การศึกษาเทคโนโลยีการรู้จำตัวบุคคลด้วยเสียงเพื่อพัฒนาแพลตฟอร์มการยืนยันตัวบุคคลด้วยเทคโนโลยีชีวมิติที่มีความมั่นคงปลอดภัย สำหรับกองทัพอากาศไทย

The Study of Speaker Recognition Technology to Develop a Secure Biometric Authentication Platform for RTAF

Received : November 22, 2022
Revised : December 6, 2022
Accepted : December 7, 2022

[1*]พายัพ ศิรินาม และ [2#]ประสงค์ ปราณีตพลกรัง
[1,2] กองการศึกษา โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช
[1*]Payap Sirinam, [2#]Prasong Praneetpolgrang
[1,2] Faculty of Academic, Navaminda Kasatriyadhiraj Royal Air Force Academy
[*]payap_siri@rtaf.mi.th, [#]prasong_pra@rtaf.mi.th

บทคัดย่อ

ในปัจจุบัน การประยุกต์ใช้งานการยืนยันตัวบุคคลด้วยเทคโนโลยีชีวมิติได้รับความนิยมเพิ่มมากขึ้น สำหรับการยืนยันตัวบุคคลแบบหลายปัจจัย ถึงกระนั้นกองทัพอากาศยังขาดการประยุกต์ใช้เทคโนโลยีชีวมิติเพื่อการยืนยันตัวบุคคล ดังนั้น คณะผู้วิจัยจึงได้วิจัยและพัฒนาองค์ความรู้ในการประยุกต์ใช้งานเทคโนโลยีชีวมิติด้วยการใช้เสียงของมนุษย์เพื่อใช้ยืนยันตัวบุคคล เนื่องจากชีวมิติดังกล่าวมีความแม่นยำ ปลอดภัย และใช้ต้นทุนในการพัฒนาระบบที่ต่ำกว่าชีวมิติประเภทอื่น ๆ โดยคณะผู้วิจัยได้พัฒนาโมเดลโดยใช้ชุดข้อมูล LibriSpeech และทำการทดสอบโดยใช้เสียงภาษาไทย ภายใต้ข้อจำกัดด้านสภาพแวดล้อมและความเสี่ยงจากเสียงปลอมแปลงจากเทคโนโลยี Deep Voice ผลการทดลองพบว่า โมเดลมีระดับความถูกต้องในการยืนยันตัวบุคคลที่ดีที่สุดอยู่ที่ 75% ณ ค่าขีดแบ่งที่ 0.9 และพบว่า ปัจจัยเกี่ยวกับสภาพแวดล้อมที่มีผลกระทบอย่างมีนัยยะสำคัญต่อประสิทธิภาพของโมเดลได้แก่ รูปแบบประโยคในการยืนยันตัวบุคคลที่แตกต่างกัน ประเภทของอุปกรณ์บันทึกเสียงที่แตกต่างกัน ในทางตรงกันข้าม การใส่หน้ากากอนามัยและเสียงรบกวนต่าง ๆ ที่ไม่ดังเกินไป กลับไม่มีผลกระทบอย่างมีนัยยะสำคัญกับประสิทธิภาพมากไปกว่านั้น ผลการวิจัยยังระบุว่า โมเดลดังกล่าวมีความเปราะบางต่อเสียงปลอมแปลงจากเทคโนโลยี Deep Voice ซึ่งเป็นการเน้นย้ำถึงความสำคัญของการพัฒนาโมเดลที่ทนทานต่อการโจมตีทางไซเบอร์ดังกล่าวก่อนที่จะนำเอาโมเดลไปประยุกต์ใช้งานจริงในอนาคต

