โครงสร้างการเรียนรู้เชิงลึกสำหรับการรู้จำการออกเสียงสระภาษาไทย

โดย
นางสาวนิยดา รักวงษ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ แบบ 1.1 ระดับปริญญาดุษฎีบัณฑิต
ภาควิชาคอมพิวเตอร์
มหาวิทยาลัยศิลปากร
ปีการศึกษา 2565

โครงสร้างการเรียนรู้เชิงลึกสำหรับการรู้จำการออกเสียงสระภาษาไทย

โดย
นางสาวนิยดา รักวงษ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ แบบ 1.1 ระดับปริญญาดุษฎีบัณฑิต
ภาควิชาคอมพิวเตอร์
มหาวิทยาลัยศิลปากร
ปีการศึกษา 2565

DEEP LEARNING STRUCTURE FOR THAI VOWEL PRONUNCIATION RECOGNITION

By

MISS Niyada RUKWONG

A Thesis Submitted in Partial Fulfillment of the Requirements

for Doctor of Philosophy INFORMATION TECHNOLOGY

Department of COMPUTER SCIENCE

Silpakorn University

Academic Year 2022

หัวข้อ	โครงสร้างการเรียนรู้เชิงลึกสำหรับการรู้จำการออกเสียงสระภาษาไทย
โดย	นางสาวนิยดา รักวงษ์
สาขาวิชา	เทคโนโลยีสารสนเทศ แบบ 1.1 ระดับปริญญาดุษฎีบัณฑิต
อาจารย์ที่ปรึกษาหลัก	ผู้ช่วยศาสตราจารย์ ดร. สุนีย์ พงษ์พินิจภิญโญ

คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร ได้รับพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต

.....คณบดีคณะวิทยาศาสตร์
(ผู้ช่วยศาสตราจารย์ ดร. นรงค์ ฉิมพาลี)

พิจารณาเห็นชอบโดย

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. รัชฎา คงคะจันทร์)

.....อาจารย์ที่ปรึกษาหลัก
(ผู้ช่วยศาสตราจารย์ ดร. สุนีย์ พงษ์พินิจภิญโญ)

.....ผู้ทรงคุณวุฒิภายใน
(ผู้ช่วยศาสตราจารย์ ดร. ทัศนวรรณ ศูนย์กลาง)

.....ผู้ทรงคุณวุฒิภายนอก
(ผู้ช่วยศาสตราจารย์ ดร. ภิญโญ แท้ประสาทสิทธิ์)

.....ผู้ทรงคุณวุฒิภายนอก
(ผู้ช่วยศาสตราจารย์ ดร. ชมนาด อินทจามรรักษ์)

60309801 : เทคโนโลยีสารสนเทศ แบบ 1.1 ระดับปริญญาดุษฎีบัณฑิต
คำสำคัญ : การเรียนรู้เชิงลึก, การรู้จำเสียง, สระภาษาไทย, โมเดล Convolutional Neural Networks

นางสาว นิยดา รักวงษ์: โครงสร้างการเรียนรู้เชิงลึกสำหรับการรู้จำการออกเสียงสระภาษาไทย อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก : ผู้ช่วยศาสตราจารย์ ดร. สุนีย์ พงษ์พินิจภิญโญ

การฝึกออกเสียงที่มีประสิทธิภาพและได้มาตรฐานเป็นสิ่งสำคัญของการออกเสียงอย่างถูกต้อง การออกเสียงสระผิดทำให้ความหมายของคำเปลี่ยนไป การฝึกออกเสียงสระสามารถก่อให้เกิดปัญหาสำหรับผู้เรียนที่ไม่ได้เป็นเจ้าของภาษาได้จึงต้องมีผู้เชี่ยวชาญให้คำแนะนำ ปัจจุบันการเรียนรู้ออนไลน์ได้รับความนิยม การนำเทคโนโลยีสำหรับการฝึกออกเสียงมาใช้เป็นเครื่องมือสามารถช่วยพัฒนาในด้านการเรียนการสอนสำหรับการเรียนรู้ภาษาเพื่อแก้ปัญหาการฝึกออกเสียงสระภาษาไทยสำหรับผู้เรียนที่ไม่ใช่เจ้าของภาษา ผู้ที่พูดภาษาไทยไม่ได้มาตรฐาน หรือผู้พิการทางการออกเสียง โดยสามารถแก้ปัญหาความไม่เพียงพอของผู้เชี่ยวชาญด้านการสอน และความซับซ้อนของกระบวนการสอนการออกเสียงสระ งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาโครงสร้างการเรียนรู้เชิงลึกสำหรับการรู้จำการออกเสียงสระภาษาไทย 18 เสียง ซึ่งเป็นเสียงสระเดี่ยวมาตรฐานของภาษาไทย โดยนำเสนอโมเดลการเรียนรู้เชิงลึกที่เป็นส่วนสำคัญในการรู้จำเสียงสระภาษาไทยสำหรับระบบการฝึกการออกเสียงโดยใช้คอมพิวเตอร์ช่วย (Computer-Assisted Pronunciation Training : CAPT) การระบุเสียงสระที่ถูกต้องเมื่อพูดในสถานการณ์จริงถือเป็นความท้าทายหลักในการรู้จำเสียงสระภาษาไทย งานวิจัยนี้มีการเปรียบเทียบประสิทธิภาพของโมเดล Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) และการรวมกันของ Convolutional Neural Network และ Long Short-Term Memory (CNN_LSTM) กับคุณสมบัติด้านเสียง Mel spectrogram (MS) และ Mel Frequency Cepstrum Coefficient (MFCC) ในการรู้จำเสียงสระภาษาไทย ชุดข้อมูลเสียงสระภาษาไทยใหม่ถูกออกแบบ รวบรวม และตรวจสอบโดยนักภาษาศาสตร์ทำให้ได้ชุดข้อมูลเสียงสระภาษาไทยที่มีความแตกต่างหลายมิติ เช่น เพศ อายุ สภาวะแวดล้อมที่ใช้ในการพูด เป็นต้น โดยมีระดับความดังของเสียงรบกวนในสภาวะแวดล้อมประมาณ 30 - 50 dB ผลลัพธ์พบว่า โมเดลที่เหมาะสมในการรู้จำเสียงสระภาษาไทยในงานวิจัยนี้คือ โมเดล CNN รวมกับคุณสมบัติด้านเสียง MS มีค่าความถูกต้อง 98.61% มีการนำเสนอวิธีการ Gradient-weighted class activation mapping (Grad-CAM) กับโมเดลการเรียนรู้เชิงลึก CNN สำหรับการรู้จำเพื่ออธิบายบริเวณที่สำคัญเมื่อโมเดลทำนายเสียงสระภาษาไทยทั้ง 18 เสียง ผลพบว่าการรู้จำเสียงสระในแต่ละสระ Grad-CAM จะพิจารณาทั้งความถี่สูงและความถี่ต่ำ งานนี้สามารถยืนยันว่าความชัดเจนและความโปร่งใสในการทำนายผลของโมเดล CNN สามารถช่วยให้ระบบการฝึกการออกเสียงโดยใช้

คอมพิวเตอร์ช่วย (CAPT) สำหรับการรู้จำเสียงสระภาษาไทยมีความถูกต้องและมีประสิทธิภาพมากยิ่งขึ้น ระบบนี้เป็นระบบที่พัฒนาเทคนิคคอมพิวเตอร์ผสมผสานกับภาษาศาสตร์ สามารถช่วยให้ผู้เรียนได้ฝึกการออกเสียงสระแบบเรียลไทม์ เสมือนมีผู้เชี่ยวชาญ คอยให้คำแนะนำเกี่ยวกับการออกเสียงสระที่ถูกต้องอย่างต่อเนื่อง เหมาะกับสถานการณ์โลกในปัจจุบันที่ต้องมีการเรียนในรูปแบบออนไลน์

60309801 : Major INFORMATION TECHNOLOGY
Keyword : Thai vowels, speech recognition, Deep Learning, Convolutional Neural Networks

MISS Niyada RUKWONG : Deep Learning Structure for Thai Vowel Pronunciation Recognition Thesis advisor : Assistant Professor sunee pongpinigpinyo, Ph.D.

Effective and proper pronunciation is essential to pronounce words correctly. Practicing Thai vowel pronunciation is difficult for non-native speakers to understand on their own. Experts are required to provide guidance. Nowadays, online learning is popular, and pronunciation training technology can help improve language teaching and learning. This technology can solve the problem of practicing Thai vowel pronunciation for non-native learners, non-standard Thai speakers, and persons with disabilities. It provides a solution to the inadequacy of instructional specialists and the complexity of teaching vowel pronunciation. The purpose of this research is to study deep learning structures for the recognition of 18 standard Thai vowel sounds. This research presents a deep learning model that plays a crucial part in recognizing Thai vowel sounds for Computer-Assisted Pronunciation Training (CAPT). Identifying the correct vowels when pronouncing them in real situations is a significant challenge in Thai vowel recognition. This present study applies deep learning models, including Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and a combination of CNN and LSTM with Mel spectrogram (MS) and Mel Frequency Cepstrum Coefficient (MFCC), to recognize Thai vowels. In the automatic recognition of Thai vowels, a new dataset for Thai vowels was designed, collected, and verified by linguists. The noise level in the environment of the sound files is between 30 - 50 dB. The results showed that the CNN model combined with the MS acoustic feature is the most suitable model for Thai vowel recognition in this research, with an accuracy of 98.61%. This work presents Gradient-weighted class activation mapping (Grad-CAM) with a CNN model for recognition to explain the importance of significant areas when the model predicted all 18 Thai vowel sounds. The results showed that Grad-CAM considers both high and low frequencies for each vowel recognition. This work confirms that the CNN model's clarity of predictions

could help CAPT systems be more accurate and efficient. This system is developed by combining computer techniques with linguistics, allowing learners to practice vowel pronunciation in real-time. It is like having experts, Thai language teachers, and linguists continuously advise learners on the correct pronunciation of vowels, making it suitable for today's world that requires online learning.

กิตติกรรมประกาศ

งานวิทยานิพนธ์นี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์จาก ผู้ช่วยศาสตราจารย์ ดร. สุนีย์ พงษ์พินิจภิญโญ อาจารย์ที่ปรึกษางานวิทยานิพนธ์ ท่านได้ให้ความรู้ คำปรึกษา ชี้แนะแนวทาง แก้ไขส่วนที่บกพร่อง และให้กำลังใจตลอดมา ขอกราบขอบพระคุณอย่างสูง

ผู้วิจัยขอกราบขอบพระคุณประธาน คณะกรรมการการสอบ คณะวิทยาศาสตร์ ภาควิชาคอมพิวเตอร์ มหาวิทยาลัยศิลปากร คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏอุตรดิตถ์ บุพการี ครอบครัว อาจารย์ เพื่อนๆ พี่ๆ น้องๆ ที่มอบการสนับสนุน การช่วยเหลือ ให้โอกาส และกำลังใจในการศึกษาจนประสบความสำเร็จ รวมถึงผู้พูด ผู้รวบรวม และผู้ตรวจสอบข้อมูลวิจัย ที่ร่วมกันสร้างชุดข้อมูลสำหรับการวิจัยครั้งนี้ หากไม่มีข้อมูลเหล่านี้เป็นไปไม่ได้เลยที่จะเกิดการเรียนรู้ที่เหมาะสมได้ และบัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร ที่สนับสนุนมอบทุนอุดหนุนการทำวิทยานิพนธ์ประจำปีงบประมาณ 2565

ท้ายที่สุดผู้วิจัยหวังเป็นอย่างยิ่งว่า งานวิทยานิพนธ์นี้จะเป็นประโยชน์ต่อผู้ที่สนใจ และสามารถนำไปใช้ต่อยอดการเรียน การสอน และการวิจัย ได้อย่างมีประสิทธิภาพและเกิดประสิทธิผล หากมีข้อผิดพลาดประการใด ผู้วิจัยขออภัยมา ณ. ที่นี้ด้วย

นางสาว นิยดา รักวงษ์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูป

หน้า

บทที่ 1

บทนำ

1.1. ความเป็นมาและความสำคัญของปัญหา

มนุษย์มีการสื่อสารด้วยภาษาหลายรูปแบบ เช่น การเขียน การพูด การใช้มือทำสัญลักษณ์ ท่าทาง หรือการใช้รูปภาพ การพูดเป็นรูปแบบพื้นฐานที่ใช้ในการสื่อสารที่ง่ายและสะดวกที่สุด ปัจจุบันการเรียนการสอนภาษาถูกจัดสรรให้กับคนหลายกลุ่ม เพื่อตอบสนองความต้องการของจุดประสงค์ที่แตกต่างกัน โดยทั่วไปการเรียนภาษาสามารถแบ่งออกเป็น การเรียนภาษาที่ 1 การเรียนภาษาที่ 2 หรือการเรียนภาษาที่ 3 นอกจากนี้ยังมีการเรียนภาษาสำหรับผู้ที่มีความบกพร่อง (disorder) ซึ่งจัดเป็นการบำบัดภาษา (language therapy) ดังนั้นกลุ่มผู้เรียนภาษาในประเทศไทยจึงมีหลายกลุ่มตามวัตถุประสงค์ของการเรียน ตัวอย่างเช่น นักเรียนในโรงเรียนเรียนภาษาไทยเป็นภาษาที่ 1 (ภาษาแม่) และเรียนภาษาอังกฤษเป็นภาษาที่ 2 กลุ่มคนวัยทำงานหรือนักศึกษาในกำลังศึกษาระดับมหาวิทยาลัยเรียนภาษาจีน ญี่ปุ่น ฝรั่งเศส เกาหลี พม่า หรือเวียดนามเป็นภาษาที่ 3 เพื่อเพิ่มความก้าวหน้าและโอกาสในสายอาชีพ กลุ่มผู้ป่วยฝึกพูดหลังจากได้รับการรักษา เช่น ผู้ป่วยที่ได้รับการผ่าตัดกล่องเสียง หรือการฝึกภาษาของผู้พิการทางสมอง เป็นต้น ดังนั้นผู้ที่ปฏิบัติการสอนภาษาอาจพบได้เป็นกลุ่มใหญ่ๆ 3 ประเภท ได้แก่ ครูสอนภาษา นักภาษาศาสตร์ และนักบำบัดภาษา

การเรียนภาษามีหลายระดับ เริ่มตั้งแต่การฝึกออกเสียง คำ และการพูดเป็นประโยค รวมไปถึงการเล่าเรื่อง หน่วยเสียงเป็นหน่วยที่เล็กที่สุดของภาษาและเป็นพื้นฐานที่สำคัญต่อการเรียนภาษาในระดับที่สูงขึ้น หน่วยเสียงในภาษามีหลายประเภทแตกต่างไปตามแต่ละภาษา เช่น เสียงสระ เสียงพยัญชนะ เสียงวรรณยุกต์ ในงานวิจัยนี้เป็นการศึกษาเสียงสระ เนื่องจากสระเป็นเสียงที่เป็นแกนกลางของพยางค์ (nucleus) และเป็นส่วนที่สำคัญในการพูด [1, 2] สระมีลักษณะเด่นที่ประกอบด้วยอย่างน้อย 3 ลักษณะคือ ลักษณะและตำแหน่งของอวัยวะที่ใช้ออกเสียง ระยะเวลา และจังหวะการพูดหรือท่วงทำนอง สระเป็นหน่วยเสียงสำคัญที่ส่งผลต่อความหมายและความชัดเจนในการพูด พฤติกรรมการพูด ภาษา และความสามารถในการได้ยิน รวมไปถึงช่วงวัยส่งผลกระทบต่อการออกเสียงสระ สระสามารถฝึกฝนได้ทั้งในด้านการผลิต (การพูด) และด้านการรับรู้ (การฟัง) ด้วยวิธีการฝึกออกเสียง อิทธิพลจากภาษาแม่และภาษาถิ่นส่งผลต่อเสียงสระได้แม้จะอยู่ในวัยเด็ก การทดสอบการออกเสียงสระไม่สามารถใช้การทดสอบ (articulation tests) เหมือนกับพยัญชนะ [3] การออกเสียงสระเป็นปัญหามากกว่าการออกเสียงพยัญชนะ [4] คุณสมบัติของสระเป็นสิ่งที่สำคัญและการฝึกออกเสียงสระสั้น-ยาวเป็นปัญหาสำหรับผู้เรียนที่ระยะเวลาของสระในภาษาแม่ไม่มีผลต่อความหมายของคำ [5]

สำหรับการฝึกออกเสียงสระบางเสียงหรือสระที่ไม่มีในภาษาแม่เป็นสิ่งที่ฝึกได้ยากกว่าการฝึกออกเสียงพยัญชนะ เนื่องจากการออกเสียงสระเป็นการเคลื่อนตำแหน่งของลิ้นเพื่อใช้ในการออกเสียง ผู้เรียนมักจะไม่สามารถทราบได้ด้วยตนเองอย่างแน่ชัดว่าตำแหน่งของลิ้นอยู่ที่ใด ดังนั้นผู้ที่ฝึกออกเสียงจึงควรมีผู้เชี่ยวชาญคอยแนะนำว่าออกเสียงถูกหรือผิดและการออกเสียงที่ถูกต้องปฏิบัติอย่างไร การอธิบายเกี่ยวกับเสียงสระที่นิยมที่สุดคือ วิธีการกำหนดตำแหน่งจุด สูง-ต่ำ และหน้า-หลังของลิ้น และการบอกพื้นที่ของลิ้นในแต่ละส่วน รวมไปถึงลักษณะของริมฝีปาก [2] แต่อย่างไรก็ตามการอธิบายลักษณะนี้ก็ยังเป็นเรื่องที่ทำความเข้าใจได้ยากและอาจเกิดความสับสนสำหรับผู้เรียนอยู่บ่อยครั้ง

โดยทั่วไปแล้วการเรียนภาษาที่ 1 ในวัยเด็กจะได้รับการฝึกฝนจากคนในครอบครัวเป็นอันดับแรก จากนั้นจะได้รับการฝึกฝนที่โรงเรียนจากครูตามบทเรียนในหนังสือที่ได้รับการรับรองจากกระทรวงศึกษาธิการ ดังนั้นงานวิจัยที่ผ่านมาแสดงให้เห็นว่า ผู้ที่ประสบปัญหาการเรียนภาษาที่ 1 มักจะเป็นผู้ที่มีปัญหาเกี่ยวกับอวัยวะที่ใช้ในการออกเสียง เช่น ผู้ที่ได้รับการรักษา เด็กออทิสติก หรือผู้ที่มีความผิดปกติตั้งแต่แรกเกิด เช่น เด็กที่มีความบกพร่องทางด้านการออกเสียง (speech sound disorder : SSD) ออกเสียงสระผิดมากกว่าเด็กปกติ [6] ผู้พูดที่เป็นดาว์ซินโดรม (Down syndrome : DS) ประสบปัญหาการออกเสียงสระมุม (/ɑ/, /æ/, /i/, /u/) [7] งานวิจัยที่ทำการวิเคราะห์และเปรียบเทียบลักษณะทางกลสัทศาสตร์ของเสียงสระภาษาไทยที่ออกเสียงโดยผู้พูดที่ใช้หลอดลม-หลอดอาหารกับผู้พูดปกติ [8]

ปัจจุบันการติดต่อสื่อสารข้ามกลุ่มหรือมีการติดต่อสื่อสารระหว่างประเทศด้วยสาเหตุหลายประการ เช่น การค้า การขนส่ง การแพทย์ การช่วยเหลือ หรือด้านการศึกษา ดังนั้นจึงทำให้มีการเรียนภาษาต่างประเทศเป็นภาษาที่ 2 (second language/L2) และภาษาที่ 3 (third language/L3) มากขึ้น ประเทศไทยมีนักศึกษาจากหลายชาติเข้ามาศึกษาภาษาไทย เช่น ชาวจีน ชาวญี่ปุ่น ชาวเกาหลี ชาวเวียดนาม ชาวพม่า ชาวยุโรป หรืออเมริกา เป็นต้น เมื่อมีการเรียนภาษาที่ 2 หรือ 3 ความแตกต่างของหน่วยเสียงระหว่างภาษาแม่และภาษาใหม่เป็นปัจจัยสำคัญที่เป็นปัญหาต่อการฝึกการออกเสียง จากการศึกษาที่ผ่านมา [9-13] พบว่ามีผู้เรียนประสบปัญหาด้านการออกเสียงสระสำหรับผู้เรียนภาษาที่ 2 และภาษาที่ 3 นอกจากนี้คนไทยที่ภาษาแม่ไม่ใช่สำเนียงมาตรฐานก็ประสบปัญหาเช่นกัน ในการศึกษาภาษาศาสตร์สังคมบางงานที่เกี่ยวกับการศึกษาเสียงสระในภาษาไทยถิ่น [14-16] พบว่าการแทรกแซงของสำเนียงจากภาษาไทยถิ่นอื่นๆ มีผลต่อการออกเสียงภาษาไทยได้ เช่น ภาษาไทยถิ่นเหนือ ภาษาไทยถิ่นใต้ แม้แต่ภาษาไทยกลางถิ่นอย่างสุพรรณบุรี หรืออยุธยา

พยางค์ในภาษาไทยจะประกอบด้วยเสียงพยัญชนะ เสียงสระและเสียงวรรณยุกต์ [17] องค์ประกอบในพยางค์หนึ่งพยางค์ถ้าไม่ปรากฏรูปวรรณยุกต์จะประกอบด้วย 1) พยัญชนะ + สระ

(Consonant + Vowel: CV) เช่น [ปา /pa:/ ‘to hurl’] หรือ [ดู / du:/ ‘to look’], 2) พยัญชนะ + สระ + พยัญชนะ (Consonant + Vowel + Consonant: CVC) เช่น [บาน /ba:n/ ‘to bloom’] หรือ [กิน /kin/ ‘to eat’] โดยที่ C สามารถเป็นได้ทั้งพยัญชนะต้น และพยัญชนะท้าย สำหรับพยัญชนะควบกล้ำจะใช้ C(C) เช่น [ปลา /pla:/ ‘fish’] หรือ [คลาน /khla:n/ ‘to crawl’] [18, 19]

สระเป็นแกนของพยางค์ที่เป็นส่วนประกอบสำคัญของคำพูด [1] แต่ละภาษามีจำนวนเสียงสระต่างกัน สระภาษาไทยมี 21 เสียง ถ้าเทียบกับภาษาอื่นๆ จะเห็นได้ว่าสระจะมีจำนวนที่แตกต่างกันไป เช่น ภาษาอังกฤษ 20 เสียง ภาษาญี่ปุ่น 5 เสียง ดังนั้นจึงเป็นปัญหากับผู้เรียนภาษาที่ 2 หรือภาษาที่ 3 ตัวอย่างเช่น ในกรณีของผู้เรียนที่ต้องฝึกออกเสียงสระใหม่ที่ไม่มีในภาษาแม่ เช่น ภาษาญี่ปุ่นไม่มีสระเออ สระออ แต่ภาษาไทยมีสระเออ สระออ เป็นต้น หรือในกรณีผู้เรียนที่ภาษาแม่ไม่มีความต่างด้านเวลาของเสียง จะประสบปัญหาเมื่อมีการออกเสียงสระเสียงสั้นและเสียงยาวในสระภาษาไทย เช่น [สด /sot/ ‘fresh’] และ [โสด /so:t/ ‘unmarried’], [ขุด /khut/ ‘to dig’] และ [ขูด /khu:t/ ‘to scrape’] [20] ซึ่งถ้าหากผู้เรียนออกเสียงสระผิดไปเพียงเล็กน้อย จะทำให้ความหมายผิดจากเดิม สระภาษาไทยพื้นฐานถูกแบ่งออกเป็น 2 แบบ คือ สระเสียงสั้นและสระเสียงยาว เช่น [ติ /ti/ ‘to censure’] หรือ [ดุ /du/ ‘to fierce’], และ [ตี /ti:/ ‘to hit’] หรือ [ดู /du:/ ‘to watch’] [21]

ในงานวิจัยนี้จะศึกษาเสียงสระเดี่ยว (pure vowels, monophthongs) เนื่องจากเสียงสระเดี่ยวเป็นเสียงสระพื้นฐานภาษาไทยซึ่งมีจำนวน 18 เสียง แบ่งออกเป็นสระเสียงสั้น 9 เสียง และสระเสียงยาว 9 เสียง โดยคู่ของสระเสียงสั้นและสระเสียงยาวเป็นเสียงสระที่มีความแตกต่างกันทางด้านระยะเวลา แต่มีคุณภาพใกล้เคียงกันทางค่าทางกลสัทศาสตร์ที่แสดงคุณลักษณะสระ (vowel quality) นั่นคือด้านความถี่ฟอร์เมินท์ (formant frequency) ในการศึกษาการแปร (เปลี่ยนแปลง) ของสระในพื้นที่สระ (vowel space) นักภาษาศาสตร์วิเคราะห์ค่าความถี่ฟอร์เมินท์ที่ 1 (first formant frequency : F1) และค่าความถี่ฟอร์เมินท์ที่ 2 (second formant frequency : F2) โดยค่าความถี่ฟอร์เมินท์ที่ 1 เกี่ยวกับความสูงต่ำของลิ้นซึ่งแปรผกผันกัน สระสูงมีค่าความถี่ฟอร์เมินท์ต่ำกว่าสระต่ำ ค่าความถี่ฟอร์เมินท์ที่ 2 เกี่ยวกับตำแหน่งหน้าหลังของลิ้นโดยแปรตามกัน สระที่ใช้ลิ้นส่วนหน้า (สระอิ สระเอะ สระแอะ) มีค่าความถี่ฟอร์เมินท์ที่ 2 สูงกว่าสระที่ใช้ลิ้นส่วนหลัง (สระอุ สระโอะ สระเอาะ) [22] การออกเสียงที่ถูกต้องหรือคล้ายคลึงกับการออกเสียงของภาษาเป้าหมายถือว่าเป็นสิ่งสำคัญอย่างหนึ่งของการเรียนภาษา เนื่องจากการออกเสียงเป็นปัจจัยสำคัญในการช่วยให้ผู้ฟังเข้าใจความหมายได้ถูกต้อง การออกเสียงที่ดีช่วยให้การสื่อสารมีประสิทธิภาพมากขึ้น [23] ในการฝึกเรียนภาษาอย่างมีประสิทธิภาพเพื่อช่วยแก้ไขปัญหาการออกเสียงและพัฒนาให้กับผู้เรียนสามารถออกเสียงภาษาได้อย่างถูกต้องมากขึ้น โดยทั่วไปจะอาศัยครูสอนภาษา ผู้เชี่ยวชาญ นักภาษาศาสตร์ (เช่น นัก

สัทศาสตร์) นักบำบัดการพูด หรือเจ้าของภาษา ในการอธิบายหรือสอนการออกเสียงให้กับผู้เรียน โดยใช้วิธีการฟังว่าผู้เรียนออกเสียงถูกต้องหรือไม่ แต่ถึงอย่างไรก็ตามในการระบุความถูกต้องจากการออกเสียงโดยการฟังนั้นความผิดพลาดย่อมเกิดขึ้นได้ เนื่องจากปัจจัยต่างๆ ของผู้ฝึกสอนและผู้เรียน

ปัจจุบันนี้เพื่อที่จะแสดงการออกเสียงได้อย่างถูกต้องและเที่ยงตรงมากขึ้น มีหลายงานที่ใช้เทคนิคเฉพาะหรือการใช้เครื่องมือพิเศษในการสอนการวิเคราะห์การออกเสียง เช่น การพัฒนาระบบการเรียนรู้การออกเสียง 3D ในภาษาจีน [24], การใช้อัลตราซาวด์กับการออกเสียงพยัญชนะ [25], แอพลิเคชันช่วยในการวิเคราะห์โทนในภาษาไทย [26], การประยุกต์ใช้โปรแกรม Praat กับการใช้หลักการทางกลสัทศาสตร์ในการวิเคราะห์การออกเสียงสระ [16, 27-31]

Praat [32] เป็นเครื่องมือยอดนิยมสำหรับนักภาษาศาสตร์ สำหรับการวิเคราะห์การออกเสียงสระจะใช้ Praat แสดงค่าความถี่ฟอร์เมินท์ ได้แก่ ค่าความถี่ฟอร์เมินท์ที่ 1 และค่าความถี่ฟอร์เมินท์ที่ 2 เพื่อตรวจสอบว่าการออกเสียงสระนั้นถูกหรือผิด ตำแหน่งของลิ้นจะถูกแสดงเป็นภาพ ซึ่งมีกระบวนการที่ซับซ้อนและต้องการผู้ที่มีทักษะในการใช้โปรแกรม โดยกระบวนการวิเคราะห์เริ่มจาก: 1) บันทึกเสียงและเปิดเสียงด้วยโปรแกรม Praat, 2) เลือกช่วงเวลาของเสียงสระในพยางค์ที่ต้องการวัด, 3) หาค่าเฉลี่ยของค่าความถี่ฟอร์เมินท์ที่ 1 และ ค่าความถี่ฟอร์เมินท์ที่ 2 ของช่วงเวลาของสระ, 4) บันทึกค่าเฉลี่ยของค่าความถี่ฟอร์เมินท์ที่ 1 และ ค่าความถี่ฟอร์เมินท์ที่ 2 ที่ได้ไปใส่ไว้ในตารางของ Microsoft Excel, 5) สร้างกราฟของผู้พูดที่เป็นเจ้าของภาษาและผู้เรียนด้วย Microsoft Excel, ภาษาโปรแกรมไพทอน หรือภาษาอาร์ 6) เปรียบเทียบกราฟของเจ้าของภาษากับผู้เรียน, 7) อธิบายกราฟให้ผู้เรียนฟังว่าออกเสียงถูกหรือผิด ซึ่งจะพบว่าวิธีการกำหนดความถูกต้องของการออกเสียงด้วยกราฟเป็นกระบวนการที่ซับซ้อน ใช้เวลานาน และไม่ใช่กระบวนการแบบเรียลไทม์ ผู้ใช้จะต้องเป็นผู้ที่เชี่ยวชาญหรือผู้ที่ศึกษาเรื่องเสียงโดยเฉพาะ หรือผู้ที่มีความรู้ทางด้านการเขียนโปรแกรม ทำให้ในปัจจุบันผู้เชี่ยวชาญเฉพาะด้านจึงมีจำนวนน้อย

ความก้าวหน้าของเทคโนโลยีสารสนเทศที่ทันสมัยและความสามารถในการประมวลผลของคอมพิวเตอร์ที่เพิ่มขึ้น ทำให้การเรียนรู้เชิงลึก (Deep leaning) ถูกนำมาประยุกต์ใช้กับงานทางด้านการรู้จำมากขึ้น เนื่องจากมีประสิทธิภาพทางการเรียนรู้และการจำแนกประเภทในด้านต่างๆ [33] เช่น การรู้จำอักขระที่เขียนด้วยลายมือ และการรู้จำเสียงพูด เป็นต้น การรู้จำเสียง (Speech Recognition) เป็นกระบวนการที่ทำให้คอมพิวเตอร์สามารถรับรู้ข้อมูลเสียง และสามารถทำการระบุหน่วยเสียง คำ วลี หรือประโยคในภาษาพูด สามารถแปลงข้อมูลไฟล์เสียงไปในรูปแบบของข้อความได้ โดยสามารถระบุเสียงพูดที่มนุษย์พูดใส่ไมโครโฟน โทรศัพท์ หรืออุปกรณ์อื่นๆ และสามารถนำไปประมวลผล วิเคราะห์ว่าเสียงนั้นเป็นเสียงของอะไร ในการรู้จำเสียงกระบวนการขั้นพื้นฐาน [34]

ประกอบด้วย สัญญาณเสียงที่ใช้เป็นข้อมูลเข้า (Speech Signal), กระบวนการประมวลผลสัญญาณเบื้องต้น (Preprocessing), การสกัดคุณลักษณะ (Feature Extraction) และสุดท้ายเป็นขั้นตอนการจำแนกประเภท (Classification) ซึ่งถูกแสดงในรูปที่ 1

รูปที่ 1 แสดงกระบวนการขั้นพื้นฐานในการรู้จำเสียง

กระบวนการเริ่มต้นโดยผู้ใช้ทำการพูดส่งสัญญาณเสียงพูด Speech Signal ซึ่งมีลักษณะสัญญาณเสียงที่เป็นสัญญาณอนาล็อก นำมาผ่านขั้นตอน Preprocessing ของการประมวลผลสัญญาณดิจิตอล แปลงสัญญาณเสียงมาเป็นสัญญาณเสียงในมิติของเวลาและความถี่ (Time-Frequency Domain) จากนั้นทำการสกัดคุณลักษณะ Feature Extraction ซึ่งเป็นการสกัดค่าที่แสดงลักษณะเฉพาะของสัญญาณเสียง และดำเนินกระบวนการจำแนกประเภท Classification โดยใช้โมเดลหรือโครงสร้างต่างๆ เช่น Artificial Neural Network (ANN), Convolutional Neural Network (CNN) เป็นต้น ในการระบุว่าเสียงนั้นเป็นเสียงของอะไร ซึ่งโดยทั่วไปคลาสเป้าหมายในการรู้จำคือ หน่วยเสียง หรือคำ ซึ่งเทคโนโลยีการรู้จำเสียงสามารถนำไปประยุกต์ใช้อย่างแพร่หลาย เช่น โรงพยาบาลสำหรับบุคลากรที่ไม่ถนัดหรือไม่ต้องการพิมพ์ ทางการทหารใช้ในการสั่งการระบบนักบินอัตโนมัติ การสั่งการรถยนต์ ใช้ร่วมกับแอพลิเคชันบนเว็บไซต์ โทรศัพท์มือถือ ใช้ในการสั่งหุ่นยนต์ การแปลเสียงให้เป็นคำพูดโทรศัพท์ตอบรับอัตโนมัติ เช่น สอบถามรอบฉายภาพยนตร์ หรือการสั่งการอุปกรณ์ไฟฟ้าต่างๆด้วยเสียง เป็นต้น

สำหรับระบบการเรียนรู้ภาษาโดยใช้คอมพิวเตอร์ช่วยสอน (Computer-Assisted Language Learning : CALL) และการฝึกการออกเสียงโดยใช้คอมพิวเตอร์ช่วย (Computer-Assisted Pronunciation Training : CAPT) [35-40] ได้รับความสนใจอย่างมากในด้านการสอนหรือการฝึกภาษา โดยถูกนำมาใช้เป็นเครื่องมือที่เพิ่มประสิทธิภาพสำหรับผู้เรียน ซึ่งระบบ CALL และ CAPT สามารถรู้จำเสียงพูด ผ่านเทคโนโลยีการรู้จำคำพูดอัตโนมัติ (Automatic Speech Recognition : ASR) ที่นิยมนำโครงสร้างการเรียนรู้เชิงลึกมาประยุกต์ใช้ นอกจากโครงสร้างการเรียนรู้เชิงลึกจะนิยมใช้กับชุดข้อมูลขนาดใหญ่ สามารถนำโครงสร้างการเรียนรู้เชิงลึกมาประยุกต์ใช้กับชุดข้อมูลขนาดเล็กได้ การปรับใช้โมเดลเชิงลึกกับชุดข้อมูลขนาดเล็กคือการใช้ Batch Normalization และการใช้ Dropout ซึ่งทำให้การบรรจบกันรวดเร็วและลดปัญหาการ overfitting ได้ [41] ตัวอย่างงานโครงสร้างการเรียนรู้เชิงลึกบนชุดข้อมูลขนาดเล็ก [42], [43] เป็นการรู้จำเสียงสระภาษาชวา ชุดข้อมูลประกอบด้วยสระกลางของภาษาชวาจำนวน 250 ไฟล์เสียงที่บันทึกโดยผู้พูดเพียงคนเดียว เอาท์พุต

ประกอบด้วย 5 คลาส ซึ่งโครงสร้างการเรียนรู้เชิงลึกมีประสิทธิภาพในการรู้จำดีกว่าโมเดล logistic regression และ multi-layer perceptron ผลลัพธ์ที่ได้คือความถูกต้องแม่นยำ 99.6% และ 94% ตามลำดับ

งานวิจัยที่ทำการศึกษาปัญหาด้านการออกเสียงสระภาษาไทยที่มีการศึกษาเกี่ยวกับระบบคอมพิวเตอร์ช่วยฝึกการออกเสียงสระภาษาไทยแบบอัตโนมัติที่ใช้ปัญญาประดิษฐ์ (AI) และมีโครงสร้างแบบการเรียนรู้เชิงลึก (Deep learning) ยังพบได้น้อย ดังนั้นงานวิจัยนี้มีจุดประสงค์เพื่อศึกษาโครงสร้างการเรียนรู้เชิงลึก (Deep learning) สำหรับการรู้จำการออกเสียงสระภาษาไทย 18 เสียง ผลลัพธ์การรู้จำเสียงที่ได้มีลักษณะเป็นหน่วยเสียงเดี่ยว (Single phoneme) ที่ใช้สำหรับการฝึกออกเสียงสระในภาษาไทยเบื้องต้น เพื่อนำไปพัฒนาระบบคอมพิวเตอร์ช่วยฝึกการออกเสียงสระภาษาไทยแบบอัตโนมัติ โดยระบบนี้ถูกพัฒนาเพื่อแก้ปัญหาการฝึกการออกเสียงสระภาษาไทยของผู้เรียนที่ไม่ใช่เจ้าของภาษา ผู้ที่เริ่มฝึกหัดพูดภาษาไทย หรือผู้ที่มีความบกพร่องทางการออกเสียง ลดปัญหาด้านการดำเนินงานที่ซับซ้อน ใช้เวลานาน ไม่แสดงผลเป็น real-time ของกระบวนการ และปัญหาการขาดแคลนผู้เชี่ยวชาญเฉพาะทางเกี่ยวการสอนการออกเสียงสระภาษาไทยอย่างถูกต้อง ในระบบคอมพิวเตอร์ช่วยฝึกการออกเสียงสระภาษาไทยแบบอัตโนมัติ จะมีการออกแบบโมเดลหรือโครงสร้างการเรียนรู้เชิงลึกในการรู้จำการออกเสียงสระภาษาไทย 18 เสียง ซึ่งเป็นเสียงสระพื้นฐานของภาษาไทย โมเดลถูกสร้างเพื่อฝึกให้คอมพิวเตอร์รู้จำเสียงสระได้เหมือนกับผู้เชี่ยวชาญที่สามารถระบุการออกเสียงสระของผู้เรียนได้ ในการวิจัยนี้ชุดข้อมูลใหม่ได้รับการออกแบบและรวบรวมเนื่องจากไม่มีข้อมูลที่เปิดเผยต่อสาธารณะ โดยชุดข้อมูลเสียงที่มีความแตกต่างกันในหลายมิติ เช่น เพศ อายุ สำเนียง สิ่งแวดล้อม เสียงรบกวน อื่น ๆ ในบริบทจริงในชีวิตประจำวัน สำหรับระบบคอมพิวเตอร์ช่วยฝึกการออกเสียงสระภาษาไทยแบบอัตโนมัติที่ใช้โครงสร้างการเรียนรู้เชิงลึกในการรู้จำเสียงสระภาษาไทยนี้ เป็นการพัฒนาระบบใหม่โดยใช้เทคนิคทางคอมพิวเตอร์บูรณาการร่วมกับภาษาศาสตร์ ที่สามารถนำไปใช้ประโยชน์ในด้านการบริการให้ความรู้กับผู้ที่เริ่มฝึกพูดภาษาไทย ผู้เรียนที่ไม่ใช่เจ้าของภาษา หรือ ผู้บกพร่องทางการออกเสียง จะเห็นได้ว่าระบบคอมพิวเตอร์ช่วยฝึกการออกเสียงสระภาษาไทยแบบอัตโนมัติ ทำให้ผู้เรียนสามารถฝึกออกเสียงสระได้แบบ real-time ตลอด 24 ชั่วโมง เปรียบเสมือนมีผู้เชี่ยวชาญ อาจารย์ภาษาไทย และนักภาษาศาสตร์ มาแนะนำการออกเสียงสระว่าถูกต้องหรือไม่ตลอดเวลา

1.2. วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีเป้าหมายเพื่อศึกษาโครงสร้างการเรียนรู้เชิงลึกสำหรับการรู้จำการออกเสียงสระภาษาไทย โดยแบ่งการศึกษาออกเป็นประเด็นย่อย ดังนี้

1. เพื่อออกแบบและเก็บตัวอย่างข้อมูลเสียงสระภาษาไทยจากเสียงของผู้ชายไทยและผู้หญิงไทย เช่น ช่วงอายุ สำเนียงการพูด สภาวะแวดล้อมที่ใช้ในการพูด เป็นต้น
2. เพื่อศึกษาโมเดลหรือโครงสร้างการเรียนรู้เชิงลึกที่เหมาะสมในการรู้จำการออกเสียงสระภาษาไทย 18 เสียงที่มีเสียงรบกวนในโลกแห่งความเป็นจริง

1.3. ขั้นตอนการดำเนินวิจัย

ในการวิจัยครั้งนี้ดำเนินตามขั้นตอนการวิจัย ดังต่อไปนี้

1. ศึกษางานวิจัยที่เกี่ยวข้อง
2. การรวบรวมข้อมูล
3. การเตรียมข้อมูล
4. การสร้างโมเดลหรือโครงสร้างการเรียนรู้เชิงลึกในการรู้จำการออกเสียงสระภาษาไทย
5. วัดประสิทธิภาพของโมเดลและระบบ

1.4. ขอบเขตงานวิจัย

ในงานวิจัยนี้ได้ทำการกำหนดขอบเขตของการวิจัยดังต่อไปนี้

1. ข้อมูล คือ ข้อมูลเสียงสระภาษาไทย 18 เสียงที่มีเสียงรบกวนในโลกแห่งความเป็นจริง เช่นเสียงรบกวนที่ได้จากรถยนต์บนท้องถนน เสียงผู้คนกำลังพูดกันในโรงอาหาร เสียงดนตรี เสียงลมในสวน และเสียงสัตว์ ซึ่งได้ทำการเก็บข้อมูลเสียงใหม่ที่ถูกบันทึกโดยโทรศัพท์มือถือ
2. งานวิจัยนี้มุ่งเน้นการจำแนกในการรู้จำเสียงสระเดี่ยวภาษาไทย 18 เสียงเท่านั้น เนื่องจากเป็นสระมาตรฐานของเสียงสระภาษาไทย
3. ผลลัพธ์ของการจำแนกเสียงสระในงานวิจัยนี้ ประกอบด้วยสระเสียงสั้น คือ อะ, อิ, อึ, อุ, เอะ, แอะ, โอะ, เอาะ และ เออะ สระเสียงยาวใช้ในงานวิจัยนี้ คือ อา, อี, อือ, อู, เอ, แอ, โอ, ออ และ เออ
4. สร้างโมเดลหรือโครงสร้างการเรียนรู้เชิงลึกสำหรับการรู้จำการออกเสียงสระภาษาไทย คือ โมเดล Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) และ Convolutional Neural Network รวมกับ Long Short-Term Memory (CNN_LSTM)

1.5. ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถจำแนกเสียงสระภาษาไทย 18 เสียงที่มีเสียงรบกวนในโลกแห่งความเป็นจริงได้
2. ได้โมเดลการจำแนกที่มีประสิทธิภาพในการรู้จำเสียงสระภาษาไทย 18 เสียงที่มีเสียงรบกวนในโลกแห่งความเป็นจริง เพื่อสามารถนำไปต่อยอดในระบบการฝึกการออกเสียงโดยใช้คอมพิวเตอร์ช่วย (Computer-Assisted Pronunciation Training : CAPT) ในการพัฒนาการเรียนภาษาไทยในอนาคต
3. ได้คลังข้อมูลเสียงสระภาษาไทยที่สามารถนำไปต่อยอดในงานการรู้จำเสียง
4. สามารถนำไปประยุกต์ใช้ในการสอนภาษาไทยสำหรับผู้ที่สนใจเรียนภาษาไทย หรือช่วยในการรักษาผู้ที่มีความบกพร่องทางการพูด ผู้ที่มีปัญหาในการออกเสียง
5. ส่งเสริมการอนุรักษ์ภาษาไทย

1.6. โครงร่างของเนื้อหาวิทยานิพนธ์

รายละเอียดของโครงร่างเนื้อหาวิทยานิพนธ์มีดังต่อไปนี้

บทที่ 1 บทนำ ประกอบด้วย ความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ของการวิจัย ขั้นตอนการดำเนินวิจัย ขอบเขตงานวิจัย ประโยชน์ที่คาดว่าจะได้รับ

บทที่ 2 งานวิจัยที่เกี่ยวข้อง ประกอบด้วย งานวิจัยทางด้านสระ งานวิจัยทางด้านการรู้จำเสียง (Speech Recognition) งานวิจัยทางด้านการรู้จำเสียงสระ (Vowel Speech Recognition) และงานวิจัย Gradient-weighted Class Activation Mapping (Grad-CAM)

บทที่ 3 ทฤษฎีที่เกี่ยวข้อง ประกอบด้วย ลักษณะของเสียงสระตามหลักสัทศาสตร์ การรู้จำเสียง (Speech Recognition) การเรียนรู้เชิงลึก (Deep Learning) วิธีการ Gradient-weighted Class Activation Mapping (Grad-CAM) และการประเมินผล

บทที่ 4 วิธีดำเนินงานวิจัยและผลการทดลองที่ 1: การรู้จำเสียงสระภาษาไทยโดยใช้ Convolutional Neural Network กับ MFCC

บทที่ 5 วิธีดำเนินงานวิจัยและผลการทดลองที่ 2: โมเดลการเรียนรู้เชิงลึกกับคุณสมบัติด้านเสียงสำหรับการรู้จำเสียงสระภาษาไทยแบบอัตโนมัติ

บทที่ 6 วิธีดำเนินงานวิจัยและผลการทดลองที่ 3: Grad-CAM สำหรับโมเดล CNN และระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง

บทที่ 7 วิธีดำเนินงานวิจัยและผลการทดลองที่ 4: โครงสร้างโมเดล CNN ร่วมกับ LSTM ในการรู้จำการออกเสียงสระภาษาไทย 18 เสียง

บทที่ 8 สรุปผลการทดลอง การอภิปรายผลและข้อเสนอแนะ

บทที่ 2

งานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงงานวิจัยที่เกี่ยวข้องกับงานวิจัยทางด้านสระทั้งภาษาต่างประเทศและภาษาไทย งานวิจัยทางด้านการรู้จำเสียง (Speech Recognition) ที่ใช้กับงานที่มีข้อมูลทั้งภาษาต่างประเทศและภาษาไทย งานวิจัยทางด้านการรู้จำเสียงสระ (Vowel Speech Recognition) ที่ใช้กับงานที่มีข้อมูลทั้งภาษาต่างประเทศและภาษาไทย และงานวิจัยที่เกี่ยวกับ Gradient-weighted Class Activation Mapping (Grad-CAM)

2.1. งานวิจัยทางด้านสระ

งานวิจัย [44] สำหรับการเรียนรู้เสียงภาษาต่างประเทศ ในงานวิจัยนี้ความสำเร็จขั้นสูงคือความถูกต้องแม่นยำในการผลิตเสียงสระในภาษาอังกฤษของผู้เรียนชาวเช็กที่เรียนภาษาอังกฤษระดับสูง ทำการวิเคราะห์สระที่มาจากความแตกต่างระหว่างคู่หมวดเสียงสระภาษาอังกฤษ “FLEECE–KIT”, “DRESS–TRAP”, และ “GOOSE–FOOT” โดยตรวจสอบความแตกต่างของความยาวของเสียง ความสูงของลิ้น และการเคลื่อนไปทางหลังของลิ้น ใช้เครื่องมือวัด Formant Measurements โดยมีผู้เรียนภาษาสองภาษาในระดับขั้นสูง (ภาษาเช็ก-ภาษาอังกฤษ) จำนวน 20 คน เป็นเพศหญิง อายุระหว่าง 19-27 ปี ซึ่งเรียนในระดับปริญญาตรีในการแปล-ล่ามภาษาอังกฤษที่กำลังศึกษาระดับมหาวิทยาลัยพาแลคกี (Palacký University) โดยดูค่าความถี่ของการออกเสียงสระ ซึ่งถ้าสระในภาษาแม่มีการออกเสียงเช่นเดียวกับสระในภาษาที่สองก็ไม่เกิดปัญหาใดๆ แม้ว่าผู้เรียนภาษาจะเรียนขั้นสูงก็ยังพบปัญหาในการผลิตสระ ปัญหาทางการถ่ายโอนสระจากภาษาแม่ไปยังภาษาที่สอง ในงานวิจัย [11] ได้ทำการศึกษาแบบจำลองการกลมกลืนการรับรู้ (Perceptual Assimilation Model: PAM) ซึ่งสามารถทำนายการเรียนรู้ของสระภาษาอังกฤษสำเนียงใต้ (Standard Southern British English : SSBE) โดยผู้เรียนชาวอาเซอร์ไบจัน (Azerbaijani : AZ) ซึ่งได้ออกแบบการทดลองทางการรับรู้การออกเสียงและทางการผลิตเสียง ในการทดสอบการรับรู้ผู้เรียนชาวอาเซอร์ไบจันได้รับการทดสอบความสามารถในการแยกความแตกต่างของสระ SSBE จำนวน 11 สระ สำหรับการทดสอบการผลิตการออกเสียง ตรวจสอบความถูกต้องของผู้เรียนในการสร้างสระ SSBE ผ่านการวัดค่าทางเสียงและการตัดสินของผู้ฟังเจ้าของภาษา ผลการวิจัยพบว่าการตัดสินของผู้ฟังเจ้าของภาษาส่วนใหญ่สอดคล้องกับการรับรู้และผลลัพธ์ความถูกต้องแม่นยำของเสียง ซึ่งผลพบว่า /ɪ/ และ /ʊ/ ส่วนใหญ่สับสนกับ /i/ และ /u/ ตามลำดับ, เสียงสระ /ʌ/ ส่วนใหญ่สับสนกับสระ /ɑ/ และ /ɒ/ ซึ่งในสระ /ɒ/ ก็สับสนกับ /ɑ/ และ /ʌ/ โดยรวมแล้วการคาดการณ์ของ PAM นั้นถูกต้องสำหรับ

ผู้เรียน ในงานวิจัย [45] สำหรับการที่รู้ระบบเสียงของภาษาที่สองหรือภาษาต่างประเทศจะช่วยในเรื่องการออกเสียง นอกจากนี้ผลกระทบของระบบเสียงภาษาแม่มีผลต่อการออกเสียงของภาษาที่สอง ดังนั้นการเปรียบเทียบระบบเสียงของทั้งสองภาษาช่วยให้ทราบถึงความแตกต่างของภาษาและแหล่งที่มาของข้อผิดพลาดของผู้เรียน ดังนั้นการศึกษานี้จึงพยายามที่จะเปรียบเทียบประสิทธิภาพของนักเรียนที่เรียนภาษาอังกฤษเป็นภาษาต่างชาติ (English as a Foreign Language : EFL) ชาวเคิร์ดและชาวเปอร์เซียในการเรียนรู้สระภาษาอังกฤษ โดยใช้สมมติฐานการวิเคราะห์เปรียบเทียบ (Contrastive Analysis Hypothesis: CAH) เพื่อเปรียบเทียบสระของชาวเคิร์ดและชาวเปอร์เซียกับภาษาอังกฤษ โดยใช้นักเรียนทั้งหมด 120 คน เพื่อศึกษาความแตกต่างที่เป็นไปได้ระหว่างการออกเสียง ซึ่งอยู่ในระดับเริ่มต้นและระดับสูง ผลการวิจัยพบความแตกต่างอย่างมีนัยสำคัญในผู้เรียนระดับเบื้องต้น มีความแตกต่างอย่างมีนัยสำคัญในการออกเสียงของสระ /ə/ มีเพียง 17% ของผู้เรียน EFL ชาวเปอร์เซียในระดับเบื้องต้นเท่านั้นที่สามารถออกเสียงสระ /ə/ ในคำพูดคำว่า “a̲bout” เพราะสระ /ə/ ไม่มีอยู่ในภาษาเปอร์เซีย ปัญหาการออกเสียงของนักเรียนไทยที่เรียนวิชาสัทศาสตร์ภาษาอังกฤษ กรณีศึกษากำลังศึกษาระดับมหาวิทยาลัยกาฬสินธุ์ [13] มีวัตถุประสงค์การวิจัยเพื่อศึกษาความคิดเห็นของนักเรียนเกี่ยวกับปัญหาในการเรียนรู้ของออกเสียง และเพื่อหาปัจจัยที่ทำให้เกิดปัญหาในการเรียนรู้ทางการออกเสียงของนักเรียน กลุ่มตัวอย่างเป็นนักเรียนระดับปริญญาตรีภาษาอังกฤษเพื่อการสื่อสารระหว่างประเทศจำนวน 12 คนที่เรียนวิชาสัทศาสตร์ภาษาอังกฤษ ซึ่งได้ทำการคัดเลือก 6 คน เพื่อทำการสัมภาษณ์ เครื่องมือที่ใช้ในการวิจัยเป็นแบบสอบถามและแบบกึ่งสัมภาษณ์ ความเห็นของนักเรียนสะท้อนให้เห็นว่าการออกเสียงสระเดี่ยวและสระผสม มีค่าเฉลี่ยอยู่ที่ระดับ 3.75 ซึ่งเป็นปัญหาในการเรียนในระดับมาก ความสามารถในการออกเสียงทำให้เกิดปัญหาในการเรียนรู้เรื่องการออกเสียง สรุปความคิดเห็นได้ว่าความแตกต่างของระบบเสียงระหว่างภาษาอังกฤษกับภาษาไทยมีปัจจัยบางประการ ได้แก่ การออกเสียงของภาษาแม่ การเรียนรู้การออกเสียงภาษาอังกฤษก่อนการเรียนการสอนและแรงจูงใจในการเรียนทำให้เกิดปัญหาในการเรียนรู้

สำหรับงานวิจัยที่วิเคราะห์ลักษณะทางกลสัทศาสตร์ของสระภาษาไทยมีจำนวนค่อนข้างน้อย ซึ่งส่วนใหญ่ศึกษาลักษณะทางกลสัทศาสตร์ของวรรณยุกต์ [46] มีการศึกษาลักษณะทางกลสัทศาสตร์ของสระภาษาไทยที่ออกเสียงโดยนักศึกษาจีนกับการออกเสียงของคนไทยที่เรียนภาษาไทยเพื่อวิเคราะห์และเปรียบเทียบลักษณะทางกลสัทศาสตร์ ผู้ให้ข้อมูลประกอบด้วย กลุ่มที่พูดภาษาจีนถิ่นยูนนาน ภาษาน่าซี และภาษาไทลื้อ เป็นภาษาแม่ ผู้ให้ข้อมูลกลุ่มคนไทยพูดภาษาไทยสำเนียงกรุงเทพฯ เป็นภาษาแม่ ใช้ค่าระยะเวลา (ความสั้น-ยาว) และค่าความถี่ฟอร์เมินท์ที่ 1 และที่ 2 (คุณสมบัติสระ) ในการวิเคราะห์ผล ผลการศึกษาพบว่านักศึกษาจีนทุกคนออกเสียงสระเสียงสั้นยาวกว่าสระเสียงสั้นที่ออกเสียงโดยคนไทย และสระเสียงยาวที่นักศึกษาจีนออกเสียงก็สั้นกว่าสระเสียงยาวของคนไทย และ

จากการศึกษายังพบอีกว่าสระบางนักศึกษาจีนมีปัญหาในการออกเสียง เช่น สระ อิ เอะ แอะ ในงาน [47] ศึกษาการออกเสียงภาษาไทยของอาจารย์ในกำลังศึกษาระดับมหาวิทยาลัยรังสิต ด้านการออกเสียงพยัญชนะ สระ วรรณยุกต์ และจังหวะในการออกเสียง โดยบันทึกเสียงขณะสอนจำนวน 1 ครั้ง นำการออกเสียงของอาจารย์แต่ละคนมาถอดเสียงโดยการฟัง และตรวจสอบความถูกต้องชัดเจนตามแนวคิดในหนังสือระบบเสียงภาษาไทย เก็บรวบรวมข้อมูลแบบการสุ่มตัวอย่างแบบง่าย (Simple random sampling) กลุ่มตัวอย่างทั้งหมด 60 คน พบว่า ถึงแม้ว่าจะเป็นชาวไทยแต่ก็ยังมีการออกเสียงผิด การออกเสียงสระที่ไม่ถูกต้องมี 2 หน่วยเสียงคือ สระอา /aa/<-า> แทนการออกเสียงเป็นเสียงสระอะ /a/ <-ะ> และการออกเสียงสระแอะ /**ɛ**/ <แ-ะ> แทนการออกเสียงสระอะ /a/ <-ะ> อีกทั้งยังพบอีกว่ามีการประสมหน่วยเสียงพยัญชนะท้าย /-j/ <ย>เป็น /aj/ <ไ-> ซึ่งไม่มีในภาษาไทยปัจจุบัน ในงานวิจัยการออกเสียงสระภาษาไทยมาตรฐานโดยผู้พูดที่ไม่ใช่เจ้าของภาษา [48] ศึกษาการวิเคราะห์ลักษณะทางกลศาสตร์ของเสียงสระภาษาไทยมาตรฐานโดยมีผู้พูดภาษาแม่ที่แตกต่างกัน 4 ภาษา ประกอบด้วย ภาษาเขมร เวียดนาม พม่า และมาเลเซีย ผู้บอกภาษาเป็นผู้พูดที่มีประสบการณ์ภาษาไทยมากจำนวนภาษาละ 3 คน ผู้พูดที่มีประสบการณ์ภาษาไทยน้อยจำนวนภาษาละ 3 คน รวมทั้งสิ้น 24 คน มีอายุระหว่าง 20-35 ปี การทดสอบประกอบด้วยสระเดี่ยวเสียงสั้นและยาว 18 หน่วยเสียง บันทึกเสียงผ่านคอมพิวเตอร์โดยตรง จำนวนคำทดสอบทั้งสิ้น 2,592 คำ วิเคราะห์ค่าทางกลสัทศาสตร์โดยใช้ค่าความถี่ฟอร์เมินท์ที่ 1 และที่ 2 ค่าระยะเวลาของเสียงสระด้วยโปรแกรม Praat เมื่อพิจารณาลักษณะทางกลศาสตร์ภาษาไทยผลวิจัยพบว่าการออกเสียงโดยผู้พูดที่พูดภาษาเขมร เวียดนาม พม่า และมาเลเซียเป็นภาษาแม่ที่มีประสบการณ์ภาษาไทยมากมีสัทลักษณะที่ดีกว่ากลุ่มผู้พูดที่มีประสบการณ์ภาษาไทยน้อย และในกลุ่มผู้พูดที่มีประสบการณ์ภาษาไทยน้อยมีการซ้อนทับกันระหว่างสระสูง กลาง และต่ำ จึงเกิดการแปรสูงของสระในด้านระดับสูงต่ำของลิ้น สำหรับด้านความสั้น-ยาวของเสียงสระ ผลพบว่าผู้บอกภาษาเกือบทั้งหมดในกลุ่มผู้พูดที่มีประสบการณ์ภาษาไทยมากและน้อย สามารถออกเสียงสระสั้นยาวในอัตราส่วนที่ใกล้เคียงกันกับผู้พูดภาษาไทย แต่ค่าระยะเวลาของเสียงสระในกลุ่มผู้พูดที่มีประสบการณ์ภาษาไทยน้อยจะพบการแปร (เปลี่ยนแปลง) ของเสียงมากกว่ากลุ่มผู้พูดที่มีประสบการณ์ภาษาไทยมาก

2.2. งานวิจัยทางด้านการรู้จำเสียง (Speech Recognition)

แบบจำลอง Convolutional Neural Networks (CNNs) ไม่ได้ถูกใช้แค่ในงานคอมพิวเตอร์วิทัศน์เท่านั้น แต่ยังสามารถนำมาประยุกต์ใช้กับงานทางด้านการรู้จำเสียงได้อีกด้วย สำหรับในงานทางด้านการรู้จำเสียง ในหลายๆงานวิจัยได้นำข้อดีของแบบจำลอง CNNs มาใช้เนื่องจาก CNNs สามารถใช้ในด้านการลดความแปรปรวนของความถี่หรือสเปกตรัมได้ สำหรับงานทางการรู้จำเสียงอัตโนมัติ (Automatic speech recognition: ASR) [49] ได้ใช้ CNN ร่วมกับกลยุทธ์ที่หลากหลาย

เช่น การพูลลิ่ง (Pooling) และการใช้ค่าน้ำหนักร่วมกัน ซึ่งทำให้ประสิทธิภาพของผลลัพธ์ดีขึ้น ในงานการรู้จำคำหลักขนาดเล็ก (Small-footprint Keyword Spotting : KWS) [50] สถาปัตยกรรม CNN ถูกใช้กับการจำแนก 14 วลี (“รับสาย”, “ปฏิเสธสาย”, “อีเมล”, “กรอไปข้างหน้า”, “รายการเพลงถัดไป”, “เพลงถัดไป”, “แทรคเพลงถัดไป”, “หยุดเพลง”, “หยุด”, “เล่นเพลง”, “ตั้งนาฬิกา”, “ตั้งเวลา”, “เริ่มจับเวลา” และ “จดบันทึก”) ให้ผลลัพธ์อัตราการปฏิเสธที่ผิด 27-44% ในงานคำพูดขนาดใหญ่ (Large-scale Speech) [51] และงานการรู้จำเสียงรบกวนที่แข็งแกร่ง (Noise Robust Speech Recognition) [52] นักวิจัยได้นำเสนอสถาปัตยกรรม CNN ที่ดีที่สุดและกลยุทธ์ที่ใช้ สำหรับในงานคำพูดขนาดใหญ่ มีอัตราความผิดพลาดของคำ (Word Error Rate: WER) ระหว่าง 12% ถึง 14% บนงานการรู้จำคำศัพท์อย่างต่อเนื่อง 3 งาน นั่นคือ งานเสียงทางด้านข่าว (Broadcast News: BN) 50 ชั่วโมงและ 400 ชั่วโมง และงานสวิตบอร์ด (Switchboard: SWB) 300 ชั่วโมง สำหรับงานการรู้จำเสียงรบกวนที่แข็งแกร่ง ที่ทดลองบนข้อมูล Aurora4 มีอัตราความผิดพลาดของคำ 8.81% นอกจากนี้ยังประสบความสำเร็จในอัตราความผิดพลาดของคำที่ลดลง 10.0% เมื่อเทียบกับ CNN แบบดั้งเดิมบนข้อมูลการถอดความการประชุม AMI (AMI meeting transcription) ความทนทานของแบบจำลอง CNN [53] ได้ใช้ร่วมกับเทคนิค 2 วิธี คือ การใช้คุณลักษณะ Autoregressive Moving Average Spectrogram และการทำ Channel Dropout เพื่อเพิ่มประสิทธิภาพของผลลัพธ์ วิธี Channel Dropout ให้ผลอัตราความผิดพลาดของคำ 16% เมื่อใช้ร่วมกับ ARMA และให้ผลอัตราความผิดพลาดของคำ 20% เมื่อใช้ร่วมกับ FBANK บนสถาปัตยกรรม CNN ขั้นพื้นฐาน ในงานวิจัย [54] ได้ทำการรวมกันของแบบจำลองใช้เพื่อเพิ่มประสิทธิภาพในการรู้จำเสียงบนงานคำศัพท์ขนาดใหญ่ สถาปัตยกรรมแบบจำลองประกอบด้วย CNN, Long Short-Term Memory (LSTM), และ Deep Neural Network (DNN) ซึ่งเรียกว่า CLDNN โดยสถาปัตยกรรม CLDNN สามารถช่วยลดอัตราความผิดพลาดของคำ 4 ถึง 6% ซึ่งมากกว่าสถาปัตยกรรม LSTM สำหรับในงาน [55] ศึกษาวิธีการดรอปเอาท์ (Dropout) ในการรู้จำเสียงอัตโนมัติที่ใช้ Long Short Term Memory (LSTM) ระบบฝึกอบรมด้วยฟังก์ชั่นการสูญเสีย Connectionist Temporal Classification (CTC) การใช้งาน Dropout ซึ่งส่งผลให้การปรับปรุงประสิทธิภาพการรู้จำเสียงพูดมีความสำคัญในชุดข้อมูล Librispeech และ ชุดข้อมูล GALE Arabic ที่มีการลดลงของอัตราข้อผิดพลาดของคำ (WER) เมื่อเทียบกับโครงสร้างพื้นฐาน 24.64% และ 13.75% ตามลำดับ งาน [56] นำเสนอการรู้จำการพูดตัวเลขภาษาอินโดนีเซีย (0-9) โดยใช้การเรียนรู้เชิงลึก Learning LongShort Term Memory (LSTM) การสกัดคุณลักษณะ Linear Predictive Coding (LPC) และคุณลักษณะ MFCC (Mel-Frequency Cepstrum) ถูกใช้เป็นข้อมูลเข้าในสถาปัตยกรรม LSTM และทำการเปรียบเทียบคุณลักษณะด้วยอัตราของความถูกต้องแม่นยำในการรู้จำ คุณลักษณะ LPC จะสกัดคุณลักษณะ

เสียงพูดตามระดับเสียง (Pitch) หรือความถี่พื้นฐาน (Fundamental Frequency) ในขณะที่ MFCC สกัดคุณลักษณะเสียงพูดตามสเปกตรัมของเสียง ในงานนี้ใช้คำพูดตัวเลข 7990 เสียง ประกอบด้วย สัมประสิทธิ์ LPC เท่ากับ 12 และ สัมประสิทธิ์ MFCC เท่ากับ 12 ผลการวิจัยพบว่าการใช้ LSTM สำหรับการรู้จำคำพูดตัวเลขภาษาอินโดนีเซีย การสกัดคุณลักษณะ MFCC มีผลความถูกต้องแม่นยำ 96.58% ซึ่งดีกว่าเมื่อเทียบกับการสกัดคุณลักษณะ LPC ที่มีความถูกต้องแม่นยำ 93.79% สำหรับงานวิจัยการจำแนกประเภทเสียงพูดที่ทำการทดลองโดยใช้ข้อมูลทางด้านอารมณ์ [57] จากคลังข้อมูล 2 คลังข้อมูล คือ คลังข้อมูล Interactive Emotional Dyadic Motion Capture (IEMOCAP) และคลังข้อมูล Emotional Tagged Corpus on Lakorn (EMOLA) ซึ่งอารมณ์ถูกแบ่งออกเป็น 4 ประเภทประกอบด้วย ความโกรธ, ความสุข, ธรรมดา และ ความโศกเศร้า พบว่าแต่ละอารมณ์ใช้คุณลักษณะที่แตกต่างกัน การใช้ MFCC ร่วมกับ Zero Crossing Rate (ZCR) ได้ผลลัพธ์ที่ดีในคลาสอารมณ์ ความโกรธและความสุข ซึ่งได้ผลความถูกต้องแม่นยำ 81.95% และ 69.86% สำหรับการใช้ CNN กับการจำแนกประเภททางอารมณ์ของการพูด [58] โดยใช้ สัญญาณเสียงพูดข้อมูลเข้าจริง ซึ่งถูกสกัดโดย Convolutional Long Short-Term Memory Recurrent Neural Network (ConvLSTM-RNN model) และทำการจำแนกประเภทโดยใช้ Support Vector Machines ซึ่งทำการทดลองบนฐานข้อมูล IEMOCAP ผลลัพธ์ของ SVM ที่มี Polynomial Kernel ใน 192 หน่วยเสียง มีอัตราความถูกต้อง 65.13% สำหรับ Mel spectrogram (MS) ซึ่งถูกแปลงจากสัญญาณเสียงพูด (16 kHz) ถูกนำไปใช้กับงานการรู้จำเสียงคำสั่ง (speech command recognition (SCR)) [59] รูปภาพ MS ที่มีขนาดคุณลักษณะเสียง 125 × 80 × 1 ถูกใช้เป็นคุณสมบัติด้านเสียง โมเดล Light Interior Search Network (LIS-Net) ถูกนำไปใช้กับงาน SCR โดยใช้ชุดข้อมูล Google Speech Command การทดลองพบว่า มีการปรับปรุงความถูกต้องแม่นยำทั้งคำสั่งจำนวน 12, 20 และ 35 คำสั่ง พร้อมทั้งใช้เวลาทำนายที่รวดเร็วและจำนวนพารามิเตอร์เพียงเล็กน้อย ผลลัพธ์ของโมเดล LIS-Net มีความถูกต้องแม่นยำถึง 98.1% โมเดลประกอบด้วยเลเยอร์ข้อมูลเข้า ตามด้วย Light Interior Search Block (LIS-Block) และ classification block แต่ละ LIS-Block ประกอบด้วย LIS-Cores หลายตัวและเลเยอร์ convolutional สองชั้น ตามด้วย batch normalization (BN) และเลเยอร์ activation โมเดลใช้ Adam optimizer ในงาน [60] ได้พัฒนาโครงสร้างที่รวม classifiers สามตัว ได้แก่ DNN, CNN และ RNN ใช้คลังข้อมูล IEMOCAP ที่บันทึกโดยนักแสดง 10 คน โมเดลนี้ใช้เพื่อรู้จำอารมณ์ 4 แบบ ได้แก่ โกรธ มีความสุข ปกติ และเศร้า ที่ระดับเฟรม low-level descriptors (LLDs) ถูกถ่ายโอนไปยัง RNN เพื่อใช้ในโมเดล LLD-RNN ที่ระดับเซ็กเมนต์ MS ถูกถ่ายโอนไปยัง CNN เพื่อใช้ในโมเดล MS-CNN ที่ระดับคำพูด เอาต์พุตของ high-level statistical functions (HSFs) ถูกถ่ายโอนไปยัง DNN เพื่อใช้ในโมเดล HSF-DNN การ

