การเปรียบเทียบโครงสร้าง CNN สำหรับการจำแนกสำเนียงไทย: VFNet กับโครงสร้าง CNN เชิงอนุกรม

Comparative Analysis of CNN Architectures for Thai Accent Classification: VFNet vs. Sequential CNN

ธีรภัทร เสิบกลิ่น * และ สุรเดช จิตประไพกุลศาล

Thiraphat Soebklin * and Suradet Jitprapaikulsarn

ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยนเรศวร

[1] Department of Electrical & Computer Engineering, Faculty of Engineering, Naresuan University, Thailand,

***Corresponding Author E-mail**: theerapatserb@gmail.com

Received: 5/09/25, **Revised**: 4/10/25, **Accepted**: 6/10/25

บทคัดย่อ

งานวิจัยนี้นำเสนอการเปรียบเทียบโครงสร้างของโมเดลโครงข่ายประสาทเทียมแบบคอนโวลูชัน (CNN) สำหรับการจำแนกสำเนียงภาษาไทย โดยเปรียบเทียบระหว่างสถาปัตยกรรมแบบขนานที่ใช้ตัวกรองหลายขนาดพร้อมกันตามแนวทางของ VFNet: A Convolutional Architecture for Accent Classification [1] กับสถาปัตยกรรมแบบอนุกรมที่เรียงตัวกรองขนาดต่างกันในแต่ละเลเยอร์ (เช่น 3⟶5⟶7) อินพุตที่ใช้คือคุณลักษณะ MFCC ที่สกัดจากชุดข้อมูลเสียง Thai Dialect Corpus [2] ผลการทดลองแสดงให้เห็นว่า โมเดลทั้งสองแบบให้ค่า Accuracy และ F1-score ใกล้เคียงกัน อย่างไรก็ตาม เมื่อพิจารณาจำนวนพารามิเตอร์และค่า Cross Entropy Loss พบว่าโครงสร้างแบบอนุกรมบางรูปแบบ เช่น 5⟶5⟶5 และ 7⟶5⟶3 ให้ผลลัพธ์ที่ดีกว่าทั้งในด้านประสิทธิภาพและความประหยัดทรัพยากร ทั้งนี้ การวิเคราะห์ขนาด Receptive Field (RF) แบบ 2 มิติ ยังช่วยอธิบายได้ว่า โครงสร้างที่มี RF ขนาดกลางจะให้ผลการจำแนกที่ดีกว่าโครงสร้างที่มี RF เล็กหรือใหญ่เกินไป สะท้อนให้เห็นถึงข้อได้เปรียบของการออกแบบสถาปัตยกรรมโมเดลอย่างเหมาะสม สำหรับการใช้งานจริงในสภาพแวดล้อมที่มีข้อจำกัดด้านหน่วยประมวลผลและหน่วยความจำ

คำสำคัญ: สำเนียงภาษาไทย, การจำแนกสำเนียง, CNN, VFNet, MFCC

Abstract

This research presents a comparative study of Convolutional Neural Network (CNN) architectures for Thai accent classification. It contrasts a parallel architecture based on [1] VFNet: A Convolutional Architecture for Accent Classification, which uses multi-size filters simultaneously, with a sequential architecture that stacks different kernel sizes across layers (e.g., 3⟶5⟶7). The input features are Mel-Frequency Cepstral Coefficients (MFCCs) extracted from the Thai Dialect Corpus [2]. Experimental results show that both models achieve comparable accuracy and F1-scores. However, further analysis reveals that sequential models such as 5⟶5⟶5 and 7⟶5⟶3 outperform the VFNet-based parallel architecture in terms of lower parameter count and cross-entropy loss. A detailed 2D receptive field (RF) analysis also indicates that architectures with moderate RF sizes tend to deliver better classification performance compared to those with very small or excessively large RFs. These findings emphasize the practical advantages of well-structured sequential CNNs for real-world deployment under computational and memory constraints.

Keywords: Thai accent, accent classification, CNN, VFNet, MFCC

1. บทนำ

ในยุคที่เทคโนโลยีรู้จำเสียงพูด (Automatic Speech Recognition: ASR) มีบทบาทสำคัญต่อการสื่อสารระหว่างมนุษย์กับเครื่องจักร เช่น ระบบผู้ช่วยอัจฉริยะ (Voice Assistant), ระบบบริการลูกค้าอัตโนมัติ (Chatbot) และระบบสั่งงานด้วยเสียง เทคโนโลยีเหล่านี้ต้องอาศัยความแม่นยำในการเข้าใจเสียงพูดของผู้ใช้จากหลากหลายภูมิภาค ความท้าทายที่สำคัญในการพัฒนาระบบ ASR ภาษาไทยคือความหลากหลายของสำเนียง ซึ่งกระตุ้นให้เกิดงานวิจัยที่มุ่งเน้นการสร้างชุดข้อมูลเสียงที่ครอบคลุมความแตกต่างของสำเนียง [2] และการพัฒนาเทคนิคเพื่อรองรับความแปรปรวนของสำเนียง

แม้ว่าสถาปัตยกรรม CNN จะถูกใช้อย่างแพร่หลายและประสบความสำเร็จอย่างสูงในงานประมวลผลภาพ แต่ก็มีงานวิจัยจำนวนมากที่แสดงให้เห็นถึงศักยภาพในการนำมาประยุกต์ใช้กับข้อมูลเสียงในรูปแบบ 2 มิติ เช่น สเปกโตรแกรม [3] ซึ่งให้ผลลัพธ์ที่ดีกว่าเทคนิคดั้งเดิม ต่อมาได้มีการพัฒนาสถาปัตยกรรมให้ซับซ้อนและมีประสิทธิภาพสูงขึ้น เช่น การนำกลไก Attention มาใช้ในชั้น Pooling เพื่อให้น้ำหนักกับข้อมูลเสียงในส่วนที่สำคัญเป็นพิเศษ [4]