คำสำคัญ: การรู้จำตัวผู้พูด, การเรียนรู้เชิงลึก, ความมั่นคงปลอดภัยทางไซเบอร์

Abstract

Applications for multi-factor authentication using biometrics are becoming increasingly popular today. However, the Royal Thai Air Force has not implemented biometric technologies for identity verification. For this reason, the research team conducted a study and created a body of knowledge on the use of biometric

technologies to identify individuals by their voice. The researchers created a model based on the LibriSpeech dataset and tested it using the Thai voice, as such biometric methods are more accurate, secure, and cost-effective than other biometric methods. The results reveal that environmental elements such as different sentence forms for verification and different types of recording devices are found to significantly affect the performance of the model. In contrast, wearing a mask and other quiet noises do not significantly decrease the performance of the model. Finally, the results also show the model's vulnerability to deep voice spoofing, highlighting the need to develop a model that is resistant to such cyberattacks before it is used in the future.

Keywords: Speaker Recognition, Deep Learning, Cyber Security

1. บทนำ

ในปัจจุบัน วิวัฒนาการด้านการประมวลผลส่งผลให้ความก้าวหน้าของเทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence) ถูกพัฒนายกระดับไปอย่างรวดเร็วโดยเฉพาะอย่างยิ่งการเรียนรู้เชิงลึก (Deep Learning) โดยความก้าวหน้าดังกล่าวได้ถูกนำมาประยุกต์ใช้เพื่อให้เกิดประโยชน์ต่อมนุษย์ เสริมสร้างความสะดวกสบาย ความมั่นคงปลอดภัยในชีวิตและความมั่นคงของชาติ (National Security) ในทางกลับกัน เทคโนโลยีปัญญาประดิษฐ์ อาจถูกนำมาใช้งานอย่างไม่ถูกต้องและอาจส่งผลกระทบในแง่ลบต่อการใช้ชีวิตของมนุษย์ เช่น การปลอมแปลงอัตลักษณ์ส่วนบุคคล (Fake Personal Identity) และที่สำคัญคือ การใช้เทคโนโลยีดังกล่าวอาจส่งผลกระทบโดยตรงต่อความมั่นคงของชาติ

การปลอมแปลงอัตลักษณ์ส่วนบุคคล ถือเป็นหนึ่งในภัยคุกคามที่มีศักยภาพ (Potential Threat) โดยเมื่อผู้เสียหาย (Victim) ถูกปลอมแปลงอัตลักษณ์ส่วนบุคคล ก็จะส่งผลเสียโดยตรงต่อชื่อเสียง และหากผู้เสียหาย เป็นบุคคลที่มีอำนาจในการสั่งการ หรือดำเนินการในงานด้านความมั่นคง อาจส่งผลกระทบต่อความมั่นคงอย่างมีนัยยะสำคัญ เช่น การปลอมแปลงตัวตนของผู้บังคับบัญชา ที่อาจส่งผลต่อภาพลักษณ์องค์กร รวมถือการออกคำสั่งทางยุทธการที่มิได้มาจากผู้ที่มีอำนาจตัวจริง ดังนั้น เทคโนโลยีการยืนยันตัวบุคคลที่มีความแม่นยำเชื่อถือได้จึงมีส่วนสำคัญในการป้องกันปัญหาดังกล่าว โดยในปัจจุบันเทคโนโลยีการยืนยันตัวบุคคลสามารถดำเนินการได้ในหลายรูปแบบ เช่น การใช้รหัสผ่านที่มีความแข็งแรง (Strong Password) การยืนยันตัวตนแบบ 2 ชั้น (Two-Factor Authentication) รวมถึงการใช้การยืนยันตัวตนด้วยข้อมูลทางชีวมิติ (Biometric Authentication) ผ่านการยืนยันตัวบุคคลด้วยการรู้จำใบหน้า (Face Recognition) การรู้จำด้วยการสแกนม่านตา (Irish Recognition) การยืนยันบุคคลด้วยการใช้เสียงบุคคล (Speaker Recognition) โดยในปัจจุบันกองทัพอากาศ มีการประยุกต์ใช้งานการยืนยันตัวตน 2 แบบ ผ่านการใช้รหัสผ่านคู่กับรหัสลับที่ถูกส่งไปยังโทรศัพท์เคลื่อนที่ของผู้ใช้งาน และถูกใช้งานเฉพาะการเข้าถึงระบบอีเมลกองทัพอากาศของผู้บังคับบัญชา รวมถึงผู้ที่มีความประสงค์จะใช้งานเท่านั้น [1] ที่สำคัญคือการใช้งานการยืนยันตัวตนด้วยข้อมูลทางชีวมิติที่มีความแม่นยำและมีความมั่นคงปลอดภัยสูงยังคงมีการประยุกต์ใช้งานในวงที่จำกัด จึงเป็นปัญหาและความท้าทายต่อการรองรับต่อความเปลี่ยนแปลงด้านเทคโนโลยีและภัยคุกคามทางไซเบอร์ที่เกิดขึ้นในอนาคต