จำแนกประเภทอารมณ์ที่ไม่ต่อเนื่องและการถดถอยของคุณลักษณะทางอารมณ์อย่างต่อเนื่องได้ ดำเนินการไปพร้อม ๆ กัน กลยุทธ์ confidence-based fusion ถูกนำมาใช้เพื่อรวมตัวจำแนกที่หลากหลายในการรับรู้สภาวะทางอารมณ์ต่างๆ โมเดลที่มีกลยุทธ์ fusion ได้รับความถูกต้องแม่นยำที่ 57.1% (weighted) และความถูกต้องแม่นยำที่ 58.3% (unweighted) ในงาน [61] โมเดล 1D และ 2D CNN LSTM ถูกนำไปใช้กับงานการรู้จำอารมณ์คำพูด โดยใช้เสียงพูด log-Mel spectrogram เป็นข้อมูลเข้าตามลำดับ โมเดลเหล่านี้ใช้สถาปัตยกรรมที่คล้ายคลึงกันซึ่งประกอบด้วย local feature learning blocks (LFLBs) สี่ชุด แต่ละบล็อกประกอบด้วยเลเยอร์ convolutional หนึ่งชั้น เลเยอร์ BN หนึ่งชั้น เลเยอร์ ELU หนึ่งชั้น เลเยอร์ max pooling หนึ่งชั้น และเลเยอร์ LSTM หนึ่งชั้น ผลการวิจัยพบว่า โมเดล 2D CNN-LSTM มีประสิทธิภาพเหนือกว่าวิธีการแบบเดิม เช่น Deep Believe Network และ CNN โครงสร้างของโมเดล 2D CNN LSTM ประกอบด้วย LFLB สี่ชุด เลเยอร์ LSTM หนึ่งชั้น และเลเยอร์ fully connected ชั้น Softmax classifier ถูกนำไปใช้กับชั้นบนสุด สำหรับ log Mel spectrogram ที่มี 251 เฟรมและ 128 Mel frequency bins ถูกใช้เป็น audio features ผลการทดลองแสดงให้เห็นว่า โดยรวมแล้วโมเดล 2D CNN LSTM มีประสิทธิภาพดีกว่าโมเดล 1D CNN LSTM ซึ่งในงานนี้ได้มีการประยุกต์ใช้ ELU กับการรู้จำอารมณ์ของคำพูด นอกจากนี้ในงาน [62] วิธี ELU ถูกใช้เพื่อให้เกิดการเรียนรู้ที่รวดเร็วและความถูกต้องแม่นยำที่สูงขึ้นใน DNN และแก้ปัญหา vanishing gradient โดยที่ ELU ให้ค่าในส่วนที่เป็นลบที่ผลักดันให้ mean unit activations เข้าใกล้ 0 ซึ่งแตกต่างจาก rectified linear units (ReLUs) ดังนั้น ELU จึงลดช่องว่างระหว่าง normal และ unit natural gradients ซึ่งนำไปสู่การเรียนรู้ที่รวดเร็ว ในชุดข้อมูลการ vision ต่างๆ ผลลัพธ์แสดงให้เห็นว่า ELU มีประสิทธิภาพเหนือกว่า ฟังก์ชันการกระตุ้น (Activation Function) อื่นๆ อย่างมีนัยสำคัญ โมเดลที่มี ELU ทำงานได้ดีกว่าโมเดลที่มี ReLU ที่มี BN

สำหรับงานทางด้านการรู้จำเสียงในภาษาไทย [63] ใช้ระบบ Neuro-Fuzzy กับภาษาไทย 8 คำ เช่นคำว่า “ไปข้างหน้า”, “หลัง”, “ซ้าย”, “ขวา” ซึ่งได้บันทึกในสภาวะแวดล้อมที่มีเสียงรบกวนที่แตกต่างกัน ผลการวิจัยแสดงให้เห็นว่าแต่ละปัจจัยมีผลกระทบที่แตกต่างกันของความถูกต้องแม่นยำในการรู้จำ ในงานวิจัยที่เกี่ยวข้องกับการสกัดคุณลักษณะ Double Filter Banks และกระบวนการรู้จำโดยใช้ Euclidian Distance [64] ใช้กับเสียงคำสั่งภาษาไทยพื้นฐานภายใต้เงื่อนไขที่หลากหลายจากอาสาสมัคร (9,000 เสียงพูด) และได้รับอัตราความถูกต้องประมาณ 96.3% ในงาน [65] มีการใช้การรู้จำเสียงในงานทางด้านการแปลภาษาไทยกับภาษาอังกฤษ โดยมีเป้าหมายเกี่ยวกับการพัฒนาพัฒนาโปรแกรมรู้จำเสียงภาษาไทยโดยใช้เวลาและทรัพยากรข้อมูลที่จำกัด ผลการทดลองประสบความสำเร็จโดยใช้วิธี rapid bootstrapping ในงาน [66] เสนอการรู้จำเสียงพูดภาษาไทย

แบบทนทานกับสัญญาณรบกวนจากสภาพแวดล้อมภายนอก ใช้วิธีการเชิงลบสเปกตรัม (Spectral subtraction: SS) ในการกำจัดเสียงรบกวน ชุดข้อมูลเป็นเสียงพูดตัวเลขภาษาไทย 0-9 จำนวน 10 คำ ผู้พูด 10 คน (เพศหญิง 5 คน และเพศชาย 5 คน) อายุระหว่าง 20-22 ปี โดยแต่ละคนพูดคนละ 10 ครั้ง ทดสอบกับสภาพแวดล้อมที่ไม่มีเสียงรบกวนและมีเสียงรบกวนระดับความดังประมาณ 40-50 dB ผลลัพธ์สามารถจำแนกเสียงตัวเลขได้ 98.0% ในสภาพแวดล้อมที่ไม่มีเสียงรบกวน และ 83.6% ในสภาพแวดล้อมที่มีเสียงรบกวน ในงาน [67] ศึกษาการรู้จำเสียงพูดภาษาไทยแบบทนทานต่อเสียงรบกวนในสภาพแวดล้อมจริง โดยการใช้ไมโครโฟนอาเรย์ (Microphone array) และอัลกอริทึม N-best LIMABEAM ใช้ห้องบันทึกเสียงขนาด 5.5 x 4 เมตร เสียงรบกวนประกอบด้วยเสียงเครื่องปรับอากาศและเสียงโทรทัศน์มีค่าความดัง 70 dB ใช้โทรศัพท์ Smartphone บันทึกเสียงประกอบด้วยเสียงพูดผู้ชาย 231 ประโยค เสียงพูดผู้หญิง 214 ประโยค จำนวนคำทั้งหมด 4,069 คำ ผลการศึกษาอัลกอริทึม N-best LIMABEAM มีค่าความถูกต้อง 27.22% ซึ่งดีกว่าอัลกอริทึม LIMABEAM ที่มีค่าความถูกต้อง 20.12% และเสียงที่มีเสียงรบกวนมีค่าความถูกต้อง 9.47% ในงาน [68] นำเสนอระบบการรู้จำอัตโนมัติสำหรับประโยคภาษาไทยโดยใช้ MFCC ร่วมกับ CNN และวิธีการ You Only Look Once (YOLO) ถูกนำมาใช้ในการจำแนกประเภท ชุดข้อมูลเกี่ยวข้องกับการให้บริการภายในสนามบิน คำพูดรวบรวมจากผู้พูด 60 คน (เพศชาย 30 คน และเพศหญิง 30 คน) จำนวนชุดข้อมูลประกกอบด้วย 2,400 ไฟล์เสียง YOLOv3 และ Tiny YOLOv3 ได้รับการฝึกฝนและประเมินประสิทธิภาพ ผลลัพธ์ Tiny YOLOv3 มีประสิทธิภาพเหมาะสม โดยระบบ ASR ที่นำเสนอที่ใช้ MFCC และ CNN มีประสิทธิภาพดีทั้งความเร็วและความถูกต้อง ซึ่งมีความถูกต้องประมาณ 82% สำหรับงานวิจัยการจำแนกประเภทเสียงพูดที่ทำการทดลองโดยใช้ข้อมูลทางด้านอารมณ์ [57] จากคลังข้อมูล 2 คลังข้อมูล คือ คลังข้อมูล Interactive Emotional Dyadic Motion Capture (IEMOCAP) และ คลังข้อมูล Emotional Tagged Corpus on Lakorn (EMOLA) ซึ่งอารมณ์ถูกแบ่งออกเป็น 4 ประเภทประกอบด้วย ความโกรธ, ความสุข, ธรรมดา และ ความโศกเศร้า พบว่าแต่ละอารมณ์ใช้คุณสมบัติที่แตกต่างกัน การใช้ MFCC ร่วมกับ Zero Crossing Rate (ZCR) ได้ผลลัพธ์ที่ดีในคลาสอารมณ์ ความโกรธและความสุข ซึ่งได้ผลความถูกต้องแม่นยำ 81.95% และ 69.86% สำหรับการใช้ CNN กับการจำแนกประเภททางอารมณ์ของการพูด [58] โดยใช้ สัญญาณเสียงพูดข้อมูลเข้าจริง ซึ่งถูกสกัดโดย Convolutional Long Short-Term Memory Recurrent Neural Network (ConvLSTM-RNN model) และทำการจำแนกประเภทโดยใช้ Support Vector Machines ซึ่งทำการทดลองบนฐานข้อมูล IEMOCAP ผลลัพธ์ของ SVM ที่มี Polynomial Kernel ใน 192 หน่วยเสียง มีอัตราความถูกต้อง 65.13%

2.3. งานวิจัยทางด้านการรู้จำเสียงสระ (Vowel Speech Recognition)

สำหรับงานเกี่ยวกับสระ สถาปัตยกรรม CNN ถูกนำไปประยุกต์ใช้กับภาษาชวาซึ่งเป็นภาษาของประเทศอินโดนีเซีย [42], [43] ซึ่ง Mel-frequency spectral coefficients (MFSC) ถูกนำมาใช้ในการสกัดคุณลักษณะ ชุดข้อมูลประกอบด้วยสระกลางของภาษาชวาจำนวน 250 ไฟล์เสียงที่บันทึกโดยผู้พูดเพียงคนเดียว เอาท์พุตประกอบด้วย 5 คลาส ผลลัพธ์ที่ได้คือความถูกต้องแม่นยำ 94% งานวิจัย [69] สระเป็นหน่วยเสียงที่มีการใช้งานมากที่สุดมีและมีความถี่ต่ำ ภาษาอัสสัม (Assamese) ซึ่งภาษาอัสสัมภาษากลางของอินเดียตะวันออกเฉียงเหนือ ทั้งหมดมีหน่วยเสียงสระ 8 เสียง คือ /i/, /e/, /ɛ/, /a/, /ɒ/, /ɔ/, /o/ และ /u/ ได้ใช้ Recurrent Neural Network (RNN) เพื่อรับรู้เสียงสระจากภาษาอัสสัม คุณลักษณะของเวกเตอร์ถูกสร้างขึ้นโดย พิจารณาคุณสมบัติการออกเสียงอะคูสติกของสระ ผลการทดลองได้รับอัตราความถูกต้องของการรู้จำ 84% สำหรับงานวิจัย [70] การรับรู้เสียงของหน่วยเสียงสระมีบทบาทสำคัญในด้านการพูด ภาษาอัสสัม (Assamese) เป็นภาษาหลักของรัฐอัสสัมและภาษาแม่ที่ใช้มากที่สุดของประชากรของรัฐอัสสัม ประเทศอินเดีย มาตรฐานภาษาอัสสัมมี 4 ภาษาหลัก คือ ภาษาถิ่นกลาง, ภาษาตะวันออก, ภาษาโกลปาริยาและภาษาคามาปี สระมี 8 เสียง และในงานนี้เป็นการวิเคราะห์เปรียบเทียบระหว่างวิธีที่ใช้ Recurrent Neural Network (RNN) และ KNearest Neighbor (KNN) การรู้จำเสียงสระโดยใช้คุณลักษณะการออกเสียงอะคูสติก (Acoustic Phonetic Features) เป็นคุณสมบัติเวกเตอร์ อัตราการรู้จำ 97% ได้มาจากการใช้ KNN สำหรับการรู้จำเสียงสระและอัตราโดยรวมวิธีที่ใช้ RNN มีความถูกต้อง 84.3% และ KNN ความถูกต้อง 87% ดังนั้นสำหรับเสียงสระอัสสัมได้รับการยอมรับว่าวิธีการที่ใช้ KNN ให้อัตราการรู้จำที่ดีกว่าวิธีการแบบ RNN ในงาน [71] ความแม่นยำของระบบรู้จำเสียงพูดขึ้นอยู่กับชุดคุณลักษณะที่ใช้ในการเป็นตัวแทนของข้อมูลเสียงพูด กระบวนการที่ต่อเนื่องเพื่อพัฒนาชุดคุณลักษณะทำให้การรู้จำเสียงพูดมีความแม่นยำยิ่งขึ้น ชุดคุณลักษณะหลายอย่างและชุดคุณลักษณะผสมที่แตกต่างกันได้พยายามทำการทดลองเพื่อให้บรรลุผลที่ดี ในงานนี้ได้ทำการศึกษาการลดชุดคุณลักษณะ MFCCs สำหรับการรู้จำเสียงสระ การศึกษามุ่งเน้นไปที่การสร้างและศึกษาพฤติกรรมของคุณลักษณะ MFCCs สำหรับเสียงสระที่แตกต่างกัน เป้าหมายคือการระบุคุณลักษณะที่สามารถแยกประเภทและปรับปรุงประสิทธิภาพของ ASR ผลการวิเคราะห์แสดงให้เห็นว่าคุณลักษณะที่ลดลงจาก 12 MFCCs เป็น 3 MFCCs ที่เสนอนั้นทำงานได้ดีและใช้เพื่อปรับปรุงความแม่นยำของระบบ ASR ได้โดยเฉพาะเมื่ออยู่ในอุปกรณ์ที่จำกัดทรัพยากร ในงาน [72] สระเสียงสั้นภาษาอาหรับถูกนำมาใช้กับโมเดล CNN สำหรับการรู้จำหน่วยเสียงภาษาอาหรับคลาสสิก 84 คลาส ที่ได้จากพยัญชนะ 28 เสียงที่สัมพันธ์กันกับสระเสียงสั้น 3 เสียง ชุดข้อมูลถูกบันทึกในรูปแบบออนไลน์จากผู้พูด 85 คน มีทั้งหมด 6,229 ไฟล์เสียง โมเดลมีถูกต้องความแม่นยำ 95.77%

มีงานวิจัยจำนวนไม่มากที่นำเสนอเกี่ยวกับการรู้จำเสียงสระภาษาไทย ในงาน [73] ได้มีการประยุกต์ใช้ Linear Predictive Coefficients (LPC) และใช้ Critical Band Intensities (CBI) การเพิ่มประสิทธิภาพถูกนำมาใช้ในการจำแนกเสียงสระเสียงสั้นและสระเสียงยาว และการรู้จำเสียงสระในภาษาไทยทั้งแบบไม่ผสมเสียงและผสมเสียง ซึ่งเสียงถูกรวบรวมจากผู้พูด 6 คน ผลลัพธ์ของการจำแนกสำหรับสระเสียงสั้นและเสียงยาวที่ไม่ได้ผสมเสียงสระ สำหรับโมเดล 3 male model ได้ผลความถูกต้องแม่นยำ 89.39% โมเดล 3 female model ได้ผลความถูกต้องแม่นยำ 89.83% และโมเดล 2 male-2 female model ได้ผลความถูกต้องแม่นยำ 87.67% สำหรับการฝึกอบรมในโมเดลชายและหญิงมีจำนวนตัวอย่างเสียง 1,134 ตัวอย่าง สำหรับโมเดลที่ผสมชายและหญิง ใช้ตัวอย่างเสียง 1,512 ตัวอย่าง ในงาน [74] ใช้โมเดล CNN ร่วมกับ MFCC ในการรู้จำเสียงสระภาษาไทยมาตฐาน ชุดข้อมูลเสียงสระภาษาไทยที่มีเสียงรบกวน รวบรวมจากผู้พูด 50 คน ชุดข้อมูลประกอบด้วยไฟล์เสียงสระ 1,800 เสียง ผลลัพธ์ประกอบด้วย 18 คลาส ประสิทธิภาพการรู้จำเสียงสระภาษาไทยได้รับความถูกต้องแม่นยำ 90.00% และ 88.89% สำหรับชุดข้อมูลเพศหญิงและเพศชายตามลำดับ

2.4. งานวิจัยทางด้าน Gradient-weighted Class Activation Mapping (Grad-CAM)

ในงาน [75, 76] เสนอเทคนิค class-discriminative localization แบบใหม่ ในการสร้าง visual explanations สำหรับการตัดสินใจจากคลาสของโมเดลที่ใช้เครือข่าย (CNN) โดยใช้ Gradient-weighted Class Activation Mapping (Grad-CAM) ที่ไม่ต้องมีการเปลี่ยนแปลงสถาปัตยกรรมหรือการฝึกอบรมซ้ำ งานวิจัยดังกล่าวได้ทำการรวม Grad-CAM กับการสร้างภาพข้อมูลที่มีความละเอียดเพื่อสร้างภาพที่มีความละเอียดสูง Guided Grad-CAM และการนำ Grad-CAM ไปประยุกต์ใช้ในโมเดลในการจำแนกรูปภาพ (image classification) การบรรยายภาพ (image captioning) และการตอบคำถามด้วยภาพ (visual question answering :VQA) รวมถึงสถาปัตยกรรมที่ใช้ ResNet ในโมเดลการจำแนกรูปภาพ การแสดงภาพของงานวิจัยช่วยระบุความเอนเอียงของชุดข้อมูล และแสดงข้อมูลเชิงลึกในโหมดความล้มเหลวของโมเดล CNN ในปัจจุบันประสิทธิภาพดีกว่าวิธีการก่อนหน้านี้ใน ILSVRC-15 มีความสามารถในการตีความและมีความเที่ยงตรงมากกว่าโมเดลพื้นฐาน สำหรับการบรรยายภาพและ VQA การแสดงภาพของงานวิจัยแสดงโมเดลที่ไม่เน้นความสนใจก็สามารถใช้ได้ การศึกษาในมนุษย์เปิดเผยว่าการแสดงภาพของงานวิจัยสามารถแยกแยะระหว่างคลาสได้อย่างแม่นยำมากขึ้น มีความน่าเชื่อถือของตัวจำแนกประเภทได้ดีขึ้นและช่วยระบุอคติในชุดข้อมูลได้ ในงาน Electroencephalogram (EEG) เป็นการตอบสนองโดยตรงต่อการทำงานของสมองสามารถใช้เพื่อตรวจจับสภาพจิตใจและสภาพร่างกาย ในการศึกษาการรับรู้อารมณ์โดยใช้ EEG เนื่องจากสัญญาณ EEG มีลักษณะที่ไม่เป็นเชิงเส้น มีลักษณะสัญญาณไม่นิ่ง และ

มีความแตกต่างกันในแต่ละบุคคล วิธีการรู้จำแบบดั้งเดิมยังคงมีข้อเสียของการสกัดคุณลักษณะที่ซับซ้อนและมีอัตราการรู้จำที่ต่ำ งานของ [77] มีจุดประสงค์เพื่อสร้างโมเดลการรู้จำอารมณ์ที่มีประสิทธิภาพ โดยไม่ขึ้นอยู่กับสิ่งเร้าต่างๆ เช่น อุปกรณ์เก็บ EEG เป็นต้น งานวิจัยได้เสนอแนวคิดใหม่ของแผนที่แจกแจงความถี่อิเล็กโทรด (electrode-frequency distribution maps: EFDMs) ที่มีการแปลงฟูเรียร์ในเวลาสั้นๆ (short-time Fourier transform: STFT) ซึ่ง convolutional neural network (CNN) เชิงลึกถูกนำเสนอสำหรับการสกัดคุณลักษณะอัตโนมัติและการจำแนกอารมณ์กับ EFDMs ซึ่งงานนี้เสนอวิธีการรู้จำอารมณ์แบบข้ามชุดข้อมูลของการเรียนรู้การถ่ายโอนโมเดลเชิงลึก การทดลองดำเนินการกับชุดข้อมูลที่เปิดเผยต่อสาธารณะสองชุด วิธีการที่เสนอนี้ได้ความถูกต้องการจำแนกเฉลี่ย 90.59% จากข้อมูล EEG ที่มีความยาวสั้น ๆ บนชุดข้อมูล SEED จากนั้นจึงนำโมเดลที่ได้รับการฝึกฝนมาใช้กับชุดข้อมูล DEAP ผ่านการเรียนรู้การถ่ายโอนโมเดลเชิงลึกซึ่งได้ความถูกต้องเฉลี่ย 82.84% งานวิจัยนี้ใช้ Gradient-weighted Class Activation Mapping (Grad-CAM) เพื่อให้ทราบว่า CNN ได้เรียนรู้คุณลักษณะใดบ้างในระหว่างการฝึกอบรมจาก EFDM และสรุปได้ว่าคลื่นความถี่สูงเหมาะสำหรับการจดจำอารมณ์มากกว่า สำหรับในงาน งาน [78] ในบทความนี้คือการขยาย conventional convolutional recurrent neural networks แบบทั่วไปโดยใช้สถาปัตยกรรมแบบ (3D) convolutional สำหรับการตรวจจับเสียงนก โดยนำเสนอ 3-dimensional (3D) convolutional recurrent neural networks ซึ่งใช้ประโยชน์จากการเรียนรู้เชิงลึกสำหรับการตรวจจับเสียงของนก โดย 3D convolutions ใช้เพื่อสกัดข้อมูลระยะยาวและระยะสั้นในความถี่พร้อมกันจากสตรีมข้อมูลเสียง และใช้ recurrent neural networks (RNN) แยกกัน โดยดำเนินการกับกับตัวกรอง (filter) แต่ละตัวของเลเยอร์ Convolutional สุดท้าย การปรับปรุงด้วยการปรับโมเดลทำให้ได้คะแนน AUC ที่ 89.58% ในการระบุรูปแบบและแสดงภาพผลกระทบของพื้นที่เฉพาะที่สำคัญสำหรับโมเดลในการทำนาย ในงานนี้ใช้ Gradient-weighted Class Activation Mapping (Grad-CAM) เพื่อแสดงให้เห็นภาพโมเดลที่ได้รับการฝึกอบรม Grad-CAM คำนวณการไล่ระดับของคะแนนที่ทำนายไว้ สำหรับคลาสใดคลาสหนึ่งโดยคำนึงถึงเอาต์พุตของ feature maps ของเลเยอร์คอนโวลูชันชั้นสุดท้าย ผลลัพธ์จะเน้นถึงความสำคัญของแผนที่คุณลักษณะ (feature maps) สำหรับคลาสเป้าหมาย ในการเปรียบเทียบ 2D convolution ใน CNN+RNN จะเน้นที่ตำแหน่งเฉพาะของเสียงนก และรวมถึงบริเวณที่มีความถี่ต่ำโดยไม่มีเสียงนก ซึ่ง 3D convolution สามารถดึงข้อมูลเวลาระยะยาว long-term time ในการร้องของนกได้มากกว่าโดยดูจากผลการแสดงภาพของ Grad-CAM ในงาน [79] ศึกษาการจำแนก Acoustic scene โดยใช้เสียงที่บันทึกจาก

รังผึ้งเพื่อตรวจจับการเปลี่ยนแปลงภายในรังผึ้ง การศึกษานี้สามารถพัฒนาเป็นระบบเฝ้าติดตามที่สามารถตรวจจับสภาวะผิดปกติในรังผึ้งได้ เนื่องจากผึ้งมีบทบาทสำคัญในอุตสาหกรรมการเกษตร และทุกปีจำนวนผึ้งยังคงลดลงอย่างต่อเนื่องจากการถูกคุกคาม การศึกษานี้มีวัตถุประสงค์หลักสองประการ จุดประสงค์แรกเกี่ยวข้องกับการใช้ข้อมูลเสียงเพื่อเปรียบเทียบประสิทธิภาพของอัลกอริธึมการเรียนรู้ของเครื่องต่างๆและกำหนดวิธีการที่เหมาะสมสำหรับการจำแนกประเภทคลาสที่เป็นเสียงของผึ้ง (Bee) และ ไม่ใช่เสียงของผึ้ง (noBee) จุดประสงค์ที่สองเกี่ยวข้องกับการใช้ gradient-weighted class activation mapping (Grad-CAM) สำหรับโมเดล CNN เพื่ออธิบายปัจจัยที่สำคัญรวมทั้งช่วงเวลา เมื่อโมเดลทำนายคลาสของ Bee และ noBee งานวิจัยนี้ใช้วิธีการสกัดคุณลักษณะกับเสียง mel spectrogram, mel-frequency ceptral coefficients (MFCCs) และการแปลงค่า constant-Q เพื่อเปรียบเทียบประสิทธิภาพของโมเดลการเรียนรู้ของเครื่องทั่วไป (machine learning) คือ support vector machine, random forest และ extreme gradient boosting กับโมเดล Convolutional neural Network (CNN) นั่นคือ shallow CNN และ VGG-13 มีการใช้ Grad-CAM เพื่อพิจารณาว่าโมเดล CNN ที่ทำงานได้ดีที่สุดสามารถรู้จำ audio scenes ได้อย่างไร นำเสนอผลการแสดงภาพโดยใช้ Grad-CAM เพื่อตรวจสอบกระบวนการฝึกอบรมของโมเดล CNN ที่ทำงานได้ดีที่สุด โดยทำการวิเคราะห์พื้นที่บน MFCC ที่มีความสำคัญสำหรับการจัดหมวดหมู่ผ่านการแสดงภาพที่ได้จากการใช้ Grad-CAM กับโมเดล VGG-13 โดยการแสดงภาพสัญญาณเสียงและแสดงให้เห็นว่าโมเดลที่มี CNN สามารถแยกแยะความแตกต่างในเสียงที่มนุษย์ไม่สามารถแยกแยะได้โดยตรง Grad-CAM ถูกนำมาใช้โดยใช้น้ำหนักของเลเยอร์ Convolutional สุดท้ายและการไล่ระดับสีที่ใช้ในการทำนายคลาสเป้าหมาย ในกรณีของคลาสที่เป็นเสียงของผึ้ง strong activation เกิดขึ้นตลอดช่วงเวลาของความถี่เฉพาะทั้งหมดในช่วงเวลา ซึ่งมีความสำคัญในการจำแนกประเภทคลาส bee สำหรับเสียงที่ไม่ใช่เสียงผึ้ง ส่วนที่เจาะจงที่มีเสียงอื่นๆ ถูกระบุว่าเป็นปัจจัยสำคัญในการทำนายคลาส noBee ซึ่งโมเดล VGG-13 สามารถกำหนดช่วงเวลาที่เกี่ยวข้องกับเสียงที่ไม่ใช่ผึ้งได้อย่างมีประสิทธิภาพ การวิเคราะห์เสียงโดยใช้โมเดล CNN ด้วยวิธีการประมวลผลล่วงหน้าที่เหมาะสมสามารถแยกแยะระหว่างเสียงของผึ้งกับเสียงที่ไม่ใช่ผึ้งได้อย่างมีประสิทธิภาพ จึงสามารถตรวจจับและระบุความผิดปกติในรังได้อย่างรวดเร็ว ในงานนี้ยังแสดงให้เห็นถึงประสิทธิภาพของ Deep CNNs ในการจำแนกเสียงของผึ้งและไม่ใช่เสียงของผึ้งที่บันทึกภายในรัง ผลการศึกษาพบว่าโมเดล VGG-13 ที่ใช้ MFCC เป็นข้อมูลข้อมูลเข้า ให้ค่าความถูกต้องสูงสุด (91.93%) สำหรับ precision, recall, และ F1-score ในแต่ละคลาส พบว่าเสียงอื่นที่ไม่ใช่เสียงของผึ้งได้รับการรู้จำอย่างมีประสิทธิภาพ

บทที่ 3

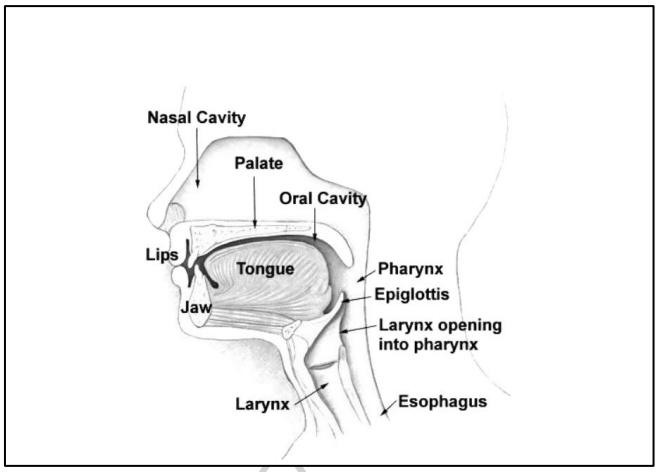
ทฤษฎีที่เกี่ยวข้อง

ในงานวิจัยนี้ได้ทำการศึกษาทฤษฎีที่เกี่ยวข้องลักษณะของเสียงสระตามหลักสัทศาสตร์ การรู้จำเสียง (Speech Recognition) การเรียนรู้เชิงลึก (Deep Learning) วิธี Gradient-weighted Class Activation Mapping (Grad-CAM) และการประเมินผล

3.1. ลักษณะของเสียงสระตามหลักสัทศาสตร์

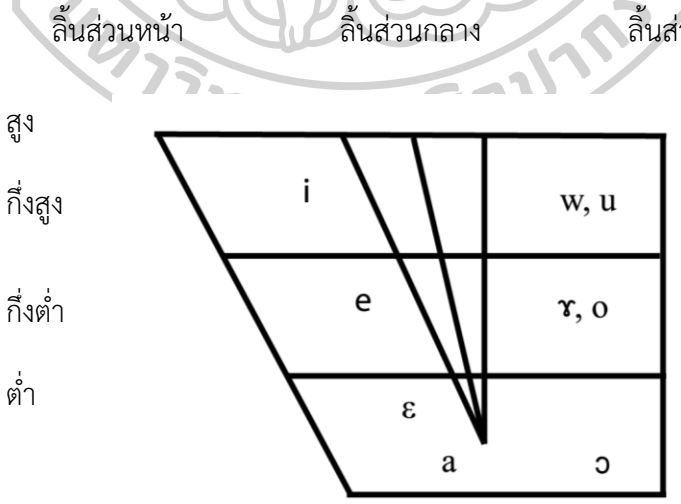
ความหมายของการออกเสียงสระ [80] สามารถให้ความหมายได้ 2 ลักษณะ คือความหมายทางสัทศาสตร์หรือลักษณะการออกเสียง และตามหน้าที่ของเสียงสระในภาษา การออกเสียงตามลักษณะทางสัทศาสตร์ เสียงสระคือเสียงที่เกิดจากลมออกมาทางปอดและผ่านทางเส้นเสียง ซึ่งโดยทั่วไป เสียงสระเป็นเสียงโฆษะ (Voiced) คือ เส้นเสียงสั่นแล้วลมเคลื่อนผ่านเหนือกลางลิ้นและออกมาจากปาก โดยไม่มีการกักลมที่จุดใดๆในช่องปาก สำหรับความหมายตามหน้าที่ของเสียงสระในภาษา เสียงสระคือเสียงที่เป็นแกนของพยางค์ (Nucleus) โดยอาจจะมีเสียงพยัญชนะอยู่ข้างหน้าหรือตามข้างหลังก็ได้ เช่น เสียงสระในคำว่า b<u>a</u>t ช<u>า</u>ม เป็นต้น การยกระดับลิ้นที่แตกต่างกันรวมถึงลักษณะของรูปริมฝีปากที่ต่างกันในเวลาที่ต่างกันจะทำให้เกิดการออกเสียงสระที่แตกต่างกันออกไป

การจำแนกเสียงสระ จะมีการพิจารณาส่วนของลิ้นที่ใช้ในการออกเสียง (Tongue Position), ความสูงต่ำของลิ้น (Tongue Height) และลักษณะของริมฝีปาก (Lip Position) โดยส่วนของลิ้นที่ใช้ในการออกเสียงสระแบ่งเป็น 3 ส่วน ดังรูปที่ 2 คือ ลิ้นส่วนหน้า (Front of the Tongue) หมายถึงลิ้นที่ตรงข้ามกับเพดานแข็ง สระในภาษาไทยที่ใช้ลิ้นส่วนหน้าในการออกเสียงเช่น สระอิ, สระแอะ เป็นต้น ลิ้นส่วนหลัง (Back of the Tongue) หมายถึงลิ้นที่อยู่ตรงข้ามกับเพดานอ่อน สระในภาษาไทยที่ใช้ลิ้นส่วนหลัง เช่น สระอุ, สระโอะ เป็นต้น ลิ้นส่วนกลาง (Central part of the Tongue) หมายถึงลิ้นที่อยู่ตรงข้ามกับส่วนต่อระหว่างเพดานแข็งกับเพดานอ่อน สำหรับด้านความสูงต่ำของลิ้น สระที่ใช้ลิ้นส่วนเดียวกันอาจมีระดับความสูงต่ำของลิ้นที่แตกต่างกัน สระในภาษาไทย เช่น สระอิ สระเอะ สระแอะ สระอะ นั้นมีการยกระดับลิ้นที่แตกต่างกัน สระอิจะมีการยกระดับลิ้นสูงที่สุดและสระอะจะมีระดับลิ้นที่ต่ำที่สุด ในการแบ่งระดับความสูงต่ำของภาษาไทย จะจำแนกระดับการยกลิ้น 4 ระดับคือ สูง, กึ่งสูง, กึ่งต่ำ และต่ำ (Close, Close-Mid, Open-Mid, Open) สำหรับลักษณะของริมฝีปาก แบ่งออกเป็น 2 ลักษณะ คือ ริมฝีปากห่อ (Rounded) และริมฝีปากไม่ห่อ (Unrounded) สระในภาษาไทย เช่น สระอิ สระเอะ สระแอะ เป็นสระที่ริมฝีปากไม่ห่อหรือที่เรียกว่าริมฝีปากเหยียด (Spread) สระอุ สระโอะ เป็นสระที่ริมฝีปากไม่ห่อ เป็นต้น



รูปที่ 2 แสดงส่วนของลิ้นที่ใช้ในการออกเสียง ลิ้นส่วนหน้า ลิ้นส่วนกลาง และ ลิ้นส่วนหลัง [81]

สระเดี่ยว (Pure Vowels, Monophthongs) หมายถึง เสียงสระที่มีคุณลักษณะการออกเสียงในพยางค์ที่คงที่ ไม่มีการเปลี่ยนแปลง เช่น สระอิ, สระเอะ, สระแอะ, สระอะ, สระเอาะ, สระโอ, สระอุ, สระเออะ และสระอุ ในภาษาไทยเป็นสระเดี่ยวเนื่องจากลักษณะการออกเสียงจะคงที่อยู่ตั้งแต่ต้นจนจบเสียงสระ สระเดี่ยวเป็นได้ทั้งสระเสียงสั้น (Short Vowels) และสระเสียงยาว (Long Vowels) คำว่าสระยาว หมายถึง มีระยะเวลา (Duration) ที่ใช้ในการออกเสียงมากกว่าสระเสียงสั้น ในการแสดงลักษณะเสียงยาว จะใช้จุด 2 จุด (:) ไว้ข้างหลังสัทอักษร เช่น สระอิ (i) เป็นสระเสียงสั้น และสระอี (i:) เป็นสระเสียงยาว ภาพเสียงสระเดี่ยวในภาษาไทยแสดงในรูปที่ 3 ซึ่งสระแต่ละสระจะมีค่าความถี่ฟอร์เมินท์ที่ 1 และ 2 แตกต่างกัน



รูปที่ 3 แสดงภาพเสียงสระเดี่ยวในภาษาไทย

3.2. การรู้จำเสียง (Speech Recognition)

การรู้จำเสียง (Speech Recognition) เป็นการที่ทำให้คอมพิวเตอร์สามารถรับรู้ข้อมูลเสียง โดยสามารถแปลงข้อมูลไฟล์เสียงไปในรูปแบบของข้อความได้ สามารถนำไปประมวลผลวิเคราะห์ว่าเสียงนั้นเป็นเสียงของอะไร ในการรู้จำเสียงกระบวนการขั้นพื้นฐาน [34] ประกอบด้วย สัญญาณเสียงที่ใช้เป็นข้อมูลเข้า (Speech Signal), กระบวนการประมวลผลสัญญาณเบื้องต้น (Preprocessing), การสกัดคุณลักษณะ (Feature Extraction) และสุดท้ายเป็นขั้นตอนการจำแนกประเภท (Classification)

3.2.1. สัญญาณเสียง (Speech Signal)

สัญญาณเสียง (Speech Signal) ที่ใช้เป็นข้อมูลเข้าเป็นเสียงพูดที่เกิดขึ้น เช่น เสียงการพูดเป็นหน่วยเสียง คำ วลี ประโยค เสียงที่เกิดจากเหตุกาณ์หนึ่งๆ เช่น เสียงนกร้อง เสียงคนไอ เป็นต้น โดยมีลักษณะเป็นสัญญาณอนาล็อก

3.2.2. กระบวนการประมวลผลสัญญาณเบื้องต้น (Preprocessing)

จากสัญญาณข้อมูลเข้าที่มีสัญญาณอนาล็อก นำมาผ่านขั้นตอนของการประมวลผลสัญญาณดิจิตอล ได้แก่ Pre-emphasis ซึ่งเป็นขั้นตอนการนำสัญญาณเสียงผ่านกระบวนการกรองเพื่อให้อัตราส่วนของสัญญาณเสียงต่อสัญญาณรบกวนมีค่าคงที่, Framing/Windowing เป็นการแบ่งสัญญาณเสียงออกเป็นส่วนย่อย ๆ (Segmentation) ขนาดความยาวประมาณ 10–30 มิลลิวินาที และ Short Time Fourier Transform เป็นการแปลงสัญญาณเสียงมาเป็นสัญญาณเสียงในมิติของเวลาและความถี่ (Time-Frequency Domain) เพื่อนำไปใช้ในขั้นตอนต่อไป

3.2.3. การสกัดคุณลักษณะ (Feature Extraction)

การสกัดคุณลักษณะ (Feature Extraction) เป็นการสกัดค่าที่แสดงลักษณะเฉพาะของสัญญาณเสียง สำหรับ Mel Frequency Cepstral Coefficients (MFCC) เป็นการสกัดคุณลักษณะทางเสียงที่เป็นที่นิยมสำหรับการรู้จำเสียง [34] ซึ่งประกอบด้วย Pre-emphasis, Framing, Windowing, Fast Fourier Transform (FFT), Mel Filter Bank, และ Discrete Cosine Transform (DCT) รูปที่ 4 แสดงกระบวนการ MFCC Feature Extraction

รูปที่ 4 แสดงกระบวนการ MFCC Feature Extraction [34]

กระบวนการแรกใน MFCC คือ ขั้นตอน Pre-emphasis สัญญาณเสียงพูดผ่านกระบวนการกรองก่อนเพื่อสร้างพลังงานในความถี่สูง จากนั้นทำการกำหนดกรอบ (Framing) เพื่อตัดสัญญาณเสียงออกเป็นส่วนเล็กๆ ข้อมูลเสียงจะมีความยาวระหว่าง 10-30 มิลลิวินาที ในระบบรู้จำเสียง การวิเคราะห์สัญญาณจะดำเนินการในช่วงเวลาสั้นๆ (Frame) ดังนั้นการตัดเฟรมให้มีขนาดที่เล็กลงเป็นสิ่งจำเป็นเพื่อให้ยังคงคุณสมบัติดั้งเดิมสำหรับการวิเคราะห์สัญญาณ การทำ Windowing ถูกใช้เพื่อหลีกเลี่ยงความไม่ต่อเนื่องของสัญญาณที่สร้างขึ้นจากกระบวนการ Framing ซึ่งฟังก์ชันที่ใช้คือ Hamming window หลังจากการทำ Windowing แต่ละเฟรมจะถูกแปลงเป็นโดเมนความถี่ การแปลงฟูริเยร์อย่างรวดเร็ว (FFT) ใช้เพื่อแปลงสัญญาณจากโดเมนเวลาเป็นโดเมนความถี่ ซึ่ง FFT เป็นอัลกอริทึมที่รวดเร็วของการแปลงฟูริเยร์แบบไม่ต่อเนื่อง (Discrete Fourier Transform: DFT) ต่อจากนั้นสัญญาณการแปลงฟูริเยร์จะถูกส่งผ่านชุดตัวกรอง Bandpass ที่มีชื่อว่า Mel Filterbank โดยกระบวนการ Mel Scaling แสดงในสมการที่ 1 ดังนี้

$$mel = 2595 \log_{10}(1+\frac{f}{700}) \quad (1)$$

โดยที่ mel หมายถึง ผลลัพธ์ของ Mel Filterbank

f หมายถึง ข้อมูลเข้าของ Filterbank

ในขณะที่ 2595 และ 700 เป็นค่าคงที่ที่ใช้กันอย่างแพร่หลายในวิธีการ MFCC และกระบวนการขั้นสุดท้าย คือการแปลงแบบไม่ต่อเนื่องโคไซน์ (DCT) ที่สร้างสัมประสิทธิ์ MFCC

สำหรับ Mel Spectrogram (MS) เป็นการสกัดคุณลักษณะทางเสียงทีมีกระบวนการเหมือนกับ MFCC แต่ไม่มีกระบวนการแปลงแบบไม่ต่อเนื่องโคไซน์ (DCT)

3.2.4. การจำแนกประเภท (Classification)

การจำแนกประเภท (Classification) เป็นกระบวนการที่เกี่ยวข้องกับการจัดหมวดหมู่ การแบ่งประเภท การจำแนกคลาส เป็นประเภท Supervised Model ซึ่งโมเดลประเภทนี้มี Class Labels ที่ชัดเจน ในการรู้จำเสียงขั้นสุดท้ายในการจำแนกเสียงเพื่อระบุว่าเป็นเสียงของใคร เสียงเหตุการณ์ใด เสียงของสระ พยัญชนะ หรือคำใด เป็นต้น

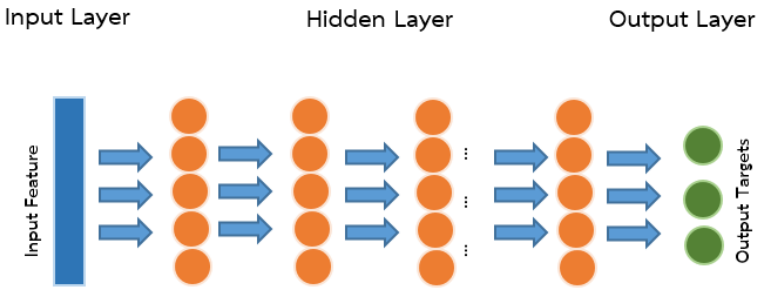
3.3. การเรียนรู้เชิงลึก (Deep Learning)

อัลกอริทึมการเรียนรู้เชิงลึก แสดงให้เห็นประสิทธิภาพทางการเรียนรู้และการจำแนกประเภทในด้านต่าง ๆ [33] เช่น การรู้จำอักขระที่เขียนด้วยลายมือ และการรู้จำเสียงพูด เป็นต้น การเรียนรู้เชิงลึกเป็นเทคนิคหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligence: AI) ซึ่งเป็นเทคนิคหนึ่งที่ใช้ในการเรียนรู้ของเครื่อง (Machine Learning: ML) ซึ่งการเรียนรู้เชิงลึกเป็นเทคนิคทางคณิตศาสตร์

สำหรับการจำแนกประเภทรูปแบบ ที่ขึ้นอยู่กับข้อมูลตัวอย่างโดยใช้เครือข่ายประสาทที่มีหลายๆชั้น [82] ซึ่งโครงข่ายประสาทในงานวิจัยที่เกี่ยวกับการเรียนรู้เชิงลึกโดยทั่วไป ประกอบด้วยชุดของหน่วยข้อมูลเข้า เช่น พิกเซล หรือคำ เป็นต้น ในชั้นข้อมูลนำเข้า (Input Layer) มีเลเยอร์ที่ซ่อนอยู่ (Hidden Layer) หลาย ๆ เลเยอร์ ที่มีหน่วยที่ซ่อนอยู่ (Hidden Units) (เรียกว่าโหนดหรือเซลล์ประสาท) และชุดเอาต์พุตในชั้นผลลัพธ์ (Output Layer) ที่มีการเชื่อมต่อระหว่างโหนดเหล่านั้น การเรียนรู้เชิงลึกสามารถช่วยในการแก้ปัญหาที่ต้องกำหนดคุณลักษณะ (Features) ด้วยตัวเองและลดระยะเวลาในการกำหนดคุณลักษณะที่ซับซ้อน โดยการเรียนรู้คุณลักษณะแบบอัตโนมัติและใช้คุณลักษณะเป็นตัวแทนของข้อมูลเข้า

3.3.1. โครงข่ายประสาทเทียมไปข้างหน้า (Feed-Forward Neural Networks หรือ Multilayer Perceptron: MLP)

โครงข่ายประสาทเทียมไปข้างหน้า (Feed-Forward Neural Network) [83] หรือที่รู้จักกันในชื่อ Multi-Layer Perceptron เป็นสถาปัตยกรรมการเรียนรู้เชิงลึกขั้นพื้นฐานและใช้กันอย่างแพร่หลายในงานการเรียนรู้ของเครื่อง ซึ่งโครงข่ายประสาทเทียมได้แรงบันดาลใจจากระบบประสาททางชีวภาพ โครงข่ายจะส่งต่อสัญญาณของเซลล์ประสาทเทียมที่เรียกว่าหน่วย (Units) ซึ่งหน่วยจะอยู่ในห่วงโซ่ของเลเยอร์เรียกว่าเลเยอร์ที่ซ่อนอยู่ (Hidden Layers) การออกแบบห่วงโซ่นี้เป็นโครงสร้างงานแบบเฉพาะในหลายระดับคล้ายกับกระบวนการทางสมอง ซึ่งโครงสร้างโครงข่ายประสาทเทียมไปข้างหน้าแสดงในรูปที่ 5 ซึ่งประกอบด้วยชั้นข้อมูลเข้า (Input Layer) ชุดของชั้นซ่อน (Hidden Layers) และชั้นส่งออก (Output Layer)



รูปที่ 5 แสดงโครงข่ายประสาทเทียมไปข้างหน้า (Feed-Forward Neural Network)

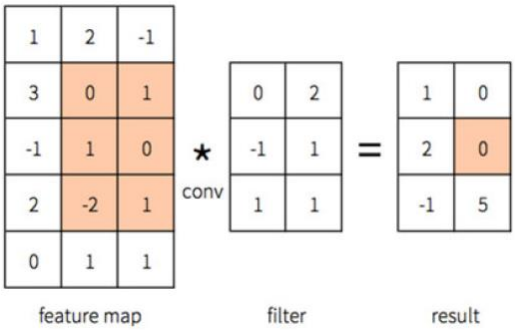
ชั้นข้อมูลเข้า (Input Layer) ได้รับคุณลักษณะข้อมูลเข้า (Input Feature) จากนั้นสัญญาณจะถูกแปลงและส่งต่อไปยังเลเยอร์ที่ซ่อนอยู่ (Hidden Layers) แต่ละเลเยอร์ในชั้นซ่อนจะมีจำนวนของหน่วยที่ซ่อนอยู่ (หรือโหนด) และฟังก์ชันการกระตุ้น สำหรับแต่ละหน่วย โดยหน่วยรับสัญญาณจากเลเยอร์ก่อนหน้าประมวลผลข้อมูลและส่งต่อสัญญาณที่แปลงแล้วไปยังหน่วยในเลเยอร์ถัดไป โดย

ปกติแล้วหน่วยระหว่างสองชั้นที่ต่อเนื่องจะเชื่อมต่ออย่างสมบูรณ์ สุดท้ายหลังจากชุดของเลเยอร์ที่ซ่อนอยู่เลเยอร์เอาท์พุต (Output Layer) จะแสดงผลลัพธ์ตามประเภทของคลาสเป้าหมาย

3.3.2. โครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network: CNN)

ในงานทางด้านคอมพิวเตอร์วิทัศน์มีการใช้งานสถาปัตยกรรม CNN กันอย่างแพร่หลาย ซึ่งโครงข่ายประสาทแบบคอนโวลูชัน (CNN) เป็นเฟรมเวิร์คเครือข่ายของการเรียนรู้เชิงลึกที่ใช้ในการสกัดคุณลักษณะจากวัตถุ ซึ่งประกอบด้วย ชั้นคอนโวลูชัน (Convolutional Layer), ชั้นพูลลิ่ง (Pooling Layer) และชั้นเชื่อมต่อสมบูรณ์ (Fully Connected Layers) ซึ่งในปัจจุบันได้มีการนำ CNN มาประยุกต์ใช้งานทางด้านการรู้จำเสียงเช่นกัน สถาปัตยกรรม CNN เป็นโครงข่ายขั้นสูงของเครือข่ายประสาทเทียมแบบมาตรฐาน [49] CNN เป็นโครงสร้างพิเศษที่ประกอบด้วยคู่ของ Convolution และ Pooling Layer แทนการใช้แค่ชั้นเชื่อมต่อสมบูรณ์ (Fully Connected Layers)

ชั้นคอนโวลูชัน (Convolution Layer) เป็นชั้นที่สกัดรูปแบบคุณลักษณะจากข้อมูลเข้า จุดประสงค์คือการสร้างแผนที่คุณลักษณะ (Feature Map) ด้วยตัวกรองคอนโวลูชัน (Convolutional Filters) และใช้ฟังก์ชันการกระตุ้นแบบไม่เป็นเชิงเส้น (Nonlinear Activation Function) เช่น tanh, sigmoid, ReLU เป็นต้น [52] ตัวอย่างการทำคอนโวลูชันแสดงในรูปที่ 6 กระบวนการคำนวณดังสมการที่ 2



รูปที่ 6 แสดงตัวอย่างของคอนโวลูชัน [52]

ซึ่งบนเลเยอร์นี้ เอาท์พุตสามารถคำนวณได้ดังนี้

$$h_l = f(W_l * h_{l-1} + b_l) \quad (2)$$

โดยที่ h_{l-1} และ h_l เป็น แผนที่คุณลักษณะ (Feature Map) ในสองเลเยอร์ติดกัน

W_l แทน ตัวกรองคอนโวลูชัน

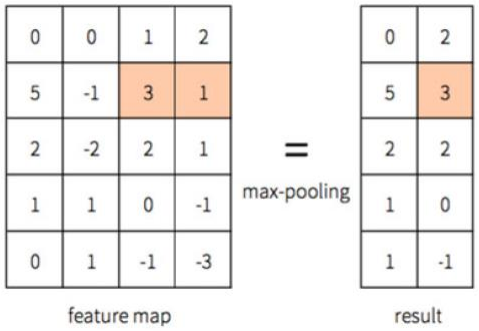
$*$ หมายถึง ตัวดำเนินการคอนโวลูชัน 2 มิติ

$f(.)$ หมายถึง ฟังก์ชันการกระตุ้นไม่เป็นเชิงเส้น

b_l หมายถึง bias

สมการที่ 2 แสดงให้เห็นสมการที่ง่ายที่สุดในเลเยอร์ก่อนหน้าที่มีเพียง 1 Feature Map เท่านั้น

ชั้นพูลลิ่ง (Pooling Layer) เป็นชั้นที่มีเป้าหมายในการลดความละเอียดของแผนที่คุณลักษณะ โครงสร้างของชั้นนี้จะตามหลังชั้นคอนโวลูชัน และฟังก์ชันการกระตุ้น การพูลลิ่งเป็นแนวคิดที่สำคัญในสถาปัตยกรรม CNN ซึ่งช่วยลดความแปรปรวนสเปกตรัมในคุณลักษณะของข้อมูลเข้า [51] ซึ่งตัวดำเนินการพูลลิ่ง (Pooling Operation) ที่นิยมใช้ คือ Max Pooling ซึ่งจะทำการเลือกองค์ประกอบที่สำคัญที่สุด ตัวอย่างการทำ 1x2 Max Pooling บน 1 Feature Map แสดงในรูปที่ 7



รูปที่ 7 แสดงตัวอย่างการทำ Max Pooling [52]

หลังจากชั้นคอนโวลูชัน (Convolution layer) และชั้นพูลลิ่ง (Pooling Layer) จะตามด้วยชั้นเชื่อมต่อ (Fully Connected Layers) ซึ่งจะทำหน้าที่รวมเอาท์พุตของเลเยอร์สุดท้ายสำหรับการจำแนกประเภท ในการจำแนกประเภทสามารถใช้ Sigmoid Function และที่นิยมใช้คือ Softmax Function ในการจำแนกคลาสแบบหลากหลาย (Multi-Class Classification)

3.3.3. โครงข่ายประสาทแบบหมุนเวียนกลับ (Recurrent Neural Network: RNN) และแอลเอสทีเอ็ม (Long Short Term Memory: LSTM)

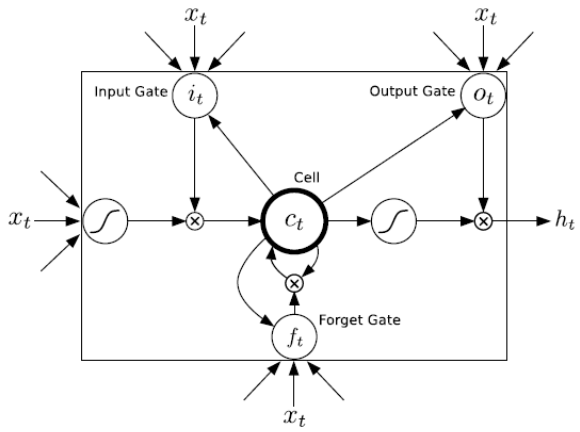
โครงข่ายประสาทแบบหมุนเวียนกลับ (RNN) [70] เป็นเครือข่ายจำลองของเซลล์ประสาทที่มีการเชื่อมต่อซึ่งถือว่าเป็นแบบจำลองของสมองมนุษย์ RNN สามารถเรียนรู้พฤติกรรมที่หลากหลายใช้กับข้อมูลที่เป็นลำดับหรือมีเรื่องของเวลามาเกี่ยวข้อง ในงานวิจัย [84] การกำหนดลำดับของข้อมูลเข้า (Input Sequence) X = (x_1; .....; x_T) RNN คำนวณลำดับของเวกเตอร์ที่ซ่อนอยู่ (Hidden Vector Sequence) h = (h_1; .....; h_T) และลำดับของเวกเตอร์เอาท์พุต (Output Vector Sequence) y = (y_1; .....; y_T) โดยมีการทำซ้ำสมการต่อไปนี้จาก t = 1 ถึง T

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (3)$$

$$y_t = W_{hy}h_t + b_y \quad (4)$$

สมการที่ 3 และ 4 แสดง W หมายถึงเมทริกซ์น้ำหนัก (เช่น W_{xh} เป็น เมทริกซ์ Input-hidden Weight) b หมายถึง เวกเตอร์ Bias (เช่น b_h คือ hidden bias vector) และ H หมายถึง ฟังก์ชัน hidden layer ซึ่งมักจะใช้ Sigmoid Function

แอลเอสทีเอ็ม (LSTM) เป็นชนิดพิเศษของ RNN ใช้ในการแก้ปัญหา Vanishing Gradient เมื่อข้อมูลเข้าข้อมูลลำดับเริ่มยาว ไม่สามารถเก็บข้อมูลก่อนหน้านั้นได้เป็นระยะเวลานาน โครงสร้าง LSTM คล้ายกับ RNN ต่างกันที่มีกลไกพิเศษในการควบคุมการไหลของข้อมูล มีความสามารถในการจดจำข้อมูลก่อนหน้าได้นานกว่า รูปที่ 8 แสดง Long Short-term Memory Cell จำนวน 1 เซลล์ [84]



รูปที่ 8 แสดง Long Short-term Memory Cell [84]

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (5)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (6)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (7)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (8)$$

$$h_t = o_t \tanh(_{ct}) \quad (9)$$

สมการที่ 5 – 9 แสดง σ คือ logistic sigmoid function และ i, f, o และ c หมายถึง input gate, forget gate, output gate และ cell

3.3.4. ฟังก์ชันการกระตุ้น (Activation Function)

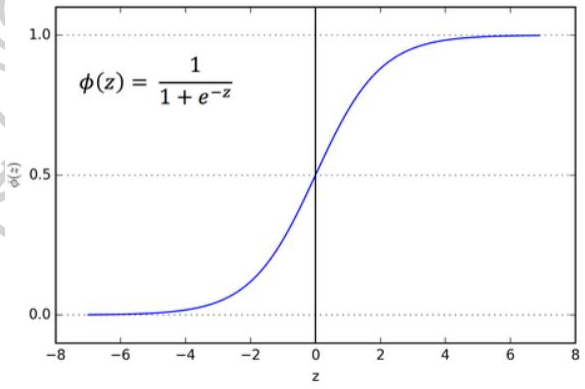
ฟังก์ชันการกระตุ้น (Activation Function) $\emptyset$ (▪) ที่ใช้กันทั่วไปในเครือข่ายประสาท เป็นฟังก์ชันการกระตุ้นบนหน่วยที่ซ่อนอยู่ซึ่งทำหน้าที่เหมือน “สวิตช์” สามารถเป็นได้ทั้ง “เปิด” (เปิดใช้งาน) หรือ“ปิด” (ปิดใช้งาน) ขึ้นอยู่กับสัญญาณข้อมูลเข้า มักจะเป็นฟังก์ชั่นที่ไม่ใช่เชิงเส้นอย่างต่อเนื่อง (Continuous NonLinear Function) เพื่อกระตุ้นการกระจายใน Hidden Layers ซึ่งสามารถทำการปรับพารามิเตอร์แบบจำลองได้ ยกตัวอย่างฟังก์ชันการกระตุ้น เช่น Sigmoid, Hyperbolic Tangent, Rectified Linear Unit, Maxout, Softmax เป็นต้น

Sigmoid Function

ฟังก์ชันการกระตุ้น Sigmoid เป็นตัวเลือกทั่วไปในโครงข่ายโครงข่ายประสาท มีสมการนิยามดังสมการที่ 10

$$\emptyset(z) = \frac{1}{1+e^{-z}} \tag{10}$$

ฟังก์ชั่นนี้มีลักษณะเส้นโค้งรูปทรงแบบ S Curve ค่า Output อยู่ระหว่าง (0–1) ดังนั้นการหาค่าความน่าจะเป็น (Probability) ของ Output โดยค่า Prob จะมีค่าตั้งแต่ 0 จนถึง 1 เมื่อ z เป็นบวกมาก $\emptyset(z)$ ใกล้กับ 1 และเมื่อ z เป็นลบมาก $\emptyset(z)$ ใกล้กับ 0 ซึ่งฟังก์ชั่น Sigmoid สามารถทำให้เครือข่ายประสาทติดขัดได้ในเวลาการฝึกอบรม รูปที่ 9 แสดงถึง Sigmoid Function



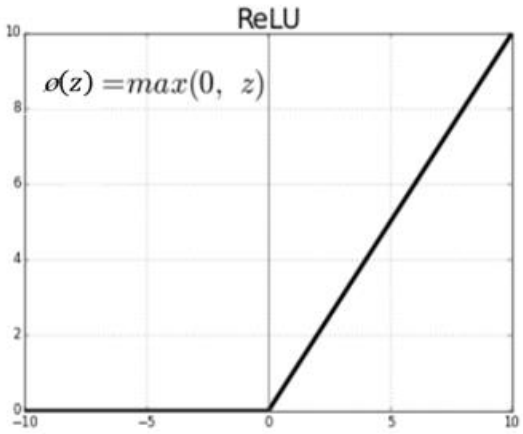
รูปที่ 9 แสดง Sigmoid Function [85]

ReLU (Rectified Linear Unit) Function

ReLU เป็นฟังก์ชันการกระตุ้นหนึ่งที่นิยมใช้ในขณะนี้ [86] เนื่องจากถูกใช้ในการเรียนรู้เชิงลึกเกือบทั้งหมด โดยมีสมการนิยามดังสมการที่ 11

$$\emptyset(z) = \max(0, z) \tag{11}$$

โดย $\phi(z)$ เป็นศูนย์เมื่อ z น้อยกว่าศูนย์และ $\phi(z)$ เท่ากับ z เมื่อ z สูงกว่าหรือเท่ากับศูนย์ ช่วงจึงอยู่ระหว่าง [0 ถึง อินฟินิตี้) ดังรูป 10



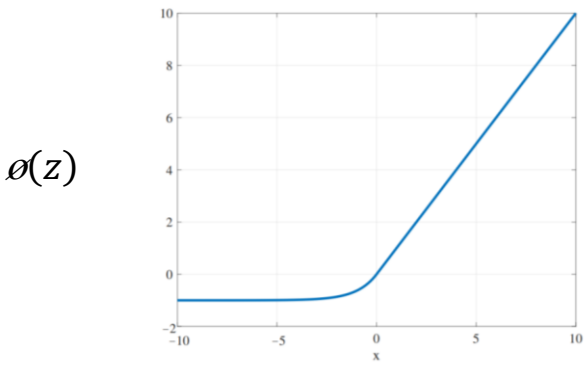
รูปที่ 10 แสดง ReLU (Rectified Linear Unit) Function

Exponential Linear Unit (ELU) Function

ELU แก้ปัญหา vanishing gradient [62, 86] ELU มีส่วนของค่าที่เป็นลบซึ่งเป็นประโยชน์สำหรับการเรียนรู้อย่างรวดเร็ว ELU ใช้ฟังก์ชันความอิ่มตัวเป็นส่วนลบ เนื่องจากฟังก์ชันความอิ่มตัวจะลดความแปรผันของยูนิตหากปิดใช้งาน จึงทำให้ ELU ทนทานต่อสัญญาณรบกวนมากขึ้น ให้ค่าลบที่ mean unit activations เข้าใกล้ 0 มากขึ้น ฟังก์ชัน ELU มีสมการนิยามดังสมการที่ 12

$$\phi(z) = \begin{cases} z & , z > 0 \\ \alpha(\exp^z - 1) & , z \leq 0 \end{cases} \tag{12}$$

โดยที่ α หมายถึงไฮเปอร์พารามิเตอร์ของ ELU ที่ควบคุมค่าที่ ELU อิ่มตัวสำหรับข้อมูลเข้าเชิงลบ สำหรับ ELU (Exponential Linear Unit) Function จะแสดงดังรูปที่ 11

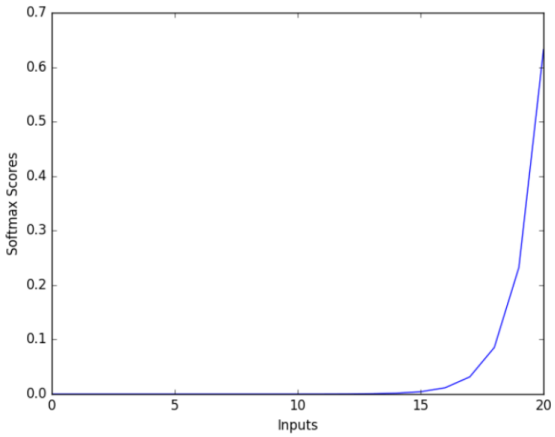


รูปที่ 11 แสดง ELU (Exponential Linear Unit) Function [87]

Softmax Function

ฟังก์ชั่น Softmax ดังรูปที่ 12 ส่วนมากใช้ในชั้นเลเยอร์เอาท์พุตของโครงข่ายประสาทเทียมในงานการจำแนกประเภทที่มีหลายคลาส โดย Output ที่ได้ออกมาเป็นค่าความน่าจะเป็น (Probability) นำไปคำนวน Negative Log Likelihood เป็น Cross Entropy Loss โดยมีสมการนิยามดังสมการที่ 13

$$\sigma(z_i) = \frac{e^{z}i}{\sum_{j=0}^{k} e^{z}j} \quad \text{เมื่อ } i = 0,1,2,\dots k \tag{13}$$



รูปที่ 12 แสดง Softmax Function [88]

3.3.5. Padding และ Stride

Padding สามารถเพิ่มความสูง (Height) และความกว้าง (Width) ของเอาต์พุตได้ ใช้เพื่อให้เอาต์พุตมีความสูงและความกว้างเท่ากันกับข้อมูลเข้า สำหรับ Stride สามารถลดความละเอียดของเอาต์พุตได้ เช่น การลดความสูงและความกว้างของเอาต์พุตให้เหลือเพียง $\frac{1}{n}$ ของความสูงและความกว้างของข้อมูลเข้า (n เป็นจำนวนเต็มที่มากกว่า 1) ซึ่ง Padding และ Stride สามารถใช้เพื่อปรับขนาดของข้อมูลได้อย่างมีประสิทธิภาพ

เมื่อข้อมูลเข้า (Input) มีความสูง (Height) และความกว้าง (Width) เท่ากับ 3 และ Convolution Kernel หรือ Filter มีความสูง (Height) และความกว้าง (Width) เท่ากับ 2 ทำให้ได้เอาต์พุต (Output) ที่มีความสูง (Height) และความกว้าง (Width) เท่ากับ 2 โดยทั่วไปสมมติว่ารูปร่างข้อมูลเข้า (Input Shape) เป็น $n_h \times n_w$ และรูปร่างหน้าต่างเคอร์เนล Convolution คือ $k_h \times k_w$ ดังนั้นนั้นรูปร่างเอาต์พุต (Output Shape) จะเป็นดังสมการ 14

$$(n_h - k_h + 1)x(n_w - k_w + 1) \tag{14}$$

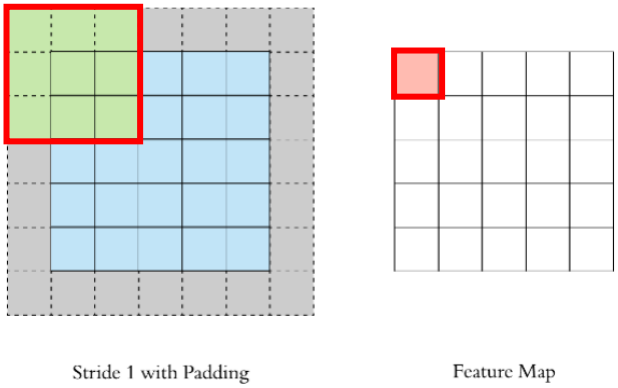
ดังนั้นรูปร่างเอาต์พุตของ Convolutional Layer จะถูกกำหนดโดยรูปร่างของข้อมูลเข้าและรูปร่างของหน้าต่างเคอร์เนล Convolution ในหลายกรณีอาจต้องการรวมเทคนิคเฉพาะเช่น Padding และ Strides เนื่องจากการเปลี่ยนไปขนาดของผลลัพธ์ โดยทั่วไป Kernels มีความกว้างและความสูงมากกว่า 1 นั่นหมายความว่าหลังจากใช้อย่างต่อเนื่องหลายครั้งจะทำให้เอาต์พุตในตอนท้ายมีขนาดเล็กกว่าข้อมูลเข้าอย่างมาก หากเริ่มต้นด้วยข้อมูลเข้าขนาด 240 x 240, 10 Layers มี 5 x 5 Convolutions จะลดขนาดเป็น 200 x 200 โดยแบ่งส่วนข้อมูลเข้า 30% และกำจัดข้อมูลที่น่าสนใจเกี่ยวกับขอบเขตของต้นฉบับไป ซึ่ง Padding สามารถจัดการกับปัญหานี้ได้ และในบางกรณีอาจต้องการลดความละเอียดลงอย่างมาก ถ้าหากพบว่าความละเอียดข้อมูลเข้าดั้งเดิมไม่เป็นไปตามที่กำหนด ซึ่งการใช้ Strides สามารถช่วยในกรณีนี้ได้

