frame lip data in this experiment. Frame sequences or full/half frames are trained and tested separately in all cases to observe the performance obtained by 1) halving the size of the input and 2) the number of frames used. As a consequence, the greatest number of frames used, which is 10 frames, is the most accomplished. In the last checkpoint of the full lip image datasets, the model achieved the best score for recognition rate from the unseen data, with 0.879% accuracy at 136 epochs from the full lip image dataset. The model was able to adjust to 0.868% accuracy at 268 epochs from half-lip image dataset in the meantime. When compared, it was discovered that 10 lip image frames may deliver a little superior performance. Other metrics are shown in Table I for the full lip image and Table II for the half lip image.

TABLE I. RESULTS OF EACH FRAME SEQUENCES WITH FULL LIP IMAGE DATASET ON PROPOSED MODEL

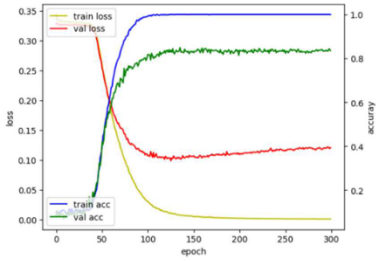
Metrics	Number of frame sequences		
	3 frames	*5 frames*	*10 frames*
Training Accuracy/Loss	1.000/0.010	1.000/0.003	1.000/0.007
Training Val Accuracy/ Val Loss	0.715/0.170	0.788/0.149	0.849/0.102
Model Last Checkpoint	140	204	136
Testing score (unseen data)	0.775	0.832	0.879

2024 International Electrical Engineering Congress (iEECON 2024)
March 6-8, 2024, Pattaya Chonburi, THAILAND

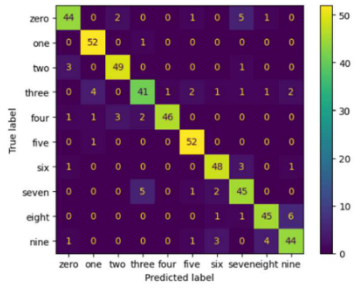
TABLE II. RESULTS OF EACH FRAME SEQUENCES WITH HALF LIP IMAGE DATASET ON PROPOSED MODEL

Metrics	Number of frame sequences		
	3 frames	*5 frames*	*10 frames*
Training Accuracy/Loss	0.998/0.010	1.000/0.004	1.000/0.002
Training Val Accuracy/ Val Loss	0.698/0.182	0.764/0.159	0.823/0.126
Model Last Checkpoint	244	241	268
Testing score (unseen data)	0.757	0.817	0.868

According to the confusion matrix in Figs. 4 and 5 for 10 full and half lip image frames, the information demonstrates that the true label and prediction label from the evaluation data contained a total of 530 utterances due to the 53 utterances in each class. Each square illustrates the number of responses. Numbers colored in purple indicate a small percentage of accurate answers, while numbers colored in yellow indicate a substantial number of correct predictions. The highest recognition accuracy was reported from 1 and 5 in 10 full lip image frames, with 52 times, whereas 51 times were reported from number 1 in 10 half lip image frames. The recognition accuracy for number 3 in 10 full lip image frames and number 0 in 10 half-lip image frames was the lowest, with 41 and 42 correct answers, respectively. The history graph clearly indicates that both graphs exhibit highly similar and closely aligned learning curves, suggesting a strong degree of comparability between them. This illustrates that a lip reading model can be effectively constructed using only half of the lip image data, as evidenced by the observed performance outcomes.



(a) The history graph obtained from model training of 10 full lip image frames



(b) The confusion matrix of 10 full lip image frames

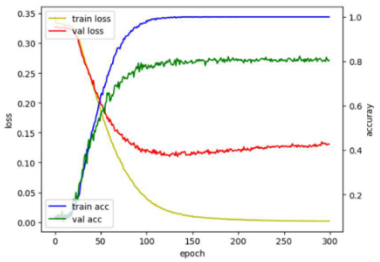
Fig. 4. The history graph obtained from model training and the confusion matrix of 10 full lip image frames