2. ขอบเขตงานวิจัย

2.1 คณะผู้วิจัยปฏิบัติงานบนทรัพยากรพื้นฐานที่มีอยู่ของห้องปฏิบัติการทางไซเบอร์ ภาควิชาคอมพิวเตอร์ กองการศึกษา โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช และที่ได้รับการจัดสรรจากโครงการวิจัยนี้ และผลงานวิจัยมีขีดความสามารถการประมวลผลอย่างจำกัดตามทรัพยากรวิจัยที่มีอยู่

2.2 การวิจัยมุ่งเน้นศึกษาแนวทางการพัฒนาและประยุกต์ใช้โมเดลที่มีประสิทธิภาพสูงสุด ณ ห้วงเวลาการทำวิจัย และสอดคล้องกับทรัพยากรที่มีอยู่จำนวน 1 โมเดล บนพื้นฐานของบริบทของกองทัพอากาศไทย

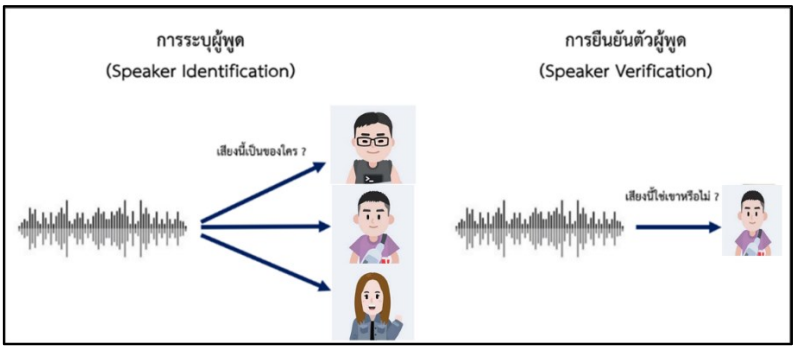
3. ทฤษฎีที่เกี่ยวข้อง

3.1 การพิสูจน์และยืนยันตัวบุคคล

การพิสูจน์และยืนยันตัวบุคคล (Authentication) เป็นกระบวนการที่ใช้ในการตรวจสอบบุคคลก่อนเข้าใช้งานระบบคอมพิวเตอร์เพื่อพิสูจน์ทราบว่าเป็นบุคคลนั้นจริง ซึ่งถือเป็นระบบที่มีความสำคัญต่อความมั่นคงปลอดภัยทางไซเบอร์ เพราะเป็นกระบวนการที่จะรับรองว่าจะไม่มีการปลอมแปลงอัตลักษณ์ของบุคคล หรืออาจกล่าวอีกนัยหนึ่งคือ สิทธิ์ในการใช้งานต้องถูกจำกัดไว้สำหรับผู้ที่มีสิทธิ์เท่านั้น โดยในปัจจุบัน กระบวนการพิสูจน์และยืนยันตัวบุคคลได้ถูกใช้งานทั่วไปและแพร่หลายในการบริการต่าง ๆ เช่น การใช้บริการผ่านระบบอินเทอร์เน็ต การเข้าใช้งานอุปกรณ์อิเล็กทรอนิกส์ การเข้าถึงบัญชีผู้ใช้งานในระบบออนไลน์ต่าง ๆ เป็นต้น สำหรับวิธีการพิสูจน์และยืนยันตัวตนแบ่งออกเป็น 3 ประเภท [2] ดังนี้ 1) สิ่งที่คุณรู้ เช่น ชื่อบัญชี รหัสผ่าน รหัส PIN หมายเลขบัตรประจำตัวประชาชน หรือคำถามคำตอบที่เป็นความลับ เป็นต้น 2) สิ่งที่คุณมี เช่น โทรศัพท์มือถือที่ใช้คู่กับ One-Time Password (OTP) หรือแอปพลิเคชันเพื่อยืนยันตัวบุคคล (Authentication Application) รวมถึง Token และบัตรต่าง ๆ เป็นต้น และ 3) สิ่งที่คุณเป็นหรือชีวมิติ (Biometric) เช่น ลายนิ้วมือ ฝ่ามือ ม่านตา ใบหน้า เสียง รวมถึงพฤติกรรมที่ทำเป็นประจำ เป็นต้น โดยส่วนใหญ่ การยืนยันตัวบุคคลจะใช้รูปแบบชื่อบัญชีผู้ใช้ (Username) และรหัสผ่าน (Password) ถึงกระนั้น รหัสผ่านที่ตั้งขึ้นในบางครั้ง อาจมีความไม่ปลอดภัย เช่น การตั้งรหัสผ่านที่มีความปลอดภัยต่ำง่ายต่อการคาดเดาง่าย และหากมีการถูกโจมตี ผู้โจมตีจะสามารถเข้าถึงทรัพยากรของระบบโดยไม่ได้รับอนุญาตได้ทันที ดังนั้น การพิสูจน์และยืนยันตัวบุคคลด้วยหลายปัจจัยจึงถูกนำมาใช้เพื่อเพิ่มความมั่นคงปลอดภัย เช่น การใช้รหัสผ่านควบคู่กับ OTP การใช้รหัสผ่านคู่กับลายนิ้วมือ เป็นต้น ซึ่งอาจมีข้อเสียคือ ต้องมีการเพิ่มขั้นตอนของการดำเนินการในการยืนยันตัวบุคคล รวมถึงการจัดหาอุปกรณ์เพิ่มเติม เช่น ที่อ่านลายนิ้วมือ ที่อ่านม่านตา ไมโครโฟน ในการรับค่าชีวมิติในการยืนยันตัวบุคคล ดังนั้น คณะผู้วิจัยจึงเลือกการยืนยันตัวบุคคลด้วยเสียง เป็นกรณีศึกษา เนื่องจากอุปกรณ์ที่ใช้ในการรับค่าชีวมิติ (ไมโครโฟน) เป็นอุปกรณ์พื้นฐานที่สามารถหาได้ง่ายสำหรับอุปกรณ์คอมพิวเตอร์และอุปกรณ์พกพาทั่วไป

3.2 การรู้จำตัวบุคคลด้วยเสียง

การรู้จำตัวบุคคลด้วยเสียง (Speaker Recognition) เป็นเทคโนโลยีการพิสูจน์และยืนยันตัวตนด้วยชีวมิติผ่านเสียงพูดของมนุษย์โดยสามารถแบ่งออกได้เป็น 2 ประเภทได้แก่ การระบุผู้พูด (Speaker Identification) และการยืนยันตัวผู้พูด (Speaker Verification) [3] โดยการระบุผู้พูด เป็นกระบวนการเพื่อระบุว่าเสียงดังกล่าวใกล้เคียงกับเสียงของบุคคลใดมากที่สุด [4] ในขณะที่การยืนยันตัวผู้พูด เป็นกระบวนการเพื่อตรวจสอบว่าเสียงดังกล่าวนั้นเป็นเสียงของบุคคลผู้นั้นจริงหรือไม่ ดังแสดงในรูปที่ 1 ถึงอย่างไรก็ตาม ขั้นตอนการระบุผู้พูดและการยืนยันตัวผู้พูดมีความสัมพันธ์กัน กล่าวคือ การระบุผู้พูด คือการยืนยันตัวผู้พูดกับคลังข้อมูลเสียงของผู้พูดหลายคน และวัดค่าความคล้าย (Similarity Score) เพื่อหาว่าเสียงดังกล่าวมีความคล้ายกับเสียงของผู้พูดคนใดมากที่สุด