Padding

จากปัญหาเมื่อใช้ Convolutional Layers คือการสูญเสียขอบเขตของข้อมูลเข้าที่เป็นต้นฉบับ เนื่องจากโดยทั่วไปจะใช้ Kernels เล็ก ๆ สำหรับ Convolution อาจสูญเสียขอบเขตของข้อมูลเข้าเพียงเล็กน้อย แต่สิ่งนี้อาจเพิ่มขึ้นเมื่อใช้ Convolutional Layers ต่อเนื่องหลายครั้ง ทางออกสำหรับปัญหานี้คือการเพิ่มพิกเซลพิเศษรอบขอบเขตของข้อมูลเข้า ซึ่งจะเป็นการเพิ่มขนาดของข้อมูลเข้าที่มีประสิทธิภาพโดยทั่วไปแล้วจะตั้งค่าของพิกเซลพิเศษเป็น 0 โดยทั่วไปถ้าเพิ่มจำนวนแถว p_h แถวของ Padding และคอลัมน์ p_w คอลัมน์ ทั้งหมดของ Padding รูปร่างเอาต์พุต (Output Shape) จะเป็นดังสมการ 15

$$(n_h - k_h + p_h + 1)x(n_w - k_w + p_h + 1) \tag{15}$$

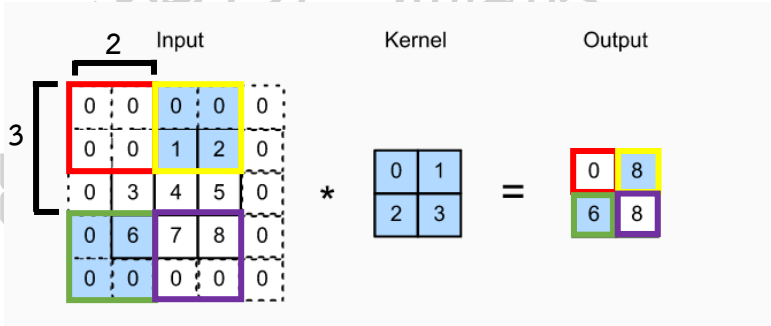
ซึ่งหมายความว่าความสูงและความกว้างของเอาต์พุตจะเพิ่มขึ้นตาม p_h และ p_w ตามลำดับ ในหลายกรณีการตั้งค่า $p_h = k_h - 1$ และ $p_w = k_w - 1$ เพื่อให้ข้อมูลเข้าและเอาต์พุตมีความสูงและความกว้างเท่ากัน การ Padding เพื่อให้ข้อมูลเข้าและเอาต์พุตยังคงมีความสูงและความกว้างเท่ากัน แสดงดังรูปที่ 13



รูปที่ 13 แสดงการ Padding เพื่อให้ข้อมูลเข้าและเอาต์พุตยังคงมีความสูงและความกว้างเท่ากัน [89]

Stride

เมื่อทำการคำนวณ Cross-Correlation เริ่มต้นด้วยหน้าต่าง Convolution ที่มุมซ้ายบนของ Input Array จากนั้นเลื่อนไปยังตำแหน่งทั้งหมดทั้งด้านล่างและด้านขวา ซึ่งในบางครั้งเพื่อประสิทธิภาพในการคำนวณหรือเนื่องจากต้องการลดขนาดลง จะทำการย้ายหน้าต่างมากกว่าหนึ่งพิกเซลในแต่ละครั้งเพื่อข้ามพื้นที่ตรงกลาง ซึ่งคือการอ้างถึงจำนวนแถวและคอลัมน์ที่เคลื่อนที่ผ่าน Stride หนึ่งๆ รูปที่ 14 แสดงการดำเนินการข้ามสหสัมพันธ์แบบสองมิติโดยมี Stride เท่ากับ 3 ในระดับแนวตั้ง (Height) และ 2 ในแนวนอน (Width) ส่วนที่แรเงาสีฟ้าเป็นส่วนของเอาต์พุต ส่วนของอินพุตและเคอร์เนลที่ใช้สำหรับการคำนวณเอาต์พุต โดยกรอบสีแดงเป็นกรณีที่เกิดจากการคำนวณผลลัพธ์จะได้ 0x0+0x1+0x2+0x3=0 สำหรับกรอบสีเหลืองเกิดจากการ Stride ในแนวนอนที่มีค่าเป็น 2 (หน้าต่างเลื่อนไป 2 คอลัมน์) ผลลัพธ์จะได้ 0x0+0x1+1x2+2x3=8 ในกรอบสีเขียวเกิดจากการ Stride ในแนวตั้งที่มีค่าเป็น 3 (หน้าต่างเลื่อนลงมา 3 แถว) ผลลัพธ์จะได้ 0x0+6x1+0x2+0x3=6 และกรอบสีม่วงเกิดจากการ Stride ในแนวนอนที่มีค่าเป็น 2 ผลลัพธ์จะได้ 7x0+8x1+0x2+0x3=8 ซึ่งจะเห็นได้ว่าการ Stride จะข้ามไปทีละแนวนอน 2 ช่องและแนวตั้ง 3 ช่อง

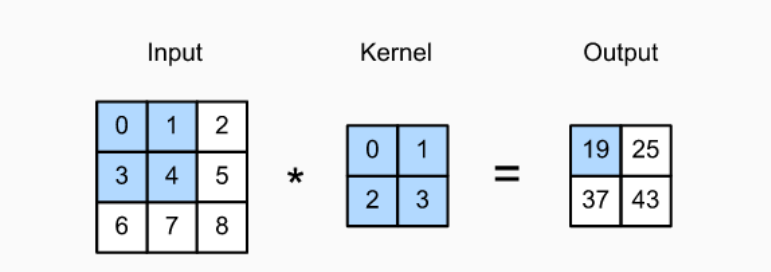


รูปที่ 14 แสดงการ Stride ของความสูงเท่ากับ 3 และความกว้างเท่ากับ 2 [90]

3.3.6. Feature Map

ผังคุณลักษณะ (Feature Map) เป็นเอาต์พุต (Output) ที่ได้หลังจากการดำเนินการ Convolution ใน Convolution Layer โดยเมื่อข้อมูลเข้าผ่านการทำ Convolution ด้วยตัวกรอง (Filter หรือ Kernal) ที่แตกต่างกัน จะได้ผลลัพธ์ที่แตกต่างกันออกไป เช่น หาขอบวัตถุ, ความเบลอ, ความคม เป็นต้น โดยใน Layer แรกๆ จะมีลักษณะเป็น เส้นตรง เส้นโค้ง จนไปถึง Layer หลัง ๆ จะ

มีลักษณะเป็นนามธรรม (Abstract) ขึ้นไปเรื่อย ๆ ซึ่งขนาดของผังคุณลักษณะมีความสัมพันธ์กับการทำ Convolution, Padding หรือ Stride เป็นต้น



รูปที่ 15 แสดงผังคุณลักษณะ (Feature Map) [90]

จากรูปที่ 15 เมื่อใช้ Input ขนาด 3 x 3 ใช้ Filter หรือ Kernal 2 x 2 ไม่มีการ Padding และ Stride เท่ากับ 1 จะได้ผังคุณลักษณะ (Feature Map) ขนาด 2 x 2 โดยส่วนที่มีสีเป็นองค์ประกอบเอาต์พุตแรกและองค์ประกอบข้อมูลเข้าและเคอร์เนลอาเรย์ที่ใช้ในการคำนวณ (0 × 0 + 1 × 1 + 1 × 1 + 3 × 2 + 4 × 3 = 19)

3.3.7. Batch Size

ขนาดแบทช์ (Batch Size) เป็น Hyperparameter ที่กำหนดจำนวนตัวอย่างเพื่อทำงานก่อนที่จะอัพเดตพารามิเตอร์โมเดลภายใน แนวคิดของ Batch เป็นการวนซ้ำในรอบการทำซ้ำสำหรับตัวอย่างหนึ่งตัวอย่างขึ้นไปและทำการคาดการณ์ ในตอนท้ายของแบทช์ การคาดการณ์จะถูกเปรียบเทียบกับตัวแปรเอาต์พุตที่คาดหวังและข้อผิดพลาดจะถูกคำนวณ จากข้อผิดพลาดนี้อัลกอริทึมการอัปเดตจะใช้ในการปรับปรุงโมเดล เช่น ปรับลงตาม Error Gradient ชุดข้อมูลการฝึกอบรมสามารถแบ่งออกเป็นชุดตั้งแต่ 1 Batch ขึ้นไป เมื่อตัวอย่างการฝึกอบรมทั้งหมดถูกใช้เพื่อสร้าง 1 Batch ขั้นตอนวิธีการเรียนรู้จะเรียกว่า Batch Gradient Descent เมื่อ Batch คือขนาดของ 1 ตัวอย่าง อัลกอริทึมการเรียนรู้จะเรียกว่า Stochastic Gradient Descent เมื่อขนาดแบทช์ (Batch Size) มากกว่า 1 ตัวอย่างและน้อยกว่าขนาดของชุดข้อมูลการฝึกอบรม อัลกอริทึมการเรียนรู้จะถูกเรียกว่า Mini-Batch Gradient Descent

Batch Gradient Descent คือ Batch Size = ขนาดของชุดการฝึกอบรม

Stochastic Gradient Descent คือ Batch Size = 1

Mini-Batch Gradient Descent คือ 1 < Batch Size < ขนาดของชุดฝึกอบรม

ในกรณีของ Mini-Batch Gradient Descent นิยมกำหนด Batch Size ได้แก่ 32 ตัวอย่าง, 64 ตัวอย่าง และ 128 ตัวอย่าง

3.3.8. Epoch

จำนวน Epochs คือ Hyperparameter ที่กำหนดจำนวนครั้งที่อัลกอริทึมการเรียนรู้ที่จะทำงานผ่านชุดข้อมูลการฝึกอบรมทั้งหมด โดย 1 Epoch หมายถึง แต่ละตัวอย่างในชุดข้อมูลการฝึกอบรมมีโอกาสอัพเดตพารามิเตอร์โมเดลภายใน ซึ่ง Epoch ประกอบด้วย 1 Batch ขึ้นไป ตัวอย่างเช่น ตามที่ได้กล่าวมาเบื้องต้น Epoch ที่มี 1 Batch เรียกว่า Batch Gradient Descent Learning Algorithm จำนวนของ Epoch มีขนาดใหญ่ ซึ่งมักจะเป็นร้อยหรือเป็นพัน ทำให้อัลกอริทึมการเรียนรู้สามารถทำงานได้จนกระทั่งข้อผิดพลาดจากแบบจำลองจะลดลงอย่างพอเพียง ซึ่งมีการกำหนดตัวอย่างของ Epoch เป็น 10, 100, 500, 1,000 และใหญ่กว่า ซึ่งส่วนใหญ่จะสร้างพล็อตกราฟแสดงผลที่แสดง Epoch ตามแกน x เป็นจำนวนครั้ง และแสดงข้อผิดพลาดหรือความถูกต้องของแบบจำลองบนแกน y ซึ่งการพล็อตกราฟแบบนี้บางครั้งเรียกว่า Learning Curves

3.3.9. Optimizer

Optimizer ทำหน้าที่ในการปรับค่าน้ำหนัก (Weight) ที่เหมาะสมที่สุด ซึ่งใช้ในโครงข่ายประสาทเทียมช่วยในเรื่องการปรับปรุงค่า Error และค่า Loss ซึ่งมีความเกี่ยวข้องกับฟังก์ชั่นการสูญเสีย (Loss Function) และพารามิเตอร์ของโมเดลโดยการอัพเดตโมเดลเพื่อตอบสนองเอาต์พุตของฟังก์ชันการสูญเสีย ซึ่ง Optimizer จะสร้างแบบจำลองให้เป็นรูปแบบที่ถูกต้องที่สุดโดยการปรับค่าน้ำหนัก (Weight)

Gradient descent

Gradient descent เป็นขั้นตอนวิธีการเรียนรู้การเพิ่มประสิทธิภาพเพื่อลดฟังก์ชันต้นทุน (Cost Function) หรือ ฟังก์ชั่นการสูญเสีย (Loss Function) ซึ่งจะช่วยให้แบบจำลองมีการคาดการณ์ที่แม่นยำ โดยเมื่อ Gradient แสดงทิศทางการเพิ่มขึ้น เมื่อต้องการค้นหาจุดต่ำสุดจำเป็นต้องไปในทิศทางตรงกันข้ามของ Gradient จะอัปเดตพารามิเตอร์ใน Gradient ทิศที่เป็นลบเพื่อลดการสูญเสีย

$$\theta = \theta - \eta \nabla J(\theta; x, y) \tag{16}$$

สมการที่ 16 θ คือ weight parameter, η คือ learning rate และ $\nabla J(\theta; x, y)$ คือ gradient ของ weight parameter θ

Stochastic Gradient Descent

Stochastic Gradient Descent มีลักษณะการใช้งานที่อาจใช้ตัวอย่างเป็น Batch ในแต่ละครั้งหรือเป็นแบบสุ่มในแต่ละรอบ ซึ่งเมื่อมีข้อมูลตัวอย่างในการฝึกอบรมที่มีขนาดใหญ่ ในการคำนวณ

Gradient แต่ละครั้งนั้นต้องใช้ตัวอย่างทั้งหมดเพื่อมาอัพเดทพารามิเตอร์ใหม่ ซึ่งใช้เวลาในการฝึกอบรมมาก สำหรับ Stochastic Gradient Descent ในแต่ละการคำนวณ Gradient จะทำการสุ่มตัวอย่างเพียงบางส่วนเพื่อใช้อัพเดทเท่านั้น ซึ่งสามารถใช้ตัวอย่างเพียงไม่มากเพื่อใช้ในการอัพเดทพารามิเตอร์ในแต่ละครั้ง แต่สามารถลู่เข้าสู่คำตอบใกล้เคียงกัน และยิ่งในกรณีที่มีพารามิเตอร์จำนวนมาก การใช้ Stochastic Gradient Descent สามารถลดปัญหา Optimization ติดอยู่ใน local ได้

Adam

Adam ย่อมาจาก Adaptive Moment Estimation และเป็นอีกวิธีหนึ่งในการใช้ Gradient ที่ผ่านมาในการคำนวณ Gradient ในปัจจุบัน Adam ใช้แนวคิดเรื่องโมเมนตัม (Momentum) ด้วยการเพิ่มเศษส่วนของ Gradient ที่ผ่านมาเข้ากับ Gradient ในปัจจุบัน Optimizer นี้ได้รับความนิยมแพร่หลายและเป็นที่ยอมรับในทางปฏิบัติสำหรับใช้ในการฝึกอบรมโครงข่ายประสาท เนื่องจากรวมจุดเด่นของ Optimizer และแก้ไขจุดด้อยต่างๆ เช่น การ Decaying Learning Rate ที่ช่วยให้โมเดลไม่หยุดเรียนได้ และลู่เข้าเร็วกว่า Gradient Descent

3.4. วิธี Gradient-weighted Class Activation Mapping (Grad-CAM)

Gradient-weighted class activation mapping (Grad-CAM) ใช้เพื่อทำให้โมเดลที่ใช้ CNN มีความโปร่งใสในการทำนายมากขึ้นโดยการสร้างคำอธิบายด้วยภาพ [76] สามารถใช้เพื่อทำความเข้าใจถึงความสำคัญของข้อมูลที่ป้อนเข้าเกี่ยวกับคลาสเป้าหมายที่สนใจ เพื่อให้ได้ class-discriminative localization map Grad-CAM สำหรับการไล่ระดับสีของคะแนนสำหรับคลาส c จะถูกคำนวณครั้งแรก (y^c) ในส่วนที่เกี่ยวกับแผนที่คุณลักษณะ (feature maps) A^k ของเลเยอร์ convolutional การไล่ระดับสีเหล่านี้ไหลย้อนกลับเป็น global-average-pooled เพื่อให้ได้น้ำหนักความสำคัญของนิวรอน ดังสมการที่ 17

$$\alpha_k^c = \frac{1}{z}\sum_{i \in w}\sum_{j \in h}\frac{\partial y^c}{\partial A_{ij}^k} \tag{17}$$

โดยที่ Z แทนจำนวนพิกเซลในแผนที่คุณลักษณะ (feature map)

Grad-CAM สามารถแสดงได้ดังสมการที่ 18

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{18}$$

เอาต์พุต $L_{Grad-CAM}^c$ ระบุว่าส่วนใดของเน็ตเวิร์กที่นำเสนอนั้นมีความน่าสนใจมากกว่า และแสดงความน่าสนใจของส่วนเหล่านั้นเป็น heat maps สำหรับแต่ละอารมณ์ ใช้ สมการที่ 19 เพื่อคำนวณ

heat maps เฉลี่ยของกลุ่มตัวอย่างทั้งหมด เพื่อทำความเข้าใจอะไรเป็นความแตกต่างเมื่อจำแนกคลาสต่างๆ

$$L_{AVE} = \frac{1}{N}\sum L^{c}_{Grad-CAM} \quad (19)$$

3.5. การประเมินผล

วิธีการวัดประสิทธิภาพในงานวิจัยนี้ใช้วิธีวัดประสิทธิภาพโดยใช้ค่าความถูกต้อง (Accuracy) ดังสมการที่ 20, ค่าความแม่นยำ (Precision) ดังสมการที่ 21, ค่าความระลึก (Recall) ดังสมการที่ 22 และการวัดประสิทธิภาพโดยรวม (F-measure) ดังสมการที่ 23

$$\text{Accuracy} = \frac{\textit{จำนวนผลลัพท์สระที่ทายถูก * 100}}{\text{จำนวนการทายผลลัพท์ทั้งหมด}} \quad (20)$$

$$\text{Precision} = \frac{tp}{tp+fp} \quad (21)$$

$$\text{Recall} = \frac{tp}{tp+fn} \quad (22)$$

$$\text{F-measure} = \frac{2(\text{Precision x Recall})}{\text{Precision + Recall}} \quad (23)$$

โดยที่ tp คือ ข้อมูลที่ทำนายแล้วถูกต้องเมื่อเทียบกับเฉลย

fp คือ ข้อมูลที่อยู่ในเฉลยแต่ไม่มีในการทำนาย

fn คือ ข้อมูลที่ทำนายแล้วไม่ถูกต้องเมื่อเทียบกับเฉลย

การประเมินความพึงพอใจของผู้ใช้ระบบ โดยสร้างแบบสอบถามประเมินความพึงพอใจผู้ใช้ระบบการรู้จำสระภาษาไทยผ่าน Web Application

ในงานวิจัยนี้ใช้การวิเคราะห์เชิงปริมาณ เพื่อวัดความพึงพอใจใช้ค่าความถี่ (Frequency) ค่าร้อยละ (Percentage) ดังสมการที่ 24, ค่าเฉลี่ย (Mean) ดังสมการที่ 25, และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ดังสมการที่ 26

1) ค่าร้อยละ (Percentage) คำนวณจากสูตร

$$P = \frac{f}{N} \times 100 \quad (24)$$

เมื่อ p แทน ค่าร้อยละ

f แทน ความถี่ที่ต้องการแปลงให้เป็นร้อยละ

N แทน จำนวนความถี่ทั้งหมด

2) ค่าเฉลี่ย (Mean) คำนวณจากสูตร

$$\overline{X} = \frac{\sum X}{N} \quad (25)$$

เมื่อ $\overline{X}$ แทน ค่าเฉลี่ย

$\sum X$ แทน ผลรวมของคะแนนในกลุ่ม

n แทน จำนวนผู้ใช้ในกลุ่มตัวอย่างที่ตอบแบบสอบถาม

3) ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation: S.D.)

$$S.D. = \sqrt{\frac{n\sum fx^2 - (\sum fx^2)}{n\,(n-1)}} \quad (26)$$

บทที่ 4

วิธีดำเนินงานวิจัยและผลการทดลองที่ 1

การรู้จำเสียงสระภาษาไทยโดยใช้ Convolutional Neural Network กับ Mel Frequency Cepstrum Coefficient

งานวิจัยนี้ออกแบบการทดลองเพื่อกำหนดค่าความเหมาะสมของพารามิเตอร์ในสถาปัตยกรรม Convolutional Neural Network (CNN) สำหรับการรู้จำการออกเสียงสระภาษาไทยที่มีเสียงรบกวน โดยได้ทำการเปรียบเทียบวิธีการโดยใช้กลยุทธ์ต่าง ๆ เช่น Padding [52] ซึ่งผลการทดลองแสดงให้เห็นว่า การขยายในแผนที่คุณสมบัติ Feature Maps สำหรับ CNN นั้นสำคัญ ซึ่งสามารถช่วยในเรื่องขนาดของคุณสมบัติแผนที่ไม่ให้มีการเปลี่ยนแปลงและช่วยให้ประสิทธิภาพของผลลัพธ์ดีขึ้น กลยุทธ์ Dropout สามารถช่วยลดปัญหา Over-fitting ได้ ซึ่ง Dropout, ReLU และ DNN ถูกนำไปใช้กับงานข่าวการออกอากาศภาษาอังกฤษ 50 ชั่วโมง (50-hour English Broadcast New) ได้ผลลัพธ์ที่ดีกว่า DNN ที่มี Sigmoid 4.2% และระบบ GMM / HMM 14.4% [91] ในงานวิจัยนี้ได้ทำการทดลองโดยเพิ่มจำนวนของ Convolution Layers และ Hidden Units เพื่อประเมินประสิทธิภาพของโมเดล สุดท้ายเปรียบเทียบ โมเดล CNN กับ โมเดล Multilayer Perceptron (MLP) และโมเดล Support Vector Machines (SVM)

4.1. ชุดข้อมูลและวิธีการ (Datasets and Methods)

การอธิบายชุดข้อมูลและวิธีการของเรียนรู้เชิงลึกในการจดจำเสียงสระภาษาไทย รายละเอียดมีดังนี้ ส่วนที่ 4.1.1. อธิบายชุดข้อมูล ในส่วนที่ 4.1.2. แสดงการแปลงสเปกโตรแกรม ส่วนที่ 4.1.3. แสดงการจำแนกประเภทของโมเดล convolutional neural networks และส่วนที่ 4.1.4. แสดงรายละเอียดการทดลอง (Implementation details)

4.1.1. ชุดข้อมูล (Dataset)

ชุดข้อมูลสระภาษาไทยแบบสาธารณะไม่เหมาะสมสำหรับการใช้งานตามวัตถุประสงค์ของการศึกษา ดังนั้นในงานวิจัยนี้ได้ทำการเก็บรวบรวมชุดข้อมูลเสียงสระภาษาไทยที่มีเสียงรบกวนที่เกิดขึ้นในสถานการณ์จริง โดยทำการรวบรวมในสภาพแวดล้อมจากหลายพื้นที่ เสียงรบกวนที่วัดได้ระหว่างการเก็บข้อมูลโดยนักภาษาศาสตร์ ซึ่งใช้โปรแกรมวัดระดับเสียง Sound Meter มีระดับเดซิเบลประมาณ 30 - 50 dB ซึ่งสามารถจัดประเภทสภาพแวดล้อมของเสียงรบกวนได้ดังนี้

- ใต้อาคารเรียน ประมาณ 30 dB
- เสียงในห้องสมุด, เสียงสัตว์ในสวนบริเวณบ้าน (เสียงสุนัขและนก) ประมาณ 40 dB
- บ้านและเครื่องใช้ภายในบ้าน ประมาณ 45 dB
- เสียงผู้คนกำลังพูดกันในโรงอาหาร, เสียงดนตรีที่วิทยาลัยดนตรี, เสียงรบกวนที่เกิดจากรถยนต์บนท้องถนน ประมาณ 50 dB

ข้อมูลเสียงพูด 44,100 Hz ถูกบันทึกจากโทรศัพท์มือถือ เสียงพูดของเพศชายและเพศหญิงจะถูกบันทึกแยกออกจากกัน ตามหลักการของ การศึกษาภาษาศาสตร์ เนื่องจากระดับเสียงของชายและหญิงมีความแตกต่างกัน โดยที่เสียงผู้ชายมีระดับต่ำ แต่เสียงของผู้หญิงมีระดับเสียงที่สูงมาก เสียงสระแบ่งออกเป็น 2 กลุ่มคือ เสียงของผู้ชายและเสียงของผู้หญิง ซึ่งได้จากเพศชาย 25 คน และเพศหญิง 25 คน โดยทั้งหมดนั้นเป็นผู้พูดภาษาไทยแบบมาตรฐาน ซึ่งมีอายุ 20-25 ปี ในการวิจัยครั้งนี้มีเอาท์พุตทั้งหมด 18 คลาส (เสียงสระเสียงสั้น 9 เสียง และเสียงสระเสียงยาว 9 เสียง) ผู้พูดแต่ละคนพูดรายการคำคนละ 2 ครั้งในแต่ละสระ ซึ่งเสียงที่รวบรวมได้ทั้งหมดมี 1,800 ไฟล์เสียง ประกอบด้วยไฟล์เสียงของเพศชาย 900 ไฟล์ (18 สระ x ชาย 25 คน x พูด 2 ครั้ง) และไฟล์เสียงของเพศหญิง 900 ไฟล์ (18 สระ x หญิง 25 คน x พูด 2 ครั้ง) โดย 80% ของไฟล์ทั้งหมดในแต่ละกลุ่มใช้สำหรับการฝึกอบรม และ 20% สำหรับการทดสอบ หลังจากการรวบรวมชุดคำพูดเสียงพูดสระภาษาไทยที่มีเสียงรบกวน นักภาษาศาสตร์ได้ทำการตัดและเลือกเสียงที่เป็นตัวแทนของข้อมูลเข้าไฟล์เสียงในแต่ละสระ โดยใช้เครื่องมือทางภาษาที่มีชื่อว่า PRAAT ซึ่งเป็นโปรแกรมคอมพิวเตอร์สำหรับวิเคราะห์สังเคราะห์และจัดการคำพูดที่พัฒนาโดย Paul Boersma และ David Weenink [32] เป็นเครื่องมือสำหรับการวิจัยและการสอนสำหรับสัทศาสตร์ที่พบบ่อย ซึ่งเป็นหนึ่งในเครื่องมือที่ครอบคลุมทางด้านการวิเคราะห์ และการเป็นตัวแทนของคำพูดถูกแสดงผลในรูปแบบทางกราฟฟิก สระภาษาไทยแสดงในรูป INTERNATIONAL PHONETIC ALPHABET (IPA) ถูกแสดงในตารางที่ 1

ตารางที่ 1 แสดงสระภาษาไทยอย่างง่ายใน INTERNATIONAL PHONETIC ALPHABET (IPA)

สระภาษาไทย			
สระเสียงสั้น		สระเสียงยาว	
Thai letter	Phonetic	Thai letter	Phonetic
อะ	/a/	อา	/a:/
อิ	/i/	อี	/i:/
อึ	/ɯ/	อือ	/ɯ:/

สระภาษาไทย			
สระเสียงสั้น		สระเสียงยาว	
Thai letter	Phonetic	Thai letter	Phonetic
อุ	/u/	อู	/u:/
เอะ	/e/	เอ	/e:/
แอะ	/ɛ/	แอ	/ɛ:/
โอะ	/o/	โอ	/o:/
เอาะ	/ɔ/	ออ	/ɔ:/
เออะ	/ɤ/	เออ	/ɤ:/

4.1.2. การแปลงสเปกโตรแกรม (Spectrogram conversion)

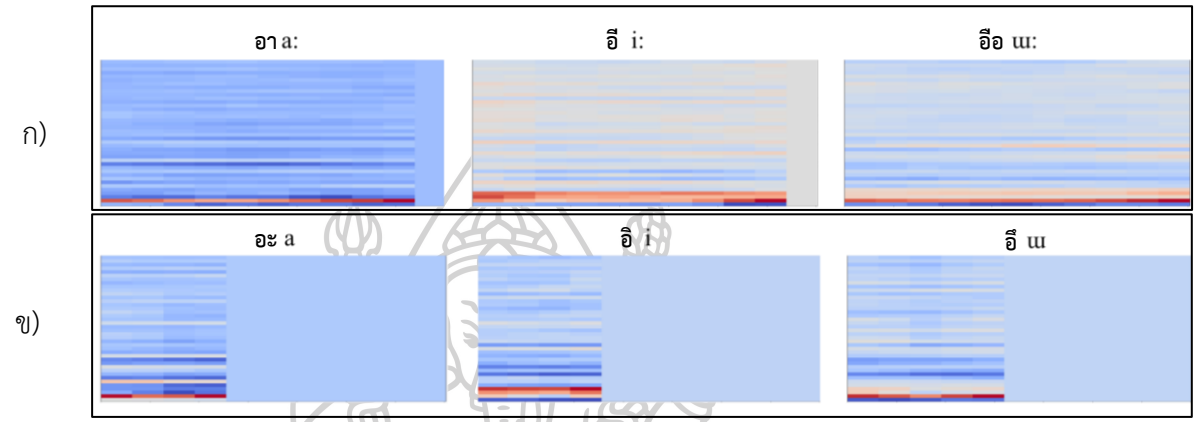
สัญญาณเสียงดิบถูกแปลงเป็น waveform แล้วแปลงเป็นสเปกโตรแกรมขนาดต่างๆ เพื่อค้นหาข้อมูลเข้าคุณสมบัติเสียงที่เหมาะสม สเปกโตรแกรมของภาพ 2 มิติประกอบด้วยแกนเวลา และแกนความถี่ ซึ่งจะถูกแสดงด้วยลำดับของสเปกตรัม สำหรับการวิเคราะห์เสียง การแสดงสเปกตรัมจะเก็บข้อมูลมากกว่าคุณลักษณะที่สร้างขึ้นด้วยมือแบบดั้งเดิม สเปกโตรแกรมมีขนาดมิติต่ำกว่าเสียงดิบ [72] การประมวลผลข้อมูลข้อมูลเข้าล่วงหน้าอย่างเหมาะสมเป็นกุญแจสำคัญอย่างหนึ่งสำหรับการเป็นตัวแทนของคุณลักษณะที่ดี สัญญาณเสียงสระไทยถูกประมวลผลล่วงหน้าโดยไลบรารี LibROSA [92] ในภาษาโปรแกรมไพทอน ซึ่งเป็นไลบารีสำหรับการวิเคราะห์เสียงและเสียงดนตรี ในขั้นตอน preprocessing สัญญาณเสียงพูดแบบโมโนโฟนิกจะลดอัตราการสุ่มตัวอย่างจาก 44,100 Hz เป็น 16,000 Hz วิธีการวิเคราะห์เสียงใช้เฟรมขนาดเล็กของสัญญาณที่เว้นระยะด้วยความยาวฮอป (hop length) ความยาวของหน้าต่างคือ 2048 ตัวอย่าง (ประมาณ 128 มิลลิวินาที) และ 512 ตัวอย่างสำหรับขนาดฮอป (ประมาณ 32 มิลลิวินาที) สัญญาณเสียงพูดจะถูกแปลงเป็นการแสดงความถี่และเวลาโดยอิงจากการแปลงฟูเรียร์ในระยะเวลาสั้น (Short-time Fourier Transform : STFT)

ในกระบวนการสกัดคุณลักษณะ ขนาดกำลังสอง (สเปกตรัมกำลัง) ของ STFT คือ linear-scaled spectrogram สเกลความถี่เชิงเส้น (linear frequency scale) ของสเปกโตรแกรมจะถูกปรับขนาดเป็น Mel scale โดยใช้ overlapping triangular filters มาตราส่วน Mel ให้มาตราส่วนเชิงเส้นสำหรับระบบการได้ยินของมนุษย์ [93] กำหนดไว้ดังสมการที่ 27

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (27)$$

โดยที่ m หมายถึง Mels และ f หมายถึง ความถี่ หน่วยเป็น เฮิรตซ์

สำหรับ MFCC ลอการิทึมของ MS จะถูกแปลงโดยใช้การแปลงโคไซน์แบบไม่ต่อเนื่อง (Discrete Cosine Transform : DCT) ผลลัพธ์ของการแปลงเรียกว่าค่าสัมประสิทธิ์ Mel Frequency Cepstrum Coefficient (MFCC) รูปที่ 16 แสดงตัวอย่างเสียงสระภาษาไทยที่แปลงเป็นคุณสมบัติเสียง MFCC ของสระเสียงยาว (ด้านบน) ได้แก่สระ อา /a:/, อี /i:/, อือ /ɯ:/ และสระเสียงสั้น (ด้านล่าง) ได้แก่ สระ อะ /a/, อิ /i/ และ อึ /ɯ/



รูปที่ 16 แสดงตัวอย่างของ MFCC audio features ของสระภาษาไทยเสียงยาว (ก) – สั้น (ข)

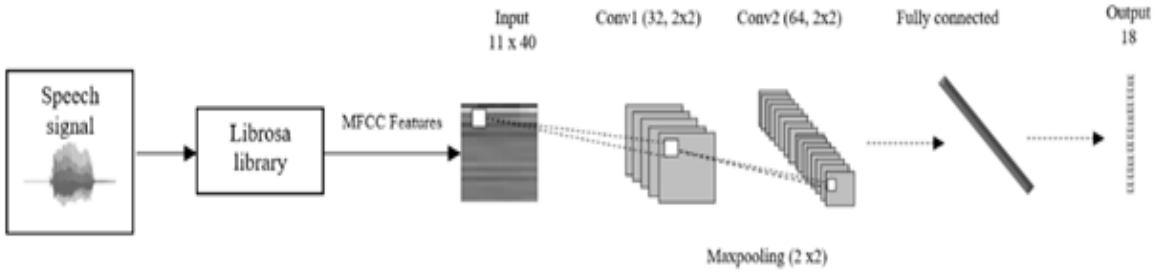
คุณลักษณะข้อมูลนำเข้า (Input Features) ในงานก่อนหน้านี้ใช้ CNN สำหรับการรู้จำเสียง [51] ได้มีการกำหนดคุณลักษณะการป้อนข้อมูลข้อมูลเข้าด้วยขนาดของเวลาและความถี่ (#times x #frequencies) คือ 11 x 40 และในการวิจัย [52] ข้อมูลเข้าเริ่มต้น ถูกกำหนดค่าเป็น 11 x 40 เช่นกัน โดยที่ #times คือขนาดของ Context Window และ #freqs คือ มิติข้อมูลของคุณลักษณะความถี่ และนักวิจัยได้ทำการทดลองขยายเวลาและความถี่ซึ่งได้รับผลลัพธ์ที่ดีขึ้นด้วยโมเดล Full-Extension (21 x 64) และได้รับค่า WER 9.8% สำหรับงานวิจัยนี้ สัญญาณเสียงพูดของสระภาษาไทยมีการประมวลผลก่อนหน้าโดยใช้แพคเกจสำหรับการวิเคราะห์เสียงและเสียงเพลงของไลบรารี LibROSA ในภาษาโปรแกรมไพทอน สำหรับการสกัดคุณสมบัติของเสียงเพื่อใช้เป็นข้อมูลเข้าได้ใช้ librosa.feature.mfcc สำหรับการแยกคุณสมบัติ MFCC โดย อัตราการสุ่มตัวอย่าง เท่ากับ 16,000 ตั้งค่าพารามิเตอร์ ให้มีจำนวน MFCC เท่ากับ 40 และเพื่อทำการทดสอบกับคุณลักษณะข้อมูลเข้าที่เหมาะสม ในงานวิจัยนี้ได้ทำการกำหนดค่าเริ่มต้นคุณสมบัติข้อมูลเข้าขนาดของเวลาและมิติข้อมูลของคุณลักษณะความถี่ เมื่อใช้ MFCC เป็น 11 x 40 และได้ทำการทดลองขยายทั้งเวลาและความถี่ในการค้นหาค่าที่เหมาะสมสำหรับการจำแนกสระภาษาไทย

4.1.3. การจำแนกประเภทด้วยโมเดล Convolutional neural networks

ข้อมูลข้อมูลเข้าจะถูกสกัดคุณสมบัติข้อมูลเข้า (Input Features) และส่งผ่านไปยังโมเดลการรู้จำเสียงสระภาษาไทยที่เป็นสถาปัตยกรรมการเรียนรู้เชิงลึก ซึ่งโมเดลนี้เป็นส่วนที่สำคัญที่สุดในการรู้จำเสียงสระของระบบการฝึกการออกเสียงอัตโนมัติสำหรับการออกเสียงสระภาษาไทย สำหรับ Convolutional Neural Network (CNN) เป็นหนึ่งในสถาปัตยกรรมการเรียนรู้เชิงลึก CNN ไม่เพียงแต่นำมาใช้ใน computer vision เท่านั้น แต่ยังรวมถึงถูกนำมาใช้ในการรู้จำคำพูดด้วย CNN เป็นสถาปัตยกรรมที่ผสมผสานระหว่างตัวสกัดคุณลักษณะและตัวจำแนกประเภท [94] ขั้นตอนการจำแนกประเภทถูกใช้เพื่อจำแนก class labels ในเลเยอร์ fully connected หลังจากการสกัดคุณลักษณะ สัญญาณเสียงพูดสระภาษาไทยจะถูกแปลงเป็นเวกเตอร์คุณสมบัติเสียง MFCC และถูกส่งไปยังโมเดล CNN เพื่อจำแนกเสียงสระภาษาไทย

1. โครงสร้างพื้นฐาน (Baseline Structure)

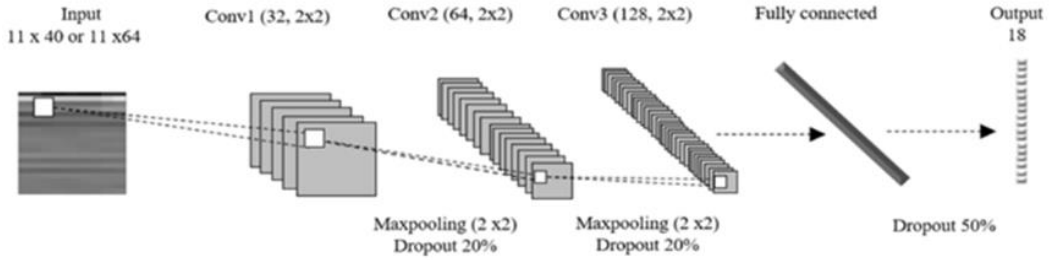
ในงานวิจัยนี้ กำหนดโครงสร้างพื้นฐานของ CNN ประกอบด้วย 2 Convolutional Layers โดย Convolutional Layer ชั้นแรกมี 32 Filters (2 x 2) ตามด้วย Max-Pooling (2 x 2) ในขณะที่ Convolutional Layer ที่สองมี 64 Filters (2 x 2) แต่ไม่มี Pooling Layers และมีการใช้ ReLU [91] ซึ่งมีการใช้กันอย่างแพร่หลายในสถาปัตยกรรม Convolutional Neural Networks หรือ Deep Learning และสามารถลดเวลาในการคำนวณ สำหรับ Pooling Layers ทั้งหมดใช้ Filter (2 x 2) และ Stride เท่ากับ 2 และ Fully Connected Layer ใช้ 64 hidden units และ Softmax Activation Function ใช้ Adam Optimizer [95] เนื่องจากช่วยในเรื่องของการบรรจบกันที่เร็วขึ้นและให้ประสิทธิภาพสูง เอาต์พุตมี 18 คลาส โดยรูปที่ 17 แสดงรายละเอียดสถาปัตยกรรมของโมเดล CNN พื้นฐานสำหรับการจำแนกเสียงสระไทย



รูปที่ 17 แสดงการสกัดคุณลักษณะของเสียงและโครงสร้างพื้นฐานของ CNN

2. Fine-Tuning the CNN model

สถาปัตยกรรม CNN ของสระภาษาไทยในงานวิจัยนี้ คือ CNN model ที่ได้มาจากผลของการทดลองในส่วนของผลลัพธ์ และสุดท้ายทำการทดลอง CNN model เพื่อเปรียบเทียบประสิทธิภาพกับ Multilayer Perceptron (MLP) model และ Support Vector Machines (SVM) model โดย CNN model ประกอบด้วย 3 Convolutional Layers โดยที่ Convolutional Layer แรกมี 32 Filters (2 x 2) ตามด้วย Max-Pooling (2 x 2) และ Dropout 20% ใน Convolutional Layer ชั้นที่ 2 มี 64 Filters (2 x 2) ตามด้วย Max-Pooling Layer และ Dropout เหมือน Convolutional Layer ชั้นแรก ในขณะที่ Convolutional Layer ชั้นที่ 3 มี 128 Filters (2 x 2) และ Dropout 20% แต่ไม่มี Pooling Layer สำหรับ Fully Connected Layer ใช้ 64 hidden units และDropout 50% และใช้ Softmax Activation Function ในโมเดลนี้ใช้กลยุทธ์ Padding, Adam Optimizer และกำหนด Batch Size ซึ่งในเพศหญิง Input Features ที่เหมาะสมคือ 11 x 40 และในเพศชายคือ 11 x 64 โดยสถาปัตยกรรม CNN model แสดงในรูปที่ 18



รูปที่ 18 แสดงสถาปัตยกรรม CNN ของเสียงสระภาษาไทยอย่างง่าย (CNN model)

4.1.4. รายละเอียดการทดลอง (Implementation details)

สำหรับการทดลองในงานวิจัยนี้ใช้ Keras Framework และทำการทดลองบนระบบปฏิบัติการ Windows 64 บิต ใช้ Intel CORE i7 CPU, 8 GB memory และ Nvidia GeForce GTX 1050 GPU

4.2. ผลการทดลอง

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาโครงสร้างที่เหมาะสม โดยใช้ CNN model สำหรับการออกเสียงสระภาษาไทย ผลการทดลองถูกนำเสนอในตารางดังต่อไปนี้

4.2.1. การขยายเวลาและความถี่ (Time and Frequency Extension)

จากการทบทวนวรรณกรรม [52] พบว่าการใช้คุณลักษณะการป้อนข้อมูลข้อมูลเข้าและการขยายเวลาและความถี่ มีประโยชน์สำหรับการพัฒนาโมเดล และได้ทำการทดลองกับเทคนิคนี้เพื่อหาค่าที่เหมาะสมสำหรับงานการรู้จำเสียงสระภาษาไทย

ตารางที่ 2 แสดงผลลัพธ์ของการขยายเวลาและความถี่

Input Features*	Accuracy (%)			
	No padding		*padding*	
	Female	*Male*	*Female*	*Male*
11 x 40	82.78	76.67	80.00	78.89
11 x 64	80.00	**80.00**	**80.56**	**80.56**
17 x 40	78.33	78.33	77.78	78.33
17 x 64	**83.89**	78.33	78.33	78.89
Avg.	81.25	78.33	79.17	79.17

*Input Features: #times x #frequencies

ตารางที่ 2 แสดงผลการทดลอง โดยคุณสมบัติข้อมูลเข้าที่เหมาะสมสำหรับทั้งเสียงของเพศชายและเพศหญิงคือ 11 x 64 ในทางตรงกันข้ามเพศหญิงที่ 17 x 64 แสดงผลลัพธ์ที่ดีที่สุดที่ 83.89% การใช้ Padding จะไม่ช่วยเพิ่มประสิทธิภาพ (ไม่ได้ใช้กับกลยุทธ์ใด ๆ)

4.2.2. ดรอปเอาท์ (Dropout)

จากผลการทดลองเกี่ยวกับการคอมพิวเตอร์วิทัศน์ แสดงให้เห็นถึงการทดลองที่มีประสิทธิภาพเมื่อใช้ Dropout และ Padding ดังนั้นในงานวิจัยนี้ได้ทำการทดลองเพิ่มเติมเพื่อให้บรรลุประสิทธิภาพที่ดีขึ้นโดยใช้เทคนิค Dropout และ Padding

ตารางที่ 3 แสดงผลลัพธ์การใช้ Dropout

Input Features*	Accuracy (%)			
	No padding		*padding*	
	Female	*Male*	*Female*	*Male*
11 x 40	**87.78**	83.89	87.22	84.44
11 x 64	87.22	85.56	**88.89**	**87.22**
17 x 40	85.56	85.56	86.11	83.89
17 x 64	87.22	**86.67**	86.67	**87.22**
Avg.	86.95	85.42	87.22	85.69

*Input Features: #times x #frequencies

ตารางที่ 3 แสดงผลลัพธ์ที่ดีขึ้นโดยใช้คุณสมบัติการป้อนข้อมูลข้อมูลเข้าที่ 11 x 40, 11 x 64 และ 17 x 64 ดังนั้นคุณลักษณะการป้อนข้อมูลทั้ง 3 นี้ จะถูกนำไปใช้ในการทดลองครั้งต่อไป

4.2.3. Batch Size

ทำการทดลองการใช้ Batch Size และไม่มีการใช้ Batch Size โดยมีการกำหนดค่าของ Batch Size ที่ 32, 64 และ 128 ตามลำดับ

ตารางที่ 4 แสดงผลลัพธ์ของ Batch Size

Batch Size	Accuracy (%)					
	11 x 40		*11 x 64*		*17 x 64*	
	Female	*Male*	*Female*	*Male*	*Female*	*Male*
no	87.22	84.44	**88.89**	87.22	86.67	87.22
32	**88.89**	83.89	87.22	**88.89**	85.56	86.67
64	**88.89**	**86.67**	86.67	86.67	86.11	**88.33**
128	88.33	84.44	86.67	87.22	**87.78**	85.00
Avg.	**88.33**	84.86	87.36	**87.50**	86.53	86.81

ตารางที่ 4 แสดงคุณสมบัติข้อมูลเข้า (11 x 40, 11 x 64 และ17 x 64) ของทั้งเพศหญิงและเพศ

ชาย เพื่อหาค่าเฉลี่ยในการพิจารณาผลลัพธ์ การทดลองสำหรับเพศหญิง คุณสมบัติการป้อนข้อมูลข้อมูลเข้าที่เหมาะสมคือ 11 x 40 โดยมีค่าเฉลี่ยความถูกต้องแม่นยำที่ 88.33% สำหรับเพศชายที่เหมาะสม คือ 11 x 64 โดยมีค่าเฉลี่ยความถูกต้องแม่นยำที่ 87.50% โดยที่ Batch Size ที่เหมาะสมของเพศหญิงและชายคือ 32 ซึ่งให้ค่าความความถูกต้องแม่นยำ 88.89%

4.2.4. จำนวนชั้น Convolution Layer

เมื่อขยายชั้นเลเยอร์ของ Convolution จาก 2 เป็น 3 ตารางที่ 5 ด้านล่างแสดงผลลัพธ์ที่ดีขึ้น โดยเฉพาะอย่างยิ่งการปรับปรุงมีความชัดเจนในเพศหญิงที่ได้รับค่าความถูกต้องแม่นยำที่ 90.00% แม้ว่าของเพศชายผลลัพธ์ของการเพิ่มชั้น Convolution Layer จะไม่แตกต่างกัน แต่ในงานวิจัยนี้ผู้วิจัยเชื่อว่าการเพิ่มชั้นเลเยอร์ Convolutional มากขึ้นจะให้ผลลัพธ์ที่ดีขึ้นดังแสดงในตารางที่ 5 การเพิ่มชั้น Convolution Layer ทำให้สามารถสกัดคุณลักษณะเสียงที่ละเอียดมากขึ้น ทำให้การรู้จำเสียงสระภาษาไทยมีประสิทธิภาพเพิ่มขึ้น

ตารางที่ 5 แสดงผลลัทธ์ของจำนวนชั้น Convolution Layer

Number of Convolution Layer	Accuracy (%)	
	Female (11 x 40)	*Male (11 x 64)*
2 layers	88.89	88.89
3 layers	**90.00**	**88.89**

4.2.5. จำนวน Hidden Units

งานวิจัยนี้ได้ทำการเปรียบเทียบผลการทดลองกับจำนวน Hidden Units ที่แตกต่างกันกับ 3 Convolutional Layers ซึ่งตารางที่ 5 แสดงผลลัพธ์ที่มีประสิทธิภาพของการเพิ่มจำนวนเลเยอร์ทั้งสองเพศ (เพศหญิงและเพศชาย)

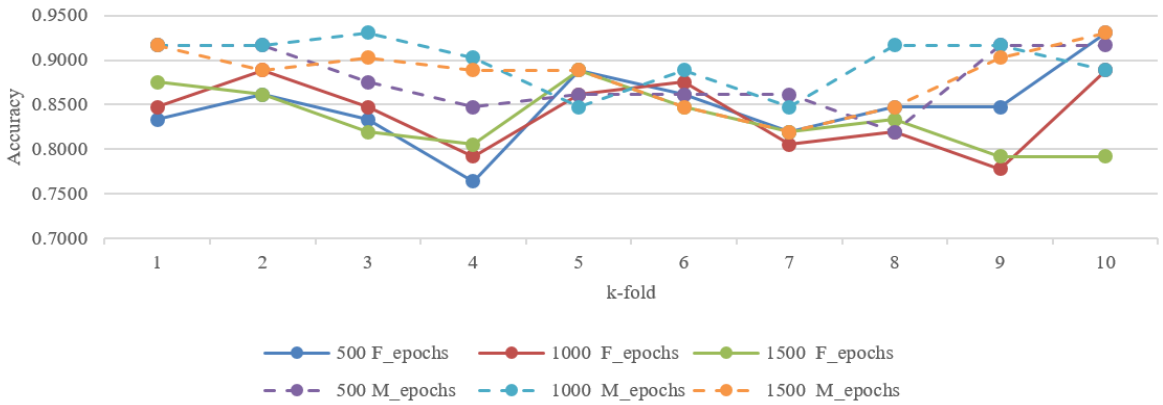
ตารางที่ 6 แสดงผลลัพธ์ความแตกต่างของจำนวน Hidden Units

Number of Hidden Units	Accuracy (%)	
	Female (11 x 40)	*Male (11 x 64)*
64 Units	**90.00**	**88.89**
256 Units	88.33	87.22
1024 Units	86.67	86.11

ตารางที่ 6 ผลลัพธ์สรุปได้ว่าการเพิ่มจำนวน Hidden Units ใน Fully Connected Layer ไม่ช่วยปรับปรุงประสิทธิภาพ ดังนั้นจึงใช้ Hidden Units เท่ากับ 64

4.2.6. การเปรียบเทียบ CNN, MLP and SVM model (k-fold = 10)

การทดลองแสดงถึงการเปรียบเทียบ CNN model กับ MLP model และ SVM model โดย CNN model ได้มาจากผลการทดลองที่ผ่านมา ผลลัพธ์จากรูปที่ 19 แสดงให้เห็นว่า Epochs ที่เหมาะสมกับเสียงของเพศหญิงคือ 500 Epochs ค่าเฉลี่ยของความถูกต้องแม่นยำคือ 84.86% และส่วนเบี่ยงเบนมาตรฐาน +/- 4.14% ในเสียงของผู้ชาย Epochs ที่เหมาะสมคือ 1,000 Epochs ค่าเฉลี่ยของความถูกต้องแม่นยำคือ 89.72% และส่วนเบี่ยงเบนมาตรฐานคือ +/- 2.79% Multilayer Perceptron Classifier (MLP Classifier) ที่ใช้ในงานวิจัยนี้ พื้นฐานของโมเดล MLP ประกอบด้วย Hidden Layer ที่มี 256 Units ใช้ RELU Activation, Adam Optimization, Batch Size ขนาด 32 และ อัตราการเรียนรู้เริ่มต้น 0.001 สำหรับ SVM model ใช้ Support Vector Classification (SVC), Linear Kernel, และ Decision Function เป็น one-vs-rest (‘ovr’) โดยทุกโมเดลใช้คุณสมบัติข้อมูลเข้าเดียวกัน



รูปที่ 19 แสดงผลลัพธ์ของ Epochs (500, 1000, 1500) ของเพศหญิงและเพศชาย

ตารางที่ 7 แสดงผลลัพธ์การเปรียบเทียบ CNN, MLP และ SVM model (k-fold = 10)

Methods	Mean Accuracy (%)	
	Female (11 x 40)	*Male (11 x 64)*
CNN	**84.86**	**89.72**
MLP	68.12	71.85
SVM	76.00	82.17

จากตารางที่ 7 ผลลัพธ์แสดงให้เห็นว่า CNN model ให้ค่าเฉลี่ยของความถูกต้องแม่นยำสูงสุดและมีประสิทธิภาพทั้งเพศหญิงและเพศชาย โดยมีค่าเฉลี่ยความถูกต้องแม่นยำ 84.86% และ 89.72% ตามลำดับ

4.2.7. Confusion matrix, Precision, Recall, และ F1-score ของ CNN model

สำหรับการวิเคราะห์ข้อผิดพลาด เมทริกซ์ความสับสนของ CNN model ของเสียงเพศหญิงและเพศชายแสดงในรูปที่ 20 และ 21 สำหรับเมทริกซ์คู่ที่สับสนที่สุดของสระภาษาไทยสำหรับเสียงของเพศหญิงคือ (‘โอ’ / oː / และ ‘โอะ’ / o /) ในเพศชาย เสียงสระคู่ที่สร้างความสับสนมากที่สุดในสระภาษาไทยคือ (‘โอ’ / oː / และ ‘โอะ’ / o /) เหมือนกับเสียงเพศหญิง และ (‘เออ’ / ɤː / และ ‘เออะ’/ ɤ /) จากการทดสอบพบว่าสระ (‘โอ’ / oː / และ ‘โอะ’ / o /) เป็นคู่ที่สับสนมากที่สุดของทั้งสองเพศ โดยคู่ที่สับสนนั้นคือสระเสียงสั้นและสระเสียงยาว ซึ่งมีความแตกต่างทางด้านของระยะเวลา

		Actual class																	
		/a:/	/i:/	/uː/	/u:/	/eː/	/ɛː/	/oː/	/ɔː/	/ɤː/	/ɑ/	/ɨ/	/ɯ/	/ʉ/	/e/	/ɛ/	/o/	/ɔ/	/ɤ/
Predicted class	/a:/	8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	/i:/	0	15	0	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0
	/uː/	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	/u:/	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	/eː/	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0
	/ɛː/	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0
	/oː/	0	0	0	1	0	0	8	0	0	0	0	0	0	0	0	0	0	0
	/ɔː/	1	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	1	0
	/ɤː/	0	0	2	0	1	0	0	0	10	0	0	0	0	0	0	0	0	0
	/ɑ/	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0
	/ɨ/	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
	/ɯ/	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0
	/ʉ/	0	0	0	0	0	0	0	0	0	0	1	0	6	0	1	0	0	0
	/e/	0	0	0	0	1	0	0	0	0	0	2	0	0	10	0	0	0	0
	/ɛ/	0	0	0	0	0	1	0	0	0	0	0	0	0	0	8	0	0	0
	/o/	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	8	0	0
	/ɔ/	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	11	0
	/ɤ/	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	7

รูปที่ 20 แสดง Confusion Matrix ของ MFCC acoustic features รวมกับโมเดล CNN ในชุดข้อมูลเพศหญิง

		Actual class																	
		/a:/	/i:/	/uː/	/u:/	/eː/	/ɛː/	/oː/	/ɔː/	/ɤː/	/ɑ/	/ɨ/	/ɯ/	/ʉ/	/e/	/ɛ/	/o/	/ɔ/	/ɤ/
Predicted class	/a:/	7	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
	/i:/	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	/uː/	0	0	7	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0
	/u:/	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	/eː/	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0
	/ɛː/	1	0	0	0	0	8	0	0	0	0	0	0	0	0	2	0	0	0
	/oː/	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	1	0	0
	/ɔː/	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0
	/ɤː/	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	1
	/ɑ/	1	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0
	/ɨ/	0	1	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0
	/ɯ/	0	0	0	0	0	0	0	0	0	0	1	7	0	0	0	0	0	0
	/ʉ/	0	0	0	0	0	0	1	0	0	0	0	0	6	0	0	0	0	0
	/e/	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
	/ɛ/	0	0	0	0	0	1	0	0	0	0	0	0	0	0	7	0	0	0
	/o/	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	8	0	0
	/ɔ/	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	12	0
	/ɤ/	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	6

รูปที่ 21 แสดง Confusion Matrix ของ MFCC acoustic features รวมกับโมเดล CNN ในชุดข้อมูลเพศชาย

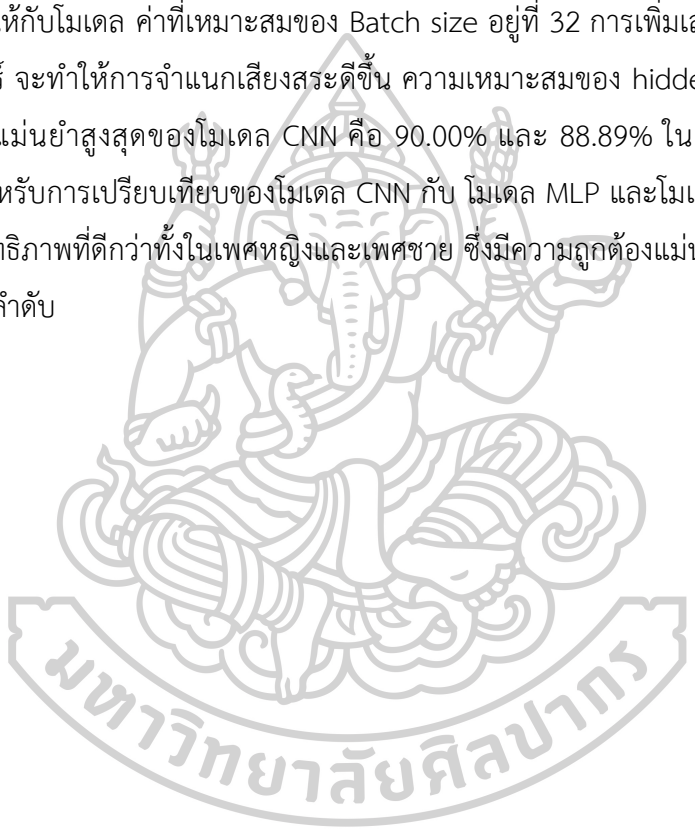
ตารางที่ 8 แสดง Precision, Recall, และ F1-score ของ CNN model

Thai Vowels	Female			Male		
	Precision	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
a:	0.89	0.89	0.89	0.78	0.78	0.78
i:	0.79	0.94	0.86	1.00	0.94	0.97
ɯ:	0.83	0.71	0.77	0.70	1.00	0.82
u:	0.83	0.83	0.83	1.00	1.00	1.00
e:	1.00	0.69	0.82	1.00	1.00	1.00
ɛ:	1.00	0.89	0.94	0.73	0.89	0.80
o:	0.89	0.67	**0.76**	0.88	0.58	**0.70**
ɔ:	0.82	0.90	0.86	1.00	0.70	0.82
ɤ:	0.77	0.83	0.80	0.88	0.58	**0.70**
a	1.00	0.90	0.95	0.90	0.90	0.90
i	1.00	0.64	0.78	0.92	0.86	0.89
ɯ	1.00	1.00	1.00	0.88	0.78	0.82
u	0.75	1.00	0.86	0.86	1.00	0.92
e	0.77	1.00	0.87	1.00	1.00	1.00
ɛ	0.89	0.89	0.89	0.88	0.78	0.82
o	0.73	0.89	0.80	0.67	0.89	0.76
ɔ	0.85	0.92	0.88	0.80	1.00	0.89
ɤ	0.78	1.00	0.88	0.60	0.86	0.71

ตารางที่ 8 นำเสนอ Precision, Recall และ F1-score ของ CNN model สำหรับการจำแนกเสียงสระภาษาไทยในแต่ละตัว ในเสียงของผู้หญิง F1-score ต่ำที่สุดคือ ‘โอ’/ o:/ (0.76) และในเสียงผู้ชายคือ ‘โอ’ / o: / (0.70) และ ‘เออ’ / ɤ: / ผลลัพธ์ F1-score สัมพันธ์กันกับเมทริกซ์ความสับสน ในทางกลับกันคะแนนสูงสุดของ F1-score (1.00) สำหรับเสียงผู้หญิง คือ ‘อึ’ / ɯ / และเสียงผู้ชายคือ ‘อู’ / u: /, ‘เอ’ / e: /, และ ‘เอะ’ / e /

4.3. สรุป

งานวิจัยนี้นำเสนอการรู้จำการออกเสียงสระภาษาไทยที่มีเสียงรบกวน โดยใช้โมเดล Convolutional Neural Network (CNN) ซึ่งงานวิจัยได้ทำการทดลองกับชุดข้อมูลที่มีสัญญาณรบกวนที่ถูกรวบรวมขึ้นมาใหม่ โดยชุดข้อมูลนี้ ประกอบด้วยการบันทึกเสียงสระของเพศหญิง 25 คน และเพศชาย 25 คน โดยเสียงสระแบ่งออกเป็น 18 เสียง เมื่อพิจารณาผลการทดลองด้านการขยายเวลาและความถี่ พบว่าคุณลักษณะการป้อนข้อมูลข้อมูลเข้าที่เหมาะสมคือ 11 x 40 ในเพศหญิง และ 11 x 64 สำหรับเพศชาย กลยุทธ์ Padding และ Dropout ที่ 20% ได้ถูกนำมาใช้เพื่อเพิ่มประสิทธิภาพให้กับโมเดล ค่าที่เหมาะสมของ Batch size อยู่ที่ 32 การเพิ่มเลเยอร์ convolutional เป็น 3 เลเยอร์ จะทำให้การจำแนกเสียงสระดีขึ้น ความเหมาะสมของ hidden units คือ 64 อัตราความถูกต้องแม่นยำสูงสุดของโมเดล CNN คือ 90.00% และ 88.89% ในเพศหญิงและเพศชายตามลำดับ สำหรับการเปรียบเทียบของโมเดล CNN กับ โมเดล MLP และโมเดล SVM พบว่า โมเดล CNN มีประสิทธิภาพที่ดีกว่าทั้งในเพศหญิงและเพศชาย ซึ่งมีความถูกต้องแม่นยำอยู่ที่ 84.86% และ 89.72% ตามลำดับ



บทที่ 5

วิธีดำเนินงานวิจัยและผลการทดลองที่ 2

โมเดลการเรียนรู้เชิงลึกกับคุณสมบัติด้านเสียงสำหรับการรู้จำเสียงสระภาษาไทยแบบอัตโนมัติ

สำหรับการออกเสียงสระภาษาไทย สิ่งสำคัญคือต้องรู้ว่าเมื่อการออกเสียงผิดเกิดขึ้น ความหมายของคำจะเปลี่ยนไปโดยสิ้นเชิง ดังนั้นการปฏิบัติที่มีประสิทธิภาพและเป็นมาตรฐานจึงเป็นสิ่งจำเป็นในการออกเสียงคำอย่างถูกต้องในฐานะเจ้าของภาษา ตั้งแต่มีการระบาดของ COVID-19 การเรียนรู้ออนไลน์ก็ได้รับความนิยม ตัวอย่างเช่น มีการแนะนำระบบแอปพลิเคชันการออกเสียงออนไลน์ที่มีครูเสมือนและกระบวนการประเมินนักเรียนที่ชาญฉลาดซึ่งคล้ายกับการฝึกอบรมที่ได้มาตรฐานโดยครูในห้องเรียนจริง งานวิจัยนี้นำเสนอการฝึกอบรมการออกเสียงโดยใช้คอมพิวเตอร์ช่วย (Computer-Assisted Pronunciation Training : CAPT) แบบออนไลน์โดยอัตโนมัติโดยใช้การเรียนรู้เชิงลึกเพื่อจดจำสระภาษาไทยในการพูด CAPT อัตโนมัติได้รับการพัฒนาเพื่อแก้ปัญหาความไม่เพียงพอของผู้เชี่ยวชาญด้านการสอนและกระบวนการสอนสระที่ซับซ้อน เป็นระบบเฉพาะที่พัฒนาเทคนิคคอมพิวเตอร์ผสมผสานกับทฤษฎีภาษาศาสตร์ โมเดลการเรียนรู้เชิงลึกเป็นส่วนที่สำคัญที่สุดในการจดจำสระที่ออกเสียงสำหรับ CAPT อัตโนมัติ ความท้าทายหลักในการจดจำสระไทยคือการระบุสระไทยที่ถูกต้องเมื่อพูดในสถานการณ์จริง Convolutional Neural Network (CNN) ซึ่งเป็นแบบจำลองการเรียนรู้เชิงลึกถูกนำไปใช้และพัฒนาในการจำแนกเสียงสระภาษาไทยที่ออกเสียง ชุดข้อมูลใหม่สำหรับสระไทยได้รับการออกแบบ รวบรวม และตรวจสอบโดยนักภาษาศาสตร์ ผลลัพธ์ของโมเดล CNN ที่เหมาะสมด้วย Mel spectrogram (MS) ให้ความแม่นยำสูงสุด 98.61% เมื่อเทียบกับ Mel Frequency Cepstrum Coefficient (MFCC) กับโมเดล baseline long short-term memory (LSTM) และ MS กับโมเดล baseline LSTM ที่มีความแม่นยำ 94.44% และ 90.00% ตามลำดับ

5.1. ชุดข้อมูลและวิธีการ (Datasets and Methods)

การอธิบายชุดข้อมูลและวิธีการของเรียนรู้เชิงลึกในการจดจำเสียงสระภาษาไทย รายละเอียดมีดังนี้ ส่วนที่ 5.1.1. อธิบายชุดข้อมูล ซึ่งประกอบด้วย การออกแบบ การรวบรวม การจัดเตรียม และการสำรวจชุดข้อมูล ในส่วนที่ 5.1.2. แสดงการแปลงสเปกโตรแกรม ส่วนที่ 5.1.3. แสดงการจำแนกประเภทของโมเดล convolutional neural networks

5.1.1. ชุดข้อมูล (Dataset)

1. การออกแบบชุดข้อมูล (Dataset design)

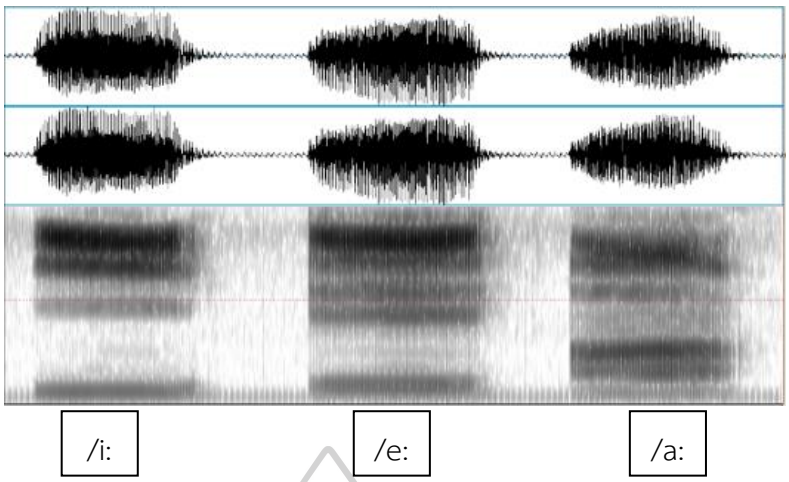
ปัจจุบันไม่มีการเผยแพร่ชุดข้อมูลเสียงสระภาษาไทยที่เป็นสาธารณะสำหรับวัตถุประสงค์การวิจัยนี้ ส่วนใหญ่ถูกรวบรวมอย่างไม่เป็นระบบและไม่ได้มาตรฐาน ดังนั้นการเตรียมชุดข้อมูลจึงได้รับการออกแบบเพื่อวัตถุประสงค์ของการศึกษาครั้งนี้ ชุดคำภาษาไทยที่ใช้บันทึกเสียงออกแบบโดยนักภาษาศาสตร์ รายการคำศัพท์นี้ได้รับการออกแบบตามทฤษฎีภาษาศาสตร์ อันเป็นผลมาจากกฎทางภาษาศาสตร์ ทุกคำมีลักษณะเฉพาะเหมือนกันซึ่งเป็นพยัญชนะเดียวกัน น้ำเสียงเดียวกัน และพยัญชนะท้ายเดียวกัน แต่ต่างกันเฉพาะในเสียงสระเท่านั้น