TABLE III. EVALUATING THE PERFORMANCE OF METHODS BASE ON THE AV DIGITS DATABASE WITH THE UNSEEN DATA

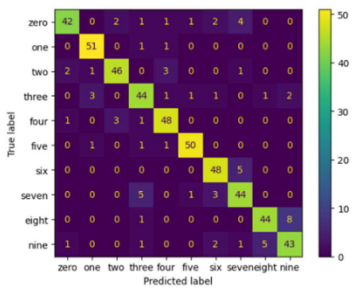
Methods	Recognition rate on number of frame sequences (full lip images)		
	3 frames	*5 frames*	*10 frames*
C3-SKI [13]	0.645	0.717	0.734
Proposed Method	**0.775**	**0.832**	**0.879**

C. Comparison

When compared to other models, the quantity of data used to train or test is the same as in the preceding section but different based on the proposed method. L. Poomhiran et al. [13] presented a concatenated image approach that introduces three keyframes by choosing the beginning, middle, and end frames of the speech. The same as in this experiment, except with the addition of 5 and 10 frames for evaluating performance. As a result, increasing the number of frames can significantly enhance the model's performance. Our proposed model approach has achieved a remarkable peak accuracy of 0.879, surpassing the performance of [13], which achieved an accuracy of 0.734 with the same number of frames. When compared to the performance of our proposed model using only half-lip image frames, as presented in Table II, the utilization of three half-lip image frames consistently proves to be the superior choice across all scenarios, as indicated in Table III. This demonstrates the effectiveness of our model in capturing critical visual information for accurate lip reading, outperforming the previous approach [13] by a significant margin.



(a) The history graph obtained from model training of 10 half lip image frames



(b) The confusion matrix of 10 half lip image frames

Fig. 5. The history graph obtained from model training and the confusion matrix of 10 half-lip image frames

V. CONCLUSION

The approach for lip reading described in this research utilizes a detailed analysis involving the examination of a specific number of frames, specifically focusing on half-lip image frames from the right side. This analysis is conducted using CNN and LSTM networks, which are well-suited for processing sequential data like lip images.

The database employed is AV Digits; although it offers multiple options, the research utilized only the straight-face, normal option, with 53 participants individually uttering the English numerals 0 to 9. This dataset contains a total of 2,650 speech samples, making it a substantial resource for the study. One of the critical aspects of achieving good performance in lip reading is the selection of an appropriate frame for analysis. This is particularly important due to the wide range of information contained in lip movement, which can vary depending on the individual's personality characteristics. Therefore, the method for frame selection used in this research is presented accordingly.

In this experiment section, we independently train and evaluate different numbers of frames, including full lip image and half lip image, based on the hypothesis that human bilateral symmetry may be relevant, to determine the optimal configuration for achieving the best performance. Moreover, we closely monitored the model's progress by examining the training graphs. Remarkably, our model exhibited highly similar learning curves for both 10 full lip image frames and 10 half lip image frames. This similarity in training progress underscores the versatility of our approach, as it demonstrates consistent performance regardless of the lip image type used. The experimental results reveal that our proposed model achieved the highest recognition rates, with an accuracy of 0.879 for 10 full lip image frames and 0.868 for 10 half lip image frames when tested with unseen datasets.

These results clearly indicate the superiority of our model in improving lip-reading performance compared to other approaches. Furthermore, the findings suggest that the performance of lip reading using only half-lip images has the potential to be comparable to traditional datasets. This highlights the significance of our approach in potentially reducing the data requirements for effective lip-reading systems while maintaining high accuracy.

ACKNOWLEDGMENT

The authors would like to acknowledge the contribution of Department of Electrical Engineering, Faculty of Engineering and Industrial Technology, Silpakorn University Scholarship for funding this research.

References

[1] W. Nittaya, K. Wetchasit and K. Silanon, "Thai Lip-Reading CAI for Hearing Impairment Student," 2018 Seventh ICT International Student Project Conference (ICT-ISPC), Nakhonpathom, Thailand, 2018, pp. 1-4

[2] S. Petridis, J. Shen, D. Cetin and M. Pantic, "Visual-Only Recognition of Normal, Whispered and Silent Speech," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 6219-6223.

[3] P. Duchnowski, U. Meier, and A. Waibel, "See Me, Hear Me: Integrating Automatic Speech Recognition and Lip-Reading," ICSLP, 1994, pp. 547-550.

[4] G. Potamianos, H. P. Graf and E. Cosatto, "An image transform approach for HMM based automatic lipreading," Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269), Chicago, IL, USA, 1998, pp. 173-177.