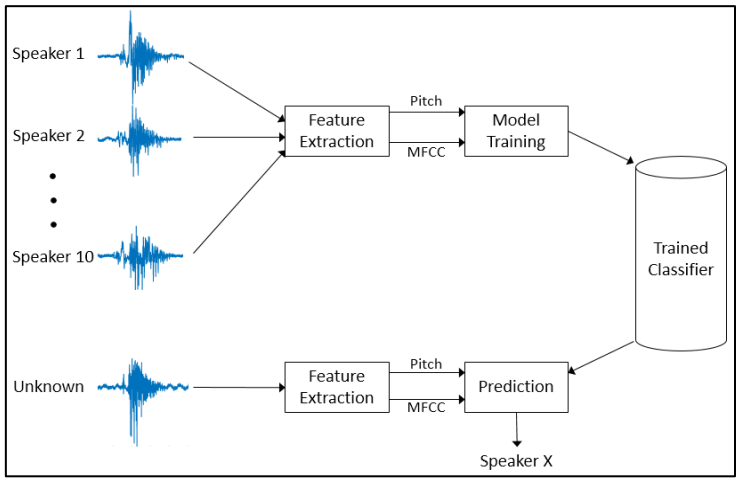
รูปที่ 1 รูปแบบการรู้จำตัวบุคคลด้วยเสียง

3.3 การรู้จำตัวบุคคลด้วยเสียงผ่านการเรียนรู้เชิงลึก

การเรียนรู้เชิงลึก (Deep Learning) ได้เข้ามามีบทบาทต่อการพัฒนาโมเดลเพื่อทำการระบุหรือยืนยันตัวบุคคลผ่านเสียงพูดของมนุษย์ โดยขั้นตอนการของการฝึกฝนโมเดล (Training Phase) และการนำไปใช้งานผ่านการทดสอบ (Testing Phase) ใช้หลักการพื้นฐานเดียวกันกับการเรียนรู้เชิงลึกในงานด้านอื่น ๆ เช่น การรู้จำวัตถุ การรู้จำภาพบุคคล เป็นต้น [5] ในรูปที่ 2 แสดงถึงขั้นตอนการฝึกฝนและทดสอบโมเดล [4] โดยมีรายละเอียด ดังนี้

3.3.1 การฝึกฝนโมเดล

ในขั้นตอนแรกคือการระบุบุคคลเป้าหมาย (Target Speakers) ที่ต้องการให้โมเดลทำการเรียนรู้ เช่น หากต้องการสร้างโมเดลเพื่อระบุว่าเสียงดังกล่าวเป็นกำลังพลกองทัพอากาศ ในกรณีนี้ สมมติมีกำลังพลอยู่ 10 คน หมายความว่า ทั้ง 10 คนเหล่านี้จะเป็นบุคคลเป้าหมายสำหรับการฝึกฝนโมเดลนี้ หลังจากนั้น ก็ทำการบันทึกเสียง n เสียง ของกำลังพลทั้ง 10 คน จากนั้นก็ทำการสกัดคุณลักษณะ (Feature Extraction) [6] ซึ่งเป็นกระบวนการในการแปลงข้อมูลคลื่นเสียง (Wave Form) ให้อยู่ในรูปแบบที่สามารถนำไปฝึกฝนโมเดลได้ ซึ่งในกรณีนี้จะอยู่ในรูปแบบของเมทริกซ์ (Matrix) ที่ประกอบด้วยค่าคุณลักษณะต่าง ๆ ของเสียง เช่น ค่า Pitch, Mel-Frequency Cepstral Coefficients (MFCCs) เป็นต้น [4] จากนั้น ก็ทำการส่งค่าคุณลักษณะดังกล่าวเพื่อทำการฝึกฝนโมเดลโดยใช้เทคนิคการเรียนรู้ของเครื่องเชิงลึกต่าง ๆ เช่น Deep Neural Networks (DNN), Convolutional Neural Networks (CNN) เป็นต้น เมื่อทำการฝึกฝนโมเดลเสร็จแล้วก็จะได้โมเดลที่พร้อมสำหรับการจำแนกว่าเสียงดังกล่าวเป็นเสียงของใคร (Trained Classifier)