ดังนั้น งานวิจัยนี้จึงนำเสนอการเปรียบเทียบประสิทธิภาพของโมเดล CNN สองแบบหลัก ได้แก่ สถาปัตยกรรมแบบ VFNet [1] ที่ใช้ฟิลเตอร์หลายขนาดพร้อมกัน และสถาปัตยกรรมแบบอนุกรมที่เรียงลำดับฟิลเตอร์ขนาดต่าง ๆ เพื่อประเมินว่าโครงสร้างแบบใด

เหมาะสมกับการใช้งานจริงมากที่สุด ทั้งในแง่ประสิทธิภาพและความประหยัดทรัพยากร

2. การแก้ปัญหา

การแก้ปัญหานี้สามารถทำได้โดยการเพิ่มความสามารถของระบบในการจำแนกสำเนียงก่อนการรู้จำคำพูด ซึ่งจะช่วยให้ระบบ ASR ปรับตัวเลือกโมเดลให้เหมาะสมกับสำเนียงผู้พูดแต่ละคนได้อย่างชาญฉลาด หรือนำไปใช้ในการทำ label อัตโนมัติ สำหรับงาน ASR งานวิจัยนี้จึงมีเป้าหมายเพื่อพัฒนาโมเดลการจำแนกสำเนียงภาษาไทยที่มีประสิทธิภาพสูง พร้อมกับเปรียบเทียบโครงสร้างโมเดลโครงข่ายประสาทแบบคอนโวลูชัน (CNN) ที่มีแนวทางการออกแบบต่างกัน ได้แก่ โครงสร้างที่อิงจากงานวิจัย VFNet [1] ซึ่งออกแบบให้ใช้ตัวกรองความถี่หลายขนาดพร้อมกัน เพื่อดึงคุณลักษณะเสียงในหลายช่วงสเกลอย่างมีประสิทธิภาพ

โครงสร้างแบบอนุกรม ซึ่งเรียงลำดับการใช้ kernel ขนาดต่าง ๆ คล้ายกับ VFNet [1] แต่นำมาจัดเรียงแบบลำดับ (เช่น 3×3 ⟶ 5×3 ⟶ 7×3) เพื่อสกัดข้อมูลจากรายละเอียดระดับล่างไปสู่บริบทกว้างอย่างเป็นลำดับชั้น รวมถึงโครงสร้างแบบอนุกรมที่ใช้ตัวกรองขนาดเท่ากันในทุก layer ลำดับ (เช่น 3×3 ⟶ 3×3 ⟶ 3×3) ทั้งสองแนวทางได้รับการประเมินด้วยข้อมูลจากชุด Thai Dialect Corpus [2] ซึ่งรวบรวมเสียงพูดจากผู้ใช้สามกลุ่มสำเนียงหลัก ได้แก่ สำเนียงไทยกลาง สำเนียงไทยคำเมือง และสำเนียงไทยโคราช โดยใช้อินพุตเป็นคุณลักษณะ MFCC ซึ่งเป็นหนึ่งในฟีเจอร์ที่นิยมในงานรู้จำเสียงแม้สถาปัตยกรรม VFNet จะมีจุดเด่นด้านการรวบรวมฟีเจอร์จากหลายขนาด แต่ก็ต้องแลกกับจำนวนพารามิเตอร์ที่มากขึ้น ในขณะที่สถาปัตยกรรมแบบอนุกรมอาจมีศักยภาพในการลดจำนวนพารามิเตอร์ได้มากโดยไม่ลดทอนประสิทธิภาพ ซึ่งจะเป็นข้อได้เปรียบในบริบทของการนำโมเดลไปใช้งานจริงที่มีข้อจำกัดด้านทรัพยากรและถ้าโมเดลมีการนำไปใช้ร่วมกับงาน ASR ซึ่งต้องใช้ทรัพยากรมากในการประมวลผล ยิ่งต้องคำนึงถึงจำนวนพารามิเตอร์ที่ใช้

นอกจากการเปรียบเทียบประสิทธิภาพของโมเดลในแง่ของ Accuracy และ F1-score แล้ว งานวิจัยนี้ยังได้วิเคราะห์ความสัมพันธ์กับขนาดของ Receptive Field (RF) และจำนวนพารามิเตอร์ เพื่อค้นหาความสมดุลระหว่างคุณภาพของโมเดลและต้นทุนการประมวลผล งานวิจัยนี้จึงมีคุณค่าทั้งในเชิงวิชาการและการประยุกต์ โดยสามารถนำผลการศึกษาไปต่อยอดในการสร้างระบบ ASR ภาษาไทยที่สามารถเข้าใจเสียงผู้ใช้จากหลากหลายสำเนียงได้อย่างมีประสิทธิภาพและคุ้มค่าต่อการใช้งานจริง