[5] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," EURASIP J. Appl. Signal Process, no.11, 2002, pp.1274-1288.

[6] Y. Fu, X. Zhou, M. Liu, M. Hasegawa-Johnson and T. S. Huang, "Lipreading by Locality Discriminant Graph," 2007 IEEE International Conference on Image Processing, San Antonio, TX, USA, 2007, pp. III - 325-III-328.

[7] S.S. Morade, and S. Patnaik, "A novel lip reading algorithm by using localized ACM and HMM: Tested for digit recognition," optik 125.18, 2014, vol. 125, pp. 5181-5186.

[8] A. A. Shaikh, D. K. Kumar, W. C. Yau, M. Z. C. Azemin and J. Gubbi, "Lip reading using optical flow and support vector machines," 2010 3rd International Congress on Image and Signal Processing, Yantai, China, 2010, pp. 327-330.

[9] N. Rathee, "Investigating back propagation neural network for lip reading," 2016 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2016, pp. 373-376.

[10] S. NadeemHashmi, H. Gupta, D. Mittal, K. Kumar, A. Nanda and S. Gupta, "A Lip Reading Model Using CNN with Batch Normalization," 2018 Eleventh International Conference on Contemporary Computing (IC3), Noida, India, 2018, pp. 1-6.

[11] T. Ozcan, and A. Basturk, "Lip reading using convolutional neural networks with and without pre-trained models." Balkan journal of electrical and computer engineering 7.2 , 2019, vol. 7, pp. 195-201.

[12] N. P. Akman, T. T. Sivri, A. Berkol and H. Erdem, "Lip Reading Multiclass Classification by Using Dilated CNN with Turkish Dataset." 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), Prague, Czech Republic, 2022, pp. 1-6

[13] L. Poomhiran, P. Meesad, and S. Nuanmeesri, "Improving the Recognition Performance of Lip Reading Using the Concatenated Three Sequence Keyframe Image Technique", Eng. Technol. Appl. Sci. Res., vol. 11, no. 2, Apr. 2021, pp. 6986–6992.

[14] W. Dweik, S. Altorman, S. Ashour, "Read my lips: Artificial intelligence word-level arabic lipreading system." Egyptian Informatics Journal, vol. 23, no. 4, 2022, Pages 1-12.

[15] A. Garg, J. Noyola, and S. Bagadia, "Lip reading using CNN and LSTM." Technical report, Stanford University, CS231 n project report 2016.

[16] L. He, B. Ding, H. Wang, and T. Zhang, "An optimal 3D convolutional neural network based lipreading method." IET Image Processing, vol. 16, no. 1, 2022, pp. 113-122.

[17] I. Fung, and B. Mak, "End-To-End Low-Resource Lip-Reading with Maxout Cnn and Lstm," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 2511-2515.

[18] P. Sindhura, S. J. Preethi, and K. B. Niranjana, "Convolutional Neural Networks for Predicting Words: A Lip-Reading System," 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Msyuru, India, 2018, pp. 929-933.

[19] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition." The Journal of the Acoustical Society of America, vol. 120, no. 5, 2006, pp. 2421-2424.

[20] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox and R. Harvey, "Extraction of visual features for lipreading." in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, 2002, pp. 198-213.

[21] A. Rekik, A. Ben-Hamadou, W. Mahdi "A New Visual Speech Recognition Approach for RGB-D Cameras." ICIAR, 2014, pp. 21-28

[22] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile GPUs." arXiv preprint arXiv:1907.06724, 2019.

[23] MediaPipe, "MediaPipe: Cross-Platform Framework for ML-Based Multi-modal (Vision, Audio, ...) Applied Research," Google, 2021 https://mediapipe.dev/. (accessed Sep. 12, 2023)

ประวัติผู้เขียน

ชื่อ-สกุล	เอกภพ จิตตโคติ
วุฒิการศึกษา	วศ.บ.วิศวกรรมอิเล็กทรอนิกส์และระบบคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์และเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยศิลปากร (2562)
ผลงานตีพิมพ์	Aekapob Jittakoti and Sopon Phumeechanya, "Temporal Keyframe Technique based on CNN and LSTM for Enhancing Lip Reading Performance," 2024 12th International Electrical Engineering Congress (iEECON), Pattaya, Thailand, 2024, pp.1-5.