2. การรวบรวมชุดข้อมูล (Dataset collection)

เก็บรวบรวมชุดข้อมูลเสียงสระจากผู้พูดภาษาไทยที่พูดภาษากลางซึ่งถือเป็นภาษาไทยอย่างเป็นทางการ ชุดข้อมูลถูกรวบรวมในสภาพแวดล้อมต่างๆ เช่น โรงเรียน โรงอาหาร สวนสาธารณะ ห้องเรียน ห้องนอน หรือบ้าน และรวบรวมจากเจ้าของภาษาในสถานการณ์จริง ซึ่งประกอบด้วยเสียงรบกวนหลายประเภทที่ 30 - 50 dB SNR (Signal to Noise Ratio) เช่น รถยนต์บนท้องถนน คนคุยกันในโรงอาหาร เสียงดนตรีที่วิทยาลัยดุริยางคศิลป์ และเสียงสัตว์ (สุนัขและนก) ดังนั้นข้อมูลที่มีอยู่จึงถูกจัดประเภทเป็นชุดข้อมูล “เสียงสระไทยที่มีเสียงรบกวน” เสียงที่บันทึกจากชุดข้อมูลรวบรวมจากเจ้าของภาษาไทยมาตรฐานจำนวน 50 คน (ชาย 25 คน หญิง 25 คน) มีคุณสมบัติตามกระบวนการคัดเลือกตัวอย่างที่ดี ตามขั้นตอนการคัดเลือกโดยนักภาษาศาสตร์ ทั้งหมดมีอายุ 20-25 ปี โดยบันทึกจากโทรศัพท์มือถือที่ 44,100 Hz (standard speech data)

3. การเตรียมชุดข้อมูล (Dataset preparation)

หลังจากการบันทึกเสียงสำเร็จ ไฟล์เสียงทั้งหมดจะถูกส่งไปยังนักภาษาศาสตร์เพื่อตรวจสอบ และได้คัดเลือกเสียงสระที่มีความสมบูรณ์ของแต่ละสระของแต่ละผู้พูด หลังจากนั้นนักภาษาศาสตร์ได้ตัดไฟล์เสียงโดยใช้ Praat [27] รูปที่ 22 แสดงสระอี /i:/, สระเอ /e:/ และสระอา /a:/ ในขณะที่ใช้ Praat เมื่อไฟล์เสียงทั้งหมดถูกตัดออก แต่ละไฟล์จะถูกตรวจสอบใหม่เพื่อทำการตรวจสอบความถูกต้องอีกครั้ง ถ้าเป็นไฟล์เสียงที่ดีก็จะถูกเลือกเก็บไว้ ถ้าไฟล์เสียงไม่ดีก็จะถูกคัดออกและจะหาเสียงใหม่ เพื่อให้ได้ไฟล์เสียงที่มีคุณภาพดีที่สุดนักภาษาศาสตร์จึงตรวจสอบไฟล์เสียงที่เลือกไว้ทั้งหมดอีกครั้งจำนวน 5 รอบ



รูปที่ 22 แสดง สระ /i:/, /e:/, และ /a:/ โดยใช้ Praat

จำนวนเสียงสระคุณภาพดีที่เหมาะสมสำหรับการนำมาใช้ในงานนวิจัยนี้มีจำนวนทั้งหมด 1,800 ไฟล์เสียง แบ่งออกเป็นเสียงผู้ชาย 900 เสียง (สระ 18 เสียง × ผู้ชาย 25 คน × พูด 2 ครั้ง) และเสียงผู้หญิง 900 เสียง (สระ 18 เสียง × ผู้หญิง 25 คน × พูด 2 ครั้ง) ชุดข้อมูลทั้ง 18 คลาสประกอบด้วยสระเสียงสั้น 9 เสียง และสระเสียงยาว 9 เสียง งานวิจัยนี้ใช้ชุดข้อมูลผสม ชุดข้อมูลแบบผสมเป็นการผสมผสานระหว่างชุดข้อมูลเสียงเพศหญิงและเพศชาย ซึ่งแตกต่างกันกับ [74] ที่ไม่มีชุดข้อมูลผสม ชุดข้อมูลถูกแบ่งออกเป็นชุดการฝึกอบรมและการทดสอบโดยใช้ K-fold cross-validation (k-fold = 5)

4. การสำรวจข้อมูล (Exploration data)

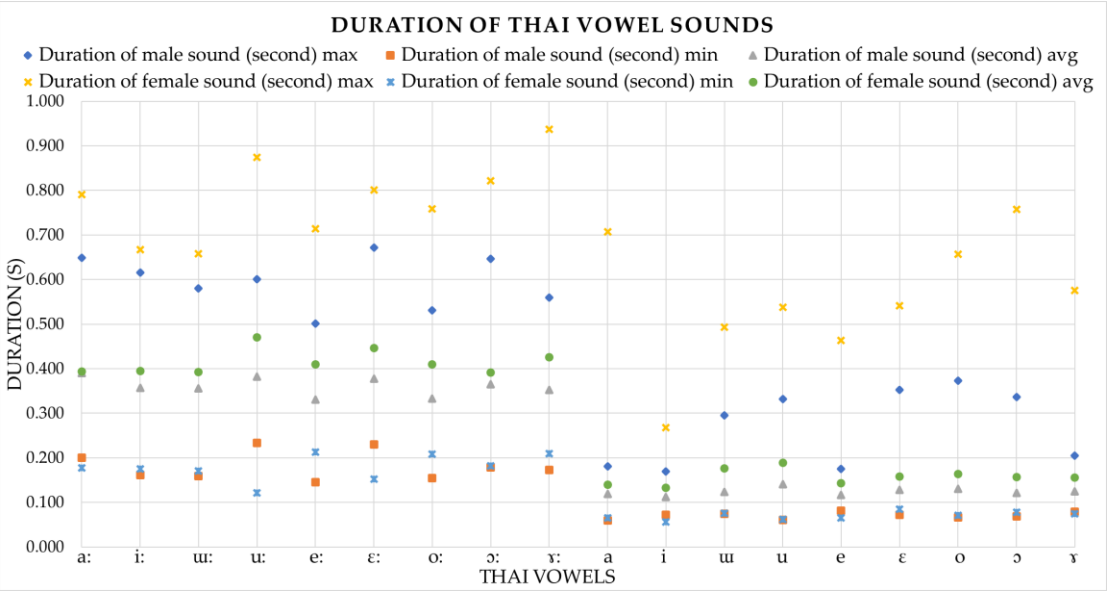
หลังจากเตรียมชุดข้อมูลสระภาษาไทย ได้ทำการสำรวจข้อมูลเสียงสระแต่ละเสียง ตารางที่ 9 แสดงระยะเวลาของสระภาษาไทยทั้งที่เป็นเสียงของเพศชายและเสียงของเพศหญิง: ค่าสูงสุด (max) ค่าต่ำสุด (min) และค่าเฉลี่ย (avg)

ตารางที่ 9 แสดงการสำรวจระยะเวลาในชุดข้อมูลเสียงสระภาษาไทย

Thai vowels	Duration of male sound (second)			Duration of female sound (second)		
	max	min	avg	max	min	avg
a:	0.649	0.200	**0.391**	0.791	0.177	0.394
i:	0.616	0.161	0.357	0.667	0.175	0.395
ɯ:	0.580	0.159	0.356	0.658	0.170	0.392
u:	0.601	0.234	0.383	0.874	0.121	**0.470**
e:	0.501	0.145	0.331	0.714	0.213	0.410

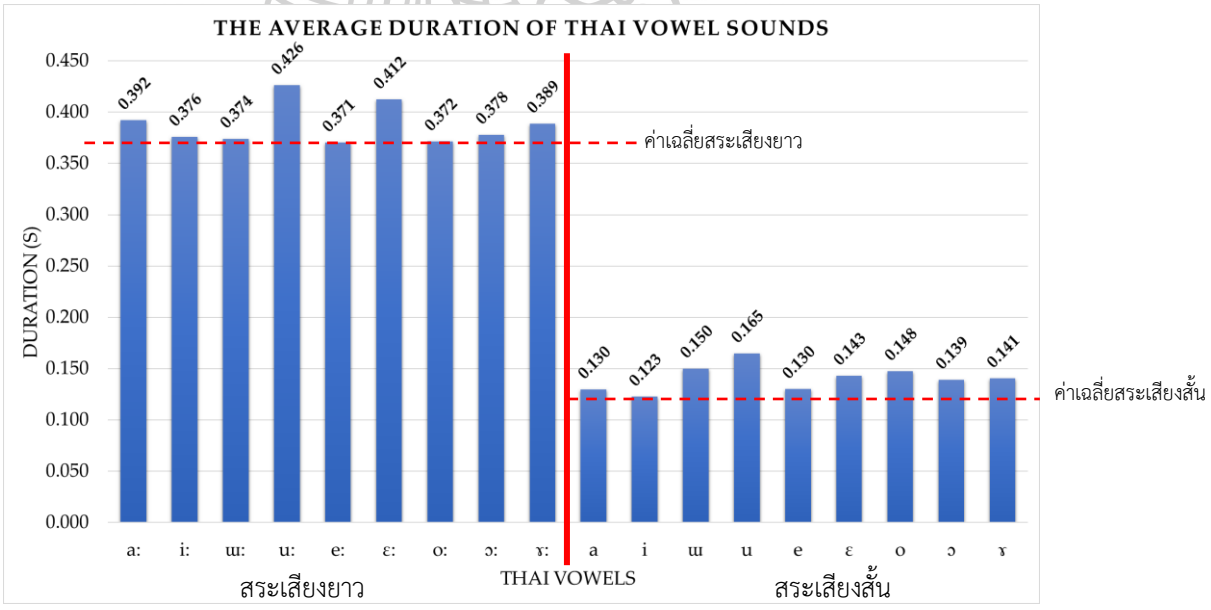
ɛ:	**0.672**	0.230	0.378	0.801	0.152	0.447
o:	0.531	0.155	0.333	0.759	0.208	0.410
ɔ:	0.646	0.179	0.365	0.821	0.182	0.391
ɤ:	0.559	0.173	0.352	**0.937**	0.210	0.425
a	0.181	**0.059**	0.120	0.707	0.065	0.140
i	0.170	0.072	**0.112**	0.268	**0.057**	**0.133**
ɯ	0.295	0.074	0.124	0.493	0.076	0.176
u	0.332	0.061	0.140	0.538	0.062	0.189
e	0.175	0.082	0.116	0.464	0.066	0.144
ɛ	0.353	0.072	0.129	0.542	0.085	0.158
o	0.373	0.067	0.131	0.657	0.071	0.164
ɔ	0.337	0.069	0.122	0.757	0.078	0.157
ɤ	0.206	0.080	0.125	0.576	0.075	0.156

การออกเสียงสระแต่ละเสียงแตกต่างกันไปในแต่ละบุคคล เพศ หรืออายุ แม้ว่าผู้พูดจะพูดสองครั้ง แต่คุณภาพเสียงก็ยังแตกต่างกัน ค่าของระยะเวลาของสระไทย 18 เสียงแสดงไว้ในตารางที่ 9 สำหรับเสียงเพศชาย ระยะเวลาสูงสุดคือ 0.672 วินาทีในสระแอ /ɛ:/ ระยะเวลาต่ำสุด 0.059 วินาทีใน สระอะ /a/ ระยะเวลาเฉลี่ยสูงสุดและต่ำสุดคือ 0.391 วินาทีในสระอา /a:/ และ 0.112 วินาทีในสระอิ /i/ ตามลำดับ สำหรับเสียงเพศหญิง ระยะเวลาสูงสุดคือ 0.937 วินาทีในสระเออ /ɤ:/ ระยะเวลาต่ำสุด 0.057 วินาทีในสระอิ /i/ ระยะเวลาเฉลี่ยสูงสุดและต่ำสุดคือ 0.470 วินาทีในสระอู /u:/ และ 0.133 วินาทีในสระอิ /i/ ตามลำดับ การออกเสียงสระอิ /i/ สำหรับทั้งชายและหญิงมีระยะเวลาที่สั้นที่สุด



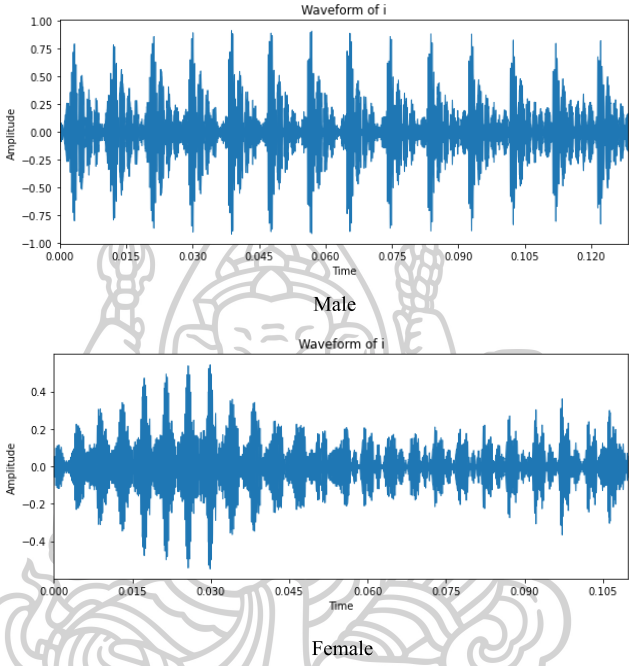
รูปที่ 23 แสดงระยะเวลาของเสียงสระภาษาไทย

รูปที่ 23 แสดงระยะเวลาสูงสุด ต่ำสุด และเวลาเฉลี่ยของเสียงสระของเพศชายและเพศหญิง สัญลักษณ์ 3 อันดับแรกแถวบน แสดงถึงระยะเวลาของเสียงเพศชาย และสัญลักษณ์ 3 แถวล่างแสดงถึงระยะเวลาของเสียงเพศหญิง ระยะเวลาที่สูงที่สุดในแต่ละสระจะพบในเพศหญิง ระยะเวลาของเพศหญิงจะมีระยะเวลายาวกว่าเพศชายในทุกสระ ระยะเวลาที่ต่ำที่สุดในแต่ละสระสำหรับทั้งชายและหญิงจะใกล้เคียงกัน ระยะเวลาเฉลี่ยของสระทั้งหมดในเพศหญิงจะมีระยะเวลาที่ยาวกว่าในเพศชาย



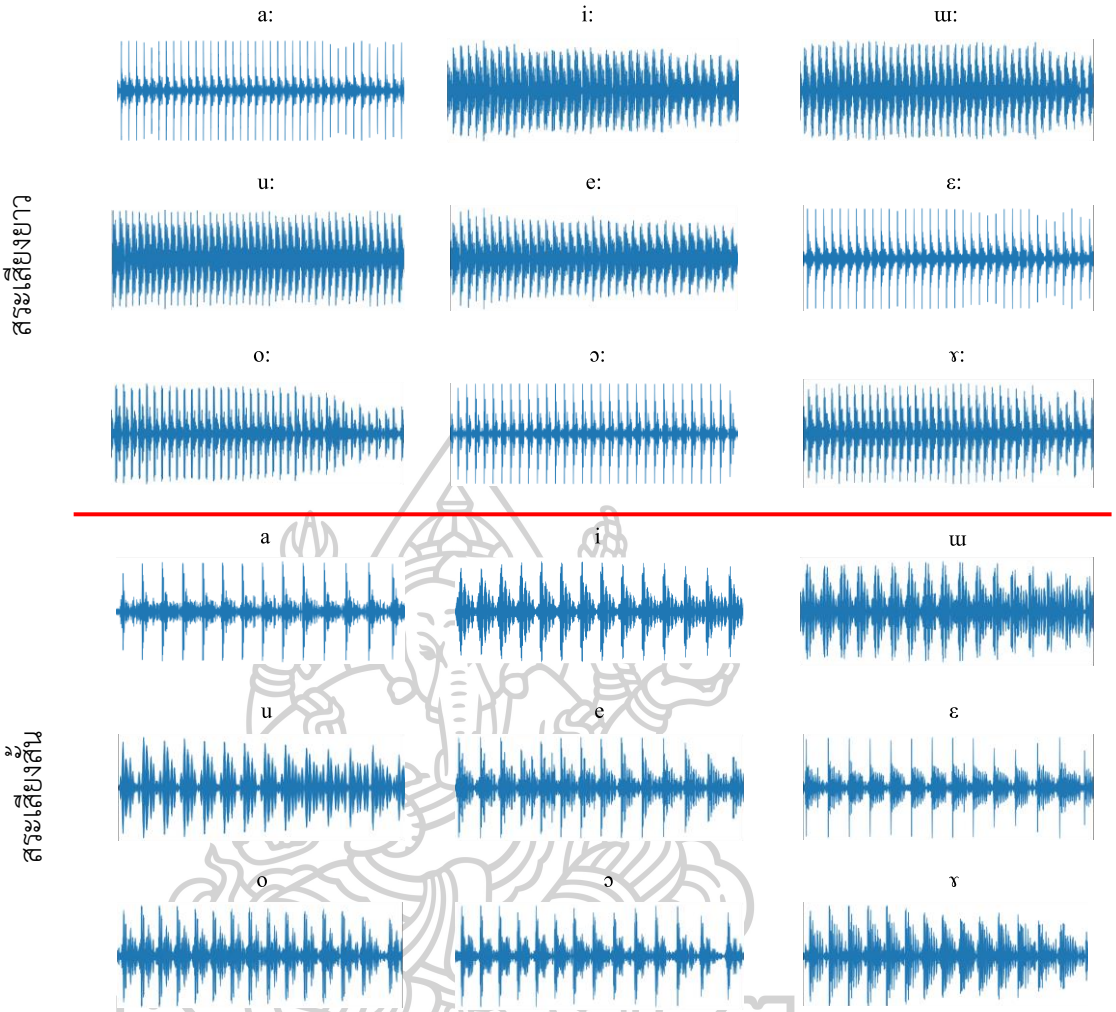
รูปที่ 24 แสดงระยะเวลาเฉลี่ยของเสียงสระภาษาไทย

รูปที่ 24 แสดงระยะเวลาเฉลี่ยของสระภาษาไทย 9 ลำดับแรกบนแกน x เป็นสระเสียงยาว และ 9 ลำดับสุดท้ายบนแกน x เป็นสระเสียงสั้น แกน y แสดงระยะเวลาของสระหน่วยเป็นวินาที จะเห็นได้ว่าสระเสียงยาวมีระยะเวลาการออกเสียงเฉลี่ยนานกว่าสระเสียงสั้น โดยที่สระเสียงยาวมีระยะเวลาเฉลี่ยประมาณ 0.3-0.4 วินาที และสระเสียงสั้นระยะเวลาเฉลี่ย 0.1-0.2 วินาที ดังนั้นจะเห็นได้ว่าความยาวของสระแต่ละสระมีความยาวต่างกัน



รูปที่ 25 แสดงระยะเวลาของสระอิในเพศหญิงและเพศชาย

รูปที่ 25 แสดงระยะเวลาของเสียงสระที่พูดโดยเพศชายและเพศหญิง รูปคลื่นเสียงของสระอิ /i/ มีระยะเวลาโดยเฉลี่ยสั้นที่สุด โดยที่แกน x คือแอมพลิจูดและแกน y คือระยะเวลา จะพบว่าเสียงพูดของทั้งสองเพศ (เพศชาย-บน เพศหญิง-ล่าง) มีความแตกต่างกัน รูปที่ 26 เป็นตัวอย่างการพูดโดยเพศชายคนหนึ่งที่ออกเสียงแต่ละสระ 18 สระครั้งเดียว พบว่าทั้ง 18 รูป มีความแตกต่างกันทั้งในด้านแอมพลิจูดและเวลา



รูปที่ 26 แสดงตัวอย่างระยะเวลาของเสียงสระภาษาไทย

จากการสำรวจไฟล์คลื่นเสียงสระภาษาไทยที่ได้รับการตรวจสอบโดยนักภาษาศาสตร์ ลักษณะของเสียงสระแต่ละเสียงจะแตกต่างกัน เนื่องจากการออกเสียงสระแต่ละเสียงมีความแตกต่างกันในหลายด้าน เช่น เพศ ลักษณะการพูด ความดัง ระยะเวลา อายุ และสิ่งแวดล้อมขณะพูด ดังนั้น การนำไฟล์เสียงสระไปใช้ควรได้รับการประมวลผลล่วงหน้า เพื่อให้ได้รูปแบบข้อมูลที่เหมาะสมสำหรับขั้นตอนต่อไป

5.1.2. การแปลงสเปกโตรแกรม (Spectrogram conversion)

สัญญาณเสียงดิบถูกแปลงเป็น waveform แล้วแปลงเป็นสเปกโตรแกรมขนาดต่างๆ เพื่อค้นหาข้อมูลเข้าคุณสมบัติอะคูสติกที่เหมาะสม สเปกโตรแกรมของภาพ 2 มิติประกอบด้วยแกนเวลาหนึ่งแกน และแกนความถี่หนึ่งแกนจะถูกแสดงด้วยลำดับของสเปกตรัม สำหรับการวิเคราะห์เสียง การแสดงสเปกตรัมจะเก็บข้อมูลมากกว่าคุณลักษณะที่สร้างขึ้นด้วยมือแบบดั้งเดิม สเปกโตรแกรมมีขนาดมิติต่ำกว่าเสียงดิบ [72]

1. การประมวลผลล่วงหน้า (Preprocessing)

การประมวลผลข้อมูลข้อมูลเข้าล่วงหน้าอย่างเหมาะสมเป็นกุญแจสำคัญอย่างหนึ่งสำหรับการเป็นตัวแทนของคุณลักษณะที่ดี สัญญาณเสียงสระไทยถูกประมวลผลล่วงหน้าโดยไลบรารี LibROSA [92] ในภาษาโปรแกรมไพทอน ซึ่งเป็นไลบารีสำหรับการวิเคราะห์เสียงและเสียงดนตรี ในขั้นตอน preprocessing สัญญาณเสียงพูดแบบโมโนโฟนิกจะลดอัตราการสุ่มตัวอย่างจาก 44,100 Hz เป็น 16,000 Hz วิธีการวิเคราะห์เสียงใช้เฟรมขนาดเล็กของสัญญาณที่เว้นระยะด้วยความยาวฮอป (hop length) ความยาวของหน้าต่างคือ 2,048 ตัวอย่าง (ประมาณ 128 มิลลิวินาที) และ 512 ตัวอย่างสำหรับขนาดฮอป (ประมาณ 32 มิลลิวินาที) สัญญาณเสียงพูดจะถูกแปลงเป็นการแสดงความถี่และเวลาโดยอิงจากการแปลงฟูเรียร์ในระยะเวลาสั้น Short-time Fourier Transform (STFT) สำหรับ discrete-time STFT คำนวณโดยใช้ Fast Fourier Transform (FFT) การแสดงทางคณิตศาสตร์ของ STFT [96] ดังสมการที่ 28

$$X[m,\omega] = \sum_{n=-\infty}^{\infty} x[n]w[n-m]exp^{-j\omega n} \quad (28)$$

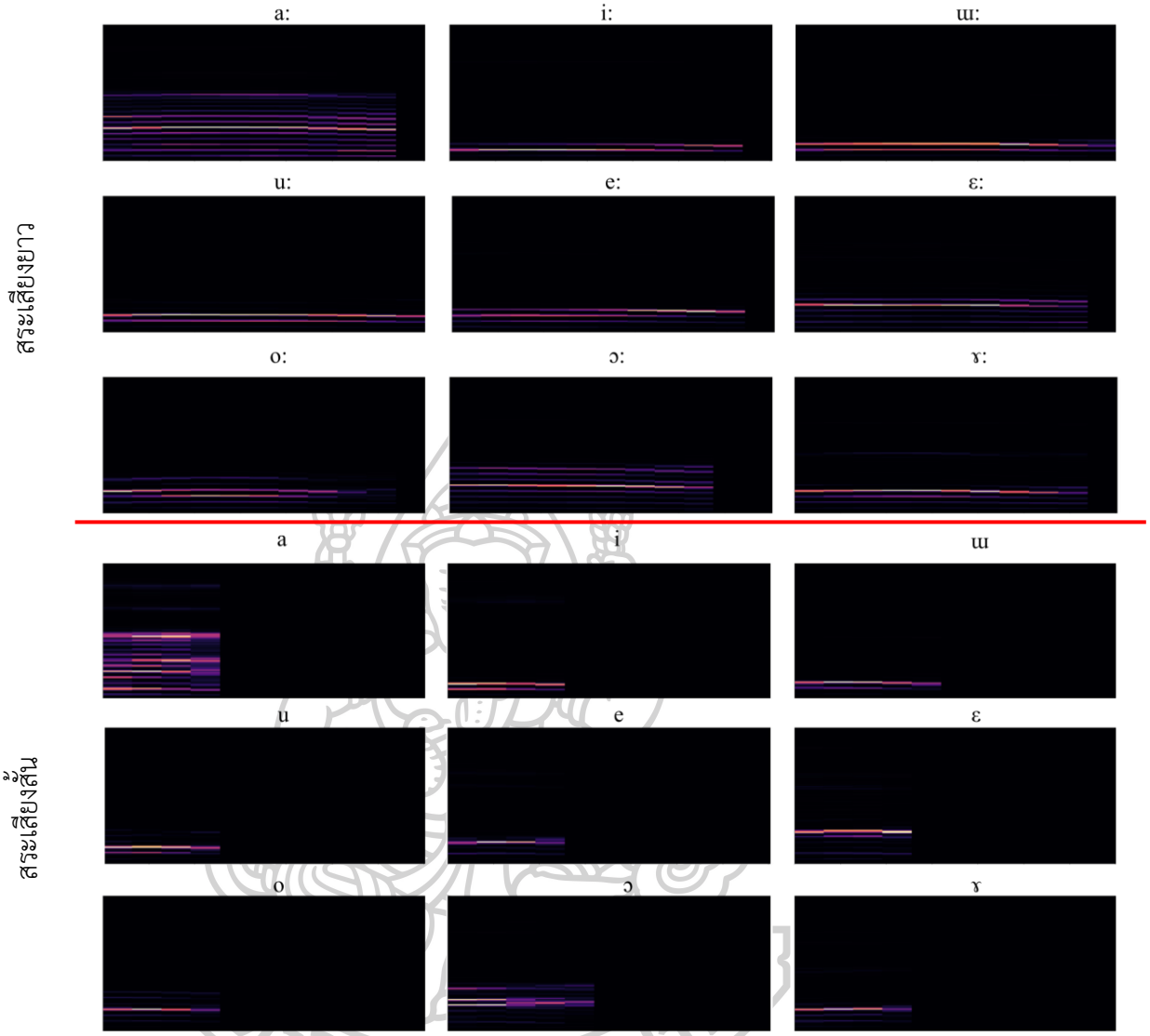
โดยที่ $x[n]$หมายถึงลำดับของสัญญาณโดเมนเวลาแบบ discretized time-domain ที่จะแปลง $w[n]$ หมายถึงฟังก์ชันหน้าต่าง m หมายถึงดัชนีเวลา ω หมายถึงความถี่ และ $X[m,\omega]$ หมายถึง STFT ลำดับของโดเมนเวลา

2. กระบวนการสกัดคุณลักษณะ (Feature extraction)

ขนาดกำลังสอง Power spectrum ของ STFT คือ linear-scaled spectrogram สำหรับ MS สเกลความถี่เชิงเส้น (linear frequency scale) ของสเปกโตรแกรมจะถูกปรับขนาดเป็น Mel scale โดยใช้ overlapping triangular filters มาตราส่วน Mel ให้มาตราส่วนเชิงเส้นสำหรับระบบการได้ยินของมนุษย์ [93] กำหนดไว้ดังสมการที่ 29

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (29)$$

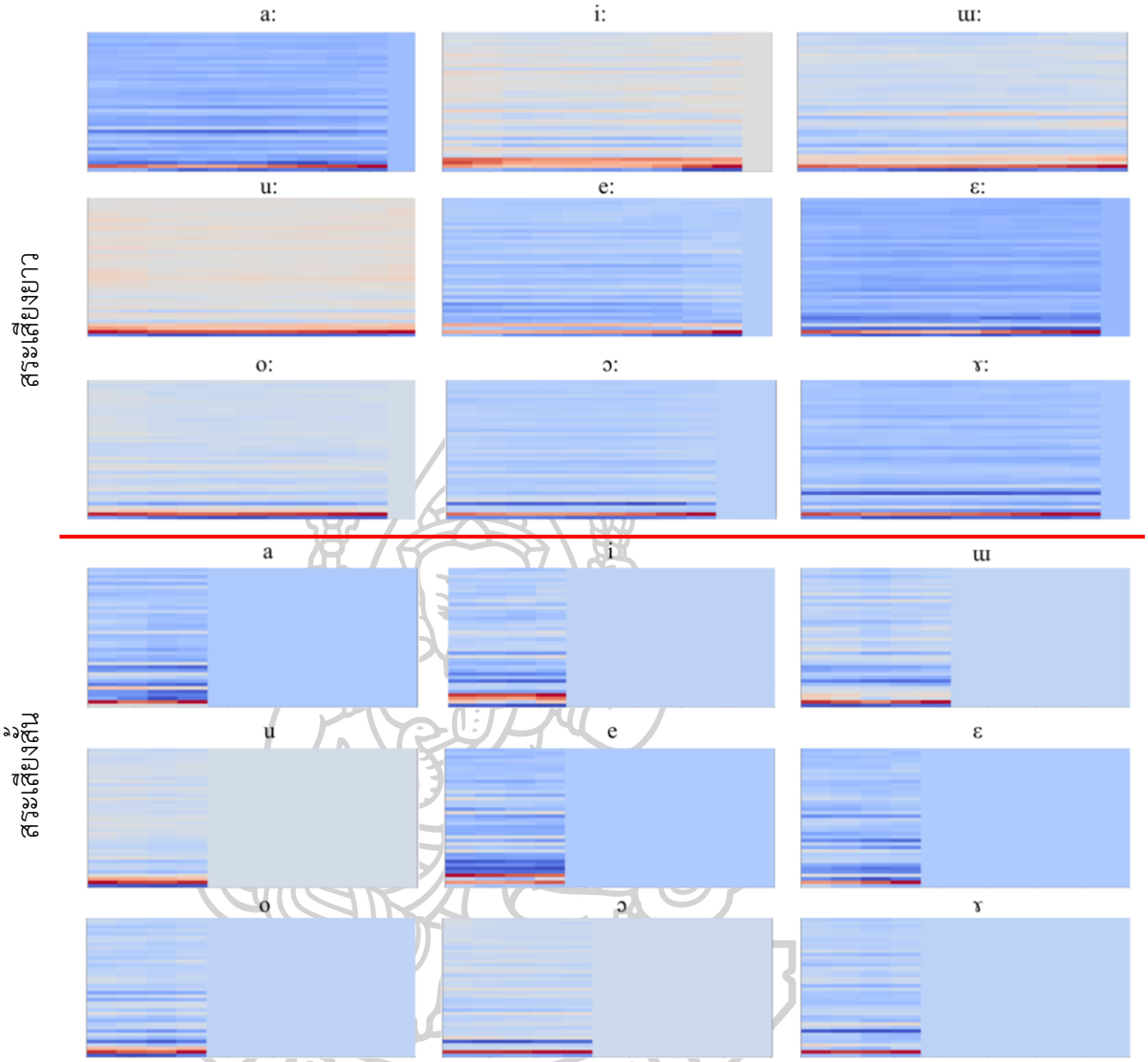
โดยที่ m หมายถึง Mels และ f หมายถึง ความถี่ เฮิรตซ์ รูปที่ 27 แสดงตัวอย่างเสียงสระภาษาไทยที่แปลงเป็นคุณสมบัติเสียง MS



รูปที่ 27 แสดงตัวอย่าง MS audio features ของสระภาษาไทย

สำหรับ MFCC ลอการิทึมของ MS จะถูกแปลงโดยใช้การแปลงโคไซน์แบบไม่ต่อเนื่อง (Discrete Cosine Transform : DCT) ผลลัพธ์ของการแปลงเรียกว่าค่าสัมประสิทธิ์ Mel Frequency Cepstrum Coefficient (MFCC) รูปที่ 28 แสดงตัวอย่างเสียงสระภาษาไทยที่แปลงเป็นคุณสมบัติเสียง MFCC

คุณลักษณะเสียงเริ่มต้นในการวิจัยนี้ ใช้ 40 Mel bands สำหรับคุณสมบัติเสียง MS, 40 MFCC สำหรับคุณสมบัติเสียง MFCC และเวกเตอร์ตามบริบท 11 [52, 74] การทดลองนี้ใช้คุณลักษณะเสียงสองแบบของกระบวนการสกัดคุณลักษณะ (MS และ MFCC) เพื่อเปรียบเทียบผลลัพธ์ด้านประสิทธิภาพจากคุณลักษณะเสียงที่แตกต่างกัน คุณลักษณะด้านเสียงถูกใช้เป็นคุณสมบัติข้อมูลเข้าสำหรับการส่งผ่านไปยังสถาปัตยกรรมการเรียนรู้เชิงลึกในกระบวนการจำแนกประเภท



รูปที่ 28 แสดงตัวอย่างของ MFCC audio features ของสระภาษาไทย

5.1.3. การจำแนกประเภทด้วยโมเดล Convolutional neural networks

ผู้เรียนฝึกออกเสียงสระไทย เสียงของผู้เรียนถูกส่งผ่านกระบวน preprocessing เพื่อแปลงคลื่นเสียงเป็นข้อมูลข้อมูลเข้าเวลาและความถี่ (time-frequency input data) จากนั้นข้อมูลข้อมูลเข้าจะถูกสกัดคุณสมบัติข้อมูลเข้า (input features) และส่งผ่านไปยังโมเดลการรู้จำเสียงสระภาษาไทยที่เป็นสถาปัตยกรรมการเรียนรู้เชิงลึก ซึ่งโมเดลนี้เป็นส่วนที่สำคัญที่สุดในการรู้จำเสียงสระของระบบการฝึกการออกเสียงอัตโนมัติสำหรับการออกเสียงสระภาษาไทย หลังจากนั้นผลการจำแนกเสียงสระจะถูกส่งไปยังขั้นตอนการเปรียบเทียบ สระที่ได้จากการรู้จำของโมเดลจะถูกนำมา

เปรียบเทียบกับสระที่เลือกโดยผู้เรียนว่าเหมือนกันหรือไม่ หากตรงกันแสดงว่าการออกเสียงของผู้เรียนนั้นถูกต้อง

Convolutional Neural Network (CNN) เป็นหนึ่งในสถาปัตยกรรมการเรียนรู้เชิงลึก CNN ไม่เพียงแต่นำมาใช้ใน computer vision เท่านั้น แต่ยังรวมถึงถูกนำมาใช้ในการรู้จำคำพูดด้วย CNN เป็นสถาปัตยกรรมที่ผสมผสานระหว่างตัวสกัดคุณลักษณะและตัวจำแนกประเภท [94] โดยทั่วไป โมเดล CNN จะประกอบด้วยเลเยอร์ Convolutional, Pooling, Normalization และ fully connected [97] Convolutional Layers ใช้เพื่อสกัดคุณลักษณะ local features จาก input data

ขั้นตอนการจำแนกประเภทถูกใช้เพื่อจำแนก class labels ในเลเยอร์ fully connected หลังจากการสกัดคุณลักษณะ สัญญาณเสียงพูดสระภาษาไทยจะถูกแปลงเป็นเวกเตอร์คุณสมบัติเสียง MS หรือ MFCC และถูกส่งไปยังโมเดล CNN เพื่อจำแนกเสียงสระภาษาไทย

1. โครงสร้างพื้นฐาน (Baseline CNN และ LSTM)

การวิจัยเริ่มต้นด้วยสถาปัตยกรรมเริ่มต้นที่ประกอบด้วยโมเดลพื้นฐาน CNN แบบตื้น (shallow CNN) ซึ่งประกอบด้วย 2 convolutional layers และ 1 pooling layer ใช้ ReLU Activation Function, Adam optimizer, และขนาด batch size เท่ากับ 32 Convolutional Layer แรกประกอบด้วย32 filters (2 × 2), ReLU Activation Function, and max pooling (2 × 2) สำหรับ Convolutional ที่สองประกอบด้วย 64 filters (2 × 2), ReLU Activation Function และไม่มี pooling layer ใน fully connected layer ประกอบด้วย 64 hidden units ในเลเยอร์สุดท้าย Softmax Activation Function ใช้สำหรับการจำแนกประเภท

ในส่วนของโมเดลพื้นฐาน LSTM เลเยอร์ LSTM ได้รับการออกแบบมาเพื่อเรียนรู้การพึ่งพาบริบทในระยะยาวของลำดับ [61] โมเดลพื้นฐาน LSTM มิติของเลเยอร์ข้อมูลเข้าถูกปรับรูปร่างใหม่ dropout (0.35) ถูกใช้หลังจากเลเยอร์ข้อมูลเข้า จากนั้นเอาท์พุตจะถูกส่งไปยังเลเยอร์ LSTM ซึ่งเลเยอร์ LSTM นั้นเหมาะสมสำหรับการสร้างแบบจำลองสัญญาณเวลา เลเยอร์ LSTM ที่หนึ่งและที่สองประกอบด้วย 512 units, tanh Activation Function, และ dropout (0.35). สุดท้าย AdamOptimizer และ Softmax Activation Function จะถูกนำไปใช้กับการจำแนกประเภท

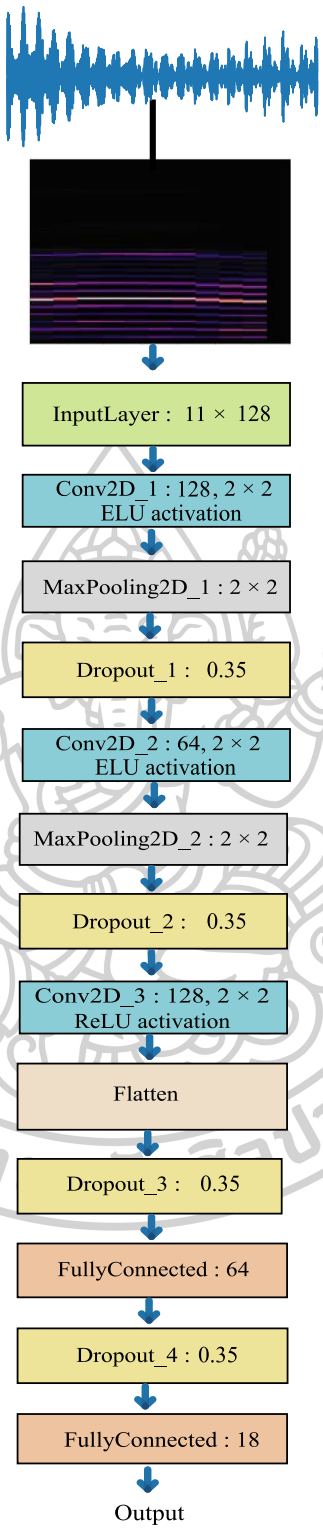
2. โครงสร้างปรับปรุง (Fine-Tuning CNN)

ในการปรับปรุงโมเดล โมเดล CNN พื้นฐานได้รับการปรับปรุงโดยเพิ่มเลเยอร์ Convolutional, max pooling, padding, และ dropout กลยุทธ์ Padding [52] สามารถรักษาขนาดของแผนที่คุณลักษณะ Pooling เป็นแนวคิดที่สำคัญในสถาปัตยกรรม CNN ที่ลดความ

แปรปรวนของสเปกตรัมในคุณสมบัติข้อมูลเข้า [51] กลยุทธ์ Dropout [91] สามารถลดปัญหาการ overfitting ได้ การกำหนดค่าไฮเปอร์พารามิเตอร์ถูกนำไปใช้กับโมเดล ใช้ Adam optimizer อัตราการเรียนรู้เริ่มต้นคือ 0.001 ขนาด batch size คือ 32 และ epoch คือ 500 ชุดข้อมูลถูกแบ่งออกเป็นชุดการฝึกอบรมและการทดสอบโดยใช้ K-fold cross-validation (k-fold = 5) แต่ละชั้นของ Convolutional ใช้ตัวกรอง Convolution กับข้อมูลเข้าคุณสมบัติเสียง ตามด้วยฟังก์ชันการกระตุ้นแบบไม่เชิงเส้น (Nonlinear Activation Function) ขนาด kernel เท่ากับ 2 ถูกกำหนดไว้สำหรับแต่ละเลเยอร์ convolutional จำนวน filters ถูกตั้งค่าเป็น 32, 64 และ 128 ในเลเยอร์ convolutional ที่แตกต่างกัน จำนวน batch sizes คือ 32, 64 และ 128 และค่า dropout ที่แตกต่างกันคือ 20%, 25%, 30%, 35%, 40%, 45%, และ 50% ถูกดำเนินการในโมเดล CNN

โมเดล Convolutional Neural Networks ที่เหมาะสมสำหรับการรู้จำสระไทย การเพิ่มเลเยอร์ Convolutional และขนาด filter การเพิ่ม max pooling การใช้ padding, dropout และ ฟังก์ชันการกระตุ้น ที่เหมาะสมสามารถปรับปรุงความแม่นยำได้ งานวิจัยนี้เสนอโมเดล CNN ที่เหมาะสมสำหรับเสียงสระภาษาไทยซึ่งประกอบด้วย 3 convolutional layers, 2 max pooling layers, 1 flatten layer, และ 2 fully connected layers โดยสถาปัตยกรรมของโมเดลจะถูกแสดงในรูปที่ 29 ซึ่งรายละเอียดของโมเดลมีดังนี้

เลเยอร์ Convolutional แรก ประกอบด้วย 128 Filters (2 x 2), Elu Activation Function, Max-Pooling (2 x 2) และ Dropout 0.35. สำหรับเลเยอร์ Convolutional ที่สอง ประกอบด้วย 64 Filters (2 x 2), Elu Activation Function, Max-Pooling (2 x 2) และ Dropout 0.35 ในเลเยอร์ Convolutional ที่สาม ประกอบด้วย 128 Filters (2 x 2), Relu Activation Function และ Dropout 0.35 แต่ไม่มี Pooling Layer สุดท้ายเลเยอร์ fully connected ประกอบด้วย 64 Hidden Units, Dropout 0.35 และ Softmax Activation Function ถูกใช้สำหรับจำแนกประเภทในเลเยอร์สุดท้าย



รูปที่ 29 แสดงสถาปัตยกรรมของข้อมูลเข้า audio features และโมเดล CNN

ข้อมูลเข้าที่เป็นคุณสมบัติเสียง MS หรือ MFCC ถูกจัดรูปแบบที่ประกอบด้วยแถว คอลัมน์ และหนึ่งช่องสัญญาณ (#frequencies, #times, 1) สำหรับป้อนเข้าสู่โมเดล CNN เวกเตอร์คุณสมบัติเสียงถูกกำหนดให้กับโหนดข้อมูลเข้าที่แตกต่างกันในเลเยอร์ 2-dimensional (2D) convolutional เลเยอร์ 2D Convolution เป็นเลเยอร์ที่สกัดรูปแบบที่สำคัญออกจากข้อมูลเข้า วัตถุประสงค์คือเพื่อสร้าง feature map ด้วย convolution filters และใช้ฟังก์ชันการกระตุ้นแบบไม่เชิงเส้น (Nonlinear Activation Function) ข้อมูลเข้าของเลเยอร์ 2D convolution คือ $x(i,j)$ ผลลัพธ์ $y(i,j)$ สามารถรับได้โดยการคอนโวลูทข้อมูลเข้า $x(i,j)$ ด้วยฟิลเตอร์คอนโวลูชั่นหรือเคอร์เนล $w(i,j)$ [61] กำหนดไว้ดังสมการที่ 30

$$y(i,j) = x(i,j) * w(i,j) \tag{30}$$

เมื่อคุณสมบัติเข้าสู่ฟังก์ชันการกระตุ้นแบบไม่เชิงเส้น ผลลัพธ์ของเลเยอร์ convolution ถูกกำหนดดังสมการที่ 31

$$y_i^l = \mathcal{O}(\sum_j y_j^{l-1} * w_{ij}^l + b_i^l) \tag{31}$$

โดยที่ y_i^l หมายถึงคุณลักษณะเอาต์พุต *i-th* ที่เลเยอร์ *l-th* และ y_j^{l-1} หมายถึงคุณสมบัติข้อมูลเข้า *j-th* ที่เลเยอร์ *(l-1)-th* และ w_{ij}^l หมายถึง convolution filter ระหว่างคุณลักษณะ *i-th* และ *j-th* และ b_i^l หมายถึง bias ที่ *i-th* ที่เลเยอร์ *l-th* และ $\mathcal{O}(\cdot)$ หมายถึง ฟังก์ชันการกระตุ้น ในการวิจัยนี้ ELU จะใช้ในการทดลองเพื่อเปรียบเทียบผลลัพธ์ด้านประสิทธิภาพกับ ReLU

หลังจากเลเยอร์ convolution และ ฟังก์ชันการกระตุ้น คุณลักษณะด้านเสียงจะถูกส่งไปยังเลเยอร์ max pooling เป้าหมายของเลเยอร์ max pooling คือการลดความละเอียดของ feature maps การพูลลิ่งเป็นแนวคิดที่สำคัญในสถาปัตยกรรม CNN ที่ลดความแปรปรวนของสเปกตรัมในคุณสมบัติข้อมูลเข้า [51] เอาต์พุตที่เลเยอร์ 2D convolutional สุดท้ายจะถูกป้อนเข้าไปในเลเยอร์ flatten และส่งผ่านไปยังเลเยอร์ fully connected สำหรับเลเยอร์ fully connected จะรวมเอาท์พุตไปยังเลเยอร์สุดท้ายสำหรับการจำแนกประเภท ในเลเยอร์สุดท้าย ฟังก์ชัน Softmax ใช้สำหรับการจำแนกประเภทหลายคลาส และเอาต์พุต Softmax ให้ความน่าจะเป็นสำหรับข้อมูลข้อมูลเข้า ผลการจำแนกหมวดหมู่ของโมเดลเป็นตัวแทนของเสียงสระภาษาไทย 18 เสียง ในงานวิจัยนี้ ผลการจำแนกประเภทมี 18 คลาส ซึ่งประกอบด้วยสระเสียงสั้น 9 เสียง และสระเสียงยาว 9 เสียง สุดท้ายมีเพียงคลาสเดียวเท่านั้นที่ถูกเลือกหลังจากขั้นตอนการจำแนกหมวดหมู่

5.1.4. รายละเอียดการทดลอง (Implementation details)

สำหรับการทดลองนี้ใช้ภาษาโปรแกรมไพทอนบนเฟรมเวิร์ก Keras และใช้ TensorFlow โดยโมเดลของงานวิจัยนี้ทดลองบน Google Colaboratory [98] ด้วย Intel (R) Xeon (R) CPU @ 2.20 GHz และ Nvidia Tesla P100 GPU

ในการวิจัยนี้เพื่อสร้างคุณสมบัติการป้อนข้อมูล MS ขนาด 11 x 128 x 1 (#times x #frequencies x #channel) กำหนดพารามิเตอร์ของสเปกโตรแกรมดังนี้ ความยาวของหน้าต่าง คือ 2,048 ความยาว Hop ระหว่างเฟรมตัวอย่างคือ 512 ช่องสัญญาณเสียง คือ 1 อัตราการสุ่มตัวอย่างเสียง 16,000 และจำนวน Mel bands คือ 128 ความถี่สูงสุดของ MS คือ 8,000 สำหรับการฝึกโมเดล จะมีการกำหนด Hyper-parameter สำหรับโมเดลโดย Optimizer คือ Adam อัตราการเรียนรู้เริ่มต้น คือ 0.001 และ Batch size คือ 32

5.2. ผลการทดลอง

ในงานวิจัยนี้ได้ศึกษาคุณลักษณะข้อมูลเข้าข้อมูลและโครงสร้างของโมเดลที่เหมาะสมสำหรับการรู้จำเสียงสระภาษาไทยของชุดข้อมูลผสม ในส่วนแรกของการทดลอง จะเป็นผลการเปรียบเทียบข้อมูลข้อมูลเข้าคุณสมบัติเสียงที่แตกต่างกับโมเดลที่แตกต่างกัน ส่วนที่สองแสดงผล confusion matrix, precision, recall, และ F1-score ของโมเดล CNN ในส่วนสุดท้ายจะแสดงการทำนายผลของโมเดล CNN กับ unseen data

5.2.1. ผลการเปรียบเทียบข้อมูลเข้าคุณสมบัติเสียงที่แตกต่างกับโมเดลที่แตกต่างกัน

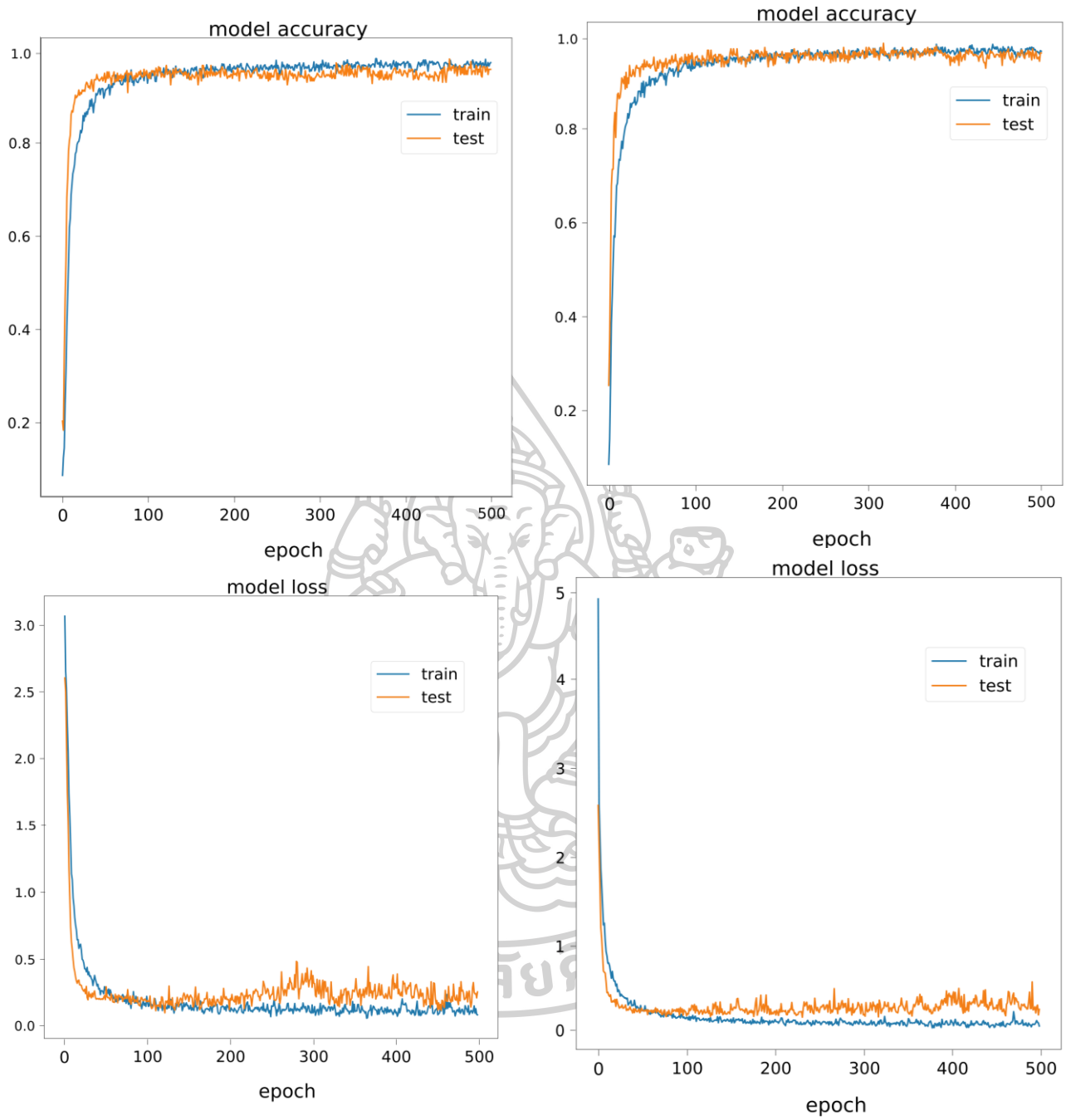
งานวิจัยนี้ใช้การผสมผสานระหว่างข้อมูลเข้าคุณสมบัติเสียง 2 แบบและโมเดล 3 โมเดล โดยมีการทดลองดังนี้ 1) คุณลักษณะด้านเสียง MFCC รวมกับโมเดลพื้นฐาน CNN 2) คุณลักษณะด้านเสียง MS รวมกับโมเดลพื้นฐาน CNN 3) คุณลักษณะด้านเสียง MFCC ที่รวมกับข้อมูลโมเดลพื้นฐาน LSTM, 4) คุณสมบัติด้านเสียง MS รวมกับโมเดลพื้นฐาน LSTM, 5) คุณสมบัติด้านเสียง MFCC รวมกับโมเดล fine-tuned CNN และ 6) คุณสมบัติด้านเสียง MS รวมกับโมเดล fine-tuned CNN

ตารางที่ 10 แสดงผลการทดลองจากการทดลองที่แตกต่างกัน (k-fold = 5)

No.	Experiment Settings	Accuracy (%)	Error of the model (Loss)
1.	MFCC + baseline CNN model	93.33	0.60
2.	MS + baseline CNN model	88.89	0.65
3.	MFCC + baseline LSTM model	94.44	0.25
4.	MS + baseline LSTM model	90.00	0.45

No.	Experiment Settings	Accuracy (%)	Error of the model (Loss)
5.	MFCC + fine-tuned CNN model	98.06	0.12
6.	MS + fine-tuned CNN model	**98.61**	**0.18**

ตารางที่ 10 แสดงผลการทดลองคุณสมบัติด้านเสียงของ MS รวมกับโมเดลพื้นฐาน CNN ทำให้ได้ค่าความถูกต้องแม่นยำต่ำสุดที่ 88.89% ในการทดลองที่สามและสี่ คุณลักษณะด้านเสียงของ MFCC หรือ MS ในโมเดลพื้นฐาน LSTM มีความถูกต้องแม่นยำต่ำที่ 94.44% และ 90.00% ตามลำดับ เลเยอร์ LSTM มีประโยชน์สำหรับการเรียนรู้การพึ่งพาบริบทในระยะยาวจากลำดับที่ยาว ในทางตรงกันข้ามการใช้เพียง LSTM กับชุดข้อมูลเสียงสระภาษาไทยที่เป็นคำพยางค์เดียวไม่โดดเด่น สำหรับงานนี้ โมเดล CNN ที่ปรับแต่งรวมกับคุณสมบัติด้านเสียงของ MS ทำให้ได้ประสิทธิภาพที่เพิ่มขึ้น โดยมีความถูกต้องแม่นยำถึง 98.61% นี่หมายความว่าการปรับจูนอย่างเหมาะสมเป็นประโยชน์ในการปรับปรุงโมเดลของเสียงสระภาษาไทย ในการทดลองนี้พบว่าผลลัพธ์ของโมเดล fine-tuned CNN แสดงให้เห็นว่าคุณสมบัติเสียงของ MS ทำงานได้ดีที่สุดสำหรับการจำแนกประเภทเสียงสระภาษาไทย



(**a**) (**b**)

รูปที่ 30 แสดง accuracy และ loss โมเดล Fine-Tuning CNN

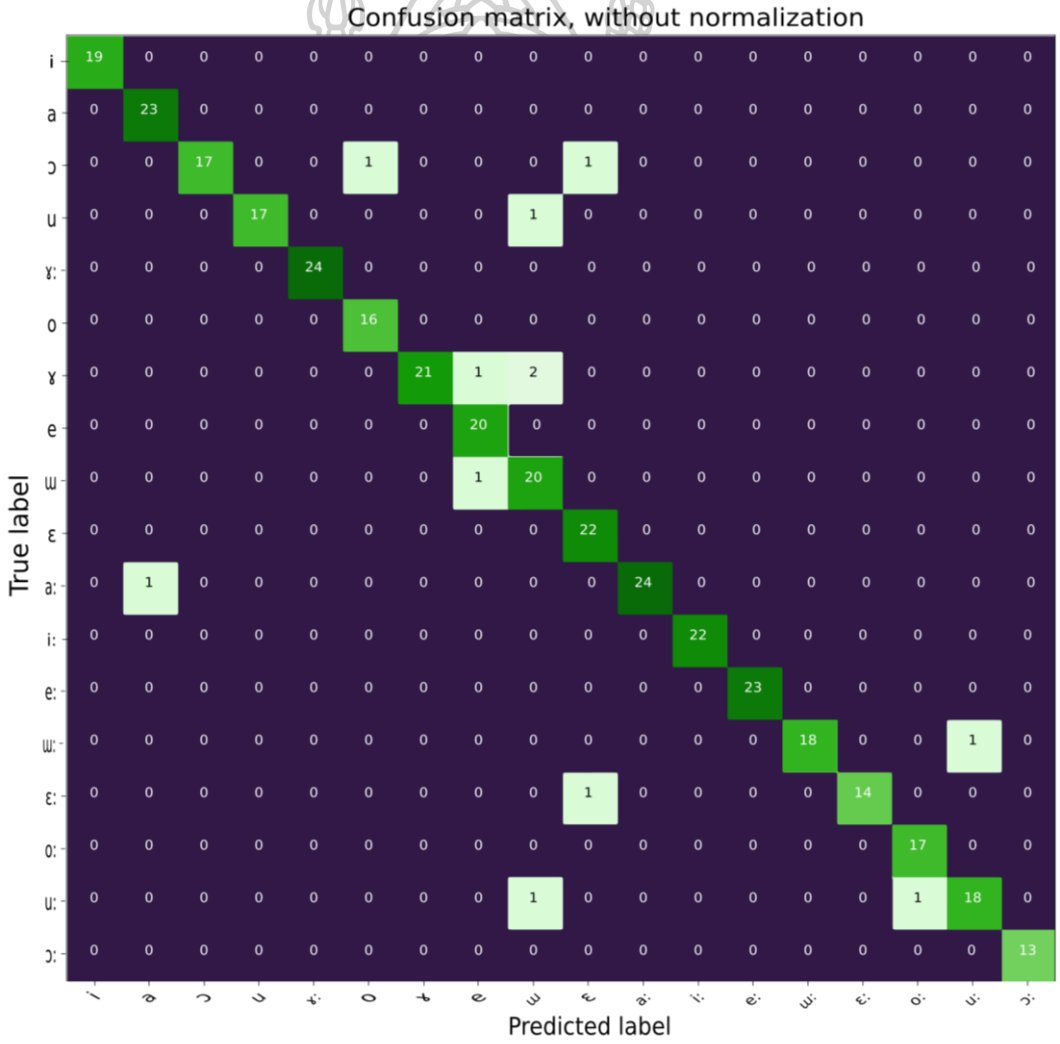
(a) MFCC + โมเดล Fine-Tuning CNN (b) MS + โมเดล Fine-Tuning CNN

กราฟเส้นของ accuracy และ loss ของโมเดล Fine-Tuning CNN รวมกับคุณสมบัติทางเสียงของ MFCC หรือ MS แสดงไว้ในรูปที่ 30 รูปภาพแสดงกราฟเส้นที่เปรียบเทียบ accuracy และ loss ของโมเดลการฝึกอบรมและการทดสอบระหว่าง 0 ถึง 500 รอบ รูปที่ 30 (a) แสดง accuracy

และ loss ของ MFCC + โมเดล Fine-Tuning CNN และค่า accuracy และ loss ของ MS + โมเดล Fine-Tuning CNN จะแสดงในรูปที่ 30 (b) สำหรับคุณสมบัติด้านเสียงของ MS โมเดลการทดสอบของโมเดล Fine-Tuning CNN จะบรรจบถึงความถูกต้องแม่นยำ 90% โมเดล Fine-Tuning CNN รวมกับ MS นั้นมีประสิทธิภาพเหนือกว่าโมเดล Fine-Tuning CNN รวมกับ MFCC และให้ความถูกต้องแม่นยำสูงสุด 98.61% โดยมีค่าการสูญเสีย 0.18 โมเดล Fine-Tuning CNN ที่ปรับแต่งมาอย่างดีสามารถลดปัญหา over-fitting ได้

5.2.2. Confusion matrix, precision, recall, และ F1-score ของโมเดล CNN

ส่วนนี้เป็นการนำเสนอการวิเคราะห์ข้อผิดพลาด โดยแสดงเมทริกซ์ความสับสนของโมเดล CNN บนชุดข้อมูลผสมซึ่งจะถูกแสดงในรูปที่ 31



รูปที่ 31 แสดง Confusion matrix ของ MS acoustic features รวมกับโมเดล CNN ในชุดข้อมูลผสม

สำหรับการวิเคราะห์ข้อผิดพลาด ใน confusion matrix ของชุดข้อมูลผสม ในโมเดล CNN สำหรับการรู้จำสระไทยแสดงไว้ในรูปที่ 31 สำหรับรายละเอียดของการจัดประเภทผิด พบว่า 10 คลาส จาก 18 คลาสมีอัตราความผิดพลาด 0% เสียงสระ ‘เออะ’ /ɤ/ เป็นคลาสที่มีการทำนายผิดสามครั้งในเมทริกซ์ความสับสน คู่เสียงสระภาษาไทยที่น่าสับสนที่สุดคือ (‘เออะ’ /ɤ/) และ (‘อึ’ /ɯ/) เสียงเหล่านี้มีความคล้ายคลึงกันซึ่งสามารถอธิบายได้ทางทฤษฎีภาษาศาสตร์เนื่องจากมีลักษณะเหมือนกัน สระ (‘เออะ’ /ɤ/) และ (‘อึ’ /ɯ/) ใช้ส่วนหลังของตำแหน่งลิ้น [99] ดังนั้นจึงอาจสร้างความสับสนให้กับโมเดลการรับรู้เสียงสระภาษาไทยได้

ตารางที่ 11 แสดง Precision, Recall, และ F1-score ของโมเดล CNN

Thai Vowels	Mixed dataset		
	Precision	Recall	F1-score
i	1.00	1.00	**1.00**
a	0.96	1.00	0.98
ɔ	1.00	0.89	0.94
u	1.00	0.94	0.97
ɤ:	1.00	1.00	**1.00**
o	0.94	1.00	0.97
ɤ	1.00	0.88	0.93
e	0.91	1.00	0.95
ɯ	0.83	0.95	0.89
ɛ	0.92	1.00	0.96
a:	1.00	0.96	0.98
i:	1.00	1.00	**1.00**
e:	1.00	1.00	**1.00**
ɯ:	1.00	0.95	0.97
ɛ:	1.00	0.93	0.97
o:	0.94	1.00	0.97
u:	0.95	0.90	0.92
ɔ:	1.00	1.00	**1.00**

ตารางที่ 11 แสดง precision, recall, และ F1-score ของโมเดล CNN สำหรับการจำแนกเสียงสระภาษาไทยแต่ละสระ สำหรับ F1-score ที่ต่ำสุดในชุดข้อมูลผสมคือ 0.89 ในสระ (‘อึ’ /ɯ/) ผล F1-score สัมพันธ์กันกับ confusion matrix ในทางกลับกัน F1-score ที่สูงที่สุดคือ 1.00 ในชุดข้อมูลแบบผสมคือสระ (‘อิ’ /i/), (‘เออ’ /ɤ:/), (‘อี’ /i:/), (‘เอ’ /e:/) และ (‘ออ’ /ɔ:/)

5.2.3. การทำนายผลของโมเดล CNN กับ unseen data

จากการวิเคราะห์ข้อผิดพลาดและการประเมินของโมเดล CNN ที่มีความแม่นยำมากกว่า 95% โมเดล CNN ที่ปรับให้เหมาะสมนี้ถูกนำมาใช้ในการฝึกการออกเสียงด้วยคอมพิวเตอร์ช่วย (CAPT) แบบอัตโนมัติซึ่งเป็นเว็บแอปพลิเคชัน ในการทดลองนี้เสียงสระภาษาไทยถูกจำแนกโดยใช้ CAPT ในสถานการณ์จริงและข้อมูลที่ใช้เป็น unseen data ผลลัพธ์ของเสียงสระที่ได้รับจากระบบถูกนำมาเปรียบเทียบกับผลการรับรู้ของนักภาษาศาสตร์และเจ้าของภาษา ชุดข้อมูล unseen data มาจากผู้ใช้ 4 คน (ชาย 2 คนและหญิง 2 คน) ทั้งหมดอายุระหว่าง 16 ถึง 30 ปี ผู้ใช้แต่ละคนฝึกออกเสียงสระ 18 เสียงและพูด 3 ครั้ง ดังนั้นชุดข้อมูล unseen data ทั้งหมดมี 216 ไฟล์เสียง (สระ 18 เสียง × 4 คน × 3 ครั้ง) จากข้อมูล unseen data ผลลัพธ์ที่รับรู้โดยระบบ พบว่ามีเสียงสระ 22 เสียง (10.19%) ไม่ตรงกับการรับรู้การฟังเสียงสระจากนักภาษาศาสตร์และเจ้าของภาษาซึ่งแสดงไว้ในตารางที่ 12 และผลลัพธ์ที่รับรู้โดยระบบ มีเสียงสระ 194 เสียง (89.81%) ตรงกันกับการรับรู้ของนักภาษาศาสตร์และเจ้าของภาษา

ตารางที่ 12 แสดงผลลัพธ์ของข้อมูลที่ unseen data ที่ไม่ตรงกับการรับรู้ของนักภาษาศาสตร์และเจ้าของภาษาของโมเดล CNN

Vowel		Perceived vowel by	
practiced pronunciation	system recognition	linguist	native speaker
ɛ:	a:	ɛ:	ɛ:
ɛ:	ɔ:	ɛ:	ɛ:
ɛ:	ɔ:	ɛ:	ɛ:
ɛ:	ɛ	ɛ:	ɛ:
ɛ:	ɤ:	ɛ:	ɛ:
e:	ɤ:	e:	e:
e:	o:	e:	e:
e:	ɤ:	e:	e:
i:	e:	i:	i:
i:	e:	i:	i:

Vowel		Perceived vowel by	
practiced pronunciation	system recognition	linguist	native speaker
ɤ:	o:	ɤ:	ɤ:
ɤ:	ɯ:	ɤ:	ɤ:
ɯ:	ɔ:	ɯ:	ɯ:
ɯ:	ɔ:	ɯ:	ɯ:
ɯ:	ɤ:	ɯ:	ɯ:
ɯ:	i:	ɯ:	ɯ:
i	e	i	i
o:	ɛ:	o:	o:
o:	o	o:	o:
u	ɯ	u	u
u:	o:	u:	u:
u:	i:	u:	u:

ตารางที่ 13 แสดงความถี่ของคู่ที่ทำนายผิดสำหรับสระภาษาไทยที่ใช้โมเดล CNN

Vowel		Frequency
practiced pronunciation	system recognition	
ɛ:	a:	1
ɛ:	ɔ:	2
ɛ:	ɛ	1
ɛ:	ɤ:	1
e:	ɤ:	2
e:	o:	1
i:	e:	2
ɤ:	o:	1
ɤ:	ɯ:	1
ɯ:	ɔ:	2