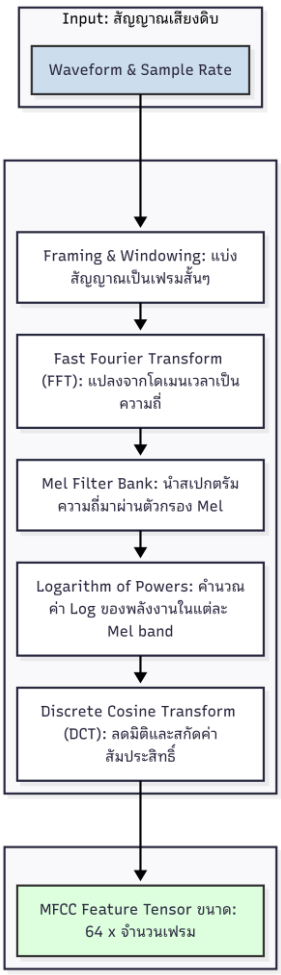
3. วิธีการทดลอง

3.1 ชุดข้อมูล

งานวิจัยนี้ใช้ชุดข้อมูลเสียงพูดจากโครงการ Thai Dialect Corpus ซึ่งเผยแพร่โดย SLSCU (Speech and Language Processing Laboratory, Chulalongkorn University) โดยเลือกมาเฉพาะ 3 กลุ่มสำเนียง ได้แก่ สำเนียงไทยกลาง คัดเลือกมา แบ่งเป็น train 50 ชั่วโมง validation 1.71 ชั่วโมง สำเนียงไทยโคราช (ตะวันออกเฉียงเหนือ) train 46.63 ชั่วโมง validation 1.50 ชั่วโมง สำเนียงไทยคำเมือง (ภาคเหนือ) train 32.43 ชั่วโมง validation 1.50 ชั่วโมง

3.2 การเตรียมข้อมูลและการสกัดคุณลักษณะ

การสกัดคุณลักษณะเสียงในงานวิจัยนี้ใช้เทคนิค Mel Frequency Cepstral Coefficients (MFCC) โดยกำหนดค่า Sampling Rate ที่ 22,050 เฮิรตซ์ และสกัดคุณลักษณะ MFCC ออกมาทั้งหมดจำนวน 64 ค่า จากสเปกตรัมที่ได้จากการประมวลผลผ่านกรองเสียง Mel จำนวน 64 แถบความถี่ (Mel filter banks) โดยจำนวน 64 ค่านี้ ถูกกำหนดให้เป็นเกณฑ์มาตรฐานที่ควบคุมให้เหมือนกันในทุกการทดลอง เพื่อให้การเปรียบเทียบประสิทธิภาพของสถาปัตยกรรม CNN แต่ละรูปแบบ ซึ่งเป็นเป้าหมายหลักของงานวิจัยนี้ เป็นไปอย่างเที่ยงตรง กระบวนการประมวลผลนี้อาศัยการใช้หน้าต่างฟูเรียร์แบบ FFT (Fast Fourier Transform) ที่มีขนาด 512 จุด และกำหนดค่าการเลื่อนหน้าต่าง (hop length) ที่ 256 จุด เพื่อควบคุมความละเอียดเชิงเวลา โดยขั้นตอนการสกัดทั้งหมดดำเนินการผ่านฟังก์ชัน torchaudio.transforms.MFCC จากไลบรารี Torchaudio แสดงดังรูปที่ 1

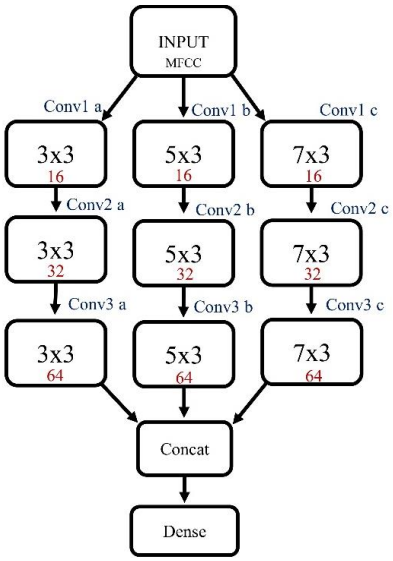


รูปที่ 1 ขั้นตอนการทำ Mel Frequency Cepstral Coefficients (MFCC) กับข้อมูลจริง

ทั้งนี้ ผู้วิจัยได้พิจารณาคุณลักษณะเสียงรูปแบบอื่น เช่น Log-mel Spectrogram ซึ่งเคยถูกนำไปใช้อย่างประสบความสำเร็จในงานจำแนกเสียงประเภทอื่น [3] อย่างไรก็ตาม เนื่องจาก MFCC เป็นคุณลักษณะที่ถูกสกัดมาเพื่อเน้นข้อมูลที่สำคัญต่อการรับรู้ของมนุษย์ ทำให้ได้ข้อมูลที่มีมิติกระชับกว่า (more compact representation) และเป็นที่ยอมรับอย่างแพร่หลายในงานวิจัยด้านเสียงพูด ผู้วิจัยจึงเลือกใช้ MFCC เป็นคุณลักษณะอินพุต เพื่อให้การเปรียบเทียบประสิทธิภาพของสถาปัตยกรรม CNN ซึ่งเป็นเป้าหมายหลักของงานวิจัยนี้ เป็นไปอย่างเที่ยงตรงและชัดเจนที่สุด

3.3 สถาปัตยกรรมของโมเดล

ในบทความในการทดลอง ได้เปรียบเทียบโมเดลหลายรูปแบบซึ่งแบ่งออกเป็น 3 กลุ่มหลัก ได้แก่ โมเดลแบบอนุกรม (Sequential CNN) ที่ใช้ kernel สามขนาดที่เรียงลำดับกันในสามเลเยอร์ เช่น 3×3 ⟶ 5×3 ⟶ 7×3 โดยใช้โดเมนเวลาเท่ากันคือ 3, โมเดลแบบขนาน VFNet [1] ที่ใช้ kernel หลายขนาดพร้อมกันโดยเลือกมา 3 ขนาดได้แก่ 3×3, 5×3 และ 7×3 (ดังแสดงโครงสร้างในรูปที่ 2) และโมเดลแบบหนึ่งเลเยอร์ (1-layer CNN) ที่ใช้ kernel เพียงขนาดเดียว เช่น 3×3 หรือ 5×5 (ดังตัวอย่างในรูปที่ 3)

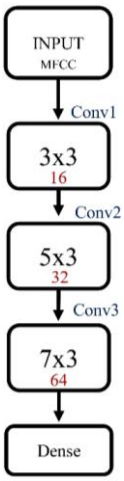


รูปที่ 2 โครงสร้าง VFNet แบบ 3 layer ที่ใช้ kernel หลายขนาดพร้อมกัน

ทั้งนี้การเลือกใช้ตัวกรองสามขนาดดังกล่าวมีเหตุผลหลักคือเพื่ออ้างอิงตามสถาปัตยกรรมของ VFNet [1] ซึ่งเป็นโมเดลต้นแบบสำหรับการเปรียบเทียบและเพื่อจับคุณลักษณะเสียงในหลายระดับสเกล (multi-scale) [1] ตั้งแต่รายละเอียดระดับต่ำด้วยฟิลเตอร์ขนาดเล็ก (3x3) ไปจนถึงบริบทที่กว้างขึ้นด้วยฟิลเตอร์ขนาดใหญ่ (7x3) นอกจากนี้ การเลือกใช้ขนาดเป็นเลขคี่ยังเป็นหลักปฏิบัติทั่วไปในการออกแบบ CNN เพื่อให้ฟิลเตอร์มีพิกเซลกลางที่ชัดเจน ซึ่งช่วยในการรักษาลำดับชั้นเชิงพื้นที่ของคุณลักษณะได้ดี [5-6] โดยทุกรูปแบบของโมเดลมีการใช้ ReLU, Batch Normalization และ MaxPooling ในแต่ละชั้น Convolution เพื่อเพิ่มประสิทธิภาพในการเรียนรู้โมเดลแบบ 3 เลเยอร์ใช้จำนวนฟิลเตอร์ในแต่ละเลเยอร์เท่ากับ 16, 32 และ 64 ตามลำดับ ส่วนโมเดลแบบ 1 เลเยอร์ใช้จำนวนฟิลเตอร์คงที่เท่ากับ 32 ฟิลเตอร์ ทุกโมเดล CNN นำมาต่อด้วย 128 Fully Connected Layers แต่ละโมเดลมีจำนวนพารามิเตอร์ของ CNN ตามสมการที่ (1) นี้

$$\text{Parameters} = (C_{in} \times K_h \times K_w + 1) \times C_{out} \qquad (1)$$

โดยที่ C_{in} คือจำนวนช่องสัญญาณเข้า $K_h \times K_w$ คือขนาดของ kernel +1 คือ bias มีค่า 1 ต่อ filter และ C_{out} คือ จำนวนฟิลเตอร์ [6]



รูปที่ 3 ตัวอย่างโครงสร้าง โมเดลเชิงอนุกรมที่ใช้ kernel หลายขนาดแบบเรียงลำดับ

3.4 การป้องกัน Overfitting และการจัดการความไม่สมดุลของข้อมูล

ในการฝึกโมเดลโครงข่ายประสาทเทียมแบบคอนโวลูชัน (CNN) ในงานวิจัยนี้ ได้กำหนดขั้นตอนและพารามิเตอร์ที่เหมาะสมเพื่อเพิ่มประสิทธิภาพการเรียนรู้และลดปัญหา Overfitting โดยใช้ตัวปรับค่าพารามิเตอร์ (Optimizer) คือ Adam ซึ่งตั้งค่าอัตราการเรียนรู้เริ่มต้นไว้ที่ 0.001

จำนวนรอบการฝึก (Epoch) สูงสุดตั้งไว้ที่ 1000 แต่ใช้กลไก Early Stopping เพื่อหยุดการฝึกหากโมเดลไม่มีแนวโน้มพัฒนา โดยกำหนดค่า Patience เท่ากับ 25 หมายความว่า หาก Validation Loss ไม่ลดลงต่อเนื่องใน 25 รอบ จะหยุดการฝึกทันที และใช้ค่า Min Delta เท่ากับ 0.001 เป็นเกณฑ์พิจารณาว่าการลดลงของ Validation Loss ถือว่ามีความหมายเพียงพอ ระบบจะบันทึกโมเดลจะกระทำเฉพาะกรณีที่ Validation Loss ต่ำที่สุดตลอดการฝึก สำหรับการเสริมความสามารถในการควบคุมอัตราการเรียนรู้ให้เหมาะสมกับช่วงเวลาต่าง ๆ ของการฝึก ได้ใช้กลยุทธ์ปรับค่าอัตราการเรียนรู้แบบ ReduceLROnPlateau โดยหาก Validation Loss ไม่มีการปรับลดลงในช่วง 5 Epoch ระบบจะลดค่า learning rate ลงครึ่งหนึ่งโดยอัตโนมัติ ได้ใช้ Dropout กับแต่ละชั้นของโมเดลเพื่อลดความเสี่ยงของการเกิด Overfitting โดยกำหนด Dropout Rate เท่ากับ 0.3 สำหรับ Convolutional Layers และ 0.5 สำหรับ Fully Connected Layers

ซึ่งช่วยป้องกันการเรียนรู้ที่จำเพาะเจาะจงเกินไป และส่งเสริมให้โมเดลสามารถ generalize ได้ดีขึ้น [5] สำหรับการแก้ปัญหาความไม่สมดุลของข้อมูลในแต่ละคลาส ได้ใช้เทคนิค Random Oversampling โดยการสุ่มเพิ่มจำนวนตัวอย่างในคลาสที่มีจำนวนน้อยให้เทียบเท่ากับคลาสที่มีจำนวนมากที่สุด ซึ่งช่วยให้โมเดลไม่เกิด Bias ต่อคลาสที่มีข้อมูลจำนวนมากกว่า แนวทางทั้งหมดนี้มีจุดมุ่งหมายเพื่อให้โมเดลมีประสิทธิภาพสูงสุดภายใต้เงื่อนไขของทรัพยากรที่จำกัด และลดความซับซ้อนของการฝึกที่อาจไม่จำเป็น