Vowel		Frequency
practiced pronunciation	system recognition	
ɯ:	ɤ:	1
ɯ:	i:	1
i	e	1
o:	ɛ:	1
o:	o	1
u	ɯ	1
u:	o:	1
u:	i:	1

ตารางที่ 13 แสดงคู่สระที่ระบบทำนายผิดมากที่สุด ซึ่งไม่ตรงกับการรับรู้เสียงสระของนักภาษาศาสตร์หรือเจ้าของภาษา ได้แก่ (‘แอ’/**ɛ**:/) และ (‘ออ’/**ɔ**:/), (‘เอ’/e:/) และ (‘เออ’ /**ɤ**:/), (‘อี’/i:/) และ (‘เอ’/ e:/) และ (‘อือ’/**ɯ**:/) และ (‘ออ’/ **ɔ**:/) ซึ่งแต่ละคู่มีความถี่การออกเสียงที่ไม่ตรงกัน 2 ครั้ง สิ่งเหล่านี้สามารถอธิบายได้ในทฤษฎีภาษาศาสตร์ว่าคู่ที่ออกเสียงผิดนั้นสัมพันธ์กับบริเวณตำแหน่งลิ้นที่คล้ายคลึงกัน ซึ่งอยู่หน้า-หลัง และ สูง-ต่ำ ของลิ้น

5.3. การฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย

การฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย แสดงในรูปที่ 32 รายละเอียดของระบบนี้นำเสนอดังนี้

1) ผู้ใช้เลือกสระภาษาไทยที่ต้องการฝึกออกเสียง เช่น สระอะ /a/

2) จากนั้นระบบจะไปที่หน้าฝึกการออกเสียงสระอะ /a/ หน้านี้มีคำอธิบายสระ 3 ประเภท ดังนี้ คลิปวิดีโอการออกเสียง 3D ของสระอะ /a/ ที่แสดงการเคลื่อนไหวของลิ้น, การบรรยายสระ เช่น ระยะเวลาของสระ (สั้นหรือยาว) รูปปาก (กลมหรือไม่กลม) และลักษณะลิ้น (ต่ำ กลาง หรือสูง), คลิปวีดิโอของคนหน้าตรงที่แสดงให้เห็นลักษณะปากที่เด่นชัด

3) เมื่อผู้ใช้ต้องการฝึกการออกเสียง สามารถกดไอคอนไมโครโฟนและพูดได้ (สัญญาณเสียงพูดเป็นสัญญาณเสียงอะนาล็อก)

4) ขั้นตอนก่อนการประมวลผลจะแปลงสัญญาณเป็นดิจิทัลและแปลงเสียงเป็นโดเมนเวลาและความถี่

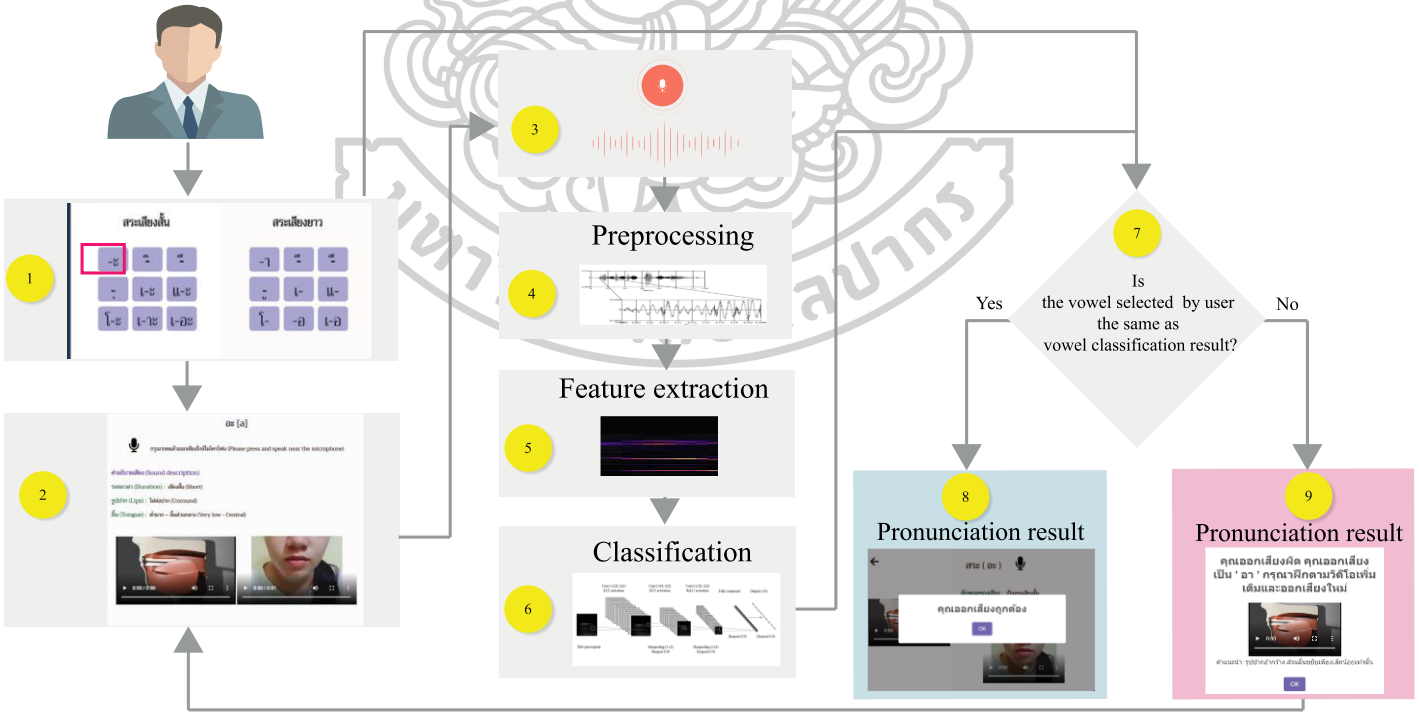
5) หลังจากนั้น คุณสมบัติต่างๆ จะถูกดึงออกมาในระหว่างขั้นตอนการสกัดคุณลักษณะ และค่าของคุณสมบัติดังกล่าวจะกำหนดลักษณะหน่วยเสียงในคุณสมบัติด้านเสียง

6) จากนั้น ใช้คุณสมบัติเสียงในกระบวนการจำแนกประเภท ในขั้นตอนนี้ ใช้สถาปัตยกรรมการเรียนรู้เชิงลึกในการจดจำเสียงสระภาษาไทย ผลลัพธ์มี 18 คลาส (สระเสียงสั้น 9 เสียง และสระเสียงยาว 9 เสียง) หลังจากขั้นตอนการจำแนกประเภท มีเพียงหนึ่งคลาสเท่านั้นที่ถูกเลือกเพื่อส่งออกเป็นผลการจำแนกประเภทเสียงสระ

7) สุดท้าย การเปรียบเทียบระหว่างสระที่เลือกโดยผู้ใช้และผลการจำแนกประเภทสระ ขั้นตอนนี้เปรียบเทียบสระที่ผู้ใช้เลือกสำหรับฝึกการออกเสียงกับผลจากการรู้จำเสียงสระภาษาไทย

8) หากสระที่ผู้ใช้เลือก (เช่น /a/) และผลการจำแนกสระ (เช่น /a/) เหมือนกัน ผลการออกเสียงจะแสดงข้อความว่า “การออกเสียงของคุณถูกต้อง”

9) ในทางตรงกันข้าม ถ้าสระที่ผู้ใช้เลือกและผลการจำแนกสระไม่เหมือนกัน เช่น ผู้ใช้เลือก /a/ แต่ผลการจำแนกสระคือ /o/ ผลการออกเสียงจะแสดงข้อความ “การออกเสียงของคุณไม่ถูกต้อง และแสดง “คุณออกเสียงคือ /o/” แทน นอกจากนี้ ข้อความจะแสดงคลิปวิดีโอเสียงสระ 3 มิติของการออกเสียงและข้อความที่อธิบายว่าออกเสียงที่ถูกต้องเป็นอย่างไร ผู้ใช้สามารถกลับไปที่หน้าฝึกการออกเสียงสระเพื่อฝึกซ้ำตามลูกศรด้านล่าง



รูปที่ 32 แสดง ระบบ Computer-Assisted Pronunciation Training สำหรับสระภาษาไทย

5.4. สรุป

สระเป็นแกนหลักของพยางค์ (นิวเคลียส) และเป็นส่วนสำคัญของคำพูด สระเกิดขึ้นในช่องปากขึ้นอยู่กับตำแหน่งของลิ้น การฝึกออกเสียงสระจึงเป็นเรื่องยากสำหรับผู้เรียนหรือผู้ที่ไม่ได้เป็นเจ้าของภาษาที่จะเข้าใจได้ง่ายด้วยตนเอง ซึ่งต้องมีผู้เชี่ยวชาญให้คำแนะนำ แต่ในปัจจุบันนี้ผู้เชี่ยวชาญในการสอนมักมีไม่เพียงพอ เพื่อแก้ปัญหาเหล่านี้ ควรมีการนำเทคโนโลยีสำหรับฝึกการออกเสียงมาใช้ งานวิจัยนี้นำเสนอคุณลักษณะด้านเสียงและโมเดล CNN ที่เหมาะสมสำหรับการรู้จำเสียงสระภาษาไทยที่มีเสียงรบกวนซึ่งใช้ในระบบ CAPT อัตโนมัติ

ระบบ CAPT ได้รับการพัฒนาสำหรับกิจกรรมการเรียนรู้ในชีวิตประจำวันที่สามารถฝึกฝนได้ทุกที่ทุกเวลา ดังนั้นชุดข้อมูลเสียงสระไทยที่มีเสียงรบกวนจึงถูกรวบรวมจากเจ้าของภาษาในสถานการณ์จริง โดยมีความแตกต่างกันในมิติต่างๆ ในบริบทชีวิตจริง เช่น เพศ อายุ สำเนียง สิ่งแวดล้อม เสียงรบกวน เป็นต้น นอกจากนี้ ชุดข้อมูลยังได้รับการออกแบบ รวบรวม และตรวจสอบโดยนักภาษาศาสตร์ตามทฤษฎีภาษาศาสตร์ โมเดล 2D-CNN ร่วมกับคุณสมบัติเสียงของ MS ช่วยเพิ่มประสิทธิภาพในการรู้จำเสียงสระภาษาไทย โดยบรรลุความแม่นยำ 98.61% โดยใช้กลยุทธ์ที่หลากหลายและการปรับแต่งไฮเปอร์พารามิเตอร์ สุดท้าย โมเดลนี้ถูกนำไปใช้กับระบบ CAPT ในสถานการณ์จริง ข้อมูลที่ป้อนเข้าจากผู้เรียนถือเป็นข้อมูล unseen data ผลลัพธ์ของเสียงสระที่ได้รับจากระบบ CAPT นั้นถูกนำมาเปรียบเทียบกับเสียงสระที่รับรู้โดยนักภาษาศาสตร์และเจ้าของภาษา และได้ผลลัพธ์ความถูกต้องแม่นยำ 89.81% การสกัดคุณสมบัติเสียงสระที่ใช้ MS ร่วมกับ CNN ให้คุณสมบัติทางเสียงที่โดดเด่นสำหรับการรู้จำเสียงสระภาษาไทย โมเดลนี้สามารถแยกแยะเสียงสระได้แม้ว่าข้อมูลจะมีเสียงรบกวน อายุ สำเนียง สภาพแวดล้อม และลักษณะทางกายภาพต่างกัน (เช่น เสียงผู้หญิงกับผู้ชาย)

ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยใช้โมเดล CNN ร่วมกับคุณสมบัติเสียง MS ที่เหมาะสมสำหรับการรู้จำเสียงของสระภาษาไทย สามารถแก้ปัญหาต่างๆ เช่น การขาดความเชี่ยวชาญ ความซับซ้อน กระบวนการที่ใช้เวลานาน ไม่มีขั้นตอนแบบครบวงจร หรือไม่ใช่กระบวนการแบบเรียลไทม์ ผลงานวิจัยนี้เป็นประโยชน์ต่อผู้มีส่วนได้ส่วนเสียที่สนใจในการพัฒนาระบบเสียงสระภาษาไทยหรือระบบการออกเสียงที่คล้ายคลึงกัน ซึ่งจะช่วยให้นักวิจัยสามารถผลิตระบบการเรียนรู้โดยประยุกต์ตามการดำเนินการที่คล้ายคลึงกัน นอกจากนี้ยังสามารถให้คำแนะนำผู้เรียน เช่น ผู้เรียนที่ไม่ใช่เจ้าของภาษา ผู้พิการทางเสียง หรือผู้ที่พูดภาษาไทยไม่ได้มาตรฐาน จึงทำให้ผู้เรียนสามารถฝึกการออกเสียงสระแบบเรียลไทม์ได้ทุกที่ทุกเวลาเหมือนมีผู้เชี่ยวชาญ ครูภาษาไทย และนักภาษาศาสตร์ให้คำแนะนำในการออกเสียงที่ถูกต้องตลอดเวลา

บทที่ 6

วิธีดำเนินงานวิจัยและผลการทดลองที่ 3

Gradient-weighted class activation mapping สำหรับโมเดล Convolutional Neural Network และระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง

Gradient-weighted class activation mapping (Grad-CAM) สามารถอธิบายถึงความสำคัญของบริเวณของพื้นที่ที่มีความสำคัญในการทำนายของโมเดล Convolutional Neural Network (CNN) โดยผลการแสดงมีลักษณะเป็นภาพ Grad-CAM เป็นกระบวนการที่ช่วยให้มนุษย์สามารถเข้าใจผลลัพธ์ที่สร้างขึ้นโดยอัลกอริทึมการเรียนรู้ของปัญญาประดิษฐ์เพื่ออธิบายโมเดล (explainable AI) ช่วยทวนสอบความถูกต้องและให้ความสมเหตุผลของการทำนายแต่ละครั้งได้ ซึ่งนิยมแพร่หลายในงานวิจัยทางด้านคอมพิวเตอร์วิทัศน์ (Computer vision) เนื่องจากสามารถทำให้ทราบว่าโมเดล CNN ใช้คุณลักษณะจากบริเวณใดในภาพมาช่วยตัดสินใจในขณะที่ใช้งาน ซึ่งใช้เฉพาะเจาะจงกับข้อมูลเข้าแต่ละอัน แต่สำหรับงานทางด้านการรู้จำเสียงยังพบได้น้อยในการนำข้อดีของ Grad-CAM มาประยุกต์ใช้ โดยเฉพาะการรู้จำเสียงสระ

โมเดลการเรียนรู้เชิงลึกที่ใช้ CNN เพื่อจดจำการออกเสียงสระภาษาไทยในงานนี้ สามารถนำไปพัฒนาระบบคอมพิวเตอร์ช่วยฝึกการออกเสียงแบบอัตโนมัติสำหรับเสียงสระภาษาไทยได้ การพัฒนาระบบนี้เพื่อแก้ปัญหาความไม่เพียงพอของผู้เชี่ยวชาญด้านการสอนและความซับซ้อนของกระบวนการสอนการออกเสียงสระ เป็นระบบใหม่ที่พัฒนาเทคนิคคอมพิวเตอร์ผสมผสานกับภาษาศาสตร์ ซึ่งระบบนี้ทำให้ผู้เรียนได้ฝึกการออกเสียงสระแบบเรียลไทม์ เสมือนมีผู้เชี่ยวชาญ ครูภาษาไทย และนักภาษาศาสตร์คอยให้คำแนะนำเกี่ยวกับการออกเสียงสระที่ถูกต้องอย่างต่อเนื่อง เหมาะกับสถานการณ์โลกในปัจจุบันที่ต้องมีการเรียนในรูปแบบลักษณะออนไลน์ ประสิทธิภาพและความถูกต้องในการพัฒนาระบบเป็นปัจจัยสำคัญที่จะทำให้ระบบสามารถนำไปใช้ได้จริงอย่างถูกต้อง

ดังนั้นในงานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอการใช้ Grad-CAM กับโมเดลการเรียนรู้เชิงลึกที่ใช้ CNN สำหรับการรู้จำเสียงสระภาษาไทยในสถานการณ์จริง เพื่ออธิบายปัจจัยที่สำคัญเมื่อโมเดลทำนายเสียงสระภาษาไทยทั้ง 18 คลาส ซึ่งประกอบด้วย สระเสียงสั้น 9 เสียงและสระเสียงยาว 9 เสียง ประโยชน์ของการวิจัยนี้สามารถนำไปพัฒนาเป็นระบบคอมพิวเตอร์ช่วยฝึกการออกเสียงแบบอัตโนมัติสำหรับเสียงสระภาษาไทยได้โดยช่วยในเรื่องความโปร่งใสในการทำนายผลของโมเดล CNN ที่ใช้ในการรู้จำเสียงสระภาษาไทย ทำให้ระบบมีประสิทธิภาพและความถูกต้องมากยิ่งขึ้น

6.1. ชุดข้อมูลและวิธีการ (Datasets and Methods)

6.1.1. ชุดข้อมูล (Dataset)

ชุดข้อมูลที่ใช้สำหรับการฝึกฝนโมเดลการเรียนรู้เชิงลึกและ Grad-CAM เป็นชุดข้อมูลที่ได้รับการการออกแบบ รวบรวม และตรวจสอบโดยนักภาษาศาสตร์ โดยชุดข้อมูลมีทั้งหมด 3 ชุด ได้แก่ ชุดข้อมูลเพศชาย (Male dataset) ชุดข้อมูลเพศหญิง (Female dataset) และชุดข้อมูลผสม (Mixed dataset) ซึ่งเป็นการผสมผสานระหว่างชุดข้อมูลเสียงเพศหญิงและเพศชาย โดยชุดข้อมูลเสียงสระภาษาไทยได้รับการรวบรวมจากผู้พูดภาษาไทยมาตรฐานอายุ 20-25 ปี จำนวน 50 คน เป็นเพศชายจำนวน 25 คน เป็นเพศหญิง 25 คน ซึ่งในชุดข้อมูลมีทั้งเสียงเงียบและเสียงรบกวน เช่น คนคุยกันในโรงอาหาร เสียงสัตว์ เสียงดนตรี เสียงลม และรถยนต์บนท้องถนน บันทึกจากโทรศัพท์มือถือที่ 44,100 Hz (standard speech data) โดยรวบรวมจากเจ้าของภาษาในสถานการณ์จริง หลังจากการบันทึกเสียงสำเร็จ ไฟล์เสียงทั้งหมดจะถูกส่งไปยังนักภาษาศาสตร์เพื่อตรวจสอบ และได้คัดเลือกเสียงสระที่มีความสมบูรณ์ โดยนักภาษาศาสตร์ได้ตัดไฟล์เสียงโดยใช้ Praat [27] จำนวนเสียงสระคุณภาพดีทั้งหมดที่ได้รับคือ 1,800 ไฟล์เสียง แบ่งออกเป็นเสียงผู้ชาย 900 เสียง (สระ 18 เสียง × ผู้ชาย 25 คน × พูด 2 ครั้ง) และเสียงผู้หญิง 900 เสียง (สระ 18 เสียง × ผู้หญิง 25 คน × พูด 2 ครั้ง) ชุดข้อมูลทั้ง 18 คลาสประกอบด้วยสระเสียงสั้น 9 เสียง และสระเสียงยาว 9 เสียง การเทรนโมเดลชุดข้อมูลถูกแบ่งออกเป็นชุดการฝึกอบรมและการทดสอบโดยใช้ K-fold cross-validation (k-fold = 5)

6.1.2. การแปลงสเปกโตรแกรม (Spectrogram conversion)

การประมวลผลล่วงหน้าและการสกัดคุณลักษณะ (Preprocessing and feature extraction) สัญญาณเสียงดิบถูกแปลงเป็นคลื่นเสียง (waveform) แล้วแปลงเป็นสเปกโตรแกรมขนาดต่างๆ เพื่อค้นหาข้อมูลเข้าคุณสมบัติอะคูสติกที่เหมาะสม สเปกโตรแกรมของภาพ 2 มิติประกอบด้วยแกนเวลาหนึ่งแกน และแกนความถี่หนึ่งแกนจะถูกแสดงด้วยลำดับของสเปกตรัม ซึ่งในงานวิจัยนี้แกน x แทนเวลา และแกน y แทนความถี่ ชุดข้อมูลที่ใช้สำหรับการฝึกฝนโมเดลการเรียนรู้เชิงลึกและ Grad-CAM จำนวน 1,800 ไฟล์เสียง ถูกนำมาผ่านการประมวลผลข้อมูลข้อมูลเข้าล่วงหน้า เพื่อให้ได้รูปแบบข้อมูลที่เหมาะสมสำหรับนำไปใช้ในขั้นตอนการจำแนกประเภทต่อไป สัญญาณเสียงสระไทยถูกประมวลผลล่วงหน้า และสกัดคุณลักษณะ MS เช่นเดียวกับการทดลองในบทที่ 5

6.1.3. การจำแนกประเภทด้วยสถาปัตยกรรม Convolutional neural networks

เวกเตอร์คุณสมบัติเสียง MS ถูกส่งไปยังสถาปัตยกรรมเชิงลึกเพื่อจำแนกเสียงสระภาษาไทยด้วยสถาปัตยกรรม Convolutional Neural Network (CNN) ขั้นตอนการจำแนกหมวดหมู่ถูกใช้เพื่อจำแนก class labels ในเลเยอร์ fully connected ซึ่งในงานวิจัยนี้ใช้โมเดล CNN สำหรับการรู้จำเสียงสระภาษาไทยที่ ประกอบด้วย Convolutional 3 เลเยอร์ และ Fully Connected 2 เลเยอร์ โดยเลเยอร์ Convolutional แรก ประกอบด้วย 128 Filters (2 x 2), Elu Activation Function, Max-Pooling (2 x 2) และ Dropout rate 0.35 สำหรับเลเยอร์ Convolutional ที่สอง ประกอบด้วย 64 Filters (2 x 2), Elu Activation Function, Max-Pooling (2 x 2) และ Dropout rate 0.35 ในเลเยอร์ Convolutional ที่สาม ประกอบด้วย 128 Filters (2 x 2), Relu Activation Function และ Dropout rate 0.35 แต่ไม่มี Pooling Layer สุดท้ายเลเยอร์ fully connected ประกอบด้วย 64 Hidden Units, Dropout rate 0.35 และ Softmax Activation Function ถูกใช้สำหรับจำแนกประเภทในเลเยอร์สุดท้ายดังรูปที่ 29

ข้อมูลเข้าที่เป็นคุณสมบัติเสียง MS ถูกจัดรูปแบบที่ประกอบด้วยแถว คอลัมน์ และหนึ่งช่องสัญญาณ (#frequencies, #times, 1) นั่นคือ 128, 11, 1 สำหรับป้อนเข้าสู่โมเดล CNN เวกเตอร์คุณสมบัติเสียงถูกกำหนดให้กับโหนดข้อมูลเข้าที่แตกต่างกันในเลเยอร์ 2-dimensional (2D) convolutional ซึ่งเป็นเลเยอร์ที่สกัดรูปแบบที่สำคัญออกจากข้อมูลเข้า วัตถุประสงค์คือเพื่อสร้าง feature map ด้วย convolution filters และใช้ฟังก์ชันการกระตุ้นแบบไม่เชิงเส้น (Nonlinear Activation Function) ข้อมูลเข้าของเลเยอร์ 2D convolution คือ $x(i,j)$ ผลลัพธ์ $y(i,j)$ สามารถรับได้โดยการคอนโวลูทข้อมูลเข้า $x(i,j)$ ด้วยฟิลเตอร์คอนโวลูชั่นหรือเคอร์เนล $w(i,j)$ [61] กำหนดไว้ดังสมการที่ 33

$$y(i,j) = x(i,j) * w(i,j) \tag{33}$$

เมื่อคุณสมบัติเข้าสู่ฟังก์ชันการกระตุ้นแบบไม่เชิงเส้น ผลลัพธ์ของเลเยอร์ convolution ถูกกำหนดดังสมการที่ 34

$$y_i^l = \phi(\sum_j y_j^{l-1} * w_{ij}^l + b_i^l) \tag{34}$$

โดยที่ y_i^l หมายถึงคุณลักษณะเอาต์พุต *i-th* ที่เลเยอร์ *l-th* และ y_j^{l-1} หมายถึงคุณสมบัติข้อมูลเข้า *j-th* ที่เลเยอร์ (*l-1)-th* และ w_{ij}^l หมายถึง convolution filter ระหว่างคุณลักษณะ *i-th* และ *j-th* และ b_i^l หมายถึง bias ที่ *i-th* ที่เลเยอร์ *l-th* และ $\phi(\cdot)$ หมายถึง ฟังก์ชันการกระตุ้น

หลังจากเลเยอร์ convolution และ ฟังก์ชันการกระตุ้น คุณลักษณะด้านเสียงจะถูกส่งไปยังเลเยอร์ max pooling เป้าหมายของเลเยอร์ max pooling คือการลดความละเอียดของ feature

maps การพูลลิ่งเป็นแนวคิดที่สำคัญในสถาปัตยกรรม CNN ที่ลดความแปรปรวนของสเปกตรัมในคุณสมบัติข้อมูลเข้า [51] เอาต์พุตที่เลเยอร์ 2D convolutional สุดท้ายจะถูกป้อนเข้าไปในเลเยอร์ flatten และส่งผ่านไปยังเลเยอร์ fully connected สำหรับเลเยอร์ fully connected จะรวมเอาท์พุตไปยังเลเยอร์สุดท้ายสำหรับการจำแนกประเภท ในเลเยอร์สุดท้าย ฟังก์ชัน Softmax ใช้สำหรับการจำแนกประเภทหลายคลาส และเอาต์พุต Softmax ให้ความน่าจะเป็นสำหรับข้อมูลข้อมูลเข้า ผลการจำแนกหมวดหมู่ของโมเดลเป็นตัวแทนของเสียงสระภาษาไทย 18 เสียง โมเดลนี้ใช้กลยุทธ์ padding, Adam Optimizer และขนาด batch เหมือนกับ [74] การกำหนดค่าไฮเปอร์พารามิเตอร์ถูกนำไปใช้กับโมเดล ใช้ Adam optimizer อัตราการเรียนรู้เริ่มต้นคือ 0.001 ขนาด batch size คือ 32 และ epoch คือ 500 ชุดข้อมูลถูกแบ่งออกเป็นชุดการฝึกอบรมและการทดสอบโดยใช้ K-fold cross-validation (k-fold = 5) ผลการจำแนกประเภทมี 18 คลาส ซึ่งประกอบด้วยสระเสียงสั้น 9 เสียง และสระเสียงยาว 9 เสียง สุดท้ายมีเพียงคลาสเดียวเท่านั้นที่ถูกเลือกหลังจากขั้นตอนการจำแนกหมวดหมู่ โมเดลผลลัพธ์ถูกบันทึกในรูปแบบ h5

6.1.4. Gradient-weighted class activation mapping (Grad-CAM) สำหรับโมเดล CNN

ในงานวิจัยทางด้านคอมพิวเตอร์วิทัศน์ได้นำ Grad-CAM มาประยุกต์ใช้เพื่อทำให้ทราบว่าโมเดล CNN ใช้คุณลักษณะจากบริเวณใดในภาพมาช่วยตัดสินใจในขณะที่ใช้งาน ซึ่งใช้เฉพาะเจาะจงกับข้อมูลเข้าแต่ละคลาส เช่น ฟีทเจอร์ใดที่ทำให้รู้สึกว่าภาพนี้คล้ายสุนัข หรือ แมว เป็นต้น โดย Grad-CAM สามารถอธิบายถึงความสำคัญของบริเวณของพื้นที่ที่มีความสำคัญ ฟีทเจอร์ที่สนับสนุนในการทำนายผลของโมเดล CNN แต่ละคลาสว่าอยู่ตำแหน่งใด โดยผลการแสดงมีลักษณะเป็นภาพ ซึ่งแต่ละคลาสก็จะได้พื้นที่ในภาพที่แตกต่างกันไป ในด้านการรู้จำเสียงของงานวิจัยนี้ เพื่อให้ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียงมีประสิทธิภาพและความถูกต้องมากยิ่งขึ้น จึงใช้ Grad-CAM เพื่ออธิบายถึงบริเวณพื้นที่ที่สำคัญเมื่อโมเดลทำนายเสียงสระภาษาไทยทั้ง 18 คลาส ซึ่ง Grad-CAM สามารถอธิบายในกรณีของเสียงที่แต่ละคนพูดออกมา โดยแสดงให้เห็นได้ว่าฟีทเจอร์ที่สนับสนุนว่าผู้พูดออกเสียงถูกอยู่ตรงตำแหน่งใด ชัดแค่ไหน ดังรูปที่ 40 และฟีทเจอร์ที่ทำให้ผู้พูดออกเสียงผิดอยู่ตรงตำแหน่งใด และชัดแค่ไหน ดังรูปที่ 41 ซึ่งสามารถนำผลที่ได้ไปเปรียบเทียบกับสถิติรวมของคนทั้งหมด หรือนำไปเทียบเฉพาะกับเพศชาย หรือเพศหญิง หรือกลุ่มภาษาถิ่นของแต่ละคนได้ ทำให้สามารถจำแนกได้ว่าในแต่ละกลุ่มภาษาถิ่นการติดสำเนียงที่ทำให้ฟังยากอยู่ที่ตรงไหน จะเห็นได้ว่าการประยุกต์ใช้ Grad-CAM จะช่วยในเรื่องความโปร่งใสของการทำนายผลของโมเดล CNN ที่ใช้ในการรู้จำเสียงสระภาษาไทย

การประยุกต์ใช้งาน Grad-CAM ตัวสร้าง (constructor) ของ Grad-CAM จะรับโมเดล (model) ซึ่งมีนิยามโครงสร้างภายใน และจะรับหมายเลขคลาส (classIdx) ที่ใช้สำหรับระบุคลาสที่กำลังสนใจที่จะสร้างจินตภาพของ Grad-CAM และชื่อเลเยอร์ convolution ในโมเดลที่กำลังสนใจ (layerName) ซึ่งหากไม่มีการระบุชื่อ ตัวสร้างจะตามหาเลเยอร์ convolution สุดท้ายในโมเดลมาให้ (last convolutional layer) ขั้นตอนแรกจะทำการสร้างโมเดลอันใหม่ (gradModel) ซึ่งใช้ข้อมูลเข้าเดียวกันกับโมเดลดั้งเดิมที่ต้องการสร้างจินตภาพ แต่แตกต่างกันตรงผลลัพธ์จะแบ่งออกมาเป็นสองส่วน โดยส่วนแรกคือผลลัพธ์จากเลเยอร์ convolution ที่ต้องการจะสร้างจินตภาพ และอีกส่วนคือผลลัพธ์ดั้งเดิมจากโมเดลที่ต้องการศึกษา คำนวณหา gradient ของความน่าจะเป็นสัมพัทธ์ ซึ่งได้จากผลลัพธ์ของโมเดล ซึ่งผลลัพธ์ของ gradModel เป็นความน่าจะเป็นสัมพัทธ์ของคลาสทุกคลาส โดยหลักการของ Grad-CAM ต้องการสำรวจเฉพาะค่าความน่าจะเป็นสัมพัทธ์ของคลาสที่สนใจ ดังนั้นจึงแยกโดยนำความน่าจะเป็นสัมพัทธ์ของคลาสที่สนใจออกมา ซึ่งก็คือ cross-entropy โดยค่าความน่าจะเป็นสัมพัทธ์ขึ้นอยู่กับค่าในฟีทเจอร์แม็พ และทำการหาอนุพันธ์ย่อยของค่าความน่าจะเป็นสัมพัทธ์โดยเทียบกับแต่ละจุดบนฟีทเจอร์แม็พ เพื่อทำให้ทราบว่าแต่ละจุดฟีทเจอร์แม็พสำคัญกับค่าความน่าจะเป็นสัมพัทธ์เพียงใด ตำแหน่งในฟีทเจอร์แม็พที่ทำให้เครื่องมีความมั่นใจในการเลือกคลาสนั้นก็คือจุดที่มีค่า gradient เป็นบวก สามารถคำนวณตำแหน่งในแผนภาพ Grad-CAM ได้ด้วยการนำตำแหน่งที่เป็นบวกของฟีทเจอร์แม็พกับ gradient ทำให้ได้จินตภาพที่เหมาะสมสำหรับฟีทเจอร์แม็พหนึ่งอัน แต่เนื่องจากมีฟีทเจอร์แม็พหลายอันที่ต้องนำมาผสานกันเพื่อให้ได้จินตภาพเดียว ซึ่งฟีทเจอร์แม็พที่สำคัญมากกว่าควรมีน้ำหนักมากกว่า จึงนำค่าเฉลี่ยของแต่ละฟีทเจอร์แม็พมาคิดเพื่อเป็นค่าถ่วงน้ำหนักของแต่ละฟีทเจอร์แม็พ จากนั้นนำค่าน้ำหนักไปคูณกับฟีทเจอร์แม็พ และคำนวณผลรวมเชิงเส้นของฟีทเจอร์แม็พ การ normalize ระดับสีใน heatmap ทำได้โดยการใช้ linear min-max scaling คือการปรับให้ตำแหน่งที่มีค่าต่ำสุดเป็น 0 และตำแหน่งที่มีค่ามากที่สุดเป็น 1 ในระบบเลขทศนิยม จากนั้นจึงปรับให้เป็นจำนวนเต็มสเกล 0 ถึง 255 เพื่อให้เหมาะสมกับการแสดงผลในลักษณะจินตภาพ ได้ heatmap ซึ่งมีลักษณะเป็นเมทริกซ์แรเงา ใน heatmap ข้อมูลจะแสดงด้วยสีหรือความสว่าง [100] มีการใช้ pseudo color โดยใช้แผนผังสีแบบ viridis ค่าอัลฟาของ heatmap ลดลงเหลือ 50% จากนั้นจึงนำ heatmap ไปใช้กับภาพต้นฉบับโดยสร้างรูปภาพแบบผสมรูปแบบสีเทา (grayscale) ซึ่ง gradient activations สามารถมองเห็นได้บนสเปกโตรแกรม

6.1.5. รายละเอียดการทดลอง (Implementation details)

ในการวิจัยนี้เพื่อสร้างคุณสมบัติการป้อนข้อมูล MS ขนาด 11 x 128 x 1 (#times x #frequencies x #channel) กำหนดพารามิเตอร์ของสเปกโตรแกรมดังนี้ ความยาวของหน้าต่างคือ 2,048 ความยาว Hop ระหว่างเฟรมตัวอย่างคือ 512 ช่องสัญญาณเสียง คือ 1 อัตราการสุ่มตัวอย่าง

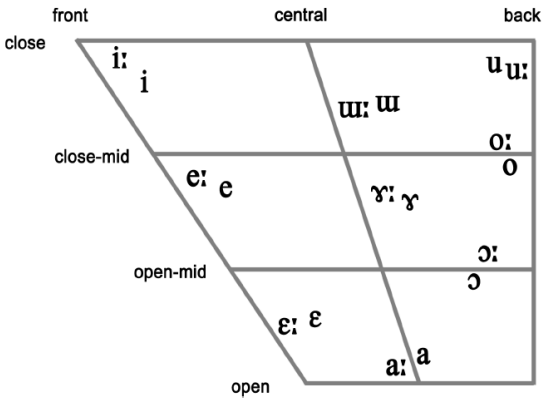
เสียง 16,000 และจำนวน Mel bands คือ 128 สำหรับการฝึกโมเดล จะมีการกำหนด Hyper-parameter สำหรับโมเดลโดย Optimizer คือ Adam อัตราการเรียนรู้เริ่มต้น คือ 0.001 และ Batch size คือ 32 สำหรับการทดลองการฝึกโมเดลใช้ภาษาโปรแกรมไพทอนบนเฟรมเวิร์ก Keras และใช้ TensorFlow โดยทดลองบน Google Colaboratory [98] ด้วย Intel (R) Xeon (R) CPU @ 2.20 GHz และ Nvidia Tesla P100 GPU

6.2. ผลการทดลอง

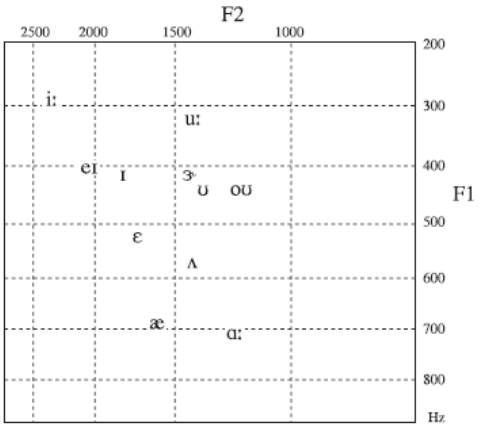
ในงานวิจัยนี้ได้ศึกษาการรู้จำเสียงสระภาษาไทยของชุดข้อมูลผสม ในส่วนแรกของการทดลองจะเป็น Grad-CAM ของโมเดล CNN กับหลักการวิเคราะห์ทางภาษาศาสตร์ ส่วนที่สองแสดงระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง ในส่วนสุดท้ายจะแสดงการประเมินความพึงพอใจของผู้ใช้ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง

6.2.1. Grad-CAM ของโมเดล CNN กับหลักการวิเคราะห์ทางภาษาศาสตร์

ผลการศึกษา Grad-CAM ของโมเดล CNN กับหลักการวิเคราะห์ทางภาษาศาสตร์โดยนำผลจาก Grad-CAM เปรียบเทียบกับกราฟแสดงตำแหน่งของสระในภาษาไทยที่ใช้หลักการทางภาษาศาสตร์ ซึ่งแบ่งตำแหน่งลิ้นเป็น 3 ส่วนคือ หน้า กลาง หลัง และแบ่งระดับของลิ้นออกเป็น 4 ระดับคือ สูง(ปิด) กึ่งสูง กึ่งต่ำ และต่ำ(เปิด) [99] ดังรูปที่ 33 โดยพิจารณาร่วมกับกราฟการเคลื่อนตัวของลิ้นในตำแหน่ง หน้า กลาง และหลัง ดังรูปที่ 33 ซึ่งจะแสดงออกในค่าความถี่ฟอร์เมินท์ที่ 2 และการเคลื่อนที่สูง-ต่ำของลิ้นจะแสดงออกในค่าความถี่ฟอร์เมินท์ที่ 1 [2]

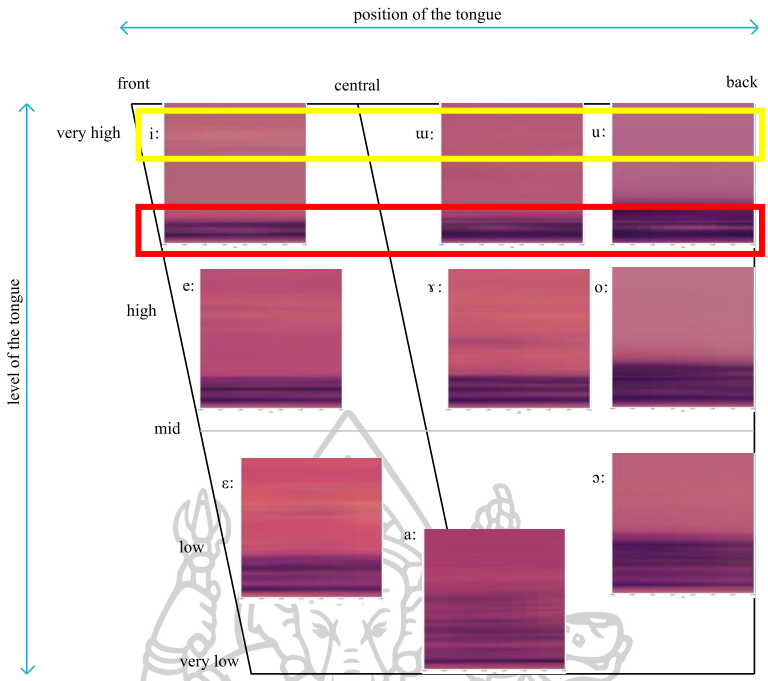


รูปที่ 33 แสดง Vowel Monophthong Phonemes [99]

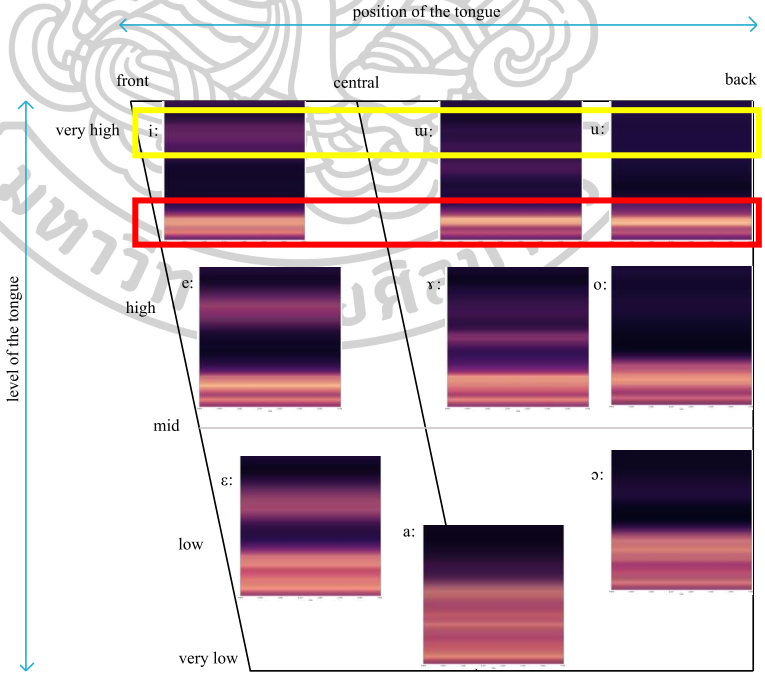


รูปที่ 34 แสดงตัวอย่างกราฟค่าความถี่ฟอร์เมินท์ที่ 1 และที่ 2 เสียงสระของกลุ่มผู้พูดภาษาอังกฤษในแคลิฟอร์เนีย [2]

จากการจัดเรียงรูปที่ 34 ตามการจัดตำแหน่งสระให้ตรงกับตารางตามการศึกษาในแบบหลักภาษาศาสตร์ จะเห็นว่าในแถว column ซ้ายสุดเป็นแถวสระที่ใช้ลิ้นส่วนหน้าในการออกเสียงซึ่งค่าค่าความถี่ฟอร์เมินท์ที่ 2 จะค่อนข้างไปทางสูง และในแถวขวาสุดเป็นแถวที่ใช้ลิ้นสวนหลังค่าความถี่ฟอร์เมินท์ที่ 2 จะมีค่าต่ำ เมื่อพิจารณารูปที่ 35 ภาพจาก Grad-CAM ที่ได้จากเลเยอร์ convolutional ที่ 2 และรูปที่ 35 ภาพจาก Grad-CAM ที่ได้จากเลเยอร์ convolutional สุดท้ายของสระเสียงยาว 9 เสียง ในชุดข้อมูลผสม (Mixed dataset) ที่นำมาจัดเรียงตามตำแหน่งของกราฟแสดงสระตามแบบของหลักภาษาศาสตร์ เพื่อสังเกตลักษณะของการเปลี่ยนแปลงในแต่ละสระ ดังนั้นเมื่อพิจารณาภาพของทั้ง 3 ตำแหน่ง (หน้า กลาง หลัง) โดยนำมาเปรียบเทียบกัน ยกตัวอย่างสระอี (/i:/), อือ (/ɯ:/) และอู (/u:/) เรียงจากซ้ายไปขวา จะเป็นว่าลักษณะของแถบสีมีความแตกต่างกันในบริเวณต่าง ๆ ของภาพ โดยทั้ง 2 ภาพที่ได้จาก Grad-CAM มีลักษณะไปในทิศทางเดียวกัน โดยจะเห็นภาพชัดเจนที่สุดใน Grad-CAM ที่ได้จากเลเยอร์ convolutional สุดท้าย นั่นคือบริเวณที่มีสีที่สว่างจะเป็นบริเวณที่มีความสำคัญกับการตัดสินใจในการทำนายแต่ละสระ โดยเมื่อพิจารณากรอบสีเหลืองของสระอี (/i:/), อือ (/ɯ:/) และอู (/u:/) จะพบว่าสระอี (/i:/) ที่ใช้ลิ้นส่วนหน้ามากกว่า Grad-CAM จะพิจารณาให้ความสำคัญบริเวณที่มีความถี่สูงมากกว่าสระอู (/u:/) ที่ใช้ลิ้นส่วนหลัง เมื่อพิจารณากรอบสีแดงของสระอี (/i:/), อือ (/ɯ:/) และอู (/u:/) จะพบว่าในส่วนที่สว่างที่สุด Grad-CAM จะพิจารณาทั้ง 3 สระโดยให้ความสำคัญบริเวณที่มีความถี่ต่ำที่ใกล้เคียงกัน ซึ่งทั้ง 3 สระมีการเคลื่อนที่ของลิ้นที่สูงเช่นกัน จะเห็นได้ว่าการพิจารณาของ Grad-CAM สอดคล้องกับหลักการทางภาษาศาสตร์ [2]



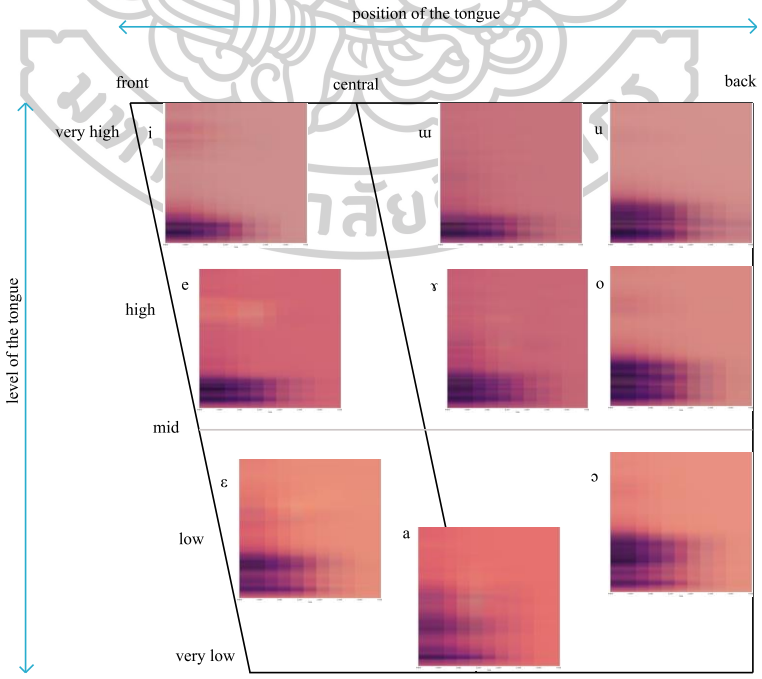
รูปที่ 35 แสดง Grad-CAM ที่ได้จากเลเยอร์ convolutional ที่ 2 ของสระเสียงยาว



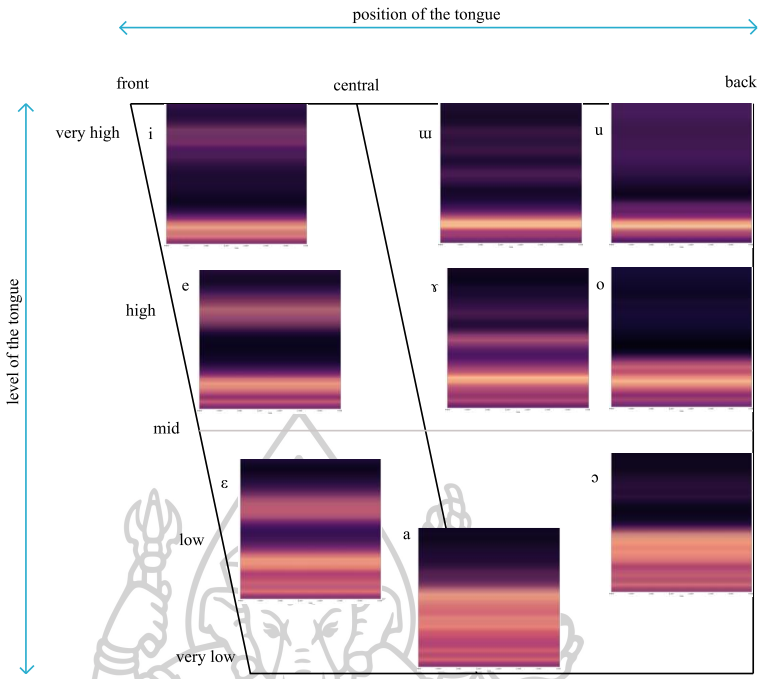
รูปที่ 36 แสดง Grad-CAM ที่ได้จากเลเยอร์ convolutional สุดท้าย ของสระเสียงยาว

จากรูปที่ 34 เมื่อพิจารณาแถว row ซึ่งเป็นการเคลื่อนที่สูงต่ำของลิ้นค่าความถี่ฟอร์เมินท์ที่ 1 จะมีการเปลี่ยนแปลง สระที่มีลักษณะการเคลื่อนที่ของลิ้นอยู่ระดับสูง ค่าความถี่ฟอร์เมินท์ที่ 1 จะค่อนข้างต่ำ และสระที่มีลักษณะการเคลื่อนที่ของลิ้นอยู่ระดับต่ำ ค่าความถี่ฟอร์เมินท์ที่ 1 มีแนวโน้มที่สูงขึ้น เมื่อพิจารณารูปที่ 36 ภาพจาก Grad-CAM ที่ได้จากเลเยอร์ convolutional สุดท้าย ของสระเสียงยาว โดยยกตัวอย่างสระ อี (/i:/), เอ (/e:/) และแอ (/ɛ:/) ซึ่งเป็นสระที่มีการใช้ลิ้นส่วนหน้าทั้ง 3 สระ แต่มีระดับการเคลื่อนที่ของลิ้นต่างกัน โดยสระอี (/i:/) เป็นสระที่ใช้ระดับการเคลื่อนที่ของลิ้นสูงที่สุด จะพบว่าในส่วนที่สว่างที่สุด Grad-CAM ของสระอี (/i:/) จะพิจารณาช่วงความถี่ที่ต่ำกว่าสระแอ (/ɛ:/) ที่เป็นสระที่ใช้ระดับการเคลื่อนที่ของลิ้นที่ต่ำกว่า ซึ่งผลลัพธ์ที่ได้จาก Grad-CAM นั้นสอดคล้องกับการเปลี่ยนแปลงของค่าความถี่ฟอร์เมินท์ที่ 1 และ 2 ตามแบบทฤษฎีทางภาษาศาสตร์ [2]

เมื่อพิจารณาสระเสียงสั้น 9 เสียงในชุดข้อมูลผสม (Mixed dataset) ดังรูปที่ 37 Grad-CAM ที่ได้จากเลเยอร์ convolutional ที่ 2 และรูปที่ 38 Grad-CAM ที่ได้จากเลเยอร์ convolutional สุดท้ายของสระเสียงสั้น โดยนำมาจัดเรียงตามตำแหน่งของกราฟแสดงสระตามแบบของหลักภาษาศาสตร์ เพื่อสังเกตุลักษณะของการเปลี่ยนแปลงในแต่ละสระ ภาพที่แสดงออกมีความแตกต่างกัน ซึ่งผลลัพธ์ที่ได้จาก Grad-CAM นั้นสอดคล้องกับการเปลี่ยนแปลงของค่าความถี่ฟอร์เมินท์ที่ 1 และ 2 ตามแบบทฤษฎีทางภาษาศาสตร์ [2, 99] เช่นเดียวกับสระเสียงยาว แต่เมื่อพิจารณารูปที่ 36 Grad-CAM ที่ได้จากเลเยอร์ convolutional ที่ 2 ของสระเสียงสั้น จะแสดงถึงการให้ความสำคัญที่เกี่ยวข้องกับระยะเวลาในการทำนายเสียงสระเสียงสั้นด้วย

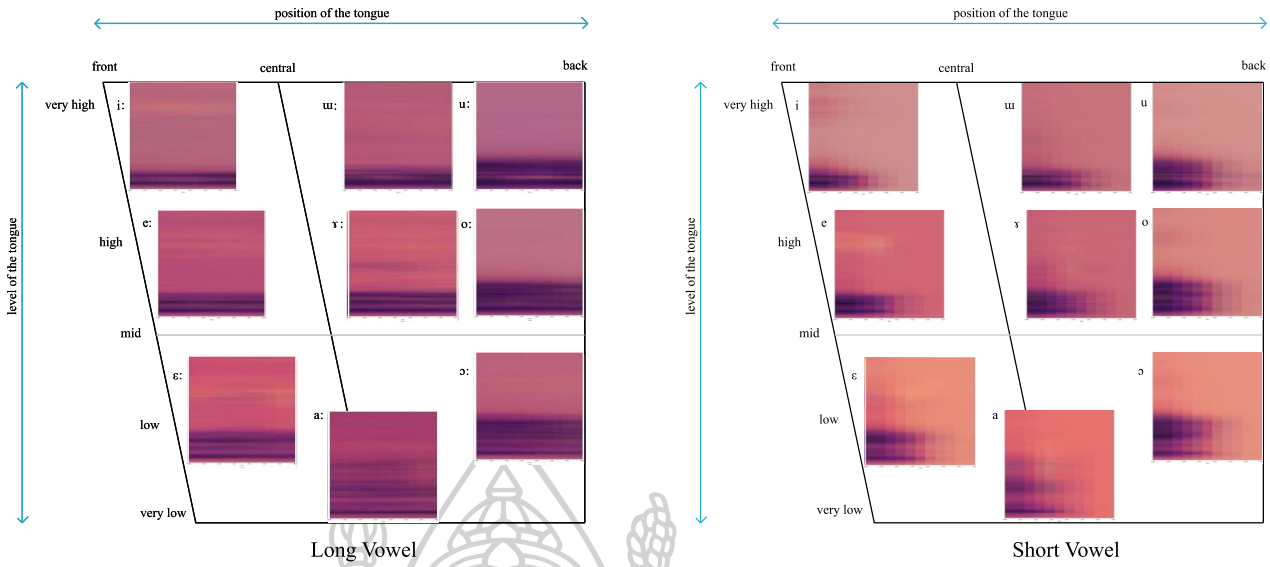


รูปที่ 37 แสดง Grad-CAM ที่ได้จากเลเยอร์ convolutional ที่ 2 ของสระเสียงสั้น



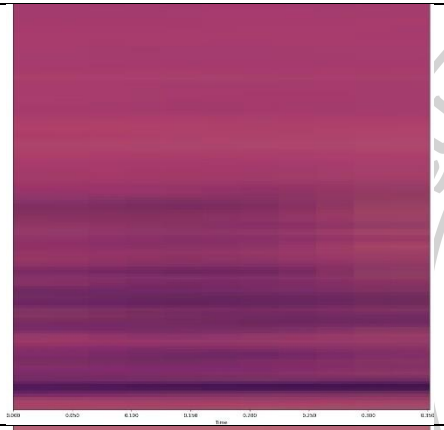
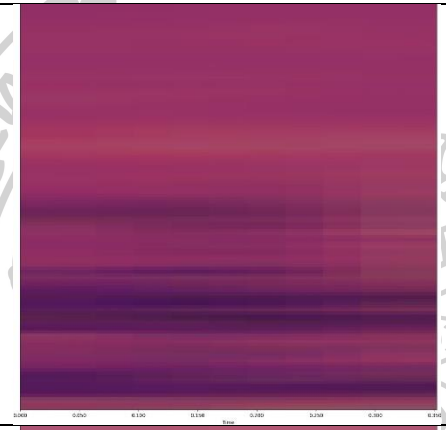
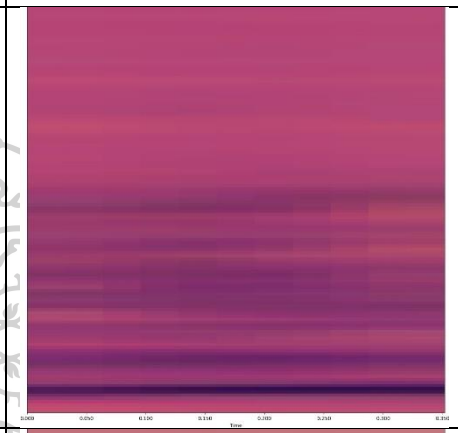
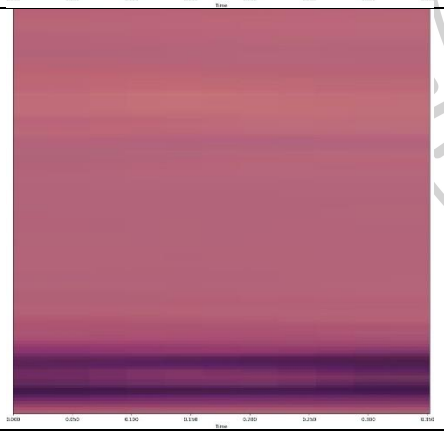
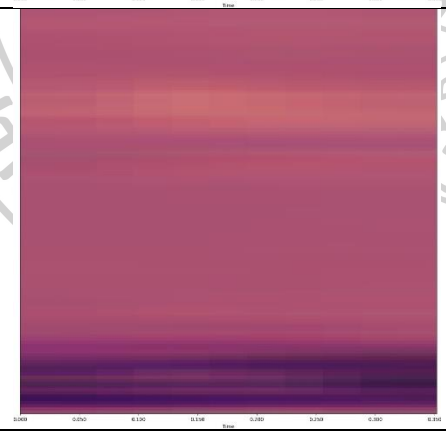
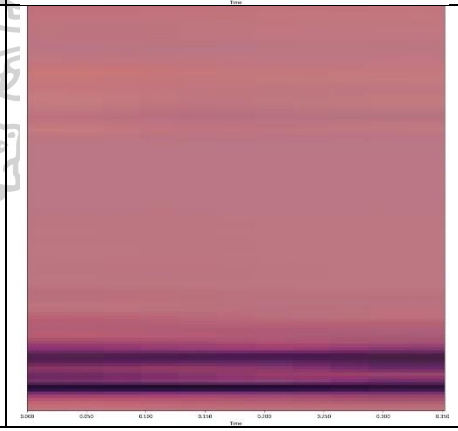
รูปที่ 38 แสดง Grad-CAM ที่ได้จากเลเยอร์ convolutional สุดท้ายของสระเสียงสั้น

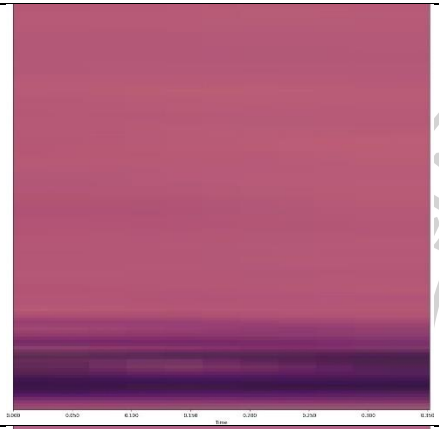
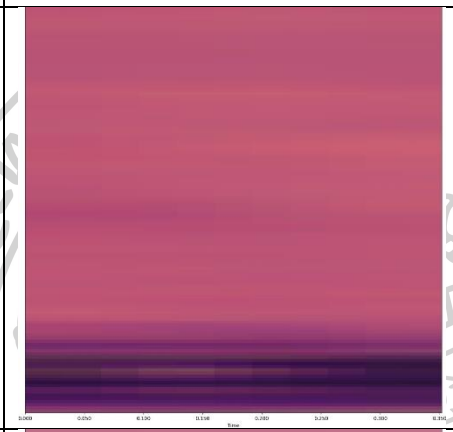
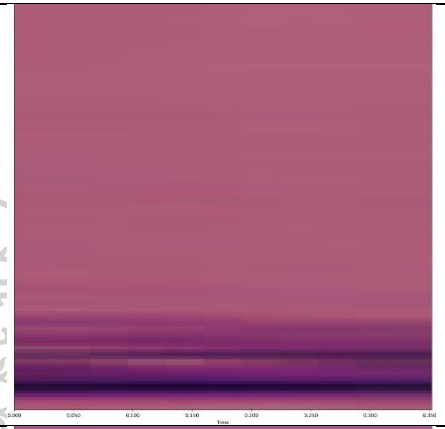
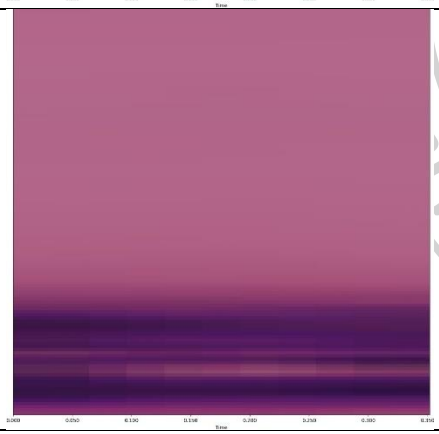
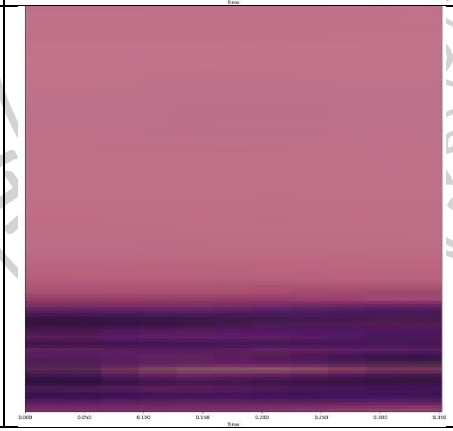
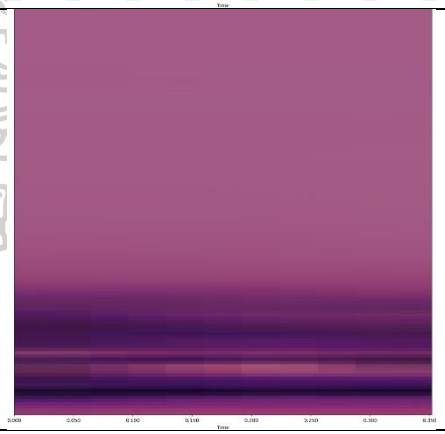
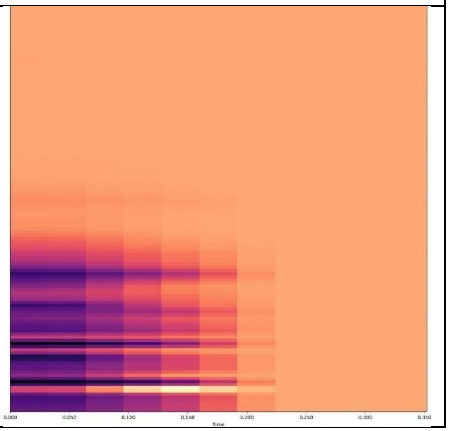
เมื่อพิจารณาจากรูปที่ 39 ที่แสดง Grad-CAM ของเสียงสระเสียงยาวและสระเสียงสั้นทั้ง 18 เสียง ในชุดข้อมูลผสม (Mixed dataset) จะเห็นได้ว่าการแสดงผลภาพของ Grad-CAM มีลักษณะไปในทิศทางเดียวกันเมื่อเรียงตามตำแหน่งของกราฟแสดงสระตามแบบของหลักภาษาศาสตร์ ในแนวแกน y จะเห็นว่าในแถว column ซ้ายสุดเป็นแถวสระที่ใช้ลิ้นส่วนหน้าในการออกเสียงทั้งสระเสียงยาวและเสียงสั้น พบว่า Grad-CAM ให้ความสำคัญกับบริเวณที่มีความถี่สูงมากกว่า column ขวาสุดที่ใช้ลิ้นสวนหลัง เมื่อพิจารณาแถว row ซึ่งเป็นการเคลื่อนที่สูงต่ำของลิ้นทั้งสระเสียงยาวและเสียงสั้น ภาพที่แถวบนสุดสระที่มีการออกเสียงโดยใช้ระดับลิ้นอยู่ที่ระดับสูง พบว่า Grad-CAM ให้ความสำคัญกับบริเวณที่มีความถี่ต่ำกว่าในแถวล่างที่สระมีการออกเสียงโดยใช้ระดับลิ้นอยู่ระดับต่ำ ในขณะที่เมื่อพิจารณาตามแนวแกน x จะเห็นได้ว่า Grad-CAM แสดงผลภาพอธิบายความสำคัญของข้อมูลเข้าเสียงสระที่ใช้ในการทำนายคลาสเป้าหมายสัมพันธ์กับการออกเสียงสระภาษาไทยในด้านของเวลาที่สระเสียงยาวจะต้องใช้ระยะเวลาในการเปล่งเสียงที่ยาวกว่าสระเสียงสั้นในคู่เสียงสระเดียวกัน เช่นภาพแรกบนสุดทางซ้ายที่แสดงบริเวณที่สำคัญในการทำนายของ สระอี-สระอิ (/i:/-/i/) หรือภาพล่างสุดที่แสดงผลของ สระอา-สระอะ (/a:/-/a/) ซึ่งสัมพันธ์กับความแตกต่างทางด้านระยะเวลา (duration) ที่เป็นปัจจัยหนึ่งของความแตกต่างของเสียงสระ [2]

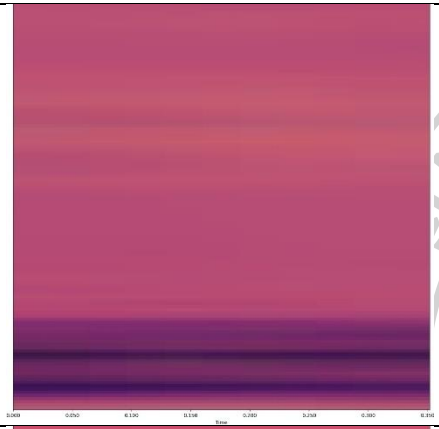
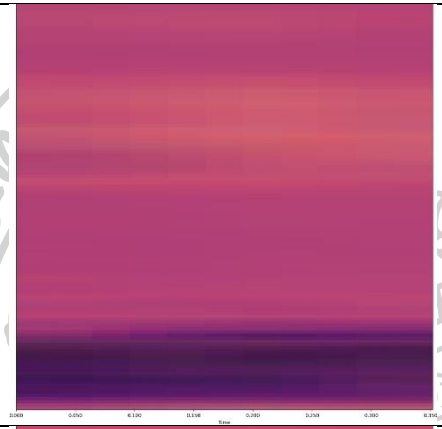
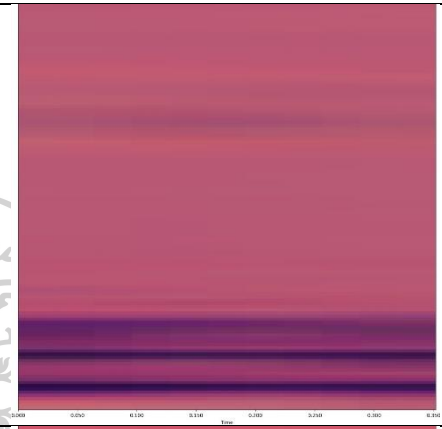
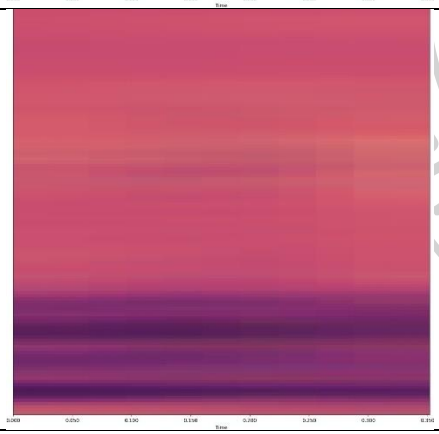
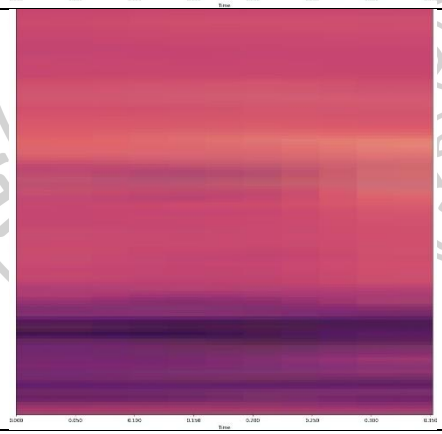
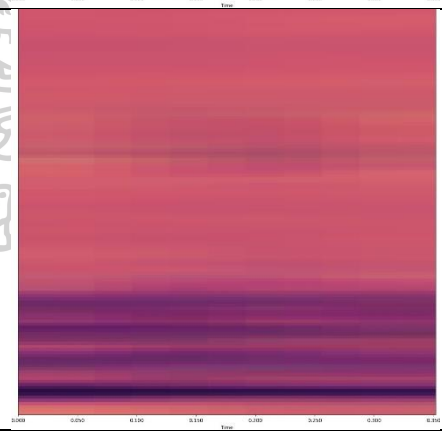
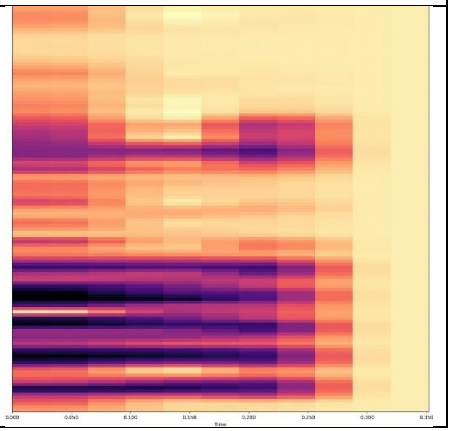