3.5 การพิจารณาเทคนิค Data Augmentation

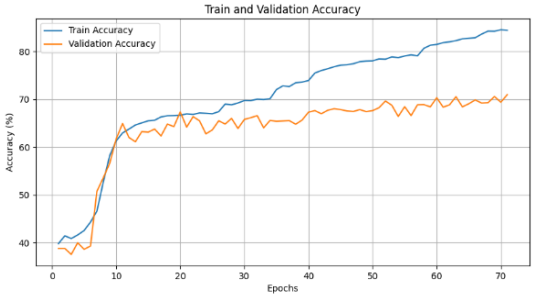
ในงานวิจัยนี้ ไม่ได้ใช้เทคนิคการเพิ่มข้อมูลเสียง (Data Augmentation) เช่น การเลื่อนระดับเสียง (Pitch Shifting) หรือการยืดหดเวลา (Time Stretching) เนื่องจากภาษาไทยเป็นภาษาที่มีวรรณยุกต์ (Tonal Language) ซึ่งโทนเสียงหรือระดับเสียง (Pitch) เป็นคุณลักษณะสำคัญอย่างยิ่งในการจำแนกสำเนียง การใช้เทคนิคดังกล่าวอาจส่งผลให้คุณลักษณะเด่นที่จำเป็นต่อการจำแนกถูกทำลายไป และอาจลดทอนประสิทธิภาพของโมเดลโดยรวม การตัดสินใจไม่ใช้เทคนิคเหล่านี้จึงเป็นการเลือกเพื่อรักษาคุณภาพและความสมบูรณ์ของข้อมูลต้นฉบับไว้ให้มากที่สุด

4. ผลการทดลอง

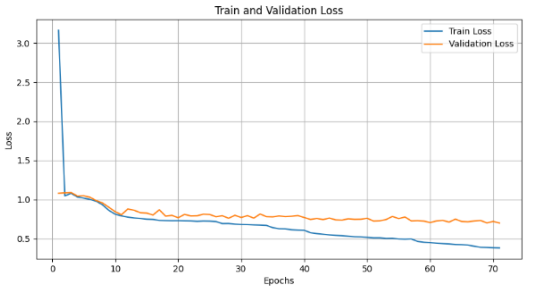
ในส่วนนี้จะนำเสนอผลการทดลอง โดยเริ่มจากการเปรียบเทียบโมเดลแบบ 1 เลเยอร์ ซึ่ง ตารางที่ 1 แสดงจำนวนพารามิเตอร์ของโมเดล 1 เลเยอร์ในแต่ละโครงสร้าง และ ตารางที่ 2 แสดงผลการทดลองด้านประสิทธิภาพ ได้แก่ ค่า Accuracy, Precision, Recall, F1-score และค่า Cross Entropy Loss ของโมเดลกลุ่มนี้ เพื่อให้เห็นภาพการเรียนรู้ที่ชัดเจนยิ่งขึ้น ตารางที่ 3 จึงแสดง Classification Report โดยละเอียดของโมเดล 5x3 ซึ่งเป็นหนึ่งในโมเดล 1 เลเยอร์ที่ให้ผลลัพธ์ดี และ รูปที่ 4 และ 5 แสดง Learning Curve ของค่า Accuracy และ Cross entropy Loss ของโมเดลดังกล่าวตามลำดับ

ถัดมาเป็นการเปรียบเทียบโมเดลแบบ 3 เลเยอร์ ซึ่งมีความซับซ้อนมากกว่า โดย ตารางที่ 4 ได้สรุปจำนวนพารามิเตอร์ของโครงสร้างแบบอนุกรม (Sequential) รูปแบบต่าง ๆ เปรียบเทียบกับโครงสร้างแบบ VFNet 3 ชั้น ส่วน ตารางที่ 5 แสดงผลการทดลองด้านประสิทธิภาพของโมเดลในกลุ่ม 3 เลเยอร์ทั้งหมด เพื่อเปรียบเทียบประสิทธิภาพเทียบกับจำนวนพารามิเตอร์ที่ใช้

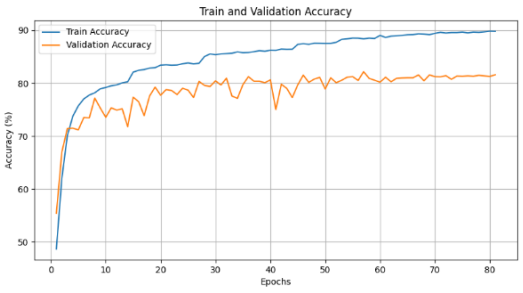
เพื่อวิเคราะห์ประสิทธิภาพของโมเดล 3 เลเยอร์ในรายละเอียด จึงได้ยกตัวอย่างโมเดล 7⟶5⟶3 ซึ่งเป็นโครงสร้างที่ให้ประสิทธิภาพสูงโดยใช้พารามิเตอร์น้อย โดยตารางที่ 6 แสดง Classification Report ของโมเดลนี้ และ รูปที่ 6 และ 7 แสดง Learning Curve ของค่า Accuracy และ Loss ตามลำดับ เพื่อยืนยันเสถียรภาพในการเรียนรู้ของโมเดล สุดท้าย เพื่อวิเคราะห์ความสัมพันธ์ระหว่างสถาปัตยกรรมโมเดลแต่ละรูปแบบและประสิทธิภาพการจำแนกสำเนียง ตารางที่ 7 ได้แสดงผลการคำนวณขนาด Receptive Field (RF) แบบ 2 มิติ ของโครงสร้าง 3 เลเยอร์แต่ละแบบ และ ตารางที่ 8 ได้นำผลลัพธ์ทั้งหมด (RF, F1-score, Loss และ พารามิเตอร์) มาจัดกลุ่มเปรียบเทียบกัน เพื่อสรุปหาโครงสร้างที่มีความสมดุลระหว่างประสิทธิภาพและทรัพยากรที่ใช้มากที่สุด



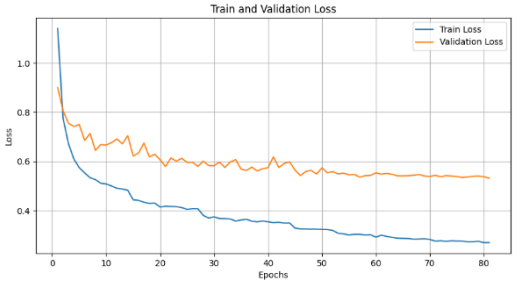
รูปที่ 4 ตัวอย่าง learning curve ค่า Accuracy ของขนาด 5x3 หนึ่งในโมเดลแบบ 1 เลเยอร์



รูปที่ 5 ตัวอย่าง learning curve ค่า Cross entropy loss ของขนาด 5x3 หนึ่งในโมเดลแบบ 1 เลเยอร์



รูปที่ 6 learning curve ค่า Accuracy ของโมเดล 7⟶5⟶3



รูปที่ 7 learning curve ค่า Cross entropy loss ของโมเดล 7⟶5⟶3

5. อภิปรายผล

จากผลการทดลองและตารางเปรียบเทียบโครงสร้างโมเดล CNN สำหรับการจำแนกสำเนียงไทย ตารางที่ 4 และ ตารางที่5 พบว่าโครงสร้างแบบ อนุกรม (sequential) เช่น 5⟶5⟶5 และ 3⟶7⟶5 ให้ผลลัพธ์ที่ดีเยี่ยมทั้งด้าน Accuracy, F1-score และ Cross entropy loss โดยไม่ต้องใช้จำนวนพารามิเตอร์สูงเกินจำเป็น โดยเฉพาะโครงสร้าง 5⟶5⟶5 ซึ่งให้ F1-score สูงที่สุด (82.70%), Loss ต่ำ (0.5212) และใช้พารามิเตอร์รวมเพียง 38,752 แสดงให้เห็นถึงสมดุลระหว่างประสิทธิภาพและต้นทุนการประมวลผลนอกจากนี้ยังพบว่าโครงสร้าง 7⟶5⟶3 ซึ่งเรียง kernel จากใหญ่ไปเล็ก ให้ผลลัพธ์ F1-score สูงถึง 82.23% โดยใช้พารามิเตอร์เพียง 26,560 ต่ำที่สุดในกลุ่มโมเดลที่ให้ผลลัพธ์ระดับสูง ชี้ให้เห็นว่าการออกแบบโครงสร้างให้เรียงลำดับ kernel อย่างมีแบบแผนสามารถลดต้นทุนการประมวลผลได้โดยไม่ลดประสิทธิภาพ

นอกจากนี้ ในการทดลองยังมีการเปรียบเทียบ Receptive Field (RF) ของแต่ละโครงสร้างในตารางที่ 7 และ นำผลมาเปรียบเทียบกับประสิทธิภาพการจำแนกสำเนียง ในตารางที่ 8 พบว่าโครงสร้างที่มี RF ในช่วง 30–38 (แกน Frequency) และ 22 (แกน Time) มักจะให้ผลลัพธ์ดีที่สุด เนื่องจากสามารถจับบริบทได้กว้างพอ โดยไม่สูญเสียรายละเอียดในพื้นที่เล็ก โครงสร้างที่มี RF ใหญ่เกินไป เช่น 7⟶7⟶7 (RF = 50×22) หรือเล็กเกินไป เช่น 3⟶3⟶3 (RF = 22×22) มีแนวโน้มให้ผลลัพธ์ด้อยลงในด้าน Cross entropy loss และ F1-score สำหรับโครงสร้าง VFNet [1] แบบ 3 ชั้น แม้ออกแบบให้ดึงฟีเจอร์หลายขนาดพร้อมกัน แต่กลับใช้พารามิเตอร์สูงถึง 116,256 และยังให้ F1-score เพียง 79.70% ต่ำกว่าหลายโมเดลแบบอนุกรม โมเดล 1 เลเยอร์แบบ VFNet [1] ที่ใช้ฟิลเตอร์ 3×3, 5×3 และ 7×3 พร้อมกัน ให้ F1-score สูงกว่า kernel เดี่ยวเพียงเล็กน้อย แม้จะใช้พารามิเตอร์เพิ่มขึ้นจาก 320 ⟶ 1,536 ซึ่งยังถือว่าไม่คุ้มค่าในการใช้งานจริง

ผลการทดลองชี้ให้เห็นว่าโครงสร้าง CNN แบบอนุกรมที่มีความซับซ้อนไม่มากนักสามารถให้ประสิทธิภาพที่สูงได้ ซึ่งสอดคล้องกับงานวิจัยก่อนหน้าที่แสดงให้เห็นว่า CNN สามารถจำแนกเสียงได้ดีแม้มีสถาปัตยกรรมที่ไม่ลึกมาก [3] อย่างไรก็ตาม โมเดลทั้งหมดในงานวิจัยนี้ใช้ MaxPooling ซึ่งมีข้อจำกัดในการที่อาจสูญเสียข้อมูลบางส่วนไปจากการเลือกเฉพาะค่าสูงสุด ดังที่ [4] ได้ชี้ให้เห็นและเสนอการใช้ Attention Pooling เพื่อให้น้ำหนักกับทุกส่วนของข้อมูลอย่างเหมาะสม ซึ่งอาจเป็นแนวทางในการปรับปรุงประสิทธิภาพของโมเดลต่อไป