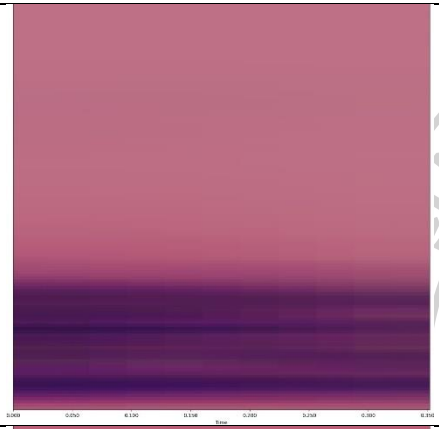
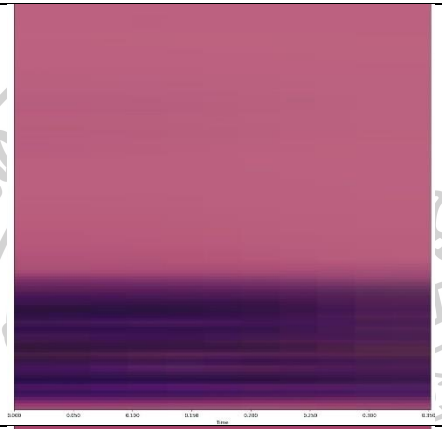
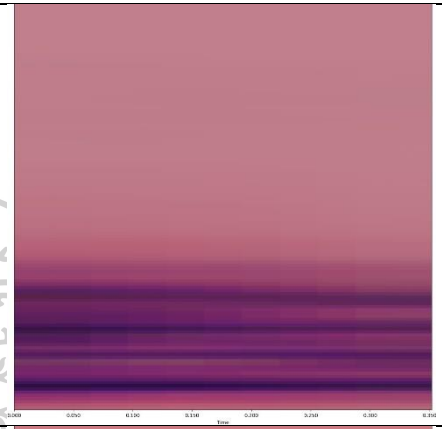
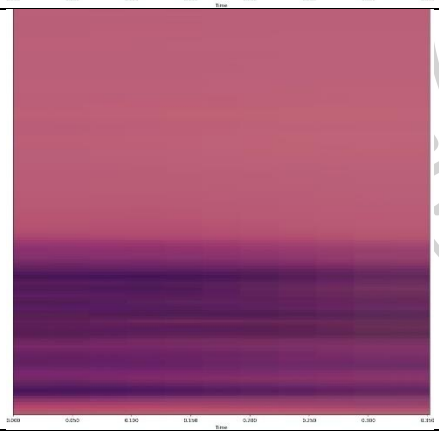
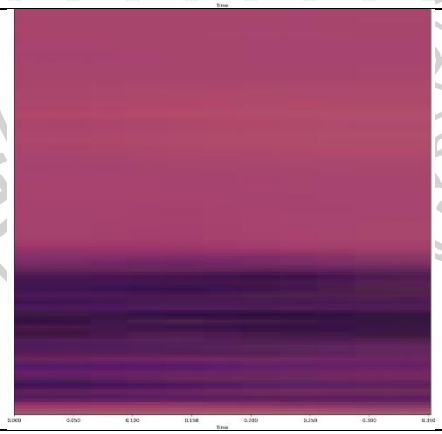
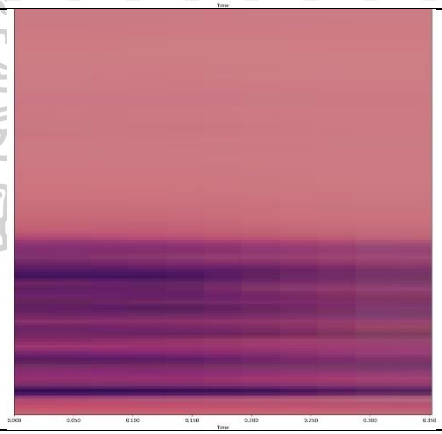
รูปที่ 39 แสดง Grad-CAM ที่ได้จากเลเยอร์ convolutional ที่ 2 ของสระเสียงยาว - สั้น

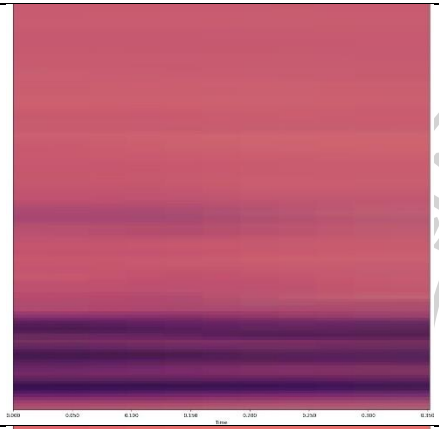
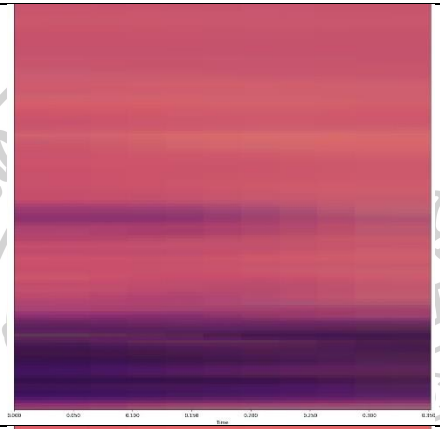
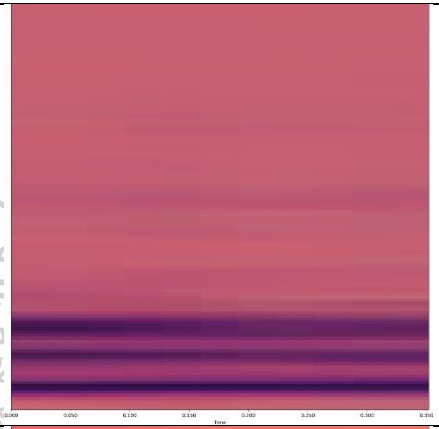
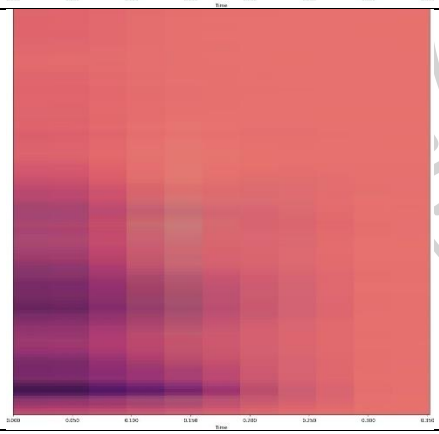
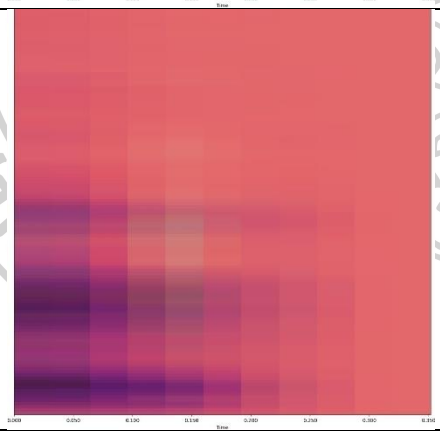
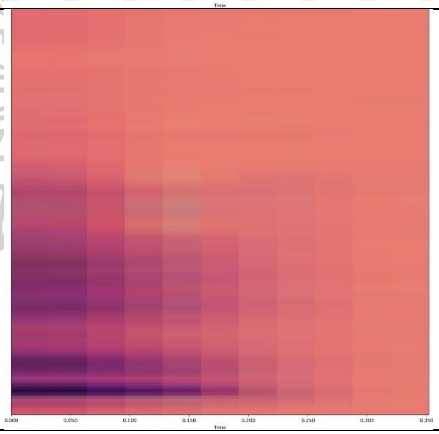
ตารางที่ 14 แสดงผลภาพอธิบายส่วนสำคัญของข้อมูลเข้า Mel spectrogram เมื่อใช้ Grad-CAM กับโมเดล CNN

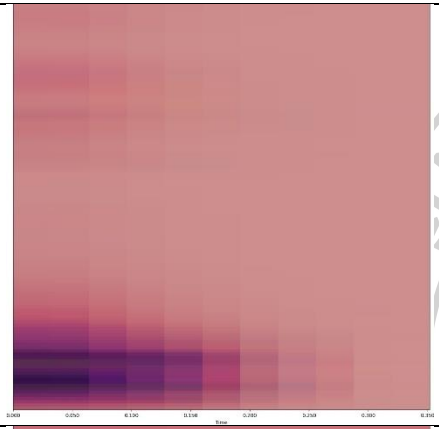
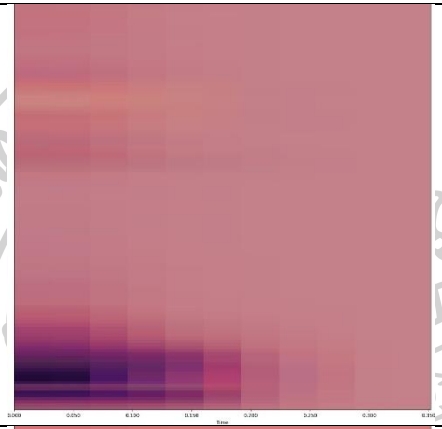
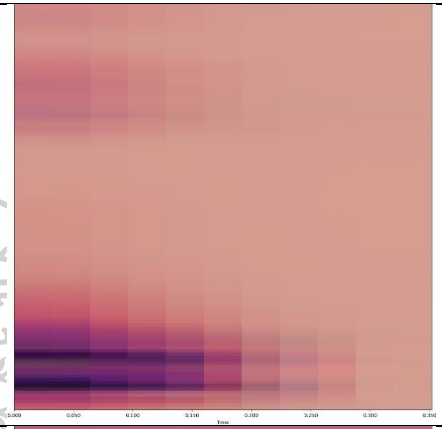
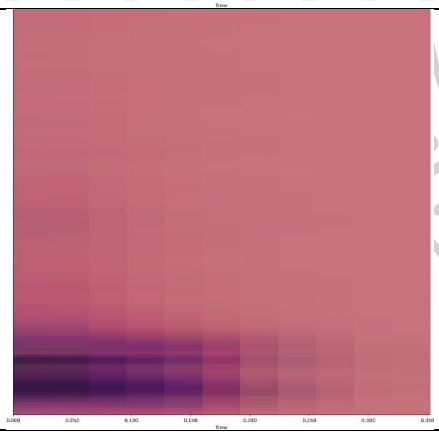
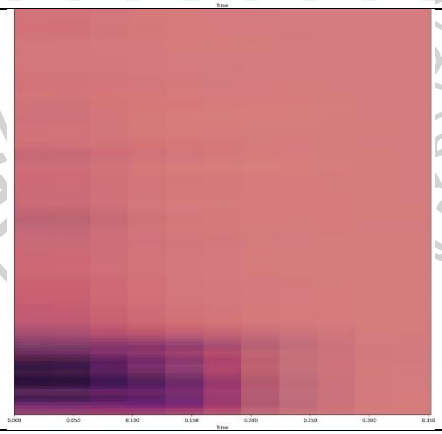
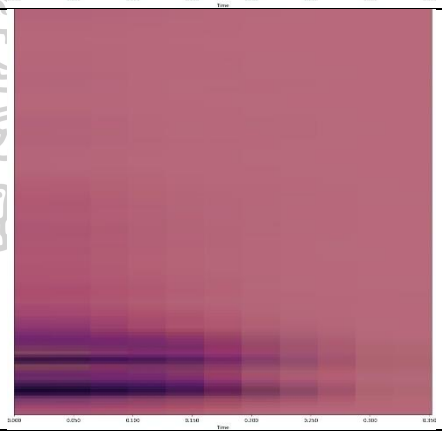
Vowel	Correct			Wrong
	Mixed dataset	Men dataset	Women dataset	Mixed dataset
01 อา				
02 อี				

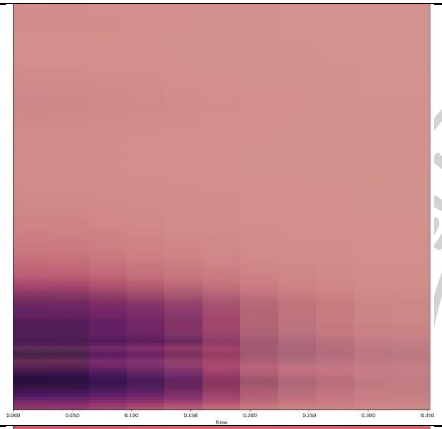
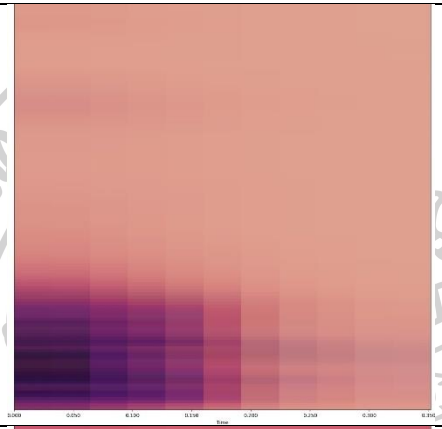
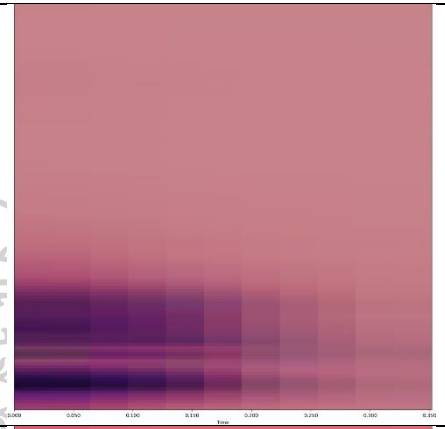
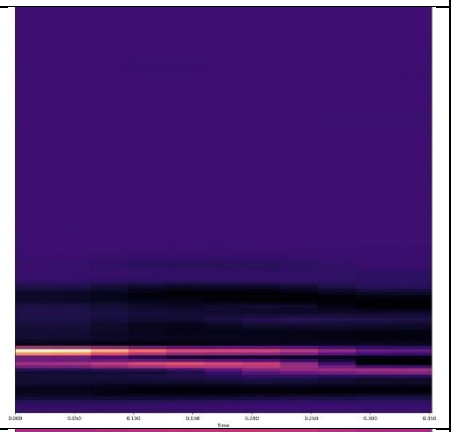
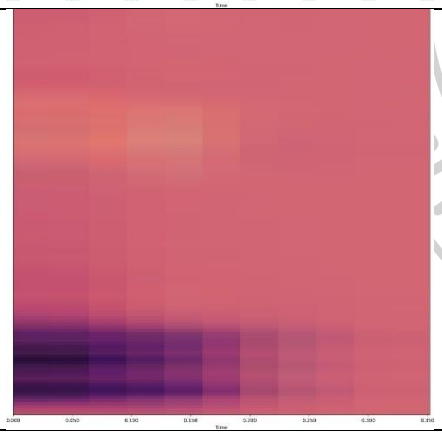
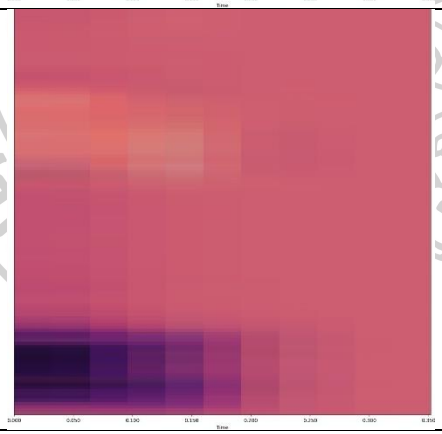
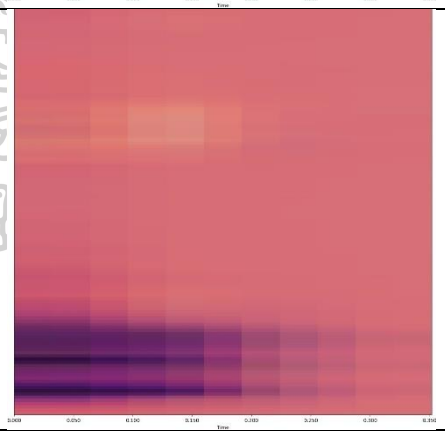
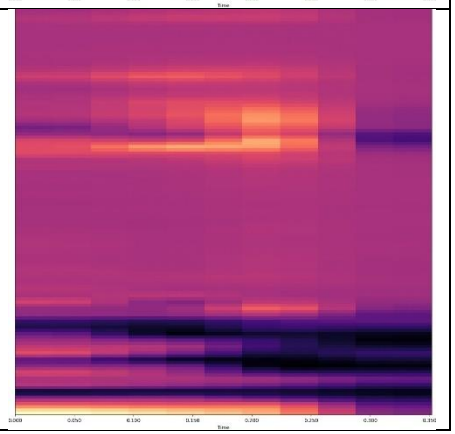
Vowel	Correct			Wrong
	Mixed dataset	Men dataset	Women dataset	Mixed dataset
03 อือ				
04 อู				

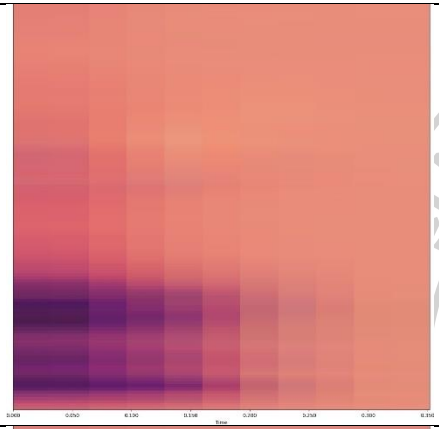
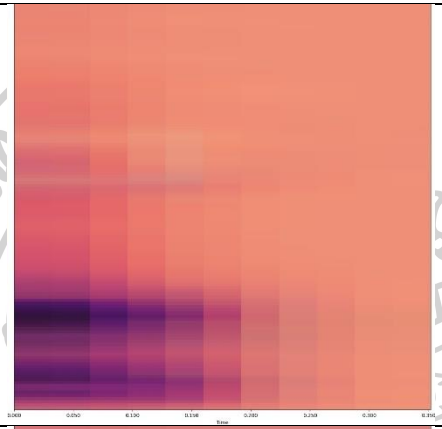
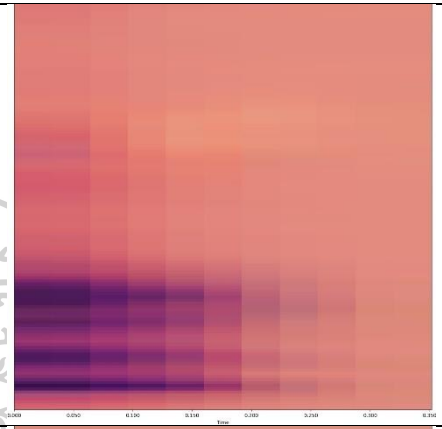
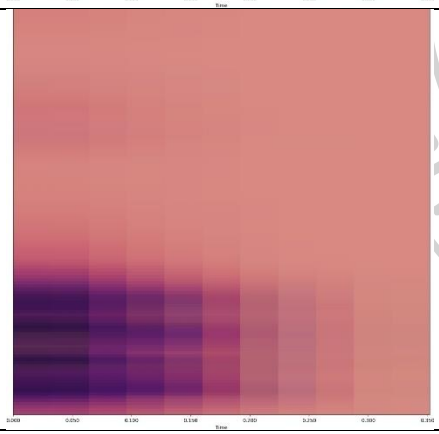
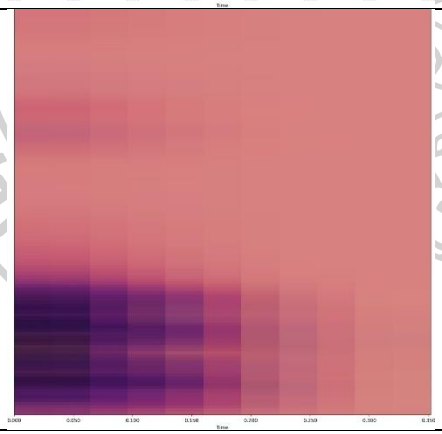
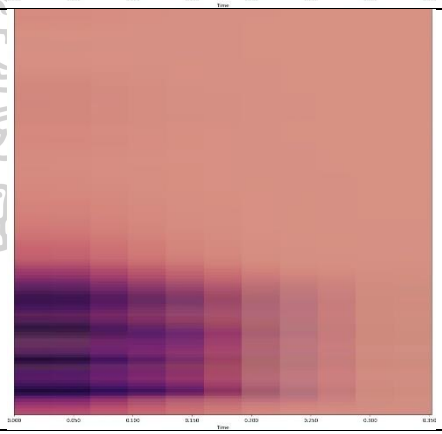
Vowel	Correct			Wrong
	Mixed dataset	Men dataset	Women dataset	Mixed dataset
05 เอ				
06 แอ				

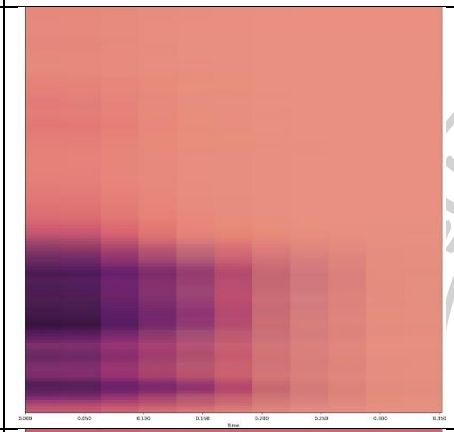
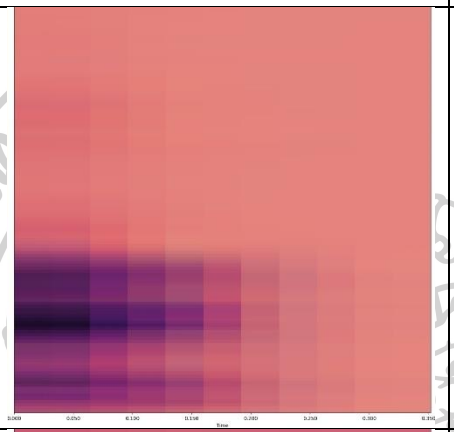
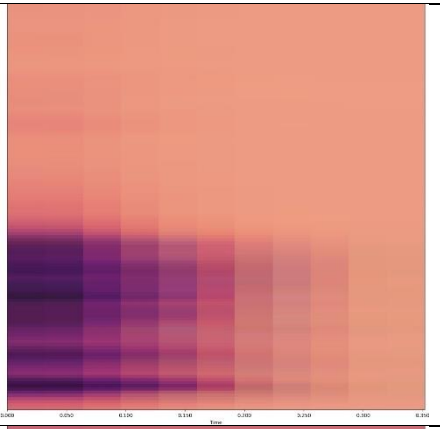
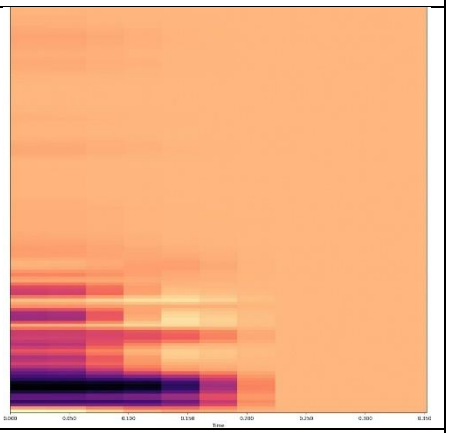
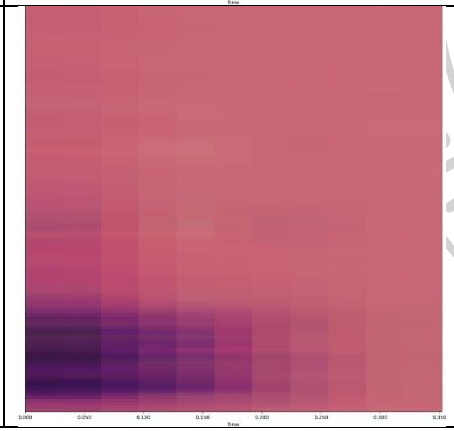
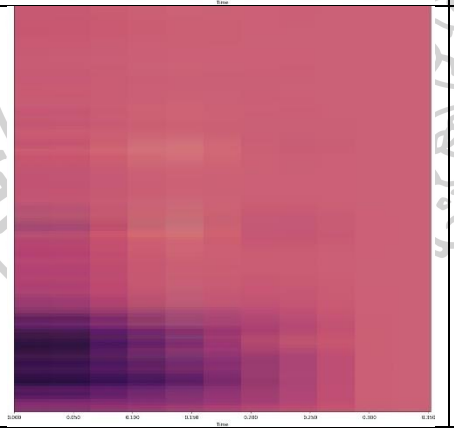
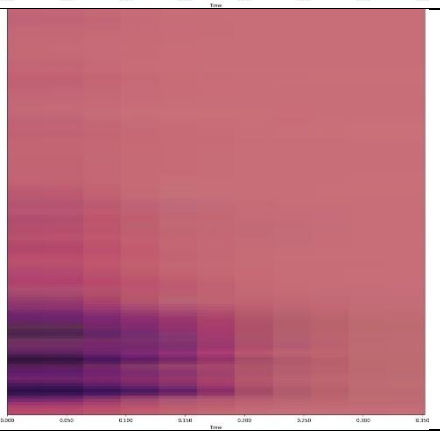
Vowel	Correct			Wrong
	Mixed dataset	Men dataset	Women dataset	Mixed dataset
07 โอ				
08 ออ				

Vowel	Correct			Wrong
	Mixed dataset	Men dataset	Women dataset	Mixed dataset
09 เออ				
S1 อะ				

Vowel	Correct			Wrong
	Mixed dataset	Men dataset	Women dataset	Mixed dataset
S2 อิ				
S3 อี				

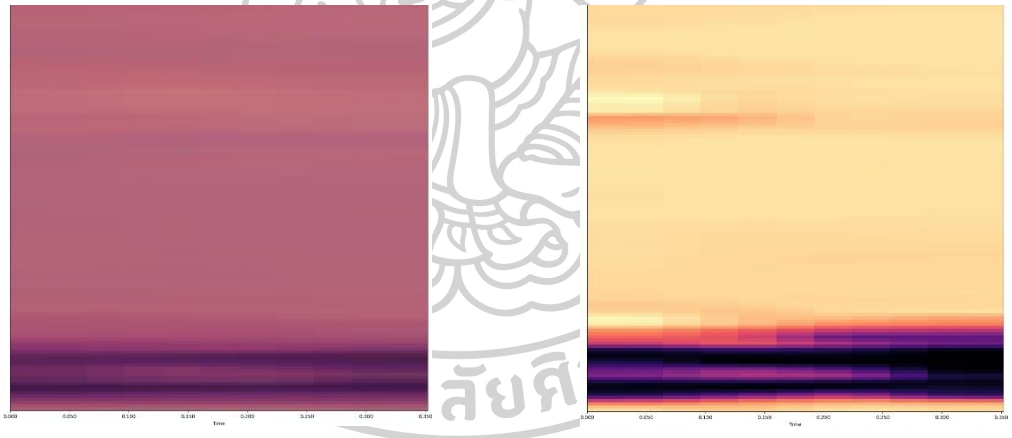
Vowel	Correct			Wrong
	Mixed dataset	Men dataset	Women dataset	Mixed dataset
S4 อุ				
S5 เอะ				

Vowel	Correct			Wrong
	Mixed dataset	Men dataset	Women dataset	Mixed dataset
S6 แอะ				
S7 โอะ				

Vowel	Correct			Wrong
	Mixed dataset	Men dataset	Women dataset	Mixed dataset
S8 เอาะ				
S9 เออะ				

ตารางที่ 14 จะแสดงผลภาพอธิบายพื้นที่ที่สำคัญของข้อมูลเข้า MS เมื่อใช้ Grad-CAM ของชุดข้อมูลผสม (Mixed dataset) ชุดข้อมูลเพศชาย (Men dataset) และชุดข้อมูลเพศหญิง (Women dataset) ในการทำนายผลของคลาสเป้าหมาย จะเห็นได้ว่า Grad-CAM อธิบายผลภาพว่าลักษณะที่สำคัญของในแต่ละคลาสมีความต่างกันทั้งความถี่และระยะเวลา

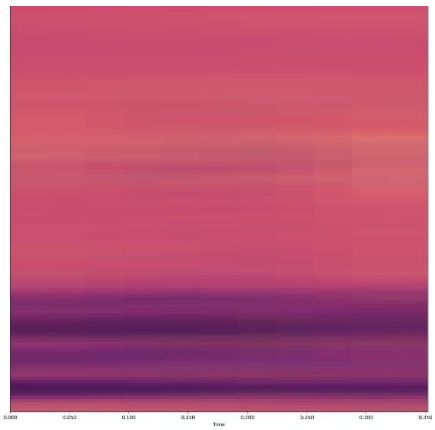
สำหรับในชุดข้อมูลผสม (Mixed dataset) เป็นชุดข้อมูลที่เกิดจากการรวมกันของชุดข้อมูลเพศชาย (Men dataset) และชุดข้อมูลเพศหญิง (Women dataset) โดยการแสดงผลภาพบริเวณที่สำคัญในการทำนายคลาสเป้าหมายในแต่ละสระโดยใช้ Grad-CAM ของเสียงสระภาษาไทยทั้ง 18 เสียง ซึ่งประกอบไปด้วยสระเสียงยาว 9 เสียง และสระเสียงสั้น 9 เสียง เพื่อให้สามารถเห็นความแตกต่างในการทำนายของด้านเวลาและความถี่ จึงแสดงภาพ Grad-CAM ที่ได้จากเลเยอร์ convolutional ที่ 2 โดยแสดงในตางรางที่ 14 ซึ่งภาพรวมของแต่ละสระที่ทำนายถูกต้องจะแสดงใน Correct โดยจะแสดงภาพรวมของเสียงสระในชุดข้อมูลผสม (Mixed dataset) ที่ทำนายถูกต้องในแต่ละสระโดยใช้ Grad-CAM แสดงผลอธิบายภาพ สำหรับข้อมูลเสียงสระที่ทำนายไม่ถูกต้องจะแสดงใน Wrong ซึ่งแสดงภาพรวมของเสียงสระที่ทำนายไม่ถูกต้องในแต่ละสระและแสดงผลอธิบายภาพโดยใช้ Grad-CAM



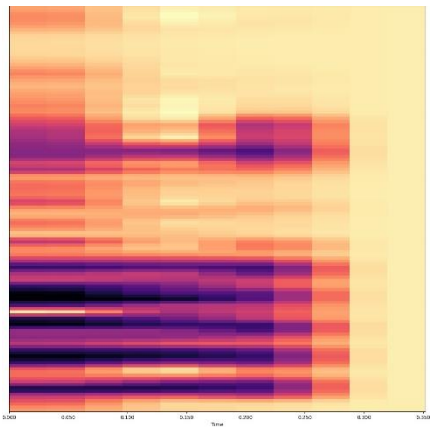
a) ภาพรวมที่ทำนายถูกต้องของเสียงสระอี (/i:/) b) ทำนายถูกต้องพูดโดยเพศหญิง 1 คน

รูปที่ 40 แสดงกรณีทำนายถูกของเสียงสระอี (/i:/)

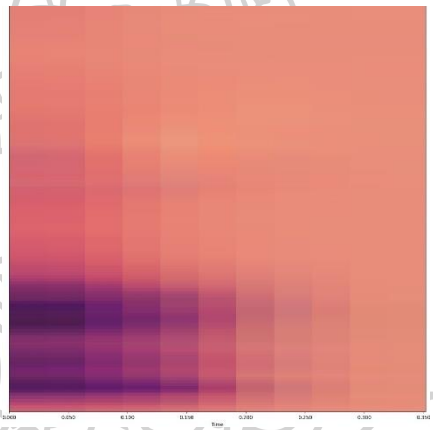
จากรูปที่ 40 แสดงการแสดงผลภาพบริเวณที่สำคัญในการทำนายคลาสเป้าหมายโดยใช้ Grad-CAM กรณีทำนายถูกของเสียงสระอี (/i:/) ซึ่งภาพ a) เป็นภาพรวมของเสียงที่ทำนายถูกที่ใช้ Grad-CAM ในการอธิบายผลภาพข้อมูลเข้า MS ของสระอีทั้งเพศชายและหญิง และภาพ b) เป็นภาพของเสียงสระอี (/i:/) ที่ทำนายถูกของผู้พูด 1 คน แสดงผลภาพโดยใช้ Grad-CAM จะเห็นได้ว่าลักษณะภาพทั้งสองมีลักษณะที่คล้ายคลึงกันทั้งในด้านความถี่และเวลา



a) ภาพรวมที่ทำนายถูกต้องของเสียงสระแอ (/**ɛ**:/)



b) ทำนายผิดพูดโดยเพศหญิง 1 คน ทำนายเป็นแอะ (/**ɛ**/)



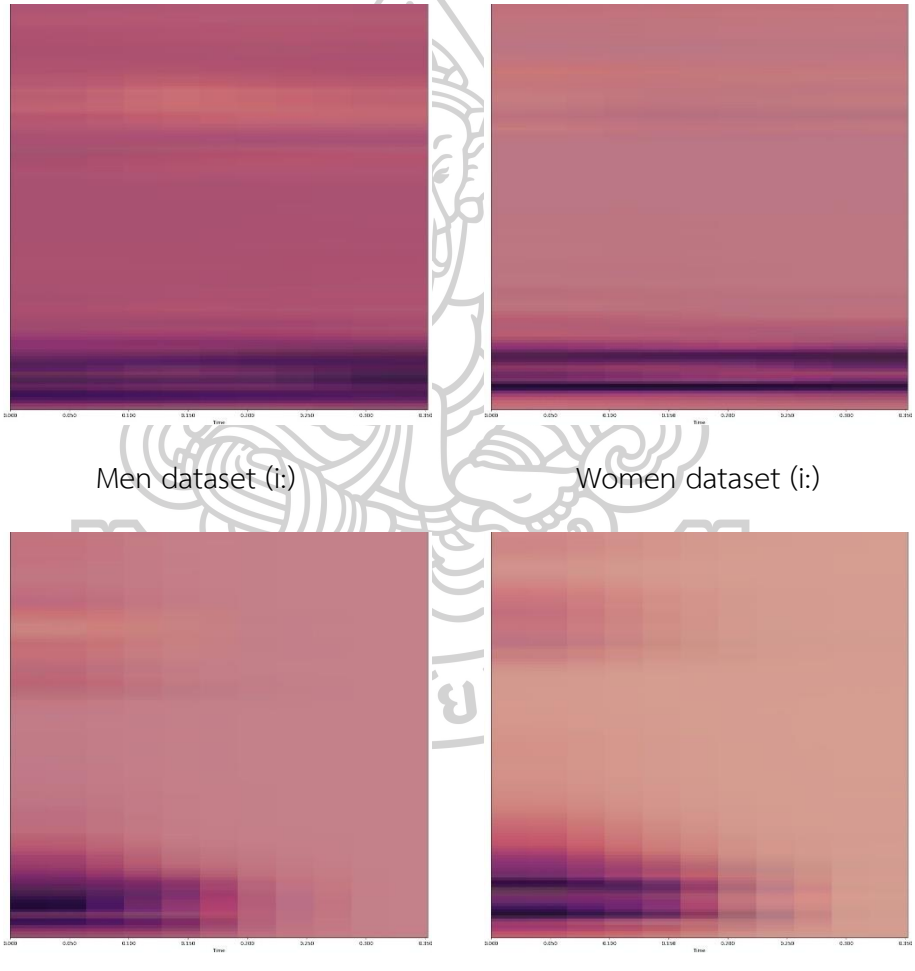
c) ภาพรวมที่ทำนายถูกต้องของเสียงสระแอะ (/**ɛ**/)

*รูปที่ 41 แสดงกรณีทำนายผิดของเสียงสระแอ (/**ɛ**:/)*

สำหรับรูปที่ 41 แสดงการแสดงผลภาพบริเวณที่สำคัญในการทำนายคลาสเป้าหมายโดยใช้ Grad-CAM กรณีทำนายผิดของเสียงสระแอ (/**ɛ**:/) โดยภาพ a) เป็นภาพรวมของเสียงที่ทำนายถูกที่ใช้ Grad-CAM ในการอธิบายผลภาพข้อมูลเข้า MS ของสระแอ (/**ɛ**:/) ทั้งเพศชายและหญิง และภาพ b) เป็นภาพของการทำนายที่ผิดที่พูดโดยผู้พูด 1 คน แสดงผลภาพโดยใช้ Grad-CAM ซึ่งการทำนายผลลัพธ์ออกมาเป็นคลาสของเสียงสระแอะ (/**ɛ**/) โดยลักษณะรูปแบบของ Grad-CAM มีส่วนคล้ายคลึงกับภาพรวมของเสียงสระแอะ (/**ɛ**/) ดังรูป c) ซึ่งจะเห็นได้ว่า Grad-CAM อธิบายภาพของการพูดของผู้พูดว่าออกเสียงสระแอะ (/**ɛ**/) ซึ่งไม่ใช่เสียงสระแอ (/**ɛ**:/) โดยพิจารณาโดยรวมจากภาพ

ของ Grad-CAM เสียงสระแอะ (/**ɛ**/) ในภาษาไทยจะใช้เวลาที่สั้นกว่าเสียงสระแอ (/**ɛ**:/) ซึ่งสอดคล้องกับหลักการของการศึกษาภาษาศาสตร์

ด้านการเปรียบเทียบการรู้จำเสียงของชุดข้อมูลเพศชายและเพศหญิง Grad-CAM แสดงผลภาพในแต่ละสระมีลักษณะคล้ายกันทั้งสองเพศ แต่ในกรณีของชุดข้อมูลเพศหญิงในแนวแกน y ที่แสดงถึงความถี่ พบว่ามีระดับที่สูงกว่าข้อมูลเพศชายเล็กน้อย ซึ่งสอดคล้องกับปัจจัยทางด้านเพศที่มีผลต่อการออกเสียง ตามหลักการของการศึกษาภาษาศาสตร์ที่เพศหญิงมีลักษณะของเสียงที่มีความถี่สูงกว่าเพศชาย ตัวอย่างเช่น สระอี (/i:/) และสระอิ (/i/) ดังรูปที่ 42



รูปที่ 42 แสดงการเปรียบเทียบการรู้จำเสียงของชุดข้อมูลเพศชายและเพศหญิงเมื่อใช้ Grad-CAM

จากการพิจารณาภาพ Grad-CAM จะพบว่ากรณีที่ภาพ Grad-CAM แสดงพื้นที่ที่มีลักษณะไม่สว่าง มีความมืด นั่นอาจหมายความว่าพบฟีทเจอร์ที่เกี่ยวข้องกับคลาสที่สนใจในการออกเสียงนั้นน้อย เป็นไปได้ว่าผู้พูดพูดไปคนละเสียงกับคลาสเป้าหมายได้ ซึ่งมีประโยชน์ในด้านการช่วยคัดกรองการลาเบลผิดได้ สำหรับในกรณีที่คลาส 2 คลาสมีลักษณะการแสดงผลของภาพ Grad-CAM สว่างเหมือนกันทั้งคู่ แสดงว่าในเสียงที่ผู้พูดพูดออกมานั้นมีฟีทเจอร์ที่คล้ายกันกับคลาสทั้ง 2 คลาส ซึ่งอาจเป็นเพราะ 2 คลาสนั้นคล้ายกันโดยธรรมชาติ สามารถพิจารณาจากข้อมูลฝึกที่นักภาษาศาสตร์ถือว่าถูกต้อง หรือว่าผู้พูดนั้นออกเสียงกำกวม ซึ่งสามารถพิจารณาจากความคลาดเคลื่อนจากฟีทเจอร์มาตรฐานของกลุ่มประชากรจำนวนมากที่ออกเสียงถูก ในการประยุกต์ใช้ Grad-CAM สามารถสร้างเฟรมเวิร์คที่สามารถจับกลุ่มเสียงในเชิงสถิติของภาษาได้แบบอัตโนมัติว่าเสียงหรือคำอะไรคล้ายกันและบอกได้ในเชิงปริมาณ ซึ่งเป็นประโยคในการวิเคราะห์ทางภาษาศาสตร์ และสามารถกำหนดในการพูดออกเสียงของแต่ละคนได้ว่าตรงกับกลุ่มใดและเป็นจุดอ่อนของกลุ่มภาษานั้น ๆ หรือไม่ ซึ่งจะเป็นประโยชน์ด้านการแนะนำในการฝึกการออกเสียงในภาษาต่าง ๆ

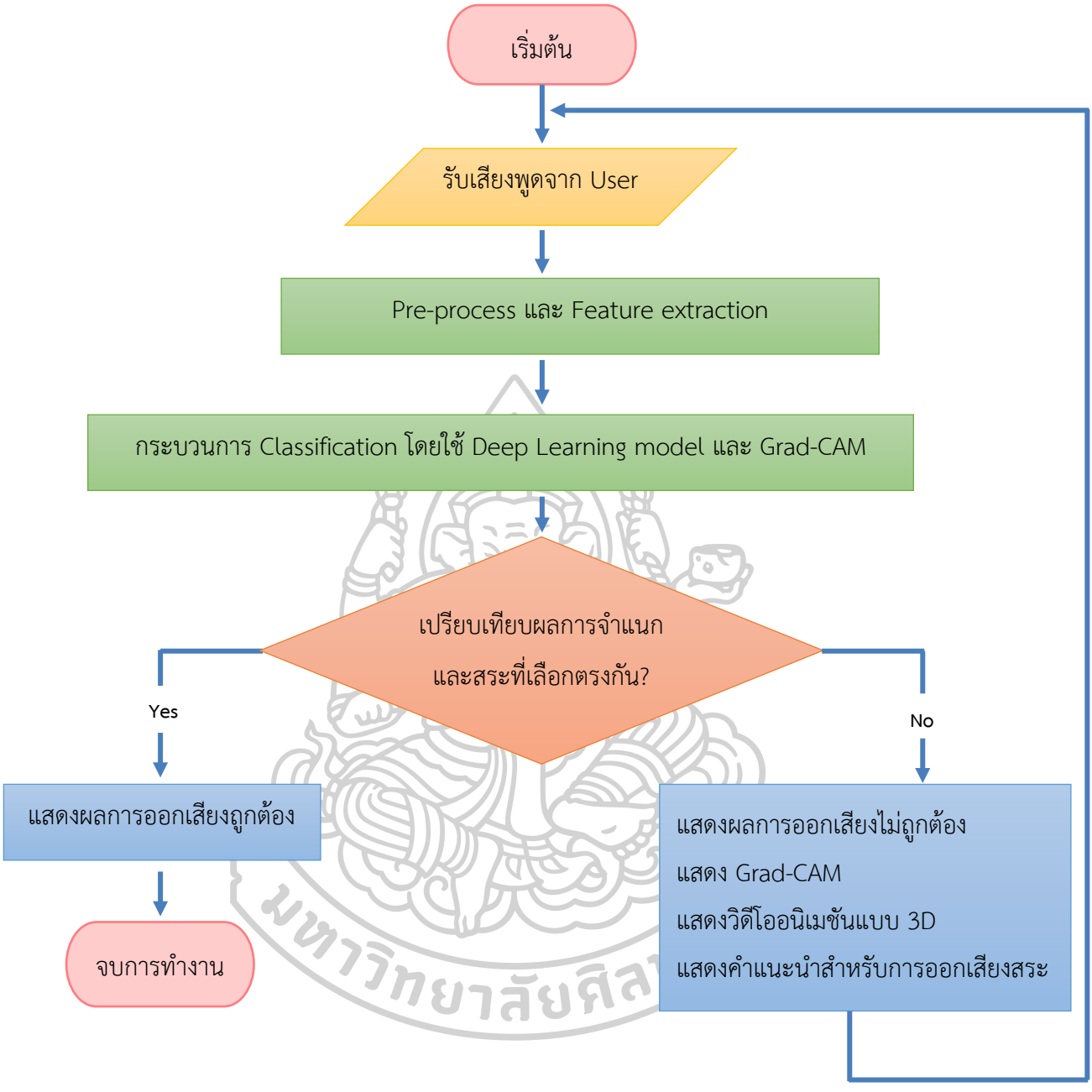
6.2.2. ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง

โครงสร้างหลักของระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง แบ่งเป็น 2 ส่วน คือ ส่วนของ Back-end และส่วนของ Front-end

Back-end

การทำงานภายในระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง เริ่มจากการรับเสียงจากผู้ใช้ระบบ (User) บันทึกเป็นไฟล์นามสกุล .wav เข้ามาในระบบ เมื่อผู้ใช้งานพูดออกเสียงสระเสร็จสิ้น ระบบจะทำการส่งไฟล์เสียงพร้อมกับค่า Identification (ID) ของสระที่ผู้ใช้เลือกที่จะทำการฝึกไปยัง back-end ไฟล์เสียงที่บันทึกได้จากการฝึกพูดของผู้ใช้จะถูกนำมาเข้าสู่กระบวนการประมวลผลล่วงหน้าและการสกัดคุณลักษณะ (Preprocessing and feature extraction) เพื่อที่จะได้ลักษณะรูปแบบของข้อมูลข้อมูลเข้า (Input data) ที่เหมาะสมที่จะนำไปจำแนกเสียงสระภาษาไทยในขั้นตอนต่อไป โดยมีขั้นตอนการตัดเสียงที่เงียบออกก่อน ซึ่งกำหนดเกณฑ์ความเงียบเป็น -30 dBFS (Decibel relative to Full Scale) เนื่องจากเป็นการกำหนดค่าให้ตัดเสียงที่เหมาะสมในการทดลองนี้ โดยเสียงเงียบที่เกิดขึ้นที่ช่วงหน้ากับช่วงหลังของไฟล์เสียงจถูกตัดออก และการบันทึกไฟล์เสียงเป็นนามสกุล .wav ใหม่ จากนั้นนำไฟล์เสียงที่ได้หลังจากการตัดเสียงเงียบมาตัดใหม่อีกครั้ง เพื่อให้ได้เสียงที่อยู่กึ่งกลางซึ่งเป็นส่วนของเสียงสระที่ชัดเจนที่สุด โดยการตัดไฟล์เสียงช่วงหน้ากับช่วงหลัง 25% ของความยาวไฟล์เสียง ซึ่งเป็น

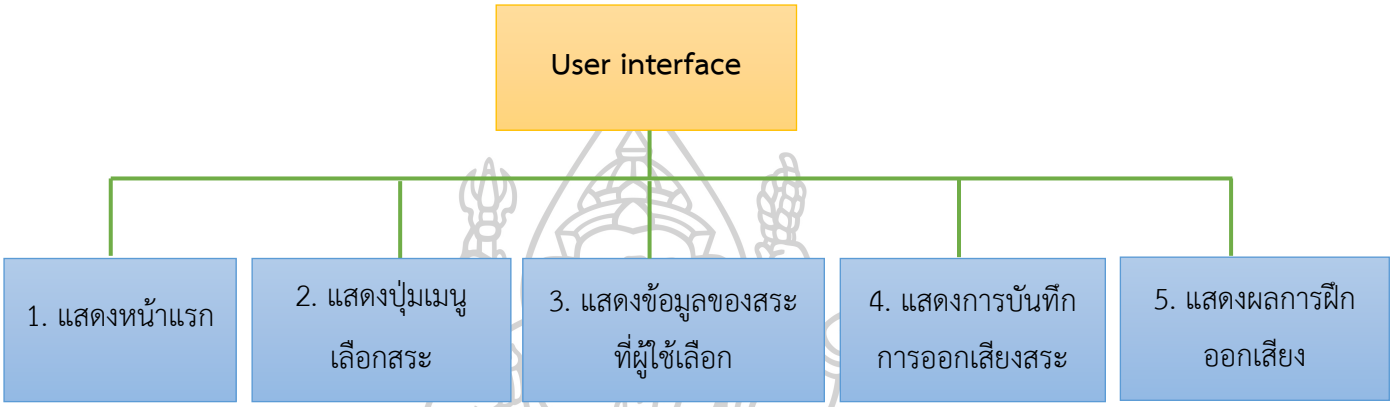
วิธีการเดียวกับนักภาษาศาสตร์ใช้ในการวิเคราะห์เสียงสระแบบดั้งเดิม และบันทึกไฟล์เสียงเป็นนามสกุล .wav ใหม่อีกครั้ง จากนั้นนำไฟล์เสียงที่ได้ไปประมวลผลเพื่อแปลงคลื่นเสียงเป็นข้อมูลข้อมูลเข้ารูปแบบเวลาและความถี่ (time-frequency input data) โดยข้อมูลข้อมูลเข้าจะถูกสกัดเป็นคุณสมบัติข้อมูลเข้า (input features) ด้วยวิธีการสกัดคุณลักษณะที่ใช้ MS ซึ่งเป็นการแปลงไฟล์ .wav ให้อยู่ในรูปแบบ MS จากนั้นจะเข้าสู่กระบวนการจำแนกประเภท (Classification) โดยใช้สถาปัตยกรรม Deep Learning ซึ่งเป็นโมเดลการรู้จำเสียงสระภาษาไทยในงานวิจัยนี้คือโมเดล Convolutional Neural Networks (CNN) เพื่อจำแนกประเภทเสียงสระภาษาไทยทั้ง 18 คลาส และประมวลผลภาพ Grad-CAM ของเสียงสระที่เกิดจากผู้พูด ซึ่งโมเดลนี้เป็นส่วนที่สำคัญที่สุดในการรู้จำเสียงสระสำหรับการออกเสียงสระภาษาไทย จากนั้นระบบจะให้ผลการทำนายเสียงว่าตรงกับคลาสของเสียงสระใด หลังจากนั้นจะเป็นขั้นตอนการเปรียบเทียบ โดยระบบจะทำการเปรียบเทียบ ID ของสระที่ผู้ใช้เลือกทำการฝึกเปรียบเทียบกับผลการจำแนกเสียงที่ได้ ถ้าค่า ID ของสระที่รับเข้าไปตรงกับผลการจำแนกเสียงแสดงว่าผู้ใช้ออกเสียงถูกต้อง ซึ่งผลลัพธ์จะถูกส่งไปแสดงที่หน้าเว็บ หากการเปรียบเทียบไม่ตรงกันจะแสดงผลให้ผู้ใช้ทราบว่าออกเสียงไม่ถูกต้อง และมีความคล้ายคลึงกับสระใด ระบบจะแสดงผลของ Grad-CAM เพื่อให้ผู้ใช้สามารถนำไปสังเกตุความแตกต่างจากภาพรวมของ Grad-CAM ของสระนั้น เพื่อนำมาปรับปรุงการฝึกออกเสียง ระบบจะแสดงวิดีโออนิเมชันแบบ 3D พร้อมทั้งแสดงคำแนะนำสำหรับการออกเสียงสระ เพื่อให้ผู้ใช้ออกเสียงได้ถูกต้องมากขึ้น ดังแสดงรูปที่ 43



รูปที่ 43 แสดงสถาปัตยกรรมของระบบ (Back-end)

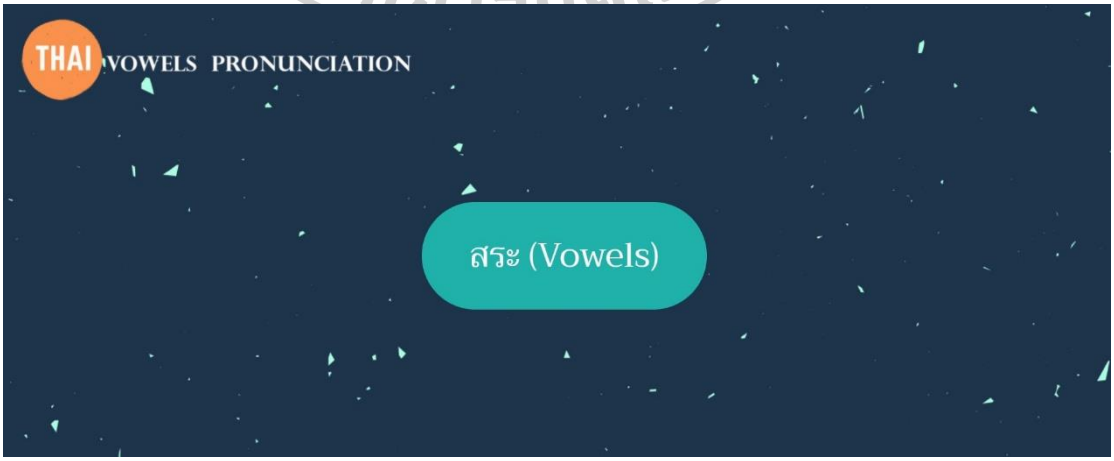
Front-end

ส่วนติดต่อกับผู้ใช้ (User interface) ของระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง สามารถแบ่งเป็น 5 ขั้นตอน ประกอบด้วย 1) แสดงหน้าแรก , 2) แสดงปุ่มเมนูเลือกสระ, 3) แสดงข้อมูลของสระที่ผู้ใช้เลือก, 4) แสดงการบันทึกการออกเสียงสระ และ 5) แสดงผลการฝึกออกเสียง ซึ่งภาพรวมของ Front-end แสดงดังรูปที่ 44



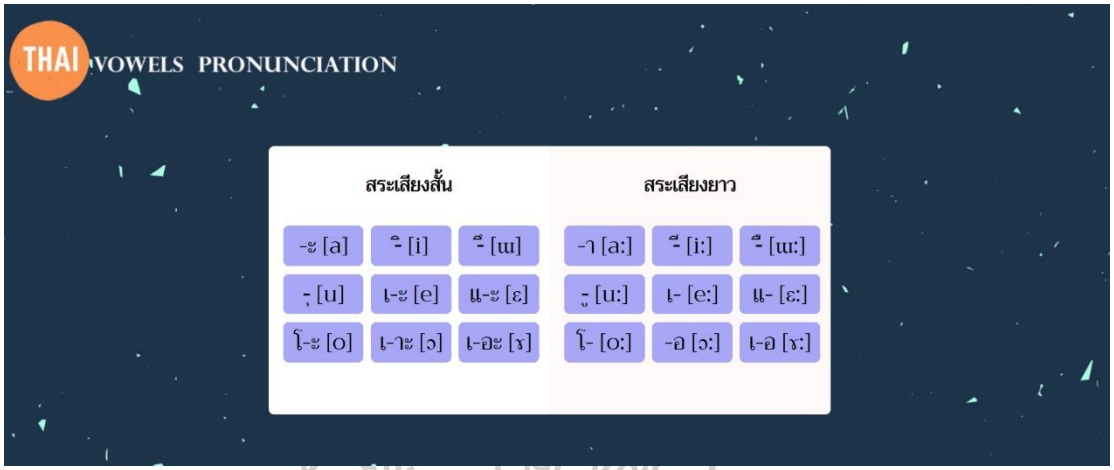
รูปที่ 44 แสดงสถาปัตยกรรมของระบบ (Front-End)

เมื่อผู้ใช้ระบบ (User) เข้าใช้งานระบบในส่วนของ Front-end ขั้นตอนที่ 1 ระบบจะแสดงหน้าแรกของระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง ซึ่งหน้าแรกจะแสดงส่วนของเมนู ซึ่งประกอบด้วยหัวข้อ “สระ (Vowels)” กดเลือกเพื่อทำการฝึกออกเสียงสระดังรูปที่ 45



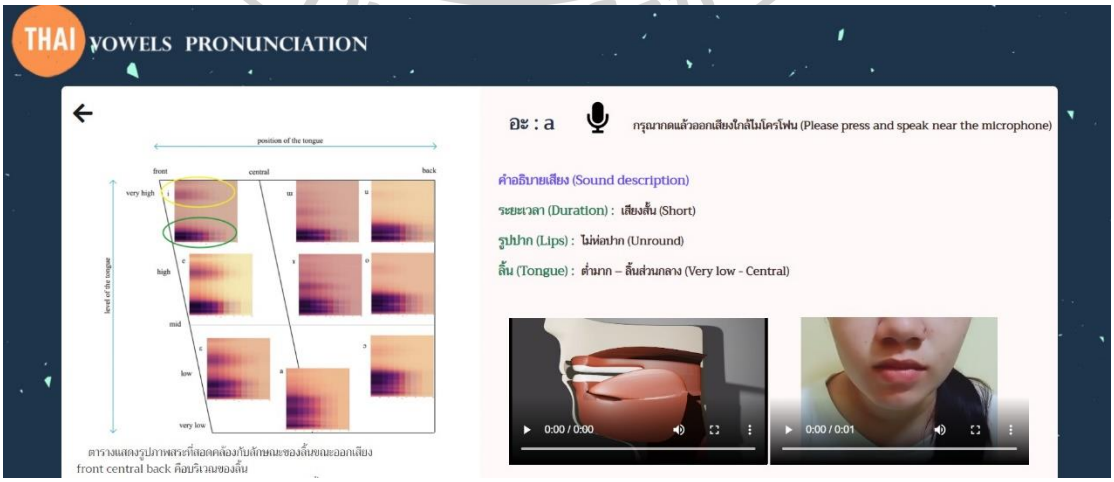
รูปที่ 45 แสดงหน้าแรกของระบบการฝึกออกเสียงอัตโนมัติสำหรับเสียงสระภาษาไทย 18 เสียง

ขั้นตอนที่ 2 ผู้ใช้สามารถเลือกคลิกปุ่มเมนูเพื่อเลือกสระที่ต้องการฝึกออกเสียง ซึ่งจะแสดงหน้าต่างของเสียงสระทั้งหมด 18 เสียง ประกอบไปด้วยสระเสียงสั้น 9 เสียงและสระเสียงยาว 9 เสียง (แสดงตัวอักษรภาษาไทย และ สัญลักษณ์ IPA) โดยผู้ใช้ทำกดการเลือกเสียงสระที่ต้องการฝึกอ่านออกเสียงดังรูปที่ 46



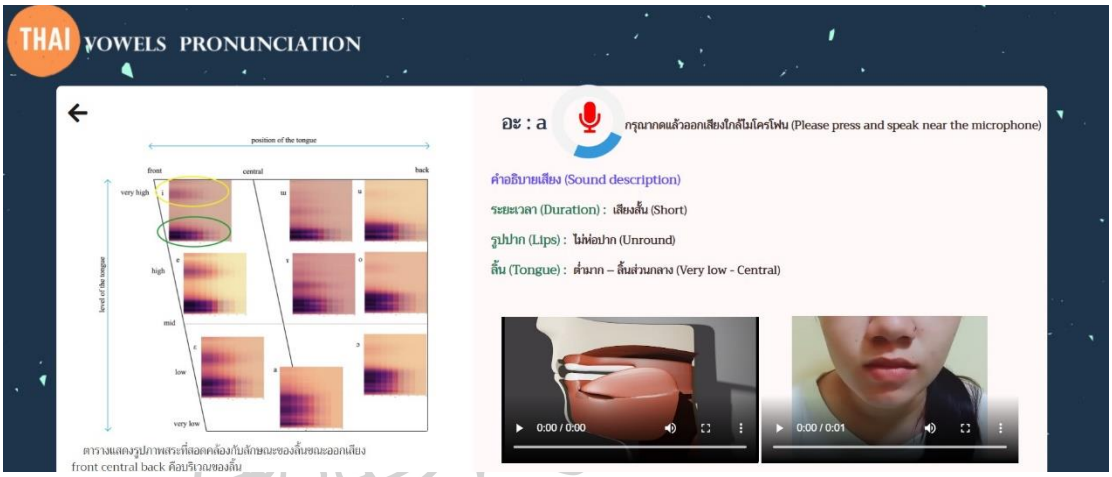
รูปที่ 46 แสดงปุ่มเมนูสระที่ต้องการฝึกออกเสียง

ขั้นตอนที่ 3 เมื่อทำการเลือกสระที่ต้องการฝึกเรียบร้อยแล้ว ระบบจะทำการดึงข้อมูลของสระนั้นมาแสดงทางหน้าเว็บ ซึ่งจะปรากฏหน้าต่างเนื้อหาของเสียงสระนั้น โดยประกอบไปด้วย ภาพรวม Grad-CAM ของสระที่ต้องการฝึกออกเสียงพร้อมคำบรรยายวิธีการสังเกต คลิปวิดีโอเสียงสระ 3 มิติ ที่แสดงการเคลื่อนที่ของลิ้น คลิปวีดีโอการขยับปาก และข้อความที่อธิบายว่าการออกเสียงที่ถูกต้องเป็นอย่างไรดังรูปที่ 47



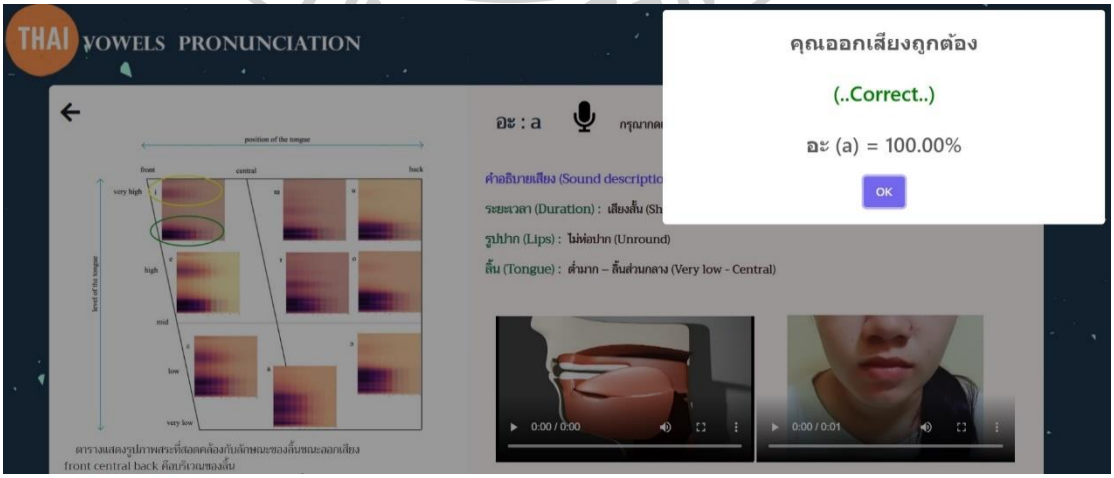
รูปที่ 47 แสดงหน้าต่างเนื้อหาของเสียงสระที่ต้องการฝึกออกเสียง

ขั้นตอนที่ 4 รูปที่ 48 แสดงการบันทึกการออกเสียงสระ เมื่อผู้ใช้ต้องการฝึกการออกเสียงสระ ผู้ใช้สามารถกดไอคอนไมโครโฟนพร้อมทั้งฝึกออกเสียงสระดังกล่าว เมื่อไมโครโฟนเปลี่ยนเป็นสีแดง ให้ผู้ใช้สามารถออกเสียงได้ โดยระบบจะรับเสียงเป็นระยะเวลา 2 วินาที เนื่องจากระยะเวลาบันทึกเสียง 2 วินาที เป็นเวลาที่เหมาะสมสำหรับการออกเสียงสระหนึ่งสระ และระบบจะหยุดบันทึกเองหลังจากวงกลมวนครบรอบและปุ่มไมค์เปลี่ยนเป็นสีดำเป็นการสิ้นสุดการบันทึกเสียง ผลลัพธ์ที่ได้จะเป็นไฟล์เสียงนามสกุล .wav และไฟล์เสียงจะถูกส่งไปยังทางด้าน back-end



รูปที่ 48 แสดงการบันทึกการออกเสียงสระ

ขั้นตอนที่ 5 ระบบจะทำการตรวจสอบเสียงที่ป้อนเข้ามาจากผู้ใช้ จากนั้นดึงผลลัพธ์ของการตรวจสอบโดยระบบจะวิเคราะห์ความถูกต้องจากการออกเสียงของผู้ใช้ และแสดงผลลัพธ์เป็นข้อความให้ผู้ใช้ทางหน้าเว็บดังรูปที่ 49 กรณีที่ผู้ใช้ออกเสียงถูก และในกรณีที่ผู้ใช้ออกเสียงผิดดังรูปที่ 50



รูปที่ 49 แสดงผลลัพธ์เป็นข้อความให้ผู้ใช้ทางหน้าเว็บกรณีที่ผู้ใช้ออกเสียงถูกต้อง

รูปที่ 49 แสดงผลลัพธ์จากการประมวลผลเสียงที่ผู้ใช้พูดฝึกออกเสียงสระ หากผู้ใช้ออกเสียงถูกต้อง จะแสดงผลลัพธ์การออกเสียงว่า “คุณออกเสียงถูกต้อง” พร้อมกับบอก % ของการออกเสียงที่เหมือนกับสระที่เลือก ซึ่งแสดงข้อความทั้งภาษาไทยและภาษาอังกฤษ

หากผู้ใช้ออกเสียงไม่ถูกต้อง จะแสดงผลลัพธ์การออกเสียงว่า ออกเสียงไม่ถูกต้องพร้อมทั้งบอกว่าผู้ใช้ออกเสียงผิดคล้ายกับเสียงสระใด พร้อมกับบอก % ของการออกเสียงคล้าย และแสดงภาพ Grad-CAM ของสระที่สัมพันธ์กับลิ้น เพื่อเปรียบเทียบกับภาพรวมของ Grad-CAM ในกราฟว่าต่างจากภาพที่ต้องการออกเสียงสระนั้นอย่างไร ซึ่งผู้ใช้สามารถฝึกตามวิดีโอภาพ 3D ของลิ้น และใช้กราฟเปรียบเทียบ เพื่อฝึกออกเสียงสระให้ดียิ่งขึ้นต่อไป โดยข้อความแสดงทั้งภาษาไทยและภาษาอังกฤษดังรูปที่ 50

รูปที่ 50 แสดงผลลัพธ์เป็นข้อความให้ผู้ใช้ทางหน้าเว็บกรณีที่ผู้ใช้ออกเสียงไม่ถูกต้อง

6.2.3. การประเมินความพึงพอใจของผู้ใช้ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง

ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง ซึ่งพัฒนาในรูปแบบของเว็บแอปพลิเคชันได้ถูกนำมาสำรวจความพึงพอใจของผู้ใช้ระบบ โดยวัตถุประสงค์ของการประเมินความพึงพอใจในงานนี้เพื่อนำไปใช้ในการปรับปรุง แก้ไข และพัฒนาระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย ซึ่งมีรายละเอียดดังนี้