6. สรุป

การเปรียบเทียบโครงสร้าง Convolutional Neural Network สำหรับการจำแนกสำเนียงไทยในครั้งนี้ แสดงให้เห็นว่าโครงสร้างแบบอนุกรมที่เลือกใช้ kernel ขนาดต่างกันในแต่ละชั้นอย่างมีแบบแผน เช่น 5⟶7⟶3, 3⟶7⟶5 และ 7⟶5⟶3 สามารถให้ประสิทธิภาพสูงเทียบเท่าหรือเหนือกว่า VFNet [1] แบบขนาน แม้ใช้พารามิเตอร์น้อยกว่าอย่างชัดเจน โครงสร้าง VFNet [1] ที่ใช้ kernel หลายขนาดพร้อมกันมีแนวโน้มใช้พารามิเตอร์สูงมาก (เช่น 116,256 ในแบบ 3 ชั้น) แต่กลับไม่ได้ให้ประสิทธิภาพที่เหนือกว่า ขณะที่โมเดลแบบอนุกรมเช่น 7⟶5⟶3 ใช้พารามิเตอร์เพียง 26,560 แต่ให้ F1-score สูงถึง 82.23% อีกทั้งผลการวิเคราะห์ Receptive Field ยังช่วยสนับสนุนว่า RF ที่เหมาะสมควรอยู่ในช่วง 30–38 (แกนความถี่) เพื่อให้จับบริบทได้ดีและไม่สูญเสียความละเอียด การออกแบบโครงสร้างโมเดล CNN ที่พิจารณาทั้งขนาด kernel,ลำดับการเรียง, RF และจำนวนพารามิเตอร์สามารถช่วยเพิ่มประสิทธิภาพในการจำแนกสำเนียง และลดต้นทุนการประมวลผลในระบบจริงได้

งานวิจัยนี้มีข้อจำกัดในการใช้ข้อมูลเพียง 3 สำเนียงหลักจาก Thai Dialect Corpus [2] ซึ่งผลลัพธ์อาจยังไม่สามารถเป็นตัวแทนของสำเนียงไทยทั้งหมดได้

7. งานวิจัยในอนาคต

สำหรับงานวิจัยในอนาคต มีแนวทางในการต่อยอดและพัฒนาหลายประการ ประการแรกคือ การขยายชุดข้อมูลให้ครอบคลุมสำเนียงภาษาไทยอื่น ๆ ที่มีความหลากหลายมากขึ้น เพื่อทดสอบประสิทธิภาพและความสามารถของโมเดลในการทำงานได้ดีกับข้อมูลใหม่ (Generalization) ประการที่สองคือ การนำโมเดลจำแนกสำเนียงไปประยุกต์ใช้ร่วมกับระบบรู้จำเสียงพูด (ASR) จริง เพื่อสร้าง Pipeline ที่สามารถเลือกแบบจำลองทางภาษา (Language Model) ที่เหมาะสมกับสำเนียงของผู้ใช้ ซึ่งจะช่วยเพิ่มความแม่นยำโดยรวมของระบบได้ ประการที่สาม คือ อาจทดสอบนัยสำคัญทางสถิติ เพื่อยืนยันความแตกต่างของประสิทธิภาพระหว่างโมเดลที่ดีที่สุดและสุดท้ายคือการปรับปรุงกลไก Pooling ของโมเดล โดยอาจทดลองแทนที่ Max Pooling ด้วย Attention-based Pooling ดังที่นำเสนอโดย [4] ซึ่งอาจช่วยให้โมเดลสามารถเรียนรู้ที่จะให้น้ำหนักกับคุณลักษณะทางเสียง (acoustic features) ที่เป็นเอกลักษณ์ของแต่ละสำเนียงได้ดีขึ้น และอาจนำไปสู่ประสิทธิภาพในการจำแนกที่สูงขึ้นได้

ตารางที่ 1 จำนวนพารามิเตอร์โมเดลแบบ 1 เลเยอร์

โครงสร้างโมเดล	Conv1	รวมทั้งหมด
1 Layer (3x3)	320	320
1 Layer (5x3)	512	512
1 Layer (7x3)	704	704
1 Layer (5x5)	832	832
1 Layer (7x7)	1600	1600
VFNet 1 Layer (3x3,5x3,7x3)	1536	1536

ตารางที่ 2 โมเดลแบบ 1 เลเยอร์

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Best Cross entropy loss
VFNet	69.81	70.09	69.75	69.71	0.7144
3 x 3	67.54	67.69	67.51	67.52	0.7369
5 x 3	70.96	70.93	70.68	70.71	0.7000
7 x 3	69.30	69.17	68.86	68.96	0.7085
5 x 5	68.22	69.32	68.70	68.30	0.7055
7 x 7	69.63	70.17	69.49	69.54	0.6962

ตารางที่ 3 ตัวอย่าง classification report ของขนาด 5x3 หนึ่งในโมเดลแบบ 1 เลเยอร์

Class	Precision (%)	Recall (%)	F1-score (%)
คำเมือง	65.89	69.90	67.83
โคราช	71.88	74.54	73.18
กลาง	75.03	67.59	71.12
Accuracy	-	-	70.96
Macro Avg	70.93	70.68	70.71
Weighted Avg	71.16	70.96	70.97

ตารางที่ 4 จำนวนพารามิเตอร์โมเดลแบบ 3 เลเยอร์

โครงสร้างโมเดล	Conv1	Conv2	Conv3	รวมทั้งหมด
3⟶3⟶3	160	4640	18496	23296
3⟶5⟶7	160	7712	43072	50944
3⟶7⟶5	160	10784	30784	41728
5⟶5⟶5	256	7712	30784	38752
5⟶3⟶7	256	4640	43072	47968
5⟶7⟶3	256	10784	18496	29536
7⟶7⟶7	352	10,784	43,072	54,208
7⟶3⟶5	352	4640	30784	35776
7⟶5⟶3	352	7712	18496	26560
VFNet 3L	768	23136	92352	116256