1. ผู้ประเมินระบบ

การประเมินที่เกี่ยวข้องกับระบบในงานนี้ มีผู้ใช้ระบบเป็นผู้ประเมิน จำนวน 29 คน ซึ่งสุ่มแบบเจาะจง (Purposive Sampling) จากเพศชายจำนวน 13 คน เพศหญิงจำนวน 16 คน แบ่งเป็นผู้เรียน 25 คน และผู้เชี่ยวชาญ 4 คน โดยผู้เรียนแบ่งเป็น 4 กลุ่มดังนี้ 1) กลุ่มคนไทยที่ใช้ภาษาพูดเป็นภาษาไทยแบบมาตรฐานที่มีอายุระหว่าง 16-30 ปี จำนวน 4 คน ประกอบด้วย เพศชาย 2 คน และเพศหญิง 2 คน 2) กลุ่มคนไทยที่กลุ่มคนไทยที่ใช้ภาษาพูดเป็นภาษาไทยแบบมาตรฐานที่มีอายุ 31 ปีขึ้นไป จำนวน 4 คน ประกอบด้วย เพศชาย 2 คน และเพศหญิง 2 คน 3) กลุ่มคนไทยที่ใช้ภาษาพูดเป็นภาษาไทยแบบสำเนียงท้องถิ่น จำนวน 8 คน ประกอบด้วย เพศชาย 4 คน และเพศหญิง 4 คน และ 4) กลุ่มผู้ใช้ที่ไม่ได้ใช้ภาษาไทยในการพูดสื่อสาร จำนวน 9 คน ประกอบด้วย เพศชาย 4 คน และเพศหญิง 5 คน และมีผู้เชี่ยวชาญ 4 คน ประกอบด้วย เพศชาย 1 คน และเพศหญิง 3 คน

2. เครื่องมือที่ใช้ในวิจัย

เครื่องมือที่ใช้ในการวิจัยนี้ได้แก่ แบบสอบถาม ประกอบได้คำถามด้วย 3 ส่วน ดังต่อไปนี้

ส่วนที่ 1 แบบสอบถามเกี่ยวกับสถานภาพทั่วไปผู้ตอบแบบสอบถาม ได้แก่ เพศ อายุ ระดับการศึกษาปัจจุบัน ภาษาพูดที่ใช้ในการสื่อสาร และสถานะผู้ใช้ เป็นแบบสอบถามแบบเลือกตอบเพียง 1 ข้อ (Check list)

ส่วนที่ 2 แบบสอบถามเกี่ยวกับความพึงพอใจของผู้ใช้ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง เป็นแบบสอบถามแบบมาตรส่วนประมาณค่า (Rating scale) ตามวิธีของลิเคิร์ท (Likert scale) [101] เนื่องจากเป็นวิธีที่ง่ายและสะดวก สามารถเก็บข้อมูลได้รวดเร็ว พิจารณาจากระดับคะแนนเฉลี่ยตามเกณฑ์ที่จัดชั้น (Class interval) ซึ่งแบ่งเป็น 5 ระดับ ให้คะแนนตามกำหนดไว้ดังนี้

พึงพอใจมากที่สุด	ให้	5	คะแนน
พึงพอใจมาก	ให้	4	คะแนน
พึงพอใจปานกลาง	ให้	3	คะแนน
พึงพอใจน้อย	ให้	2	คะแนน
พึงพอใจน้อยที่สุด	ให้	1	คะแนน

ส่วนที่ 3 ข้อเสนอแนะทั่วไป เป็นแบบสอบถามแบบปลายเปิด (Open-ended)

3. การเก็บรวบรวมข้อมูล

1. นำแบบสอบถามฉบับสมบูรณ์ที่อยู่ในรูปแบบออนไลน์ที่ใช้ Google form ส่งให้กับผู้ใช้ระบบ

2. ทำการวิเคราะห์และประเมินผลข้อมูล

4. การวิเคราะห์ข้อมูล

ส่วนที่ 1 ข้อมูลทั่วไปผู้ตอบแบบสอบถาม เป็นแบบสอบถามแบบเลือกตอบเพียง 1 ข้อ (Check – list) ใช้การวิเคราะห์ข้อมูลโดยหาค่าความถี่และค่าร้อยละ (Percentage)

ส่วนที่ 2 ความพึงพอใจ เป็นแบบสอบถามแบบมาตราส่วนประมาณค่า (Rating scale) สถิติที่ใช้ในการวิเคราะห์ข้อมูล คือ ค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐาน ใช้การแปลความหมายของข้อมูลที่ได้จากแบบสอบถาม สำหรับเกณฑ์การแปลค่าเฉลี่ยได้กำหนดน้ำหนักคะแนนในแบ่งชั้น 5 ระดับเท่า ๆ กัน ดังนี้

ค่าเฉลี่ย	4.21 – 5.00	ความพึงพอใจอยู่ในระดับ มากที่สุด
ค่าเฉลี่ย	3.41 – 4.20	ความพึงพอใจอยู่ในระดับ มาก
ค่าเฉลี่ย	2.61 – 3.40	ความพึงพอใจอยู่ในระดับ ปานกลาง
ค่าเฉลี่ย	1.81 – 2.60	ความพึงพอใจอยู่ในระดับ น้อย
ค่าเฉลี่ย	1.00 – 1.80	ความพึงพอใจอยู่ในระดับ น้อยที่สุด

ส่วนที่ 3 ข้อเสนอแนะทั่วไป เป็นแบบสอบถามแบบปลายเปิด (Open-ended) ใช้การวิเคราะห์ข้อมูลเชิงเนื้อหา (Content analysis)

5. สถิติที่ใช้ในการวิเคราะห์ข้อมูล

ในงานวิจัยนี้ใช้การวิเคราะห์เชิงปริมาณ เพื่อวัดความพึงพอใจใช้ค่าความถี่ (Frequency) ค่าร้อยละ (Percentage) ดังสมการที่ 35, ค่าเฉลี่ย (Mean) ดังสมการที่ 36 และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ดังสมการที่ 37

1) ค่าร้อยละ (Percentage) คำนวณจากสูตร

$$P = \frac{f}{N} \times 100 \quad (35)$$

เมื่อ p แทน ค่าร้อยละ

f แทน ความถี่ที่ต้องการแปลงให้เป็นร้อยละ

N แทน จำนวนความถี่ทั้งหมด

2) ค่าเฉลี่ย (Mean) คำนวณจากสูตร

$$\overline{X} = \frac{\sum X}{N} \quad (36)$$

เมื่อ $\overline{X}$ แทน ค่าเฉลี่ย

$\sum X$ แทน ผลรวมของคะแนนในกลุ่ม

n แทน จำนวนผู้ใช้ในกลุ่มตัวอย่างที่ตอบแบบสอบถาม

3) ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation: S.D.)

$$S.D. = \sqrt{\frac{n\sum fx^2 - (\sum fx^2)}{n\,(n-1)}} \quad (37)$$

6. ผลการประเมินความพึงพอใจของผู้ใช้ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง

งานวิจัยนี้ได้สำรวจความพึงพอใจของผู้ใช้ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง ซึ่งได้แบ่งการวิเคราะห์ผลการวิจัยดังนี้

ตอนที่ 1 ข้อมูลทั่วไปของผู้ตอบแบบสอบถาม

ตอนที่ 2 ความพึงพอใจของผู้ใช้ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง

ตอนที่ 3 ข้อเสนอแนะอื่น ๆ

ตอนที่ 1 ข้อมูลทั่วไปของผู้ตอบแบบสอบถาม

ตารางที่ 15 แสดงจำนวนและค่าร้อยละของจำนวนผู้ใช้ที่ตอบแบบสอบถาม

ข้อมูลทั่วไป	จำนวน (n = 29)	ร้อยละ
1. เพศ		
ชาย	13	44.83
หญิง	16	55.17
2. อายุ		
ต่ำกว่า หรือเท่ากับ 15	0	0.00
16-30 ปี	19	65.52
31 ปี ขึ้นไป	10	34.48
3. ระดับการศึกษาปัจจุบัน		
อนุบาล	0	0.00
ประถมศึกษา	0	0.00
มัธยมศึกษา	0	0.00
กำลังศึกษาระดับมหาวิทยาลัย	10	34.48
สำเร็จการศึกษา	19	65.52
4. ภาษาพูดที่ใช้สื่อสาร		
ภาษาไทย สำเนียงมาตรฐาน	11	37.93
ภาษาไทย สำเนียงท้องถิ่น	9	31.03
ภาษาอื่นๆ ที่ไม่ใช่ภาษาไทย	9	31.03
5. สถานะผู้ใช้ระบบ		
ผู้ใช้งานระบบ หรือ ผู้เรียน	25	86.21
ผู้เชี่ยวชาญ (ครู-อาจารย์ ภาษาไทย, นักภาษาศาสตร์)	4	13.79

จากตารางที่ 15 พบว่า ผู้ตอบแบบสอบถามส่วนใหญ่เป็นเพศหญิง จำนวน 16 คน คิดเป็นร้อยละ 55.17 ส่วนใหญ่มีอายุ 16-30 ปี จำนวน 19 คน คิดเป็นร้อยละ 65.52 ระดับการศึกษาปัจจุบันส่วนใหญ่สำเร็จการศึกษา จำนวน 19 คน คิดเป็นร้อยละ 65.52 ภาษาพูดที่ใช้สื่อสารส่วนใหญ่ใช้ภาษาไทย สำเนียงมาตรฐาน จำนวน 11 คน คิดเป็นร้อยละ 37.93 และส่วนใหญ่มีสถานะผู้ใช้ระบบเป็นผู้ใช้งานระบบ หรือ ผู้เรียน จำนวน 25 คน คิดเป็นร้อยละ 86.21

ตอนที่ 2 ความพึงพอใจของผู้ใช้ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง

ตารางที่ 16 แสดงผลการประเมินความพึงพอใจ

รายการประเมิน	ระดับความพึงพอใจ				
	มากที่สุด 5	มาก 4	ปานกลาง 3	น้อย 2	น้อยที่สุด 1
การออกแบบส่วนของผู้ใช้งาน					
1. ความสวยงาม ความทันสมัย น่าสนใจของหน้าเว็บไซต์	11	15	3	-	-
2. การจัดรูปแบบในเว็บไซต์ง่ายต่อการอ่านและการใช้งาน	16	11	2	-	-
3. สีสันในการออกแบบเว็บไซต์มีความเหมาะสม	14	12	3	-	-
4. เมนูง่ายต่อการใช้งาน	18	9	2	-	-
5. ขนาดตัวอักษร และรูปแบบตัวอักษร อ่านได้ง่ายและสวยงาม	16	11	2	-	-
การใช้งานของระบบ					
1. ความถูกต้องครบถ้วนของข้อมูล	19	8	2	-	-
2. ภาพกับเนื้อหามีความสอดคล้องกันและสามารถสื่อความหมายได้	22	6	1	-	-
3. ความเหมาะสมของข้อมูลภายในเว็บไซต์	13	15	1	-	-
4. ความสะดวกในการเชื่อมโยงข้อมูลภายในเว็บไซต์	16	8	4	1	-
ประโยชน์และการนำไปใช้งาน					
1. เพิ่มทักษะออกเสียงสระภาษาไทย	21	7	1	-	-
2. ลดปัญหาการออกเสียงสระภาษาไทยที่ไม่ถูกวิธี	21	7	1	-	-
3. เข้าใจหลักการออกเสียงสระภาษาไทยที่ถูกต้อง	21	8	-	-	-
4. สามารถนำทักษะไปปรับใช้ในการออกเสียงในชีวิตประจำวัน	23	5	1	-	-

ตารางที่ 17 แสดงผลการวัดระดับความพึงพอใจ

รายการประเมิน	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน	ระดับความพึงพอใจ
การออกแบบส่วนของผู้ใช้งาน			
1. ความสวยงาม ความทันสมัย น่าสนใจของหน้าเว็บไซต์	4.28	0.65	มากที่สุด
2. การจัดรูปแบบในเว็บไซต์ง่ายต่อการอ่านและการใช้งาน	4.48	0.63	มากที่สุด
3. สีสันในการออกแบบเว็บไซต์มีความเหมาะสม	4.38	0.68	มากที่สุด
4. เมนูง่ายต่อการใช้งาน	4.55	0.63	มากที่สุด
5. ขนาดตัวอักษร และรูปแบบตัวอักษร อ่านได้ง่ายและสวยงาม	4.48	0.63	มากที่สุด
รวม	**4.43**	**0.64**	**มากที่สุด**
การใช้งานของระบบ			
1. ความถูกต้องครบถ้วนของข้อมูล	4.59	0.63	มากที่สุด
2. ภาพกับเนื้อหามีความสอดคล้องกันและสามารถสื่อความหมายได้	4.72	0.53	มากที่สุด
3. ความเหมาะสมของข้อมูลภายในเว็บไซต์	4.41	0.57	มากที่สุด
4. ความสะดวกในการเชื่อมโยงข้อมูลภายในเว็บไซต์	4.34	0.86	มากที่สุด
รวม	**4.52**	**0.65**	**มากที่สุด**
ประโยชน์และการนำไปใช้งาน			
1. เพิ่มทักษะออกเสียงสระภาษาไทย	4.69	0.54	มากที่สุด
2. ลดปัญหาการออกเสียงสระภาษาไทยที่ไม่ถูกวิธี	4.69	0.54	มากที่สุด
3. เข้าใจหลักการออกเสียงสระภาษาไทยที่ถูกต้อง	4.72	0.45	มากที่สุด
4. สามารถนำทักษะไปปรับใช้ในการออกเสียงในชีวิตประจำวัน	4.76	0.51	มากที่สุด
รวม	**4.72**	**0.51**	**มากที่สุด**
รวมความพึงพอใจในการใช้ระบบภาพรวม	**4.55**	**0.60**	**มากที่สุด**
ค่าเฉลี่ยรวมความพึงพอใจ คิดเป็นร้อยละ (4.55 / 5) x 100 = 91.00			

ตารางที่ 17 เป็นการแสดงผลการประเมินความพึงพอใจของผู้ใช้ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง จำนวน 29 คน และการแปลผลการประเมินระบบโดยวัดจากความพึงพอใจของผู้ใช้งานระบบได้ผลดังตารางที่ 17 พบว่าผู้ตอบแบบสอบถามส่วนใหญ่มีความพึงพอใจในภาพรวมอยู่ในระดับมากที่สุด (ค่าเฉลี่ย = 4.55 ส่วนเบี่ยงเบนมาตรฐาน = 0.60) โดยคิดเป็นร้อยละ 91.00 เมื่อพิจารณารายด้าน พบว่าพึงพอใจมากที่สุดทั้ง 3 ด้าน ได้แก่ ด้านประโยชน์และการนำไปใช้งาน (ค่าเฉลี่ย = 4.72 ส่วนเบี่ยงเบนมาตรฐาน = 0.51) คิดเป็นร้อยละ 94.40 ด้านการใช้งานของระบบ (ค่าเฉลี่ย = 4.52 ส่วนเบี่ยงเบนมาตรฐาน = 0.65) คิดเป็นร้อยละ 90.40 และด้านการออกแบบส่วนของผู้ใช้งาน (ค่าเฉลี่ย = 4.43 ส่วนเบี่ยงเบนมาตรฐาน = 0.65) คิดเป็นร้อยละ 88.60

ตอนที่ 3 ข้อเสนอแนะอื่น ๆ

1. เว็บไซต์ควรมีสีสันและดึงดูดตา

2. การออกแบบและการจัดรูปแบบควรมีความสะดวก โดยเฉพาะเวลาเข้าที่ options ต่างๆ อย่างเช่น เวลาเข้าที่เมนูสระ อยากให้ตัวสระทั้งหมดยังอยู่ไม่ว่าจะกดสระใด เพื่อทำให้ผู้ใช้งานมีความสะดวกขึ้นเวลาใช้ระบบเพราะไม่จำเป็นต้องกด Back เวลาที่อยากทดสอบสระอื่นๆ

3. เป็นประโยชน์สำหรับผู้ที่ออกเสียงสระมาตรฐานไม่ชัดเจน รูปแบบเว็บไซต์ไม่ซับซ้อนและเข้าใจง่าย

4. ควรเพิ่มตัวเลือกบนเว็บไซต์ เช่น next (ปุ่มถัดไป) และ previous (ปุ่มก่อนหน้า) และควรใส่วิดีโอภาพเคลื่อนไหวของการออกเสียง

6.3. สรุป

งานวิจัยนี้นำเสนองานการรู้จำเสียงสระภาษาไทยโดยใช้โมเดล Convolutional Neural Network (CNN) กับการใช้เทคนิค gradient-weighted class activation mapping (Grad-CAM) เพื่อช่วยในเรื่องความโปร่งใสในการทำนายผลของโมเดล และอธิบายบริเวณของพื้นที่ที่มีความสำคัญในการทำนายของโมเดล สำหรับเสียงสระภาษาไทยทั้ง 18 เสียง ชุดข้อมูลถูกรวบรวมและได้รับการตรวจสอบโดยนักภาษาศาสตร์ ซึ่งประกอบด้วยสระเสียงสั้น 9 เสียงและสระเสียงยาว 9 เสียง โดยรวบรวมจากผู้พูดภาษาไทย 50 คน (เพศชาย 25 คนและเพศหญิง 25 คน) จำนวน 1,800 ไฟล์เสียง มีการนำเสนอระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง ซึ่งพัฒนาในรูปแบบของเว็บแอปพลิเคชัน จากนั้นทำการสำรวจความพึงพอใจของผู้ใช้ระบบ ผลการประเมินความพึงพอใจของผู้ใช้ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง จำนวน 29 คน พบว่าผู้ตอบแบบสอบถามส่วนใหญ่มีความพึงพอใจใน

ภาพรวมอยู่ในระดับมากที่สุด (ค่าเฉลี่ย = 4.55 ส่วนเบี่ยงเบนมาตรฐาน = 0.60) โดยคิดเป็นร้อยละ 91.00

งานวิจัยนี้มีส่วนช่วยในการปรับปรุงประสิทธิภาพของการรู้จำเสียงสระภาษาไทย ผลการศึกษาของ Grad-CAM สามารถช่วยในเรื่องความโปร่งใสในการทำนายผลของโมเดล ถูกนำไปเปรียบเทียบหาความสัมพันธ์กับหลักการภาษาศาสตร์ ซึ่งนำไปพัฒนาระบบคอมพิวเตอร์ช่วยฝึกการออกเสียงแบบอัตโนมัติสำหรับเสียงสระภาษาไทยได้ ทำให้ระบบมีประสิทธิภาพและความถูกต้องมากยิ่งขึ้น ผลการประเมินความพึงพอใจในระบบสามารถนำไปใช้ในการปรับปรุง แก้ไข และพัฒนาระบบ โดยระบบสามารถนำไปประยุกต์ใช้กับการฝึกออกเสียงสระภาษาไทยกับผู้ที่ไม่ใช่เจ้าของภาษา หรือบุคคลที่มีความผิดปกติในการพูดได้

บทที่ 7

วิธีดำเนินงานวิจัยและผลการทดลองที่ 4

โครงสร้างโมเดล Convolutional Neural Network ร่วมกับ Long Short-Term Memory ในการรู้จำการออกเสียงสระภาษาไทย 18 เสียง

การออกเสียงผิดสามารถทำให้ความหมายของคำจะเปลี่ยนไป ดังนั้นการออกเสียงที่มีประสิทธิภาพและเป็นมาตรฐานเป็นสิ่งสำคัญในการออกเสียงอย่างถูกต้อง งานวิจัยนี้นำเสนอการเรียนรู้เชิงลึกที่ประยุกต์ใช้โมเดล Convolutional Neural Network (CNN) ร่วมกับ Long Short-Term Memory (LSTM) เพื่อจดจำเสียงสระภาษาไทย โมเดลการเรียนรู้เชิงลึกเป็นส่วนที่สำคัญในการจดจำเสียงสระที่ออกเสียงสำหรับระบบการฝึกการออกเสียงโดยใช้คอมพิวเตอร์ช่วย (Computer-Assisted Pronunciation Training : CAPT) การระบุสระภาษาไทยที่ถูกต้องเมื่อพูดในสถานการณ์จริงถือเป็นความท้าทายหลักในการจดจำเสียงสระภาษาไทย ชุดข้อมูลสำหรับสระไทยได้รับการออกแบบ รวบรวม และตรวจสอบโดยนักภาษาศาสตร์ ผลลัพธ์ของโมเดล CNN_LSTM ร่วมกับ Mel spectrogram (MS) ให้ความถูกต้องแม่นยำ 99.17%

7.1. ชุดข้อมูลและวิธีการ (Datasets and Methods)

การอธิบายชุดข้อมูลและวิธีการมีรายละเอียดมีดังนี้ ส่วนที่ 7.1.1. อธิบายชุดข้อมูล ในส่วนที่ 7.1.2. แสดงการแปลงสเปกโตรแกรม ส่วนที่ 7.1.3. แสดงการจำแนกประเภทของโมเดล CNN_LSTM ส่วนสุดท้ายแสดงรายละเอียดการทดลอง

7.1.1. ชุดข้อมูล (Dataset)

การออกแบบชุดข้อมูล ชุดคำภาษาไทยที่ใช้บันทึกเสียงออกแบบโดยนักภาษาศาสตร์ รายการคำศัพท์นี้ได้รับการออกแบบตามทฤษฎีภาษาศาสตร์ โดยทุกคำมีลักษณะเฉพาะเหมือนกันซึ่งเป็นพยัญชนะเดียวกัน เสียงวรรณยุกต์เดียวกัน และพยัญชนะท้ายเดียวกัน แต่ต่างกันเฉพาะในเสียงสระเท่านั้น สำหรับการรวบรวมชุดข้อมูล เก็บรวบรวมชุดข้อมูลเสียงสระจากผู้พูดภาษาไทยที่พูดภาษากลางในสภาพแวดล้อมต่างๆ มีวิธีการเก็บและจำนวนข้อมูลเสียงที่ใช้ในงานวิจัยเช่นเดียวกับบทที่ 5

7.1.2. การแปลงสเปกโตรแกรม (Spectrogram conversion)

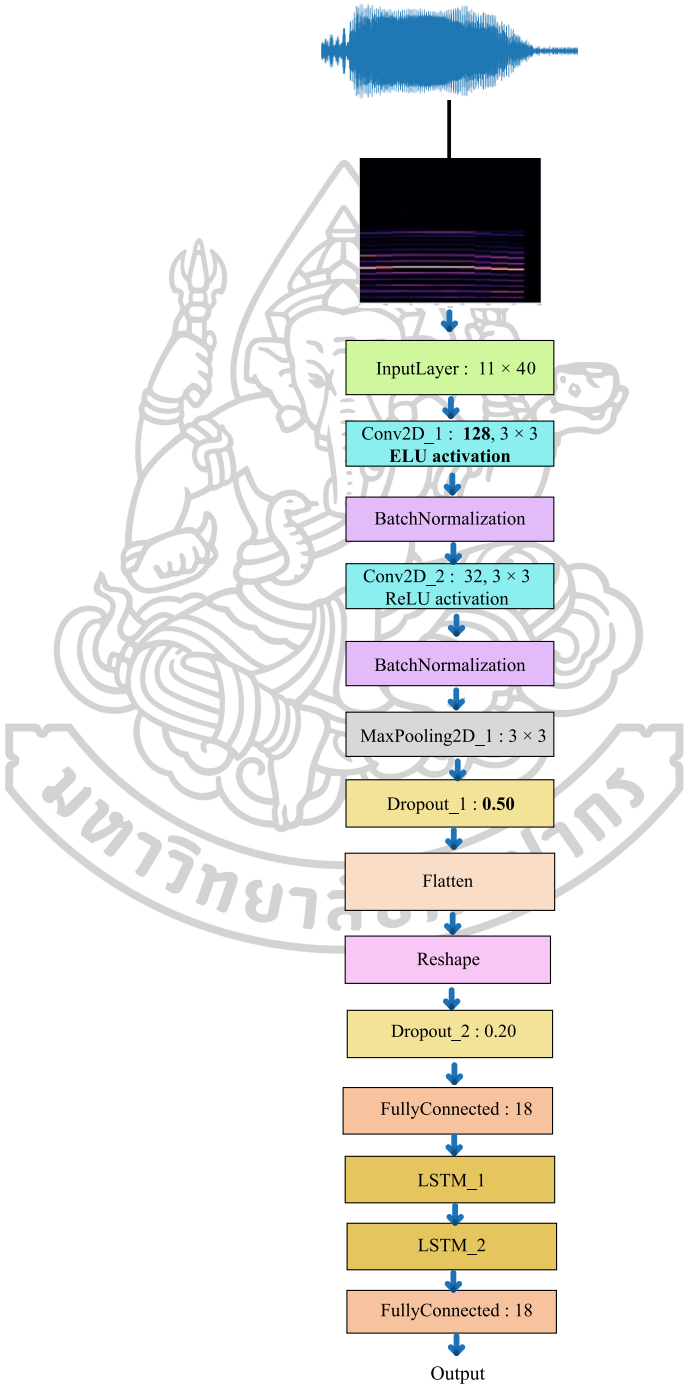
สัญญาณเสียงดิบถูกแปลงเป็น waveform แล้วแปลงเป็นสเปกโตรแกรมขนาดต่างๆ เพื่อค้นหาข้อมูลเข้าคุณสมบัติเสียงที่เหมาะสม สเปกโตรแกรมของภาพ 2 มิติประกอบด้วยแกนเวลา และ

แกนความถี่จะถูกแสดงด้วยลำดับของสเปกตรัม สัญญาณเสียงสระภาษาไทยถูกประมวลผลล่วงหน้า คุณลักษณะเสียงเริ่มต้นในการวิจัยนี้ ใช้ 40 Mel bands สำหรับคุณสมบัติเสียง MS, 40 MFCC สำหรับคุณสมบัติเสียง MFCC และเวกเตอร์ตามบริบท 11 [52, 74] การทดลองนี้ใช้คุณลักษณะเสียงสองแบบของกระบวนการสกัดคุณลักษณะ (MS และ MFCC) เพื่อเปรียบเทียบผลลัพธ์ด้านประสิทธิภาพจากคุณลักษณะเสียงที่แตกต่างกัน คุณลักษณะด้านเสียงถูกใช้เป็นคุณสมบัติข้อมูลเข้าสำหรับการส่งผ่านไปยังสถาปัตยกรรมการเรียนรู้เชิงลึกในกระบวนการจำแนกประเภท

7.1.3. การจำแนกประเภทด้วยโมเดล CNN ร่วมกับ LSTM

เสียงถูกส่งผ่านกระบวน preprocessing เพื่อแปลงคลื่นเสียงเป็นข้อมูลข้อมูลเข้าเวลาและความถี่ (time-frequency) จากนั้นจะถูกสกัดคุณสมบัติข้อมูลเข้า (input features) และส่งผ่านไปยังโมเดลการรู้จำเสียงสระภาษาไทย ซึ่งเป็นส่วนที่สำคัญในระบบการฝึกการออกเสียงอัตโนมัติสำหรับการออกเสียงสระภาษาไทย ขั้นตอนการจำแนกประเภทถูกใช้เพื่อจำแนก class labels ในเลเยอร์ fully connected หลังจากการสกัดคุณลักษณะ สัญญาณเสียงพูดสระภาษาไทยจะถูกแปลงเป็นเวกเตอร์คุณสมบัติเสียง MS หรือ MFCC และถูกส่งไปจำแนกเสียงสระภาษาไทยโดยใช้โมเดล CNN ที่รวมกับ LSTM ซึ่งได้รับแรงบันดาลใจจากการทบทวนวรรณกรรม [54] ที่นำเสนอการผสมผสานระหว่างโมเดล CNN และ LSTM สำหรับงานการรู้จำเสียง ในงานวิจัยนี้ใช้ Convolutional 2 เลเยอร์ และ LSTM 2 เลเยอร์ ที่คล้ายกับ [54] ขั้นแรกลดความแปรปรวนของความถี่ในข้อมูลข้อมูลเข้า โดยการถ่ายโอนข้อมูลเข้าผ่านเลเยอร์ Convolutional ซึ่งมีการใช้กลยุทธ์ Padding [52] ที่สามารถรักษาขนาดของแผนที่คุณลักษณะ กลยุทธ์ Pooling [51] ที่ลดความแปรปรวนของสเปกตรัมในคุณสมบัติข้อมูลเข้า กลยุทธ์ Dropout [91] ที่สามารถลดปัญหาการ overfitting จากนั้นมิติเอาต์พุตของ CNN สุดท้ายจะถูกเปลี่ยนรูปร่างและส่งผ่านเลเยอร์ LSTM ซึ่งเลเยอร์ LSTM นั้นเหมาะสำหรับการสร้างแบบจำลองสัญญาณเวลา แต่ละชั้นของ Convolutional ใช้ตัวกรอง Convolution กับข้อมูลเข้าคุณสมบัติเสียงตามด้วยฟังก์ชันการกระตุ้นแบบไม่เชิงเส้น (Nonlinear Activation Function) แต่ละเลเยอร์ convolutional จำนวน filters ถูกตั้งค่าเป็น 32, 64 และ 128 ในเลเยอร์ convolutional ที่แตกต่างกัน จำนวน batch sizes คือ 32, 64 และ 128 ค่า dropout ที่แตกต่างกันคือ 20%, 25%, 30%, และ 35% epoch เป็น 500 และ 1,000 ถูกดำเนินการในโมเดล ซึ่งรายละเอียดของโมเดล CNN_LSTM ที่เหมาะสมมีดังต่อไปนี้ เลเยอร์ Convolutional แรกประกอบด้วย 128 Filters (3 x 3), Elu Activation Function, ตามด้วยเลเยอร์ Batch Normalization และในเลเยอร์ Convolutional ที่สอง ประกอบด้วย 32 Filters (3 x 3), Relu Activation Function, Batch Normalization, Max-pooling (3 x 3) และ Dropout 0.50 สำหรับเลเยอร์ LSTM ทั้ง 2 เลเยอร์ ประกอบด้วย 512 Units, Tanh Activation Function และ

Dropout 0.20 มีการใช้ Adam Optimizer อัตราการเรียนรู้ 0.001 ขนาด batch size คือ 32 และ epoch คือ 1,000 และ Softmax Activation Function ในการจำแนกประเภท ชุดข้อมูลถูกแบ่งออกเป็นชุดการฝึกอบรมและการทดสอบโดยใช้ K-fold cross-validation (k-fold = 5) ข้อมูลข้อมูลเข้าที่เหมาะคือ 11 x 40 (#times, #frequencies) MS ซึ่งแสดงโครงสร้างโมเดล CNN_LSTM ดังรูปที่ 51



รูปที่ 51 แสดงสถาปัตยกรรมโมเดล CNN_LSTM

ข้อมูลเข้าที่เป็นคุณสมบัติเสียง MS หรือ MFCC ถูกจัดรูปแบบที่ประกอบด้วยแถว คอลัมน์ และหนึ่งช่องสัญญาณ (#frequencies, #times, 1) สำหรับป้อนเข้าสู่โมเดล CNN เวกเตอร์คุณสมบัติเสียงถูกกำหนดให้กับโหนดข้อมูลเข้าที่แตกต่างกันในเลเยอร์ 2-dimensional (2D) convolutional เลเยอร์ 2D Convolution เป็นเลเยอร์ที่สกัดรูปแบบที่สำคัญออกจากข้อมูลเข้า วัตถุประสงค์คือเพื่อสร้าง feature map ด้วย convolution filters และใช้ฟังก์ชันการกระตุ้นแบบไม่เชิงเส้น (Nonlinear Activation Function) หลังจากเลเยอร์ convolution, ฟังก์ชันการกระตุ้น และ Batch Normalization คุณลักษณะด้านเสียงจะถูกส่งไปยังเลเยอร์ max pooling เป้าหมายของเลเยอร์ max pooling คือการลดความละเอียดของ feature maps การพูลลิ่งเป็นแนวคิดที่สำคัญในสถาปัตยกรรม CNN ที่ลดความแปรปรวนของสเปกตรัมในคุณสมบัติข้อมูลเข้า [51] เอาต์พุตที่เลเยอร์ 2D convolutional สุดท้ายจะถูกป้อนเข้าไปในเลเยอร์ flatten เพื่อส่งต่อไปยังเลเยอร์ LSTM สำหรับเลเยอร์ LSTM มีกลไกพิเศษในการควบคุมการไหลของข้อมูลโดยใช้ส่วนประกอบสี่ส่วน Cells ที่มีการเชื่อมต่อแบบ self-recurrent, input gate, forget gate และ output gate ถูกใช้เป็นส่วนประกอบในเลเยอร์ LSTM สำหรับ LSTM unit ได้รับการอัปเดตทุกช่วงเวลา t อธิบายโดยสมการ (38)–(42) [61]

$$i_t = \sigma(W_i y_t^{l-1} + U_i y_{t-1}^{l} + b_i) \tag{38}$$

$$f_t = \sigma(W_f y_t^{l-1} + U_f y_{t-1}^{l} + b_f) \tag{39}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_c y_t^{l-1} + U_c y_{t-1}^{l} + b_c) \tag{40}$$

$$o_t = \sigma(W_o y_t^{l-1} + U_o y_{t-1}^{l} + b_o) \tag{41}$$

$$y_t^{l} = o_t \tanh c_t \tag{42}$$

โดยที่ y_t^{l-1} หมายถึงข้อมูลเข้าของ LSTM unit และ y_t^{l} หมายถึงเอาต์พุตของ LSTM unit $i_t, f_t,$ และ o_t หมายถึงเวกเตอร์เกท c_t หมายถึงสถานะของ LSTM unit และ *W, U,* และ *b* แสดงถึงเมทริกซ์พารามิเตอร์และเวกเตอร์ และ σ หมายถึงฟังก์ชัน sigmoid และ tanh หมายถึงไฮเปอร์โบลิกแทนเจนต์ ในสมการ (38)–(42) ตัวยก *l–1* และ *l* แสดงถึงดัชนีของคุณสมบัติข้อมูลเข้าและเอาต์พุต ตัวห้อย *i, f, o,* และ *c* ในสมการ (38)–(41) หมายถึง input gate, forget gate, output gate, และ cell ตามลำดับ เลเยอร์ LSTM ได้รับการออกแบบมาเพื่อเรียนรู้การพึ่งพาบริบทในระยะยาวของลำดับ [61] เอาต์พุตของเลเยอร์ 2D Convolution สุดท้ายได้รับการเปลี่ยนรูปแบบให้เป็นลำดับ local feature และส่งต่อไปยังเลเยอร์ LSTM หลังจากนั้นจะเรียนรู้การพึ่งพาตามบริบทจาก local feature ผลลัพธ์จากเลเยอร์ LSTM ประกอบด้วยความสัมพันธ์ local และคุณลักษณะตาม

บริบท global ในเลเยอร์สุดท้ายเอาต์พุต softmax ให้ความน่าจะเป็นสำหรับการทำนาย ผลการจำแนกหมวดหมู่ของโมเดลเป็นตัวแทนของเสียงสระภาษาไทย 18 คลาส

7.1.4. รายละเอียดการทดลอง (Implementation details)

สำหรับการทดลองนี้ใช้ภาษาโปรแกรมไพทอนบนเฟรมเวิร์ก Keras และใช้ TensorFlow โดยโมเดลของงานวิจัยนี้ทดลองบน Google Colaboratory [98] ด้วย Intel (R) Xeon (R) CPU @ 2.20 GHz และ Nvidia Tesla P100 GPU ในการวิจัยนี้เพื่อสร้างคุณสมบัติการป้อนข้อมูล MS ขนาด 11 x 40 x 1 (#times x #frequencies x #channel) กำหนดพารามิเตอร์ของสเปกโตรแกรมดังนี้ ความยาวของหน้าต่างคือ 2,048 ความยาว Hop ระหว่างเฟรมตัวอย่างคือ 512 ช่องสัญญาณเสียงคือ 1 อัตราการสุ่มตัวอย่างเสียง 16,000 และจำนวน Mel bands คือ 40 ความถี่สูงสุดของ MS คือ 8,000 สำหรับการฝึกโมเดล จะมีการกำหนด Hyper-parameter สำหรับโมเดลโดย Optimizer คือ Adam อัตราการเรียนรู้เริ่มต้น คือ 0.001 และ Batch size คือ 32

7.2. ผลการทดลอง

ในงานวิจัยนี้ได้ศึกษาคุณลักษณะข้อมูลเข้าข้อมูลและโครงสร้างของโมเดล CNN ร่วมกับ LSTM สำหรับการรู้จำเสียงสระภาษาไทยของชุดข้อมูลผสม ในส่วนแรกของการทดลอง จะเป็นผลการเปรียบเทียบข้อมูลข้อมูลเข้าคุณสมบัติเสียงที่แตกต่าง ส่วนที่สองแสดงผล confusion matrix, precision, recall, และ F1-score ของโมเดล CNN_LSTM ในส่วนสุดท้ายจะแสดงการทำนายผลของโมเดล CNN_LSTM กับ unseen data

7.2.1. ผลการเปรียบเทียบข้อมูลเข้าคุณสมบัติเสียงที่แตกต่างกับโมเดล CNN_LSTM

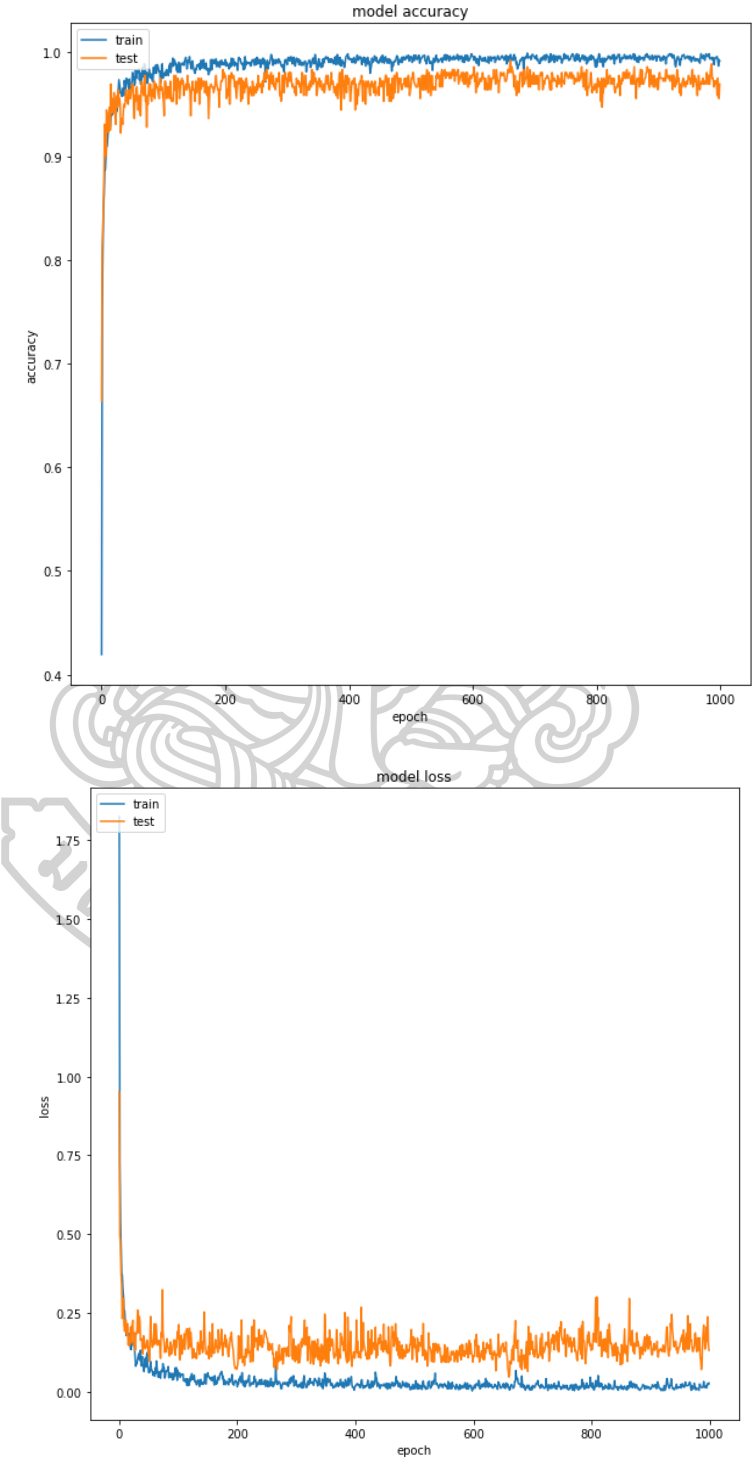
งานวิจัยนี้ใช้ข้อมูลเข้าคุณสมบัติเสียง 2 แบบกับโมเดล CNN_LSTM โดยมีการทดลองดังนี้ 1) คุณลักษณะด้านเสียง MS กับโมเดล CNN_LSTM และ 2) คุณลักษณะด้านเสียง MFCC กับโมเดล CNN_LSTM ซึ่งผลลัพธ์แสดงในตารางที่ 18

ตารางที่ 18 แสดงผลคุณสมบัติเสียงกับโมเดล CNN_LSTM (k-fold = 5)

No.	Experiment Settings	Accuracy (%)	Error of the model (Loss)
1.	MS + CNN_LSTM	**99.17**	**0.05**
2.	MFCC + CNN_LSTM	97.78	0.13

ตารางที่ 18 แสดงผลการทดลองคุณสมบัติด้านเสียงของ MS รวมกับโมเดล CNN_LSTM ในชุดข้อมูลผสม (Mixed data) ทำให้ได้ประสิทธิภาพที่เพิ่มขึ้น โดยมีความถูกต้องแม่นยำถึง 99.17% ซึ่งมากกว่าคุณสมบัติด้านเสียงของ MFCC รวมกับโมเดล CNN_LSTM ที่มีความถูกต้องแม่นยำ

97.78% คุณสมบัติด้านเสียงและการปรับจูนโมเดลที่เหมาะสมเป็นประโยชน์ในการปรับปรุงโมเดลของเสียงสระภาษาไทย ในการทดลองนี้พบว่าผลลัพธ์คุณสมบัติเสียงของ MS ทำงานได้ดีสำหรับการจำแนกประเภทเสียงสระภาษาไทย

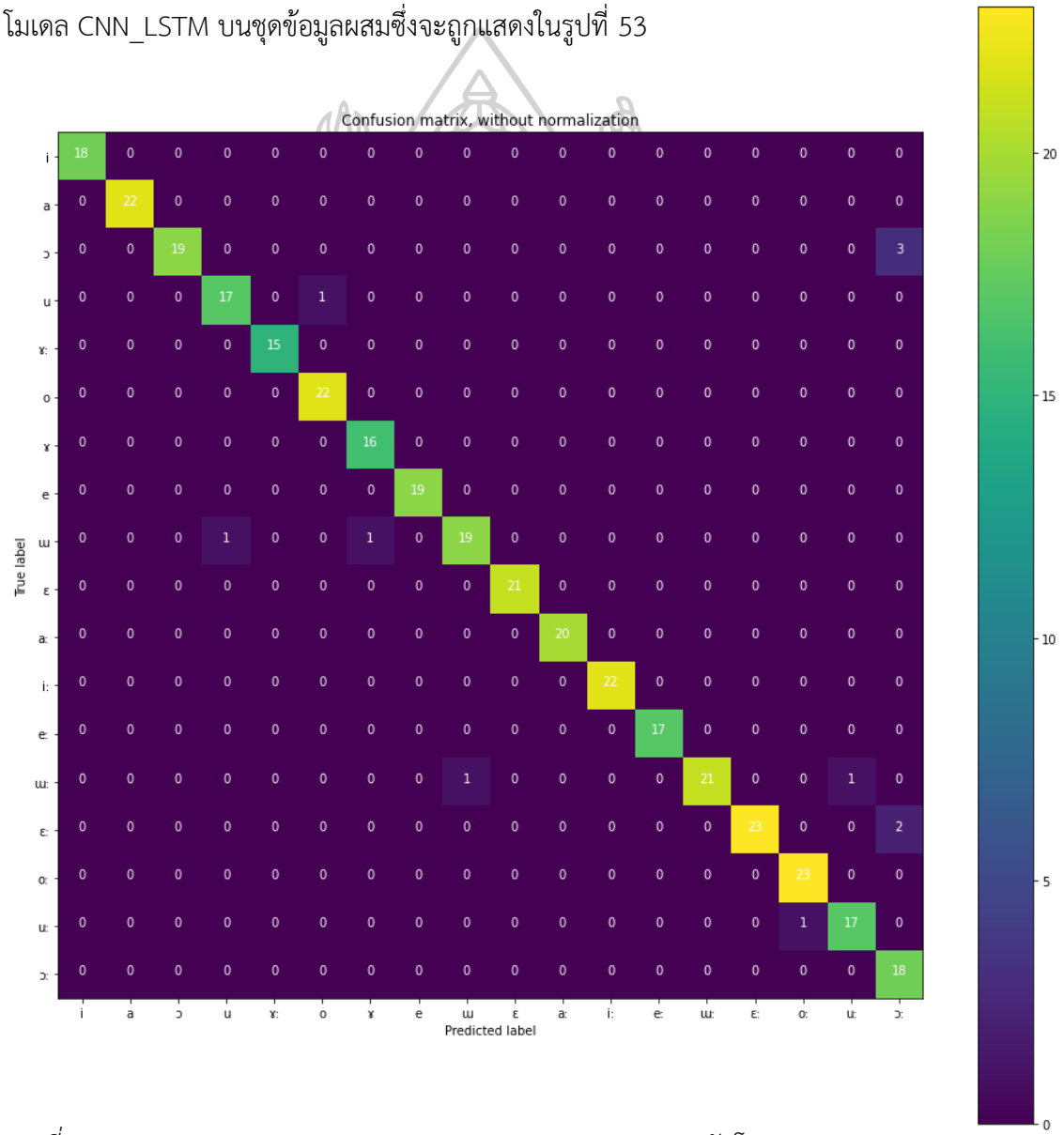


รูปที่ 52 แสดง accuracy และ loss ของโมเดล CNN_LSTM

กราฟเส้นของ accuracy และ loss ของโมเดล CNN_LSTM รวมกับคุณสมบัติทางเสียงของ MS แสดงไว้ในรูปที่ 52 รูปภาพแสดงกราฟเส้นที่เปรียบเทียบ accuracy และ loss ของโมเดลการฝึกอบรมและการทดสอบระหว่าง 0 ถึง 1,000 รอบ คุณสมบัติด้านเสียงของ MS กับโมเดลการทดสอบของโมเดล CNN_LSTM ให้ความถูกต้องแม่นยำสูงสุด 99.17% โดยมีค่าการสูญเสีย 0.05

7.2.2. Confusion matrix, precision, recall, และ F1-score ของโมเดล CNN_LSTM

ส่วนนี้เป็นการนำเสนอการวิเคราะห์ข้อผิดพลาด โดยแสดงเมทริกซ์ความสับสนของโมเดล CNN_LSTM บนชุดข้อมูลผสมซึ่งจะถูกแสดงในรูปที่ 53



รูปที่ 53 แสดง Confusion matrix ของ MS acoustic features กับโมเดล CNN_LSTM ในชุดข้อมูลผสม

สำหรับการวิเคราะห์ข้อผิดพลาด ใน confusion matrix ของชุดข้อมูลผสม ในโมเดล CNN_LSTM สำหรับการรู้จำสระภาษาไทยแสดงไว้ในรูปที่ 52 เสียงสระ ‘เอาะ’ /ɔ/ เป็นคลาสที่มีการทำนายผิดสามครั้งในเมทริกซ์ความสับสน คู่เสียงสระภาษาไทยที่น่าสับสนที่สุดคือ (‘เอาะ’ /ɔ/) และ (‘ออ’ /ɔ:/) เสียงเหล่านี้มีความคล้ายคลึงกันซึ่งสามารถอธิบายได้ทางทฤษฎีภาษาศาสตร์ เนื่องจากมีลักษณะการเคลื่อนที่และระดับของลิ้นคล้ายกันในสระ (‘เอาะ’ /ɔ/) กับ (‘ออ’ /ɔ:/) แต่การใช้ระยะเวลาแตกต่างกัน โดยสระ(‘เอาะ’ /ɔ/) ใช้เวลาในการออกเสียงที่สั้นกว่าสระ (‘ออ’ /ɔ:/) [99] ดังนั้นจึงอาจสร้างความสับสนให้กับโมเดลการรับรู้เสียงสระภาษาไทยได้

ตารางที่ 19 แสดง Precision, Recall, และ F1-score ของโมเดล CNN_LSTM

Thai Vowels	Mixed dataset		
	Precision	Recall	F1-score
i	1.00	1.00	**1.00**
a	1.00	1.00	**1.00**
ɔ	1.00	0.86	0.93
u	0.94	0.94	0.94
ɤ:	1.00	1.00	**1.00**
o	0.96	1.00	0.98
ɤ	0.94	1.00	0.97
e	1.00	1.00	**1.00**
ɯ	0.95	0.90	0.93
ɛ	1.00	1.00	**1.00**
a:	1.00	1.00	**1.00**
i:	1.00	1.00	**1.00**
e:	1.00	1.00	**1.00**
ɯ:	1.00	0.91	0.95
ɛ:	1.00	0.92	0.96
o:	0.96	1.00	0.98
u:	0.94	0.94	0.94
ɔ:	0.78	1.00	0.88

ตารางที่ 19 แสดง precision, recall, และ F1-score ของโมเดล CNN_LSTM สำหรับการจำแนกเสียงสระภาษาไทยแต่ละสระ สำหรับ F1-score ที่ต่ำสุดในชุดข้อมูลผสมคือ 0.88 ในสระ (‘ออ’ /ɔ:/) ผล F1-score สัมพันธ์กันกับ confusion matrix ในทางกลับกัน F1-score ที่สูงที่สุดคือ 1.00 ในชุดข้อมูลแบบผสมคือสระ (‘อิ’ /i/), (‘อะ’ /a/), (‘เออ’ /ɤ:/), (‘เอะ’ /e/), (‘แอะ’ /ɛ/), (‘อา’ /a:/), (‘อี’ /i:/) และ (‘เอ’ /e:/)

7.2.3. การทำนายผลของโมเดล CNN_LSTM กับ unseen data

จากการวิเคราะห์ข้อผิดพลาดและการประเมินของโมเดล CNN_LSTM ที่มีความแม่นยำมากกว่า 95% โมเดล CNN_LSTM ที่ปรับให้เหมาะสมถูกนำมาใช้กับข้อมูลที่เป็น unseen data ผลลัพธ์ของเสียงสระที่ได้รับจากการรู้จำของโมเดลถูกนำมาเปรียบเทียบกับผลการรับรู้ของนักภาษาศาสตร์และเจ้าของภาษา ชุดข้อมูล unseen data มาจากผู้ใช้ 4 คน (ชาย 2 คนและหญิง 2 คน) ทั้งหมดอายุระหว่าง 16 ถึง 30 ปี ผู้ใช้แต่ละคนฝึกออกเสียงสระ 18 เสียงและพูด 3 ครั้ง ดังนั้นชุดข้อมูล unseen data ทั้งหมดมี 216 ไฟล์เสียง (สระ 18 เสียง × 4 คน × 3 ครั้ง) จากข้อมูล unseen data ผลลัพธ์ที่รับรู้โดยระบบ พบว่ามีเสียงสระ 34 เสียง (15.74%) ไม่ตรงกับการรับรู้การฟังเสียงสระจากนักภาษาศาสตร์และเจ้าของภาษา และผลลัพธ์ที่รับรู้โดยระบบมีเสียงสระ 182 เสียง (84.26%) ตรงกันกับการรับรู้ของนักภาษาศาสตร์และเจ้าของภาษา

ตารางที่ 20 แสดงความถี่ของคู่ที่ทำนายผิดสำหรับสระภาษาไทยที่ใช้โมเดล CNN_LSTM

Vowel		Frequency
practiced pronunciation	system recognition	
a:	ɤ:	1
a:	a	2
ɛ:	ɛ	3
i:	e:	**5**
ɯ:	ɤ:	3
ɯ:	i:	2
ɯ:	e:	1
ɤ	a	1
ɯ	u	2
ɯ	a	1
ɯ	ɤ	1

Vowel		Frequency
practiced pronunciation	system recognition	
i	e	1
i	o	1
i	i:	1
o:	ɛ:	1
o:	e	1
o:	o	1
o:	u	1
u:	i:	1
u:	e:	2
ɔ:	ɔ	1
ɔ	i	1

ตารางที่ 20 แสดงคู่สระที่ระบบทำนายผิดมากที่สุด ซึ่งไม่ตรงกับการรับรู้เสียงสระของนักภาษาศาสตร์หรือเจ้าของภาษา ได้แก่ (‘อี’/i:/) และ (‘เอ’/e:/) ซึ่งมีความถี่การออกเสียงที่ไม่ตรงกัน 5 ครั้ง สิ่งเหล่านี้สามารถอธิบายได้ในทฤษฎีภาษาศาสตร์ว่าคู่ที่ออกเสียงผิดนั้นคล้ายคลึงกันที่บริเวณตำแหน่งลิ้นอยู่ด้านหน้าเช่นเดียวกัน แต่ระดับความสูง-ต่ำของลิ้นต่างกัน

7.3. สรุป

งานวิจัยนี้นำเสนอคุณลักษณะด้านเสียงและโมเดล CNN_LSTM สำหรับการรู้จำเสียงสระภาษาไทยที่มีเสียงรบกวน โดยประยุกต์ใช้ประโยชน์ของ CNN ด้านการลดความแปรปรวนของสเปกตรัมในข้อมูลข้อมูลเข้าร่วมกับการใช้ LSTM ที่เหมาะสำหรับการสร้างโมเดลที่เกี่ยวกับสัญญาณเวลาชุดข้อมูลถูกรวบรวมจากเจ้าของภาษาในสถานการณ์จริง โมเดล CNN_LSTM ร่วมกับคุณสมบัติเสียงของ MS ช่วยเพิ่มประสิทธิภาพในการรู้จำเสียงสระภาษาไทย มีความถูกต้องแม่นยำ 99.17% โมเดลนี้ถูกนำไปใช้กับข้อมูล unseen data ผลลัพธ์ของเสียงสระที่ได้รับจากระบบนั้นถูกนำมาเปรียบเทียบกับเสียงสระที่รับรู้โดยนักภาษาศาสตร์และเจ้าของภาษาได้ผลลัพธ์ความถูกต้องแม่นยำ 84.26% การสกัดคุณสมบัติเสียงสระที่ใช้ MS ร่วมกับ CNN_LSTM ให้ประสิทธิภาพที่ดีสำหรับการรู้จำเสียงสระภาษาไทย ผลงานวิจัยนี้เป็นประโยชน์ต่อผู้มีส่วนได้ส่วนเสียที่สนใจในการพัฒนาระบบเสียงสระภาษาไทยหรือระบบการออกเสียงที่คล้ายคลึงกัน ซึ่งจะช่วยให้นักวิจัยสามารถ

ผลิตระบบการเรียนรู้โดยประยุกต์ตามการดำเนินการที่คล้ายคลึงกัน นอกจากนี้ยังสามารถให้คำแนะนำผู้เรียนทำให้ผู้เรียนสามารถฝึกการออกเสียงสระเหมือนมีผู้เชี่ยวชาญ ครูภาษาไทย และนักภาษาศาสตร์ให้คำแนะนำในการออกเสียงที่ถูกต้องตลอดเวลา

บทที่ 8

สรุปผลการทดลอง การอภิปรายผล และข้อเสนอแนะ

8.1. สรุปผลการทดลอง

การออกเสียงให้ถูกต้องเป็นเป้าหมายสำคัญอย่างหนึ่งของการเรียนรู้ภาษา การออกเสียงที่ดีทำให้การสื่อสารมีประสิทธิภาพมากขึ้น การออกเสียงที่ถูกต้องและชัดเจนเป็นสิ่งสำคัญในการสร้างความมั่นใจว่าผู้ฟังสามารถเข้าใจความหมายได้อย่างถูกต้อง การออกเสียงที่ผิดสามารถทำให้ความหมายเปลี่ยนแปลงได้ สระเป็นแกนหลักของพยางค์ (นิวเคลียส) และเป็นส่วนสำคัญของคำพูด สระเกิดขึ้นในช่องปากขึ้นอยู่กับตำแหน่งของลิ้น การฝึกออกเสียงสระจึงเป็นเรื่องยากสำหรับผู้เรียนหรือผู้ที่ไม่ได้เป็นเจ้าของภาษาที่จะเข้าใจได้ง่ายด้วยตนเอง ซึ่งต้องมีผู้เชี่ยวชาญให้คำแนะนำ การนำเทคโนโลยีสำหรับฝึกการออกเสียงมาใช้เพื่อเป็นเครื่องมือที่สามารถช่วยพัฒนาในด้านการเรียนการสอนสำหรับการเรียนรู้ภาษาเพื่อแก้ปัญหาการฝึกออกเสียงสระไทยสำหรับผู้เรียนที่ไม่ใช่เจ้าของภาษาหรือผู้ที่พูดภาษาไทยไม่ได้มาตรฐาน หรือผู้พิการทางการออกเสียง แก้ปัญหาการขาดแคลนผู้เชี่ยวชาญในการสอนออกเสียงสระไทย แก้ปัญหากระบวนการเดิมที่ยุ่งยาก ใช้เวลานาน และคิดค้นเครื่องมือใหม่สำหรับการเรียนภาษาออนไลน์ที่เหมาะสมกับสถานการณ์ปัจจุบันได้

โครงสร้างการเรียนรู้เชิงลึกที่ประยุกต์ใช้ในระบบการฝึกการออกเสียงโดยใช้คอมพิวเตอร์ช่วย (Computer-Assisted Pronunciation Training : CAPT) สำหรับเสียงสระภาษาไทยได้รับการออกแบบมาให้จดจำเสียงสระพื้นฐาน 18 เสียงในภาษาไทย การออกเสียงสระไทย 18 เสียงนั้นซับซ้อนและเนื่องจากหน่วยเสียงบางตัวมีลักษณะคล้ายคลึงกัน ดังนั้นผู้เรียนที่ไม่ใช่เจ้าของภาษาจึงไม่สามารถแยกแยะได้และต้องการความช่วยเหลือจากผู้เชี่ยวชาญ ในงานวิจัยนี้ได้นำเสนอการออกแบบและเก็บตัวอย่างข้อมูลเสียงสระภาษาไทย เนื่องจากคลังข้อมูลเสียงภาษาไทยที่มีอยู่ไม่สามารถนำมาใช้ตามวัตถุประสงค์ของการวิจัยนี้ได้ ดังนั้นเพื่อให้ได้ข้อมูลสระเชิงคุณภาพในทางทฤษฎีตามหลักการทางภาษาศาสตร์ ชุดข้อมูลใหม่จึงได้รับการออกแบบและรวบรวม ชุดข้อมูลที่รวบรวมจากผู้พูดภาษาไทยมาตรฐานในมิติต่างๆ ชุดข้อมูลได้รับการออกแบบ รวบรวม และตรวจสอบคุณภาพข้อมูลโดยนักภาษาศาสตร์ งานวิจัยนี้นำเสนอวิธีการรู้จำเสียงสระภาษาไทยโดยใช้โมเดล Convolutional Neural Network (CNN) ซึ่งเป็นหนึ่งในโครงสร้างการเรียนรู้เชิงลึกที่เป็นที่นิยม โมเดล CNN ถูกนำมาใช้กับชุดข้อมูลการรู้จำเสียงสระภาษาไทย ซึ่งโมเดลนี้สร้างขึ้นเพื่อฝึกคอมพิวเตอร์ให้รู้จักเสียงสระเหมือนผู้เชี่ยวชาญที่สามารถระบุการออกเสียงสระของผู้เรียนได้ โมเดล CNN ถูกนำไปประยุกต์ใช้กับการรู้จำเสียงอัตโนมัติใน CAPT สำหรับสระภาษาไทย โมเดล CNN ที่

เหมาะสมที่สุดถูกนำมาใช้เพื่อเรียนรู้คุณลักษณะสเปกตรัมของสระภาษาไทย 18 เสียง งานวิจัยนี้นำเสนอข้อมูลเข้าคุณสมบัติอะคูสติกเสียงที่แตกต่างกัน 2 แบบสำหรับโมเดล CNN โมเดล Long Short-Term Memory (LSTM) และการรวมกันของโมเดล CNN และ LSTM ซึ่งถูกใช้ร่วมกับคุณสมบัติเสียง Mel Spectrogram (MS) และ Mel Frequency Cepstral Coefficients (MFCC) งานวิจัยนี้นำเสนอระบบ CAPT อัตโนมัติสำหรับเสียงสระภาษาไทยที่สามารถแสดงผลการออกเสียงสระของผู้เรียนได้ หากการออกเสียงไม่ถูกต้อง ระบบจะแนะนำวิธีปฏิบัติที่ถูกต้องด้วยข้อความ วิดีโอที่ออกเสียงด้วยผู้พูด และวิดีโอ 3 มิติ และนำเสนอผลความพึงพอใจการใช้ระบบของผู้ใช้ระบบ และงานวิจัยนี้นำเสนองานการรู้จำเสียงสระภาษาไทยโดยใช้โมเดล CNN ร่วมกับเทคนิค Gradient-weighted class activation mapping (Grad-CAM) เพื่อตรวจสอบความถูกต้องและอธิบายบริเวณของพื้นที่ที่มีความสำคัญในการทำนายของโมเดลสำหรับเสียงสระภาษาไทย

ในการพัฒนาระบบ CAPT สำหรับกิจกรรมการเรียนรู้ในชีวิตประจำวันที่สามารถปฏิบัติตลอดเวลา ดังนั้นชุดข้อมูลเสียงสระไทยจึงรวบรวมจากเจ้าของภาษาในสถานการณ์จริงในมิติต่างๆ เช่น เพศ อายุ สิ่งแวดล้อม เป็นต้น ด้านการออกแบบชุดข้อมูล ชุดคำภาษาไทยที่ใช้บันทึกเสียงออกแบบโดยนักภาษาศาสตร์ รายการคำศัพท์นี้ได้รับการออกแบบตามทฤษฎีภาษาศาสตร์ อันเป็นผลมาจากกฎทางภาษาศาสตร์ ทุกคำมีลักษณะเฉพาะเหมือนกันซึ่งเป็นพยัญชนะเดียวกัน น้ำเสียงเดียวกัน และพยัญชนะท้ายเดียวกัน แต่ต่างกันเฉพาะในเสียงสระเท่านั้น ด้านการรวบรวม งานวิจัยนี้ได้ทำการเก็บรวบรวมชุดข้อมูลเสียงสระจากผู้พูดภาษาไทยที่พูดภาษากลางซึ่งถือเป็นภาษาไทยอย่างเป็นทางการ เป็นชุดข้อมูลภาษาไทยที่มีเสียงรบกวนที่เกิดขึ้นในสถานการณ์จริง โดยทำการรวบรวมในสภาพแวดล้อมต่างๆ ซึ่งประกอบด้วยเสียงที่มีสัญญาณการรบกวนหลายประเภทอยู่ระหว่าง 30 - 50 dB เช่น เสียงรบกวนที่เกิดจาก รถยนต์บนท้องถนน เสียงผู้คนกำลังพูดกันในโรงอาหาร เสียงดนตรีที่วิทยาลัยดนตรี เป็นต้น เสียงที่บันทึกจากชุดข้อมูลรวบรวมจากเจ้าของภาษาไทยมาตรฐานจำนวน 50 คน (ชาย 25 คน หญิง25 คน) มีคุณสมบัติตามกระบวนการคัดเลือกตัวอย่างที่เหมาะสม ตามขั้นตอนการคัดเลือกโดยนักภาษาศาสตร์ ทั้งหมดมีอายุ 20-25 ปี โดยบันทึกจากโทรศัพท์มือถือที่ 44,100 Hz (standard speech data) เสียงสระประกอบด้วยสระเสียงสั้น 9 เสียง และสระเสียงยาว 9 เสียง เสียงที่รวบรวมได้ทั้งหมดมี 1,800 ไฟล์เสียง ประกอบด้วยไฟล์เสียงของเพศชาย 900 ไฟล์ (18 สระ x ชาย 25 คน x พูด 2 ครั้ง) และไฟล์เสียงของเพศหญิง 900 ไฟล์ (18 สระ x หญิง 25 คน x พูด 2 ครั้ง) และเสียงทุกเสียงในชุดข้อมูลถูกตรวจสอบโดยนักภาษาศาสตร์ตามทฤษฎีภาษาศาสตร์

โมเดลการรู้จำเสียงสระภาษาไทย 18 เสียงที่เหมาะสมในงานวิจัยนี้ คือ โมเดล 2-dimensional (2D)-CNN ที่ใช้ร่วมกับการสกัดคุณสมบัติอะคูสติกของเสียงสระที่ใช้ MS ขนาด 11 x 128 (#times x #frequencies) เป็นคุณสมบัติอะคูสติกที่โดดเด่นสำหรับการรู้จำเสียงสระภาษาไทย ช่วยเพิ่มประสิทธิภาพในการรู้จำเสียงสระภาษาไทย โดยให้ค่าความถูกต้องแม่นยำ 98.61% ซึ่งโมเดลนี้ได้ถูกนำมาประยุกต์ใช้ในระบบ CAPT บนสถานการณ์จริง โดยข้อมูลนำเข้าที่ได้รับจากผู้เรียนถือเป็นข้อมูลที่ไม่เคยผ่านการฝึกอบรมหรือทดสอบมาก่อน (unseen data) จะถูกนำมาทดสอบกับระบบ โดยผลลัพธ์ของเสียงสระที่รับรู้จากระบบ CAPT จะถูกเปรียบเทียบกับเสียงสระที่นักภาษาศาสตร์และเจ้าของภาษารับรู้ ซึ่งผลลัพธ์ที่รับรู้โดยระบบตรงกันกับการรับรู้ของนักภาษาศาสตร์และเจ้าของภาษามีความถูกต้องแม่นยำ 89.81% โดยโมเดลนี้สามารถแยกแยะเสียงสระได้แม้ว่าข้อมูลจะมีเสียง อายุ สำเนียง สภาพแวดล้อม และลักษณะทางกายภาพที่แตกต่างกัน (เช่น เสียงผู้หญิงกับผู้ชาย)

ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง ถูกพัฒนาในรูปแบบของเว็บแอปพลิเคชันได้นำมาใช้กับผู้เรียน จากนั้นทำการสำรวจความพึงพอใจของผู้ใช้ระบบจำนวน 29 คน พบว่าผู้ตอบแบบสอบถามส่วนใหญ่เป็นเพศหญิง จำนวน 16 คน คิดเป็น 55.17% ส่วนใหญ่มีอายุ 16-30 ปี จำนวน 19 คน คิดเป็น 65.52% ระดับการศึกษาปัจจุบันส่วนใหญ่สำเร็จการศึกษา จำนวน 19 คน คิดเป็น 65.52% ภาษาพูดที่ใช้สื่อสารส่วนใหญ่ใช้ภาษาไทยสำเนียงมาตรฐาน จำนวน 11 คน คิดเป็น 37.93% และส่วนใหญ่มีสถานะผู้ใช้ระบบเป็นผู้ใช้งานระบบ หรือ ผู้เรียน จำนวน 25 คน คิดเป็น 86.21% ผลการประเมินความพึงพอใจของผู้ใช้ระบบการฝึกออกเสียงอัตโนมัติโดยใช้คอมพิวเตอร์ช่วยสำหรับเสียงสระภาษาไทย 18 เสียง พบว่าผู้ตอบแบบสอบถามส่วนใหญ่มีความพึงพอใจในภาพรวมอยู่ในระดับมากที่สุด (ค่าเฉลี่ย = 4.55 ส่วนเบี่ยงเบนมาตรฐาน = 0.60) โดยคิดเป็น 91.00%

การรู้จำเสียงสระภาษาไทยโดยใช้โมเดล CNN ร่วมกับ Grad-CAM เพื่อตรวจสอบความถูกต้องและอธิบายบริเวณของพื้นที่ที่มีความสำคัญในการทำนายของโมเดล เพื่อช่วยในการปรับปรุงประสิทธิภาพของการรู้จำเสียงสระภาษาไทยสำหรับเสียงสระภาษาไทย 18 เสียง ผลการศึกษาของ Grad-CAM สามารถช่วยในเรื่องความโปร่งใสในการทำนายผลของโมเดล ผลลัพธ์ที่ได้จากการทำนายของ Grad-CAM จะถูกนำไปเปรียบเทียบหาความสัมพันธ์กับหลักการภาษาศาสตร์ พบว่าการรู้จำเสียงสระภาษาไทยในแต่ละสระ Grad-CAM จะพิจารณาทั้งความถี่สูงและความถี่ต่ำ โดยการออก

เสียงสระที่ใช้ลิ้นส่วนหน้าพบว่า Grad-CAM จะพิจารณาให้ความสำคัญบริเวณที่มีความถี่สูงมากกว่าสระที่ใช้ลิ้นส่วนหลัง และการออกเสียงสระมีการเคลื่อนที่ของลิ้นที่สูงกว่า Grad-CAM จะพิจารณาให้ความสำคัญบริเวณที่มีความถี่ต่ำกว่าการออกเสียงสระมีการเคลื่อนที่ของลิ้นที่ต่ำ สามารถนำประโยชน์เรื่องความโปร่งใสในการทำนายผลของโมเดลไปพัฒนาระบบคอมพิวเตอร์ช่วยฝึกการออกเสียงแบบอัตโนมัติสำหรับเสียงสระภาษาไทยได้ ทำให้ระบบมีประสิทธิภาพและความถูกต้องมากยิ่งขึ้น

การพัฒนา CAPT อัตโนมัติสำหรับเสียงสระภาษาไทยโดยใช้โครงสร้างการเรียนรู้เชิงลึก เป็นระบบใหม่ที่พัฒนาโดยการประยุกต์ใช้เทคนิคทางคอมพิวเตอร์ผสมผสานกับทฤษฎีหลักการทางภาษาศาสตร์ ระบบสามารถใช้เพื่อให้คำแนะนำแก่ผู้เรียน สามารถช่วยให้ผู้เรียนได้ฝึกออกเสียงสระแบบเรียลไทม์เสมือนมีผู้เชี่ยวชาญ ครูภาษาไทย และนักภาษาศาสตร์คอยให้คำแนะนำในการออกเสียงสระให้ถูกต้อง ในอนาคตผลลัพธ์ของงานวิจัยนี้ สามารถใช้เป็นประโยชน์ต่อการพัฒนานวัตกรรมของระบบการฝึกการออกเสียงภาษาไทยต่อไป ตัวอย่างเช่น การรู้จำวรรณยุกต์ คำ วลี และประโยคเพื่ออำนวยความสะดวกในการเรียนรู้การออกเสียงภาษาไทยมาตรฐาน อีกทั้งผลงานวิจัยนี้สามารถเป็นประโยชน์ต่อผู้มีส่วนได้ส่วนเสียที่มีความสนใจในการพัฒนาระบบเสียงสระภาษาไทยหรือระบบการออกเสียงที่คล้ายกัน ซึ่งจะช่วยให้นักวิจัยสามารถสร้างแอปพลิเคชันการเรียนรู้โดยทำตามการดำเนินการที่คล้ายคลึงกันได้

8.2. การอภิปรายผล

งานวิจัยนี้เกี่ยวข้องกับการรู้จำเสียงสระเดี่ยวภาษาไทยมาตรฐาน 18 เสียง ซึ่งเป็นเสียงพูดภาษาไทยที่มีลักษณะเป็นเสียงสระเดี่ยว โดยประยุกต์ใช้ AI กับโมเดลการเรียนรู้เชิงลึกในการฝึกการออกเสียงด้วยคอมพิวเตอร์ช่วย (Computer-Assisted Pronunciation Training : CAPT) ของเสียงสระภาษาไทยแบบอัตโนมัติ โมเดลการรู้จำเสียงสระภาษาไทยมาตรฐาน 18 เสียงเป็นสิ่งสำคัญสำหรับการรู้จำเสียงพูดอัตโนมัติ โมเดลนี้ถูกใช้เพื่อจดจำการออกเสียงสระภาษาไทยของผู้เรียน เพื่อให้มีประสิทธิภาพในการรู้จำของโมเดล จึงจำเป็นต้องมีข้อมูลและโครงสร้างที่เหมาะสมสำหรับการฝึกโมเดล

คลังเสียงภาษาไทยที่มีอยู่ไม่สามารถนำไปใช้กับวัตถุประสงค์ของงานวิจัยนี้ได้ การสร้างชุดข้อมูลที่ตรงตามสภาพแวดล้อมการใช้งาน การควบคุมกำกับคุณภาพของข้อมูลอย่างถูกต้องและเหมาะสม เพื่อเป็นจุดการทำงานร่วมกันระหว่างผู้เชี่ยวชาญเฉพาะสาขาอาชีพ (Domain experts) และผู้เชี่ยวชาญทางด้านปัญญาประดิษฐ์ (AI experts) ดังนั้นเพื่อให้ได้ข้อมูลเสียงสระเชิงคุณภาพตาม

ทฤษฎีในหลักการทางภาษาศาสตร์ ชุดข้อมูลใหม่จึงได้รับการออกแบบ รวบรวม และตรวจสอบตามทฤษฎีภาษาศาสตร์โดยนักภาษาศาสตร์ ชุดข้อมูลถูกรวบรวมจากเจ้าของภาษาในสถานการณ์จริง โดยใช้เจ้าของภาษาไทยมาตรฐานจำนวน 50 คน (ชาย 25 คน หญิง 25 คน) ทั้งหมดมีอายุ 20-25 ปี โดยบันทึกจากโทรศัพท์มือถือที่ 44,100 Hz (standard speech data) เสียงสระประกอบด้วยสระเสียงสั้น 9 เสียง และสระเสียงยาว 9 เสียง เสียงที่รวบรวมได้ทั้งหมดมี 1,800 ไฟล์เสียง มีการรวบรวมในสภาพแวดล้อมจากหลายพื้นที่ เสียงรบกวนที่วัดได้ระหว่างการเก็บข้อมูลโดยนักภาษาศาสตร์ ซึ่งใช้โปรแกรมวัดระดับเสียง Sound Meter มีระดับความดังประมาณ 30 - 50 dB จัดประเภทสภาพแวดล้อมของเสียงรบกวนได้ดังนี้

- ใต้อาคารเรียน ประมาณ 30 dB
- เสียงในห้องสมุด, เสียงสัตว์ในสวนบริเวณบ้าน (เสียงสุนัขและนก) ประมาณ 40 dB
- บ้านและเครื่องใช้ภายในบ้าน ประมาณ 45 dB
- เสียงผู้คนกำลังพูดกันในโรงอาหาร, เสียงดนตรีที่วิทยาลัยดนตรี, เสียงรบกวนที่เกิดจากรถยนต์บนท้องถนน ประมาณ 50 dB

ข้อมูลที่มีอยู่จึงถูกจัดประเภทเป็นชุดข้อมูลเสียงสระไทยที่มีเสียงรบกวนนี้มีลักษณะคล้ายกับงาน [66] ที่ศึกษาวิธีการรู้จำเสียงพูดภาษาไทยแบบทนทานต่อเสียงรบกวนภายนอก ที่ชุดข้อมูลมีสภาพแวดล้อมที่มีเสียงรบกวนระดับความดังประมาณ 40-50 dB แต่แตกต่างกันที่ชุดข้อมูลมีสภาพแวดล้อมจริงบริเวณข้างถนนที่มีรถสัญจรไปมาตลอดเท่านั้น และชุดข้อมูลเป็นชุดเสียงพูดตัวเลขจำนวน 10 คำ ได้จากผู้พูดจำนวน 10 คน (เพศหญิง 5 คน เพศชาย 5 คน) อายุระหว่าง 20 – 22 ปี โดยผู้ทดสอบพูดเสียงตัวเลขภาษาไทยตั้งแต่ 0 ถึง 9 คำละ 10 ครั้ง และชุดข้อมูลในงานวิจัยนี้มีความแตกต่างกับงาน [67] ที่ชุดข้อมูลมีสภาพแวดล้อมของห้องมีเสียงรบกวนเป็นเสียงเครื่องปรับอากาศและเสียงโทรทัศน์ ซึ่งวัดค่าความดังได้ 70 dB การบันทึกเสียงใช้เสียงพูดจำนวนคำทั้งหมด 4,069 คำ แบ่งเป็น เสียงผู้ชายจำนวน 231 ประโยค เสียงพูดของผู้หญิงจำนวน 214 ประโยค บันทึกในห้องขนาด 5.5 x 4 เมตร

โมเดล CNN ในงานวิจัยนี้สามารถลดความแปรผันของความถี่ในคุณสมบัติเสียงและแยกคุณลักษณะเสียงที่เหมาะสม รักษาขนาดของแผนที่คุณลักษณะด้วยกลยุทธ์ padding เช่นเดียวกับในการศึกษา [52] โมเดล CNN ลดความแปรปรวนของสเปกตรัมในคุณสมบัติข้อมูลเข้าด้วย max pooling เช่นเดียวกับในงาน [51] ลดปัญหาการ overfitting โดยกลยุทธ์ dropout เช่นเดียวกับที่ทำใน [91] เพิ่มการมาบรรจบกันเร็วขึ้นและเพิ่มประสิทธิภาพด้วย Adam Optimizer เช่นเดียวกับในการศึกษา [95] ELU ที่นำมาใช้ในแบบจำลอง CNN ให้ผลลัพธ์ที่ดี สอดคล้องกับงาน [62, 86] ในการศึกษา [61] ELU ถูกนำไปใช้ในแต่ละ LFLB โมเดล CNN_LSTM ถูกสร้างขึ้นเพื่อเรียนรู้

คุณลักษณะที่เกี่ยวข้องกับอารมณ์จากสเปกโตรแกรมของคำพูดและ log-Mel โมเดล 2D CNN LSTM ที่ใช้ฟังก์ชัน ELU ได้รับความแม่นยำในการจดจำที่ 95.33% และ 95.89% ใน Berlin EmoDB ของการทดลองที่ขึ้นกับผู้พูดและไม่ขึ้นกับผู้พูด ตามลำดับ ELU ช่วยให้การเรียนรู้ CNN ได้รวดเร็วขึ้นและนำไปสู่ความแม่นยำในการจำแนกประเภทที่สูงขึ้น ประโยชน์ของกลยุทธ์เหล่านี้ช่วยปรับปรุงประสิทธิภาพของโมเดล CNN เสียงสระภาษาไทย

ด้วยคุณสมบัติด้านเสียงของโมเดล เทคนิคการขยายเวลาและความถี่เช่นเดียวกับในการศึกษา [52, 74] สามารถใช้กับโมเดล CNN นี้ได้ ผลลัพธ์ระบุว่าคุณสมบัติด้านเสียงของ MS ที่เหมาะสมที่สุดคือ 11 × 128 (เวลา × ความถี่) ผลการศึกษานี้แตกต่างจาก [74] ที่ใช้โมเดล CNN กับคุณสมบัติเสียง MFCC ชุดข้อมูลสองชุด (ชุดข้อมูลเพศหญิงและเพศชาย) ถูกใช้ในงานนั้น ขนาดคุณสมบัติเสียงที่เหมาะสมสำหรับชุดข้อมูลทั้งสองมีความแตกต่างกัน 11 × 40 และ 11 × 64 ตามลำดับ อัตราที่แม่นยำของโมเดล CNN คือ 90.00% และ 88.89% สำหรับเสียงเพศหญิงและเพศชายตามลำดับ ในงานดังกล่าว MFCC ใช้มาตราส่วนความถี่ลอการิทึมและ DCT ในขณะที่งานวิจัยนี้ MS ใช้มาตราส่วนความถี่เชิงเส้น ความแตกต่างอีกประการหนึ่งชุดข้อมูลในงานวิจัยนี้คือการรวมกันของข้อมูลเสียงเพศหญิงและเพศชาย ดังนั้นข้อมูลเสียงที่ต่างกันและคุณสมบัติด้านเสียงที่แตกต่างกันนำไปสู่การกำหนดค่าคุณสมบัติเสียงที่มีขนาดแตกต่างกันทั้งในแง่ของความถี่และโครงสร้างโมเดล

จากผลการวิจัยพบว่า โมเดล CNN ที่ปรับให้เหมาะสมใช้ร่วมกับ MS ให้ประสิทธิภาพที่ดีด้วยความแม่นยำ 98.61% ซึ่งสูงกว่า MFCC เมื่อใช้กับโมเดลพื้นฐาน LSTM และ MS ที่ใช้กับโมเดลพื้นฐาน LSTM ที่มีความแม่นยำ 94.44% และ 90.00% ตามลำดับ เลเยอร์ LSTM มีประโยชน์สำหรับการเรียนรู้การพึ่งพาบริบทระยะยาว แต่ LSTM ที่ใช้กับชุดข้อมูลเสียงสระภาษาไทยที่เป็นคำพยางค์เดียว ไม่ใช่ประโยคยาวๆ การใช้ LSTM ยังไม่โดดเด่นในงานนี้ สำหรับ MS ใช้ร่วมกับ CNN สามารถแยกแยะเสียงสระภาษาได้ดี โดยสามารถสกัดคุณลักษณะที่สำคัญของแต่ละสระได้ดีกว่า แม้ว่าชุดข้อมูลผสมจะมีความซับซ้อนกันในหลายมิติ เช่น ความหลากหลายของเสียงรบกวน อายุ สำเนียง สภาพแวดล้อม และลักษณะทางกายภาพที่แตกต่างกัน (เช่น เสียงผู้หญิงกับผู้ชาย) ในทำนองเดียวกัน [59] MS ถูกนำไปใช้กับงานการรู้จำคำสั่งเสียง (speech command recognition : SCR) และบรรลุผลการปฏิบัติงานที่ดี MS ที่มีขนาดคุณลักษณะ 125 × 80 × 1 ถูกนำมาใช้เป็นคุณลักษณะด้านเสียง โมเดล Light Interior Search Network (LIS-Net) ถูกนำไปใช้กับงาน SCR โดยใช้ชุดข้อมูล Google Speech Command ผลลัพธ์ของโมเดล LIS-Net มีความถูกต้องแม่นยำ 97%

งานวิจัยนี้ใช้ชุดข้อมูลเสียงสระคล้ายกับการศึกษา [42, 43] ที่ใช้ชุดข้อมูลเสียงพูดสระสำหรับการจำแนกประเภทด้วย CNN ผลการจำแนกประเภทมีความถูกต้อง 99.6% และ 94% ตามลำดับ งานวิจัยนี้แตกต่างจาก [42, 43] ตรงที่งานดังกล่าวเป็นเสียงสระภาษาชวา ชุดข้อมูลไม่ได้กล่าวถึง

ทฤษฎีภาษาศาสตร์ในการออกแบบ รวบรวม ตรวจสอบข้อมูล ไม่ได้กล่าวถึงว่าเป็นชุดข้อมูลที่มีเสียงรบกวน ชุดข้อมูลนั้นถูกบันทึกจากผู้พูดเพียงคนเดียว เป็นไฟล์เสียงสระกลางภาษาชวา (middle vowels) 250 ไฟล์เสียง ซึ่งแบ่งออกเป็น 5 คลาส (/e/, /ɛ/, /ə/, /o/, และ /ɔ/) แม้ว่าชุดข้อมูลของงานวิจัยนี้จะมีมิติที่หลากหลายมากกว่างานดังกล่าว ซึ่งก็คือจำนวนผู้พูด ไฟล์เสียงทั้งหมด ความหลากหลายของคลาส และสภาพแวดล้อมที่หลากหลายของเสียง โมเดลการรับรู้ของงานวิจัยยังให้ผลลัพธ์เป็นที่น่าพอใจ การออกเสียงสระถูกนำไปใช้กับหน่วยเสียงภาษาอาหรับคลาสสิก [72] ซึ่งแตกต่างจากงานวิจัยนี้เช่นกัน ในการศึกษานั้นข้อมูลประกอบด้วยพยัญชนะ 28 เสียงกับสระเสียงสั้น 3 เสียง และใช้ CNN เพื่อจัดหมวดหมู่ 84 คลาส ไฟล์เสียงที่บันทึกมีทั้งหมด 6,229 รายการ ในชุดข้อมูลได้รับการบันทึกออนไลน์จากผู้พูด 85 คน ซึ่งเป็นผู้พูดภาษาอาหรับเป็นภาษาแม่ 81 คน และคนที่ไม่ใช่เจ้าของภาษา 4 คน ผลลัพธ์ได้รับความถูกต้องแม่นยำ 95.77% การรู้จำเสียงสระภาษาไทยโดยใช้การเรียนรู้เชิงลึกถูกเปรียบเทียบกับงานการรู้จำเสียงสระโดยใช้การเรียนรู้เชิงลึกในตารางที่ 21

ตารางที่ 21 งานวิจัยการรู้จำสียงสระที่ใช้โครงสร้างการเรียนรู้เชิงลึก

งานวิจัยสระที่ใช้การเรียนรู้เชิงลึก	**วิธีการและผลลัพธ์**
Javanese vowels [44,45]	ใช้โมเดล CNN กับ MFSC ในการรู้จำสระภาษาชวา 5 คลาส ชุดข้อมูลเสียงสระกลางภาษาชวา (middle vowels) 250 ไฟล์เสียง บันทึกโดยผู้พูด 1 คน ผลลัพธ์ได้ค่าความถูกต้อง 99.6% และ 94% ตามลำดับ
Thai vowels [43]	ประยุกต์ใช้โมเดล CNN กับเสียงสระภาษาไทยมาตรฐานร่วมกับ MFCC มีการใช้ชุดข้อมูล 2 ชุด ได้แก่ ชุดข้อมูลเพศหญิงและเพศชาย รวบรวมจากผู้ให้ข้อมูล 50 คน แต่ละชุดข้อมูลประกอบด้วยไฟล์เสียง 900 ไฟล์ ผลลัพธ์ประกอบด้วย 18 คลาส ได้ค่าความถูกต้อง 90.00% และ 88.89% ในชุดข้อมูลเพศหญิงและเพศชายตามลำดับ
Arabic short vowels [52]	ศึกษาการรู้จำสระเสียงสั้นภาษาอาหรับโดยใช้โมเดล CNN กับหน่วยเสียงภาษาอาหรับคลาสสิกในการจำแนก 84 คลาส ที่ได้มาจากเสียงพยัญชนะ 28 เสียง ที่สัมพันธ์กับสระเสียงสั้น 3 เสียง บันทึกชุดข้อมูลในรูปแบบออนไลน์จากผู้พูด 85 คน จำนวน 6,229 รายการ โมเดลมีความถูกต้อง 95.77%
Thai vowels [งานวิจัยที่นำเสนอ]	มีการประยุกต์ใช้โมเดล CNN ร่วมกับ MS ในการรู้จำเสียงสระภาษาไทยมาตรฐาน ชุดข้อมูลผสมระหว่างเพศหญิงและเพศชาย ประกอบด้วยไฟล์เสียงสระ 1,800 ไฟล์ที่บันทึกจากเจ้าของภาษา 50 คน โดยชุดข้อมูลได้รับการออกแบบ การรวบรวม และตรวจสอบโดยนักภาษาศาสตร์และ

งานวิจัยสระที่ใช้การเรียนรู้เชิงลึก	วิธีการและผลลัพธ์
	เจ้าของภาษา ผลลัพธ์ประกอบด้วย 18 คลาส ได้ความถูกต้อง 98.61%

ในหลายงานโมเดลไม่ได้ถูกประยุกต์ใช้ในสถานการณ์จริง ดังนั้นเมื่อนำไปใช้กับระบบ จึงไม่สามารถระบุผลกระทบที่แท้จริงได้ เพื่อตรวจสอบความทนทาน โมเดล CNN ถูกนำไปใช้กับระบบ CAPT ในสถานการณ์จริง ข้อมูลนำเข้าที่ได้รับจากผู้เรียนถือเป็นข้อมูลที่ไม่เคยผ่านการฝึกหรือทดสอบเมื่อเปรียบเทียบกับการรับรู้เสียงสระของนักภาษาศาสตร์และเจ้าของภาษา การตรวจจับเสียงสระของระบบ CAPT มีความถูกต้องแม่นยำ 89.81% สิ่งนี้บ่งชี้ว่าการใช้มิติข้อมูลที่หลากหลายและการออกแบบ รวบรวม และตรวจสอบข้อมูลจะมีประโยชน์อย่างมากสำหรับการสร้างข้อมูลข้อมูลเข้าคุณภาพสำหรับโมเดลการรู้จำเสียง แต่ถึงอย่างไรก็ตาม เนื่องจากการใช้เสียงของผู้พูดกับสภาพแวดล้อมในการทดสอบระบบการรู้จำแตกต่างกับเสียงที่ใช้ในการฝึกฝน การออกเสียงที่ไม่ชัดเจนของผู้พูด การใช้ไมโครโฟนในอุปกรณ์ที่แตกต่างกัน ทำให้ได้คุณภาพเสียงที่แตกต่างกัน สามารถส่งผลต่อประสิทธิภาพของการรู้จำได้ ซึ่งน่าจะเป็นผลมาจากทรัพยากรที่มีอย่างจำกัด ไม่ว่าจะเป็นทางด้านงบประมาณ เวลา จำนวนอาสาสมัคร เป็นต้น ทำให้ข้อมูลที่ใช้ในการฝึกเพื่อรู้จำเสียงมีปริมาณและลักษณะความหลากหลายของเสียงที่ยังไม่เพียงพอ

ระบบอัตโนมัติสำหรับการฝึกออกเสียงที่ใช้โมเดล CNN สำหรับการรู้จำเสียงสระภาษาไทยสามารถแก้ปัญหาต่างๆ เช่น การขาดแคลนผู้เชี่ยวชาญ กระบวนการที่ใช้เวลานาน มีความซับซ้อนและไม่ใช่แบบเรียลไทม์ การเรียนรู้เชิงลึกถูกนำไปใช้กับโมเดลการรู้จำ ASR เพื่อจดจำการออกเสียงของผู้เรียน ระบบใช้ข้อมูลข้อมูลเข้าเสียงพูดดิบและใช้คุณสมบัติเสียงของ MS ร่วมกับ CNN เพื่อสกัดคุณสมบัติ ลักษณะเด่นของเสียงสระและแยกประเภทเสียงสระ หลังจากนั้น สระที่ได้จากการจำแนกประเภทจะถูกเปรียบเทียบกับสระที่ผู้เรียนเลือก หากการเปรียบเทียบตรงกันแสดงว่าผู้เรียนออกเสียงถูกต้อง

เพื่อเพิ่มความถูกต้องของการออกเสียง นักภาษาศาสตร์ใช้วิธีสัทศาสตร์อะคูสติกสำหรับการวิเคราะห์การออกเสียง โดยทั่วไปงานวิจัยเหล่านี้จำเป็นต้องมีการสกัดค่าความถี่ฟอร์เมินท์ที่ 1 และที่ 2 ที่สร้างขึ้นด้วยมือ ซึ่งแสดงถึงความท้าทายที่สำคัญเมื่อมีข้อมูลจำนวนมากและมีความแปรผันของเสียงของผู้พูด งานวิจัยนี้แตกต่างจากการศึกษาก่อนหน้านี้ที่ใช้ Praat กับหลักสัทศาสตร์ในการวิเคราะห์การออกเสียง [16, 27-31] โดยทั่วไปแล้ว นักภาษาศาสตร์มักใช้ค่าความถี่ฟอร์เมินท์ที่ 1 และที่ 2 โดยใช้โปรแกรม Praat จากนั้นค่าความถี่ฟอร์เมินท์ที่ 1 และที่ 2 ของเจ้าของภาษาและที่ไม่ใช่เจ้าของภาษาจะถูกพล็อตโดยใช้ภาษาโปรแกรม Microsoft Excel, ภาษาไพทอน หรือ ภาษาอาร์ จากนั้นนำภาพมาเปรียบเทียบเพื่อหาข้อแตกต่างระหว่างเจ้าของภาษากับผู้เรียน ถ้าต่างกันแสดง

ว่าออกเสียงไม่ถูกต้อง ขั้นตอนปกติเหล่านี้เป็นวิธีการแบบหลายขั้นตอน ไม่ใช่แบบเรียลไทม์ และต้องการผู้เชี่ยวชาญในการใช้โปรแกรม Praat หรือผู้ที่สามารถเขียนโปรแกรมในภาษาอาร์ หรือภาษาไพทอนได้ วิธีการดั้งเดิมของนักภาษาศาสตร์อาจซับซ้อนสำหรับคนอื่น ๆ ในขณะที่งานวิจัยนี้มีวิธีการที่ไม่ซับซ้อน ทำงานแบบเรียลไทม์และสามารถแก้ปัญหาเหล่านั้นได้

การประยุกต์ใช้เทคนิค Gradient-weighted Class Activation Mapping (Grad-CAM) กับโมเดล Convolutional Neural Network (CNN) ในการรู้จำเสียงสระภาษาไทย สามารถช่วยอธิบายบริเวณของพื้นที่ที่มีความสำคัญในการทำนายของโมเดล โดยสามารถแสดงภาพเสียงสระแต่ละสระซึ่งสามารถแยกแยะการทำนายระหว่างคลาส จึงช่วยในเรื่องความน่าเชื่อถือของตัวจำแนกประเภทได้ดีขึ้นเช่นเดียวกับในงาน [75, 76] ที่นำเสนอเทคนิค Grad-CAM สำหรับการตัดสินใจคลาสของโมเดลที่ใช้โมเดล CNN ในงานวิจัยนี้ศึกษาการรู้จำการออกเสียงสระภาษาไทยนี้เป็นงานที่ใช้สัญญาณเสียงที่มีลักษณะไม่นิ่ง มีความแตกต่างกันในแต่ละบุคคล ซึ่งคล้ายกับงาน [77] ที่ศึกษาโมเดล CNN ในการรู้จำอารมณ์ มีลักษณะเป็นสัญญาณ Electroencephalogram (EEG) ซึ่งเป็นการตอบสนองโดยตรงต่อการทำงานของสมองที่สามารถใช้เพื่อตรวจจับสภาพจิตใจและสภาพร่างกาย โดยสัญญาณ EEG เป็นสัญญาณไม่นิ่ง มีลักษณะที่ไม่เป็นเชิงเส้น และมีความแตกต่างกันในแต่ละบุคคลเช่นกัน มีการใช้ Grad-CAM และการสกัดคุณลักษณะอัตโนมัติและการจำแนกอารมณ์กับแผนที่แจกแจงความถี่อิเล็กโทรด (electrode-frequency distribution maps: EFDMs) เพื่อให้ทราบว่า CNN ได้เรียนรู้คุณลักษณะใดบ้างในระหว่างการฝึกอบรม พบว่าคลื่นความถี่สูงเหมาะสำหรับการรู้จำอารมณ์มากกว่า ต่างจากงานวิจัยนี้ที่การรู้จำเสียงสระภาษาไทยในแต่ละสระ Grad-CAM จะพิจารณาทั้งความถี่สูงและความถี่ต่ำ โดยการออกเสียงสระที่ใช้ลิ้นส่วนหน้าพบว่า Grad-CAM จะพิจารณาให้ความสำคัญบริเวณที่มีความถี่สูงมากกว่าสระที่ใช้ลิ้นส่วนหลัง และการออกเสียงสระมีการเคลื่อนที่ของลิ้นที่สูงกว่า Grad-CAM จะพิจารณาให้ความสำคัญบริเวณที่มีความถี่ต่ำกว่าการออกเสียงสระมีการเคลื่อนที่ของลิ้นที่ต่ำ ผลลัพธ์ของ Grad-CAM นี้สอดคล้องกับหลักการทางภาษาศาสตร์ [2] ในงานวิจัยการรู้จำเสียงสระภาษาไทยนี้ใช้ 2-dimensional (2D) convolutional neural network ร่วมกับ Grad-CAM เพื่อแสดงให้เห็นภาพโมเดลที่ได้รับการฝึกอบรม Grad-CAM คำนวณการไล่ระดับของคะแนนที่ทำนายไว้สำหรับคลาสใดคลาสหนึ่งเช่นเดียวกับงาน [78] ที่ผลลัพธ์จะเน้นถึงความสำคัญของแผนที่คุณลักษณะ (feature maps) สำหรับคลาสเป้าหมาย แต่ต่างกันที่งานดังกล่าวใช้ 3-dimensional (3D) conventional convolutional recurrent neural network สำหรับการตรวจจับเสียงนก ผลการแสดงภาพของ Grad-CAM ของ 3D convolution สามารถดึงข้อมูลเวลาระยะยาว long-term time

ในการร้องของนก ในงานวิจัยนี้ใช้โมเดล CNN และ Grad-CAM เพื่ออธิบายปัจจัยหรือบริเวณที่สำคัญ รวมทั้งช่วงเวลาเช่นเดียวกับงาน [79] แต่ในงานดังกล่าวเป็นการตรวจจับการเปลี่ยนแปลงภายในรังผึ้ง สำหรับพัฒนาเป็นระบบเฝ้าติดตามที่สามารถตรวจจับสภาวะผิดปกติในรังผึ้ง ที่ใช้โมเดลทำนายคลาสที่เป็นเสียงของผึ้ง (Bee) และไม่ใช่เสียงของผึ้ง (noBee) ผลลัพธ์แสดงให้เห็นว่า CNN สามารถแยกแยะความแตกต่างในเสียงที่มนุษย์ไม่สามารถแยกแยะได้ และสามารถแยกแยะระหว่างเสียงของผึ้งกับเสียงที่ไม่ใช่ผึ้งได้อย่างมีประสิทธิภาพ จึงสามารถตรวจจับและระบุความผิดปกติในรังได้ และ Grad-CAM ถูกนำมาใช้โดยการใช้น้ำหนักของเลเยอร์ Convolutional สุดท้ายและการไล่ระดับสีที่ใช้ในการทำนายคลาสเป้าหมาย ในการทำนายคลาส noBee สำหรับเสียงที่ไม่ใช่เสียงผึ้ง พบว่าบริเวณที่เฉพาะเจาะจงที่มีเสียงอื่นๆ ถูกระบุว่าเป็นปัจจัยสำคัญในการทำนาย และในกรณีของคลาส bee ที่เป็นเสียงของผึ้ง strong activation เกิดขึ้นตลอดช่วงเวลาของความถี่เฉพาะทั้งหมดในช่วงเวลานั้น ซึ่งคล้ายกับงานวิจัยนี้ที่ Grad-CAM ในแต่ละสระแสดงภาพบริเวณพื้นที่ที่สำคัญของความถี่เฉพาะที่เกิดขึ้นตลอดช่วงเวลานั้น

8.3. ข้อเสนอแนะ

โดเมนชุดข้อมูลเสียงที่แตกต่างกัน ทำให้สถาปัตยกรรมโครงสร้างเชิงลึกหรือโมเดลที่ใช้ในการรู้จำมีความแตกต่างกัน คุณสมบัติทางเสียงที่ต่างกันทำให้ผลลัพธ์ที่ได้มีความต่างกัน การมีโครงสร้างและชุดข้อมูลนำเข้าหรือคุณสมบัติเสียงที่เหมาะสม จะช่วยให้ประสิทธิภาพการรู้จำมีประสิทธิภาพที่เพิ่มขึ้น แม้ว่าชุดข้อมูลที่มีอยู่ในขณะนี้จะสามารถวิเคราะห์และหาข้อสรุปได้ แต่การวิเคราะห์นั้นยังคงต้องอาศัยข้อมูลและปัจจัยอื่นๆที่เกิดขึ้นเมื่อมีการทำนายผลข้อมูลใหม่ที่ไม่เคยถูกฝึกฝนมาก่อน การเพิ่มจำนวนตัวอย่างข้อมูลในการฝึกเรียนรู้ในหลายๆมิติเป็นอีกทางหนึ่งที่จะช่วยให้ระบบมีประสิทธิภาพและมีความทนทานเพิ่มขึ้นได้ ทรัพยากรต่างๆในการทำวิจัย เช่น ความสามารถของเครื่องคอมพิวเตอร์ งบประมาณ ความเพียงพอของข้อมูล เป็นต้น เป็นปัจจัยที่สำคัญที่จะเอื้อต่อการทำวิจัยไม่ว่าจะเป็นด้านการกำหนดโครงสร้างโมเดล ประสิทธิภาพของโมเดลเมื่อนำไปใช้งานจริง สำหรับโมเดลที่ใช้ทรัพยากรมากบางครั้งไม่สามารถนำมาประยุกต์ใช้กับงานวิจัยที่มีลักษณะ Low resource ได้มากนัก เทคนิคและวิธีการใหม่ๆ สามารถช่วยพัฒนาในเรื่องการรู้จำเสียงสระในภาษาไทยหรือการรู้จำคำในภาษาไทยเพื่อเป็นการปรับปรุงการฝึกการออกเสียงโดยใช้คอมพิวเตอร์ช่วย (Computer-Assisted Pronunciation Training : CAPT) ได้ ดังนั้นการพัฒนาประสิทธิภาพของระบบเพื่อต่อยอดประโยชน์ทางการเรียนเป็นสิ่งที่พึงกระทำ หากได้รับวิธีการที่เหมาะสมและข้อมูลที่เพียงพอต่อการศึกษาย่อมสร้างองค์ความรู้ที่มีประโยชน์ในอนาคตได้

ภาคผนวก

ภาคผนวก ก. ผลประเมินความพึงพอใจผู้ใช้ระบบ CAPT

ตารางที่ 22 แสดงข้อมูลทั่วไปผู้ประเมินความพึงพอใจผู้ใช้ระบบ CAPT

ลำดับ	เพศ	อายุ	ระดับการศึกษาปัจจุบัน	ภาษาพูดที่ใช้สื่อสาร	สถานะผู้ใช้
1	หญิง	16-30	สำเร็จการศึกษา	ไทย สำเนียงท้องถิ่น	ผู้ใช้งานระบบ หรือ ผู้เรียน
2	ชาย	16-30	กำลังศึกษาระดับมหาวิทยาลัย	ไทย สำเนียงมาตรฐาน	ผู้ใช้งานระบบ หรือ ผู้เรียน
3	ชาย	16-30	สำเร็จการศึกษา	ไทย สำเนียงมาตรฐาน	ผู้ใช้งานระบบ หรือ ผู้เรียน
4	หญิง	16-30	กำลังศึกษาระดับมหาวิทยาลัย	ภาษาอื่นๆ ที่ไม่ใช่ภาษาไทย	ผู้ใช้งานระบบ หรือ ผู้เรียน
5	หญิง	16-30	สำเร็จการศึกษา	ภาษาอื่นๆ ที่ไม่ใช่ภาษาไทย	ผู้ใช้งานระบบ หรือ ผู้เรียน
6	หญิง	มากกว่า หรือเท่ากับ 31	สำเร็จการศึกษา	ภาษาอื่นๆ ที่ไม่ใช่ภาษาไทย	ผู้ใช้งานระบบ หรือ ผู้เรียน
7	ชาย	16-30	กำลังศึกษาระดับมหาวิทยาลัย	ภาษาอื่นๆ ที่ไม่ใช่ภาษาไทย	ผู้ใช้งานระบบ หรือ ผู้เรียน
8	ชาย	มากกว่า หรือเท่ากับ 31	สำเร็จการศึกษา	ไทย สำเนียงมาตรฐาน	ผู้เชี่ยวชาญ (ครู-อาจารย์ ภาษาไทย, นักภาษาศาสตร์)
9	หญิง	16-30	กำลังศึกษาระดับมหาวิทยาลัย	ไทย สำเนียงมาตรฐาน	ผู้ใช้งานระบบ หรือ ผู้เรียน
10	หญิง	16-30	สำเร็จการศึกษา	ไทย สำเนียงมาตรฐาน	ผู้ใช้งานระบบ หรือ ผู้เรียน

ลำดับ	เพศ	อายุ	ระดับการศึกษาปัจจุบัน	ภาษาพูดที่ใช้สื่อสาร	สถานะผู้ใช้
11	ชาย	16-30	สำเร็จการศึกษา	ไทย สำเนียงท้องถิ่น	ผู้ใช้งานระบบ หรือ ผู้เรียน
12	หญิง	มากกว่า หรือเท่ากับ 31	สำเร็จการศึกษา	ไทย สำเนียงมาตรฐาน	ผู้เชี่ยวชาญ (ครู-อาจารย์ ภาษาไทย, นักภาษาศาสตร์)
13	หญิง	มากกว่า หรือเท่ากับ 31	สำเร็จการศึกษา	ไทย สำเนียงท้องถิ่น	ผู้เชี่ยวชาญ (ครู-อาจารย์ ภาษาไทย, นักภาษาศาสตร์)
14	ชาย	16-30	สำเร็จการศึกษา	ภาษาอื่นๆ ที่ไม่ใช่ภาษาไทย	ผู้ใช้งานระบบ หรือ ผู้เรียน
15	ชาย	16-30	กำลังศึกษาระดับมหาวิทยาลัย	ภาษาอื่นๆ ที่ไม่ใช่ภาษาไทย	ผู้ใช้งานระบบ หรือ ผู้เรียน
16	หญิง	16-30	กำลังศึกษาระดับมหาวิทยาลัย	ภาษาอื่นๆ ที่ไม่ใช่ภาษาไทย	ผู้ใช้งานระบบ หรือ ผู้เรียน
17	หญิง	16-30	กำลังศึกษาระดับมหาวิทยาลัย	ไทย สำเนียงท้องถิ่น	ผู้ใช้งานระบบ หรือ ผู้เรียน
18	หญิง	มากกว่า หรือเท่ากับ 31	สำเร็จการศึกษา	ไทย สำเนียงท้องถิ่น	ผู้ใช้งานระบบ หรือ ผู้เรียน
19	หญิง	16-30	สำเร็จการศึกษา	ไทย สำเนียงท้องถิ่น	ผู้ใช้งานระบบ หรือ ผู้เรียน
20	หญิง	มากกว่า หรือเท่ากับ 31	สำเร็จการศึกษา	ไทย สำเนียงมาตรฐาน	ผู้เชี่ยวชาญ (ครู-อาจารย์ ภาษาไทย, นักภาษาศาสตร์)
21	ชาย	16-30	สำเร็จการศึกษา	ไทย สำเนียงท้องถิ่น	ผู้ใช้งานระบบ หรือ ผู้เรียน

ลำดับ	เพศ	อายุ	ระดับการศึกษาปัจจุบัน	ภาษาพูดที่ใช้สื่อสาร	สถานะผู้ใช้
22	หญิง	มากกว่า หรือเท่ากับ 31	กำลังศึกษาระดับมหาวิทยาลัย	ไทย สำเนียงมาตรฐาน	ผู้ใช้งานระบบ หรือ ผู้เรียน
23	ชาย	มากกว่า หรือเท่ากับ 31	สำเร็จการศึกษา	ไทย สำเนียงมาตรฐาน	ผู้ใช้งานระบบ หรือ ผู้เรียน
24	หญิง	มากกว่า หรือเท่ากับ 31	สำเร็จการศึกษา	ไทย สำเนียงมาตรฐาน	ผู้ใช้งานระบบ หรือ ผู้เรียน
25	ชาย	16-30	กำลังศึกษาระดับมหาวิทยาลัย	ไทย สำเนียงท้องถิ่น	ผู้ใช้งานระบบ หรือ ผู้เรียน
26	ชาย	16-30	สำเร็จการศึกษา	ภาษาอื่นๆ ที่ไม่ใช่ภาษาไทย	ผู้ใช้งานระบบ หรือ ผู้เรียน
27	ชาย	มากกว่า หรือเท่ากับ 31	สำเร็จการศึกษา	ไทย สำเนียงมาตรฐาน	ผู้ใช้งานระบบ หรือ ผู้เรียน
28	หญิง	16-30	กำลังศึกษาระดับมหาวิทยาลัย	ภาษาอื่นๆ ที่ไม่ใช่ภาษาไทย	ผู้ใช้งานระบบ หรือ ผู้เรียน
29	ชาย	16-30	สำเร็จการศึกษา	ไทย สำเนียงท้องถิ่น	ผู้ใช้งานระบบ หรือ ผู้เรียน

ตารางที่ 23 แสดงผลประเมินความพึงพอใจผู้ใช้ระบบ CAPT ด้านการออกแบบส่วนของผู้ใช้งาน

ลำดับ	[ความสวยงาม ความทันสมัย น่าสนใจของหน้าเว็บไซต์]	[การจัดรูปแบบในเว็บไซต์ง่ายต่อการอ่านและการใช้งาน]	[สีสันในการออกแบบเว็บไซต์มีความเหมาะสม]	[เมนูง่ายต่อการใช้งาน]	[ขนาดตัวอักษร และรูปแบบตัวอักษร อ่านได้ง่ายและสวยงาม]
1	4	4	5	4	4
2	5	5	5	5	5
3	4	4	5	5	5
4	5	5	5	5	5
5	5	5	5	5	5
6	5	5	5	5	5
7	3	3	3	4	3
8	5	5	5	5	5
9	4	4	4	4	4
10	5	5	5	5	5
11	5	5	5	5	5
12	4	4	4	5	4
13	4	5	4	5	4
14	4	4	4	5	4

ลำดับ	[ความสวยงาม ความทันสมัย น่าสนใจของหน้าเว็บไซต์]	[การจัดรูปแบบในเว็บไซต์ง่ายต่อการอ่านและการใช้งาน]	[สีสันในการออกแบบเว็บไซต์มีความเหมาะสม]	[เมนูง่ายต่อการใช้งาน]	[ขนาดตัวอักษร และรูปแบบตัวอักษร อ่านได้ง่ายและสวยงาม]
15	4	5	4	4	4
16	4	4	4	4	5
17	4	4	4	4	4
18	5	5	5	5	5
19	5	5	5	5	5
20	4	5	4	5	4
21	5	4	5	5	5
22	4	5	4	5	5
23	5	5	5	5	5
24	4	4	4	4	4
25	4	5	4	4	5
26	4	4	4	4	4
27	4	4	5	5	4
28	3	5	3	3	5
29	3	3	3	3	3

ตารางที่ 24 แสดงผลประเมินความพึงพอใจผู้ใช้ระบบ CAPT ด้านการใช้งานของระบบ

ลำดับ	[ความถูกต้องครบถ้วนของข้อมูล]	[ภาพกับเนื้อหามีความสอดคล้องกันและสามารถสื่อความหมายได้]	[ความเหมาะสมของข้อมูลภายในเว็บไซต์]	[ความสะดวกในการเชื่อมโยงข้อมูลภายในเว็บไซต์]
1	5	5	5	4
2	5	5	5	5
3	5	4	4	4
4	5	5	5	5
5	5	5	5	5
6	5	5	5	5
7	5	5	4	3
8	5	5	4	4
9	5	5	5	5
10	5	5	5	5
11	5	5	5	5
12	3	5	5	4
13	5	5	5	5
14	4	4	4	4
15	5	4	4	4

ลำดับ	[ความถูกต้องครบถ้วนของข้อมูล]	[ภาพกับเนื้อหามีความสอดคล้องกันและสามารถสื่อความหมายได้]	[ความเหมาะสมของข้อมูลภายในเว็บไซต์]	[ความสะดวกในการเชื่อมโยงข้อมูลภายในเว็บไซต์]
16	4	5	4	5
17	4	4	4	3
18	5	5	5	5
19	5	5	5	5
20	4	5	5	5
21	5	5	4	5
22	5	5	4	5
23	4	4	4	5
24	4	4	4	4
25	5	5	4	3
26	4	5	4	4
27	5	5	4	5
28	4	5	4	2
29	3	3	3	3

ตารางที่ 25 แสดงผลประเมินความพึงพอใจผู้ใช้ระบบ CAPT ด้านประโยชน์และการนำไปใช้งาน

ลำดับ	[เพิ่มทักษะออกเสียงสระภาษาไทย]	[ลดปัญหาการออกเสียงสระภาษาไทยที่ไม่ถูกวิธี]	[เข้าใจหลักการออกเสียงสระภาษาไทยที่ถูกต้อง]	[สามารถนำทักษะไปปรับใช้ในการออกเสียงในชีวิตประจำวัน]
1	5	4	5	5
2	5	5	5	5
3	4	4	4	5
4	5	5	5	5
5	5	5	5	5
6	5	5	5	5
7	5	5	5	5
8	5	5	5	4
9	4	4	5	5
10	5	5	5	5
11	5	5	5	5
12	5	5	4	5
13	5	5	5	5
14	5	5	4	5
15	5	5	4	4

ลำดับ	[เพิ่มทักษะออกเสียงสระภาษาไทย]	[ลดปัญหาการออกเสียงสระภาษาไทยที่ไม่ถูกวิธี]	[เข้าใจหลักการออกเสียงสระภาษาไทยที่ถูกต้อง]	[สามารถนำทักษะไปปรับใช้ในการออกเสียงในชีวิตประจำวัน]
16	5	5	5	4
17	4	4	4	4
18	5	5	5	5
19	5	5	5	5
20	5	5	5	5
21	5	5	5	5
22	5	5	5	5
23	5	5	5	5
24	4	4	4	4
25	4	5	5	5
26	5	5	4	5
27	4	4	5	5
28	4	4	5	5
29	3	3	4	3

รายการอ้างอิง

[1] R. B. Noss, *Thai reference grammar*. Foreign Service Institute, Department of State, 1964.

[2] P. Ladefoged and K. Johnson, *A course in phonetics*. Cengage learning, 2005.

[3] R. D. Kent and C. Rountrey, "What acoustic studies tell us about vowels in developing and disordered speech," *American Journal of Speech-Language Pathology,* vol. 29, no. 3, pp. 1749-1778, 2020.

[4] B. G. Evans and W. Alshangiti, "The perception and production of British English vowels and consonants by Arabic learners of English," (in English), *Journal of Phonetics,* vol. 68, pp. 15-31, May 2018, doi: 10.1016/j.wocn.2018.01.002.

[5] L. Rallo Fabra and J. Romero, "Native Catalan learners' perception and production of English vowels," (in English), *Journal of Phonetics,* vol. 40, no. 3, pp. 491-508, May 2012, doi: 10.1016/j.wocn.2012.01.001.

[6] E. Roepke and F. Brosseau-Lapré, "Vowel errors produced by preschool-age children on a single-word test of articulation," *Clinical Linguistics & Phonetics,* pp. 1-23, 2021.

[7] M. Carl, R. D. Kent, E. S. Levy, and D. Whalen, "Vowel acoustics and speech intelligibility in young adults with down syndrome," *Journal of Speech, Language, and Hearing Research,* vol. 63, no. 3, pp. 674-687, 2020.

[8] C. Intajamornrak, "The acoustic characteristics of vowels produced by Thai tracheoesophageal and normal speakers, and the perception of tracheoesophageal vowels," Master. Arts (Linguistics), Chulalongkorn University. Bangkok (Thailand). Graduate School., 2002.

[9] S. Lee and M.-H. Cho, "The impact of L2-learning experience and target dialect on predicting English vowel identification using Korean vowel categories," *Journal of Phonetics,* vol. 82, p. 100983, 2020.

[10] Y.-A. Lu and S.-I. Lee-Kim, "The effect of linguistic experience on perceived vowel duration: evidence from Taiwan Mandarin speakers," *Journal of Phonetics,* vol. 86, p. 101049, 2021.

[11] P. Ghaffarvand Mokari and S. Werner, "Perceptual assimilation predicts acquisition of foreign language sounds: the case of Azerbaijani learners' production and perception of Standard Southern British English vowels," *Lingua,* vol. 185, pp. 81-95, 2017, doi: 10.1016/j.lingua.2016.07.008.

[12] N. Kartushina and C. D. Martin, "Third-language learning affects bilinguals' production in both their native languages: a longitudinal study of dynamic changes in L1, L2 and L3 vowel production," (in English), *Journal of Phonetics,* vol. 77, p. 100920, Nov 2019, doi: ARTN 10092010.1016/j.wocn.2019.100920.

[13] S. Sahatsathatsana, "Pronunciation problems of Thai students learning english phonetics: a case study at Kalasin University," *Journal of Education,* vol. 11, pp. 67-84, 2017.

[14] S. Maspong and P. Pittayaporn, "Length contrast of high vowels in the Thai language of the Sukhothai period: What do the inscriptions say?," *Cahiers de Linguistique Asie Orientale,* vol. 48, no. 1, pp. 30-60, 2019.

[15] P. Teeranon, "Initial Consanant Voicing Perturbation of the Fundamental Frequency of Oral Vowels and Nasal Vowels: A Controversial Case from Ban Doi Pwo Karen," *MANUSYA: Journal of Humanities,* vol. 15, no. 2, pp. 39-59, 2012.

[16] C. Intajamornrak, "Variation and change of the phrae pwo karen vowels and tones induced by language contact with the Tai Languages," *Manusya: Journal of Humanities,* vol. 15, no. 2, pp. 1-20, 2012.

[17] S. Kanokphara, "Syllable structure based phonetic units for context-dependent continuous Thai speech recognition," in *Eighth European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003.

[18] L. Jeerapradit, A. Suchato, and P. Punyabukkana, "HMM-based Thai singing voice synthesis system," in 22*nd International Computer Science and Engineering Conference (ICSEC)*, Chiang Mai, Thailand, 2019: IEEE, pp. 1-4, doi: 10.1109/icsec.2018.8712801.

[19] S. Aunkaew, M. Karnjanadecha, and C. Wutiwiwatchai, "Constructing a phonetic transcribed text corpus for Southern Thai dialect speech recognition," in *12th International Joint Conference on Computer Science and Software Engineering*

(JCSSE), Songkhla, Thailand, 2015: IEEE, pp. 69-73.

[20] A. S. Abramson and N. Reo, "Distinctive vowel length: duration vs. spectrum in Thai," *Journal of Phonetics,* vol. 18, no. 2, pp. 79-92, 1990.

[21] A. Munthuli, C. Tantibundhit, C. Onsuwan, K. Kosawat, and C. Wutiwiwatchai, "Frequency of occurrence of phonemes and syllables in Thai: analysis of spoken and written corpora," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS* 2015), Glasgow, U.K., 2015, pp. 3-7.

[22] ศ. จิตวิริยนนท์, "การปรับค่าความถี่ฟอร์เมินท์เพื่อศึกษาการแปรของสระในภาษาไทย," วารสารภาษาและวัฒนธรรม, 2562.

[23] J. Anderson-Hsieh and K. Koehler, "The effect of foreign accent and speaking rate on native speaker comprehension," *Language learning,* vol. 38, no. 4, pp. 561-613, 1988.

[24] X. L. Peng, H. Chen, L. Wang, and H. A. Wang, "Evaluating a 3-D virtual talking head on pronunciation learning," (in English), *Int J Hum-Comput St,* vol. 109, pp. 26-40, Jan 2018, doi: 10.1016/j.ijhcs.2017.08.001.

[25] M. Tabain and R. Beare, "An ultrasound study of coronal places of articulation in Central Arrernte: apicals, laminals and rhotics," (in English), *Journal of Phonetics,* vol. 66, pp. 63-81, Jan 2018, doi: 10.1016/j.wocn.2017.09.006.

[26] P. Teeranon, "Thai tones in chinese students after using the tone application and their attitudes," *Journal of Language and Linguistic Studies,* vol. 16, no. 4, pp. 1680-1697, 2020.

[27] P. Boersma and V. Van Heuven, "Speak and unspeak with PRAAT," *Glot International,* vol. 5, no. 9/10, pp. 341-347, 2001.

[28] L. Ling and H. Wei, "A research on guangzhou dialect's negative transfer on british english pronunciation by speech analyzer software Praat and ear recognition method," in *2nd International Conference on Computers, Information Processing and Advanced Education*, Ottawa, Canada, 2021, pp. 1123-1132.

[29] G. P. Georgiou, "Discrimination of L2 Greek vowel contrasts: evidence from learners with arabic L1 background," *Speech Communication,* vol. 102, pp. 68-

77, 2018.

[30] H. Liu, J. Liang, V. J. van Heuven, and W. Heeringa, "Vowels and tones as acoustic cues in Chinese subregional dialect identification," *Speech Communication,* vol. 123, pp. 59-69, 2020.

[31] K. Nimz, "Vowel perception and production of late Turkish learners of L2 German," in *ICPhS*, 2011, pp. 1494-1497.

[32] P. Boersma and D. Weenink, "Praat: doing phonetics by computer " *Glot International,* vol. 5, no. 9-10, pp. 341--347, 2001, doi: 10.1126/science.6287572.

[33] F. Q. Lauzon, "An introduction to deep learning," in *11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, Montreal, Quebec, Canada, 2012: IEEE, pp. 1438-1439.

[34] R. Hidayat, A. Bejo, S. Sumaryono, and A. Winursito, "Denoising Speech for MFCC Feature Extraction Using Wavelet Transformation in Speech Recognition System," in *10th International Conference on Information Technology and Electrical Engineering (Icitee),* Bali, Indonesia, 2018, pp. 280-284. [Online]. Available: <Go to ISI>://WOS:000455749500051.

[35] J. Fu, Y. Chiba, T. Nose, and A. Ito, "Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models," *Speech Communication,* vol. 116, pp. 86-97, 2020.

[36] L. Ferrer, H. Bratt, C. Richey, H. Franco, V. Abrash, and K. Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Communication,* vol. 69, pp. 31-45, 2015.

[37] G. Short, K. Hirose, M. Kondo, and N. Minematsu, "Automatic recognition of Japanese vowel length accounting for speaking rate and motivated by perception analysis," *Speech Communication,* vol. 73, pp. 47-63, 2015.

[38] J. Gamper and J. Knapp, "A review of intelligent CALL systems," *Computer Assisted Language Learning,* vol. 15, no. 4, pp. 329-342, 2002.

[39] W. L. Martens and R. Wang, "Applying adaptive recognition of the learner's vowel space to English pronunciation training of native speakers of Japanese," in *SHS Web of Conferences*, Aizuwakamatsu, Fukushima, Japan, 2021, vol. 102: EDP

Sciences, p. 01004.

[40] P. M. Rogerson-Revell, "Computer-Assisted Pronunciation Training (CAPT): current issues and future directions," *RELC Journal,* vol. 52, no. 1, pp. 189-205, 2021.

[41] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *3rd IAPR Asian conference on pattern recognition (ACPR)*, Kuala Lumpur, Malaysia, 2015: IEEE, pp. 730-734.

[42] C. K. Dewa and Afiahayati, "Suitable CNN weight initialization and activation function for Javanese vowels classification," *Procedia Computer Science,* vol. 144, pp. 124-132, 2018, doi: 10.1016/j.procs.2018.10.512.

[43] C. K. Dewa, "Javanese vowels sound classification with Convolutional Neural Network," in *International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Mataram, Indonesia, 2016: IEEE, pp. 123-128, doi: 10.1109/ISITIA.2016.7828645.

[44] Š. Šimáčková and V. J. Podlipský, "Production accuracy of L2 vowels: Phonological parsimony and phonetic flexibility," *Research in Language,* vol. 16, no. 2, pp. 169-191, 2018, doi: 10.2478/rela-2018-0009.

[45] K. Mirzaei, H. Gowhary, A. Azizifar, and Z. Esmaeili, "Comparing the Phonological Performance of Kurdish and Persian EFL Learners in Pronunciation of English Vowels," *Procedia - Social and Behavioral Sciences,* vol. 199, pp. 387-393, 2015, doi: 10.1016/j.sbspro.2015.07.523.

[46] ล. ยี่, "การเปรียบเทียบลักษณะทางกลสัทศาสตร์ของสระภาษาไทยที่ออกเสียงโดยนักศึกษาจีนที่เรียนภาษาไทยกับผู้พูดภาษาไทย," วารสารมนุษยศาสตร์, 2560.

[47] จ. รักษาพล, "การศึกษาการออกเสียงภาษาไทยของอาจารย์ในมหาวิทยาลัยรังสิต," วารสารศิลปศาสตร์, 2564.

[48] C. Intajamornrak, "Pronunciation of Standard Thai Vowels by Non-native Speakers," *Journal of Liberal Arts Prince of Songkla University,* vol. 13, no. 2, pp. 42-65, 2021.

[49] S. Newatia and R. K. Aggarwal, "Convolutional Neural Network for ASR," in *Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2018: IEEE, pp. 638-642, doi:

10.1109/ICECA.2018.8474688.

[50] T. N. Sainath and C. Parada, "Convolutional Neural Networks for small-footprint keyword spotting," in *Interspeech*, 2015, pp. 1478-1482.

[51] T. N. Sainath *et al.*, "Deep Convolutional Neural Networks for large-scale speech tasks," *Neural Networks,* vol. 64, pp. 39-48, 2015, doi: 10.1016/j.neunet.2014.08.005.

[52] Y. Qian and P. C. Woodland, "Very deep Convolutional Neural Networks for robust speech recognition," *Ieee/Acm Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 12, pp. 2263--2276, 2016, doi: 10.1039/C3DT52500G.

[53] G. Kovács, L. Tóth, D. Van Compernolle, and S. Ganapathy, "Increasing the robustness of CNN acoustic models using autoregressive moving average spectrogram features and channel dropout," *Pattern Recognition Letters,* vol. 100, pp. 44-50, 2017, doi: 10.1016/j.patrec.2017.09.023.

[54] T. N. Sainath, O. Vinyals, A. Senior, and N. York, "Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, 2015: IEEE, pp. 4580-4584, doi: 10.1109/ICASSP.2015.7178838.

[55] J. Billa, "Dropout approaches for LSTM based speech recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018: IEEE, pp. 5879-5883.

[56] E. R. Swedia, A. B. Mutiara, and M. Subali, "Deep Learning Long-Short Term Memory (LSTM) for Indonesian Speech Digit Recognition using LPC and MFCC Feature," in *Third International Conference on Informatics and Computing (ICIC)*, Palembang, Indonesia, 2018: IEEE, pp. 1-5.

[57] P. Sukhummek, S. Kasuriya, T. Theeramunkong, C. Wutiwiwatchai, and H. Kunieda, "Feature selection experiments on emotional speech classification," in *12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Rangsit

University, Thailand, 2015: IEEE, pp. 1-4.

[58] N. Kurpukdee, T. Koriyama, T. Kobayashi, S. Kasuriya, C. Wutiwiwatchai, and P. Lamsrichan, "Speech emotion recognition using convolutional long short-term memory neural network and support vector machines," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Kuala Lumpur, Malaysia, 2017: IEEE, pp. 1744-1749.

[59] N. T. Anh, Y. J. Hu, Q. H. He, T. N. L. Tran, T. K. D. Hoang, and C. Guang, "LIS-Net: an end-to-end light interior search network for speech command recognition," (in English), *Computer Speech and Language,* vol. 65, p. 101131, Jan 2021, doi: ARTN 10113110.1016/j.csl.2020.101131.

[60] Z. W. Yao, Z. H. Wang, W. H. Liu, Y. Q. Liu, and J. H. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," (in English), *Speech Communication,* vol. 120, no. February, pp. 11-19, Jun 2020, doi: 10.1016/j.specom.2020.03.005.

[61] J. F. Zhao, X. Mao, and L. J. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," (in English), *Biomedical Signal Processing and Control,* vol. 47, pp. 312-323, Jan 2019, doi: 10.1016/j.bspc.2018.08.035.

[62] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *ICLR* 2016, 2015.

[63] K. Srijiranon and N. Eiamkanitchat, "Thai speech recognition using Neuro-fuzzy system," in *12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Rangsit University, Thailand, 2015: IEEE, pp. 1-6.

[64] P. Phokharatkul, K. Nantanitikorn, and S. Phaiboon, "Thai speech recognition using Double filter banks for basic voice commanding," in *International Conference on Computer, Mechatronics, Control and Electronic Engineering*, Changchun, China, 2010, vol. 6: IEEE, pp. 33-36, doi: 10.1109/CMCE.2010.5609930.

[65] S. Suebvisai, P. Charoenpornsawat, A. Black, M. Woszczyna, and T. Schultz, "Thai automatic speech recognition," in *International Conference on Acoustics,*

Speech, and Signal Processing, Philadelphia, PA, USA, 2005, vol. 1: IEEE, pp. I/857-I/860 Vol. 1.

[66] ม. โพธิโสโนทัย และ ฉ. ฟองสมุทร, "วิธีการรู้จำเสียงพูดภาษาไทยแบบทนทานต่อเสียงรบกวนภายนอก," วารสารเทคโนโลยีสารสนเทศ, 2554.

[67] ร. กอปัญญาพิพัฒน์ และ ร. คงคะจันทร์, "การรู้จำเสียงพูดแบบทนทานต่อเสียงรบกวนสำหรับภาษาไทย โดยใช้อัลกอริทึม N-best LIMABEAM," วารสารวิทยาศาสตร์และเทคโนโลยี, 2559.

[68] K. Sukvichai, C. Utintu, and W. Muknumporn, "Automatic speech recognition for Thai sentence based on MFCC and CNNs," in *second international symposium on instrumentation, control, artificial intelligence, and robotics (ICA-SYMP)*, Bangkok, Thailand, 2021: IEEE, pp. 1-4.

[69] M. Sharma, M. Sarma, and K. K. Sarma, "Recurrent Neural Network based approach to recognize assamese vowels using experimentally derived acoustic-phonetic features," in *1st International Conference on Emerging Trends and Applications in Computer Science*, Shillong, India, 2013: IEEE, pp. 140-143.

[70] M. Sharma and K. K. Sarma, "Dialectal Assamese vowel speech detection using acoustic phonetic features, KNN and RNN," in *2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, New Delhi NCR, India, 2015: IEEE, pp. 674-678.

[71] S. Sharma and P. K. Das, "Reduced feature sets for vowel recognition," in *8th International Conference on Electrical and Computer Engineering*, Pan Pacific Sonargaon Dhaka, Dhaka, Bangladesh, 2014: IEEE, pp. 116-119.

[72] A. Asif, H. Mukhtar, F. Alqadheeb, H. F. Ahmad, and A. Alhumam, "An approach for pronunciation classification of classical Arabic phonemes using deep learning," *Applied Sciences,* vol. 12, no. 1, p. 238, 2022.

[73] N. Suktangman, K. Khanthavivone, and K. Songwatana, "Optimizing vowel recognition in Thai spoken language using reduced LPC spectrum and reduced feature set of critical band intensities," in *International Symposium on Communications and Information Technologies, ISCIT*, Bangkok, Thailand, 2006, pp. 128-132, doi: 10.1109/ISCIT.2006.339901.

[74] N. Rukwong and S. Pongpinigpinyo, "Thai vowels speech recognition using

Convolutional Neural Networks," in *14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, Chiang Mai, Thailand, 2019: IEEE, doi: 10.1109/iSAI-NLP48611.2019.9045520. [Online]. Available: https://ieeexplore.ieee.org/document/9045520/

[75] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, Venice, Italy, 2017, pp. 618-626.

[76] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision,* vol. 128, no. 2, pp. 336-359, 2020/02/01 2020, doi: 10.1007/s11263-019-01228-7.

[77] F. Wang *et al.*, "Emotion recognition with convolutional neural network and EEG-based EFDMs," *Neuropsychologia,* vol. 146, p. 107506, 2020.

[78] I. Himawan, M. Towsey, and P. Roe, "3D convolution recurrent neural networks for bird sound detection," in *Proceedings of the 3rd Workshop on Detection and Classification of Acoustic Scenes and Events*, Surrey, UK, 2018: Detection and Classification of Acoustic Scenes and Events, pp. 1-4.

[79] J. Kim, J. Oh, and T.-Y. Heo, "Acoustic Scene Classification and Visualization of Beehive Sounds Using Machine Learning Algorithms and Grad-CAM," *Mathematical Problems in Engineering,* vol. 2021, 2021.

[80] น. รณเกียรติ, "สัทศาสตร์ภาคทฤษฎีและภาคปฏิบัติ," (เสียงสระ: มหาวิทยาลัยธรรมศาสตร์, 2548.

[81] wikipedia. "Human mouth." https://en.wikipedia.org/wiki/Human_mouth (accessed 3 June, 2023).

[82] G. Marcus, "Deep learning: A critical appraisal," 2018.

[83] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature,* vol. 323, no. 6088, pp. 533-536, 1986.

[84] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE international conference on acoustics,*

speech and signal processing, Vancouver, BC, Canada, 2013: IEEE, pp. 6645-6649.

[85] S. SHARMA. "Activation Functions in Neural Networks." https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6 (accessed 2021).

[86] J. X. Gu *et al.*, "Recent advances in Convolutional Neural Networks," (in English), *Pattern Recognition,* vol. 77, pp. 354-377, May 2018, doi: 10.1016/j.patcog.2017.10.013.

[87] S. Sheng, P. Chen, Y. Yao, L. Wu, and Z. Chen, "Atomic network-based DOA estimation using low-bit ADC," *Electronics,* vol. 10, no. 6, p. 738, 2021.

[88] H. Sharma. "Activation Functions : Sigmoid, ReLU, Leaky ReLU and Softmax basics for Neural Networks and Deep Learning." https://medium.com/@himanshuxd/activation-functions-sigmoid-relu-leaky-relu-and-softmax-basics-for-neural-networks-and-deep-8d9c70eed91e (accessed 2021).

[89] A. Dertat. "Applied Deep Learning - Part 4: Convolutional Neural Networks." https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2 (accessed 4 June, 2023).

[90] A. S. a. M. Li. "Convolutional Neural Networks." https://d2l.ai/chapter_convolutional-neural-networks/padding-and-strides.html (accessed 2021).

[91] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings*, Vancouver, BC, Canada, 2013: IEEE, pp. 8609-8613, doi: 10.1109/ICASSP.2013.6639346.

[92] B. McFee *et al.*, "librosa: audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, Austin, Texas, U.S., 2015, pp. 18-25.

[93] B. Thornton, "Audio recognition using mel spectrograms and Convolution Neural Networks," 2019.

[94] Y. Han, J. Kim, and K. Lee, "Deep Convolutional Neural Networks for predominant instrument recognition in polyphonic music," (in English), *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 1, pp. 208-221, Jan 2017, doi: 10.1109/taslp.2016.2632307.

[95] S. K. Gouda, S. Kanetkar, D. Harrison, and M. K. Warmuth, "Speech recognition: keyword spotting through image recognition," *arXiv preprint arXiv:*1803.03759*,* 2018.

[96] I. Papadimitriou, A. Vafeiadis, A. Lalas, K. Votis, and D. Tzovaras, "Audio-based event detection at different SNR settings using two-dimensional spectrogram magnitude representations," *Electronics,* vol. 9, no. 10, p. 1593, 2020.

[97] F. Demir, M. Turkoglu, M. Aslan, and A. Sengur, "A new pyramidal concatenated CNN approach for environmental sound classification," (in English), *Applied Acoustics,* vol. 170, p. 107520, Dec 15 2020, doi: ARTN 10752010.1016/j.apacoust.2020.107520.

[98] T. Carneiro, R. V. M. Da Nobrega, T. Nepomuceno, G. B. Bian, V. H. C. De Albuquerque, and P. P. Reboucas, "Performance analysis of Google Colaboratory as a tool for accelerating deep learning applications," (in English), *Ieee Access,* vol. 6, pp. 61677-61685, 2018, doi: 10.1109/Access.2018.2874767.

[99] G. Slayden, "Central Thai phonology," ed, 2009.

[100] L. Wilkinson and M. Friendly, "The history of the cluster heat map," *The American Statistician,* vol. 63, no. 2, pp. 179-184, 2009.

[101] I. E. Allen and C. A. Seaman, "Likert scales and data analyses," *Quality progress,* vol. 40, no. 7, pp. 64-65, 2007.

มหาวิทยาลัยศิลปากร

ประวัติผู้เขียน

ชื่อ-สกุล	นางสาวนิยดา รักวงษ์
วัน เดือน ปี เกิด	28 พฤศจิกายน 2527
สถานที่เกิด	จังหวัดลพบุรี
วุฒิการศึกษา	วิทยาศาสตร์มหาบัณฑิต (วท.ม.)
ที่อยู่ปัจจุบัน	286/36 หมู่บ้านมลชญา 4 ต.ท่าเสา อ.เมือง จ.อุตรดิตถ์ 53000