ตารางที่ 5 โมเดลแบบ 3 เลเยอร์

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Best Cross entropy loss
3⟶3⟶3	80.34	80.57	80.67	80.58	0.5581
3⟶5⟶7	80.66	81.86	80.00	80.46	0.5693
3⟶7⟶5	82.57	83.58	82.15	82.60	0.5333
5⟶5⟶5	82.71	82.74	82.86	82.70	0.5212
5⟶3⟶7	80.27	80.69	80.26	80.40	0.5424
5⟶7⟶3	81.52	81.70	81.58	81.62	0.5196
7⟶7⟶7	80.66	81.37	80.46	80.71	0.5485
7⟶3⟶5	80.66	80.93	81.36	80.75	0.5276
7⟶5⟶3	82.17	83.05	81.93	82.23	0.5325
VFNet	79.76	81.09	79.39	79.70	0.5572

ตารางที่ 6 ตัวอย่าง classification report ของโมเดล 7⟶5⟶3 หนึ่งในโมเดลแบบ 3 เลเยอร์ ที่มีจำนวนพารามิเตอร์เพียง 26560 แต่ให้ผลลัพธ์ที่ดีกว่าในบางโมเดลที่มีจำนวนพารามิเตอร์มากกว่า

Class	Precision (%)	Recall (%)	F1-score (%)
คำเมือง	82.09	83.83	83.36
โคราช	77.53	86.57	81.80
กลาง	88.73	75.39	81.52
Accuracy	-	-	82.17
Macro Avg	83.05	81.93	82.23
Weighted Avg	82.70	82.17	82.16

ตารางที่ 7 Receptive Field (RF) 2D สำหรับแต่ละโครงสร้าง CNN

โครงสร้างโมเดล	RF หลัง Layer 1	RF หลัง Layer 2	RF หลัง Layer 3	RF สุดท้าย
3⟶3⟶3	4×4	10×10	22×22	22×22
3⟶5⟶7	4×4	14×10	42×22	42×22
3⟶7⟶5	4×4	18×10	38×22	38×22
5⟶3⟶7	6×4	12×10	40×22	40×22
5⟶5⟶5	6×4	16×10	36×22	36×22
5⟶7⟶3	6×4	20×10	32×22	32×22
7⟶3⟶5	8×4	14×10	34×22	34×22
7⟶5⟶3	8×4	18×10	30×22	30×22
7⟶7⟶7	8×4	22×10	50×22	50×22

ตารางที่ 8 เปรียบเทียบ RF F1-score และ Cross Entropy Loss

โครงสร้าง	RF	F1-score (%)	Loss	พารามิเตอร์	ข้อสังเกต
3⟶3⟶3	22×22	80.58	0.5581	23,296	baseline, พารามิเตอร์ต่ำ แต่ loss สูง
3⟶5⟶7	42×22	80.46	0.5693	50,944	RF ใหญ่ , พารามิเตอร์ สูง
3⟶7⟶5	38×22	82.60	0.5333	41,728	F1 สูง, พารามิเตอร์ ปานกลาง
5⟶5⟶5	36×22	82.70	0.5212	38,752	มีความสมดุล ระหว่าง ประสิทธิภาพและ พารามิเตอร์ดีที่สุด
5⟶3⟶7	40×22	80.40	0.5424	47,968	ประสิทธิภาพไม่ สอดคล้องกับ พารามิเตอร์ที่สูง
5⟶7⟶3	32×22	81.62	0.5196	29,536	ประสิทธิภาพดี Loss และ พารามิเตอร์ต่ำ
7⟶7⟶7	50×22	80.71	0.5485	54,208	RF ใหญ่มาก, พารามิเตอร์สูง
7⟶3⟶5	34×22	80.75	0.5276	35,776	สมดุลดี , F1 ปานกลาง

โครงสร้าง	RF	F1-score (%)	Loss	พารามิเตอร์	ข้อสังเกต
7⟶5⟶3	30×22	82.23	0.5325	26,560	F1-score สูง โดยใช้พารามิเตอร์น้อยที่สุด (ในกลุ่มประสิทธิภาพสูง)
VFNet (3L)	-	79.70	0.5572	116,256	แม้ใช้ multi-scale แต่แพ้ทุกตัวด้าน efficiency

เอกสารอ้างอิง

[1] A. Ahmed, P. Tangri, A. Panda, D. Ramani, and S. Karmakar, “VFNet: A Convolutional Architecture for Accent Classification,” in Proc. IEEE 16th India Council Int. Conf. (INDICON), Nov. 1-4 2019, pp. 1–4.

[2] A. Suwanbandit, B. Naowarat, O. Sangpetch, and E. Chuangsuwanich, “Thai Dialect Corpus and Transfer-based Curriculum Learning for Dialect ASR,” in Proc. Interspeech 2023, Dublin, Ireland, Aug. 20–24 2023, pp. 4069–4073, doi: 10.21437/Interspeech.2023-1828.

[3] K. J. Piczak, “Environmental sound classification with convolutional neural networks,” in 2015 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2015, pp. 1-6.

[4] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley, and B. W. Schuller, "Attention-based convolutional neural networks for acoustic scene classification," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), 2018, pp. 39-43.

[5] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.

[6] Stanford CS231n, “Convolutional Neural Networks for Visual Recognition,” [Online]. Available: http://cs231n.stanford.edu [Accessed: 19-Jun-2025].