รูปที่ 2 ขั้นตอนการฝึกฝนและทดสอบโมเดล [4]

3.3.2 การทดสอบโมเดล

เมื่อมีเสียงของบุคคลที่ต้องการระบุหรือยืนยันตัวบุคคล ขั้นตอนแรก คือการการสกัดคุณลักษณะของเสียงเช่นเดียวกับการฝึกฝนโมเดล จากนั้นก็ทำการส่งค่าคุณลักษณะที่สกัดออกมาได้ไปยังโมเดลที่ถูกฝึกฝนมาในขั้นแรก โดยโมเดลจะทำการวิเคราะห์เพื่อวัดค่าความคล้าย จากนั้นทำการตัดสินใจว่าเสียงดังกล่าวเป็นเสียงของบุคคลใด หรือเป็นเสียงของบุคคลเป้าหมายหรือไม่

3.4 ผลกระทบด้านความมั่นคงปลอดภัยไซเบอร์จากการใช้งาน Deep Voice

ในห้วงที่ผ่านมา มีการศึกษาผลกระทบจากการใช้งานการโคลนเสียงด้วยเครือข่ายประสาทเทียมว่ามีผลกระทบในเชิงลบอย่างไร โดยในปี พ.ศ.2564 นักวิจัยของโรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช ได้ทำการวิจัยโดยการนำเอาเรียนรู้เชิงลึกมาประยุกต์ใช้ในการสร้างเทคโนโลยีที่เกี่ยวข้องกับการเลียนเสียงของบุคคลหนึ่งผ่านโมเดล เช่น โมเดลการสังเคราะห์เสียงของมนุษย์จากประโยคข้อความต่าง ๆ ที่เป็นภาษาไทย [7] โดยงานวิจัยนี้ คณะผู้วิจัยได้ศึกษาและนำเสนอแนวทางการพัฒนาและประยุกต์ใช้โมเดลการเลียนเสียงเชิงลึกสำหรับงานด้านสงครามไซเบอร์ ผ่านการใช้โมเดล GAN [8] โดยผลการศึกษาพบว่า โมเดล StarGAN-VC และ CycleGAN-VC สามารถนำมาใช้ในการแปลงเสียงของบุคคลทั่วไปให้กลายเป็นบุคคลเป้าหมาย เช่น นักการเมือง ผู้บริหารประเทศ เพื่อสร้างข่าวปลอมในสงครามไซเบอร์ได้ และมีศักยภาพในการหลอกลวงผู้ใช้งานอินเทอร์เน็ตให้หลงเชื่อว่าเสียงที่ถูกปลอมแปลงเป็นเสียงของบุคคลเป้าหมายจริง โดยในกรณีที่ร้ายแรงที่สุด กลุ่มตัวอย่างกว่า 40 % ถูกหลอกลวงด้วยเสียงปลอมแปลง ซึ่งผลการค้นพบดังกล่าวได้เน้นย้ำและสร้างความตระหนักต่อภัยคุกคามรูปแบบใหม่ รวมถึงการแสวงหาแนวทางการตรวจจับและป้องกัน

4. การดำเนินการวิจัย

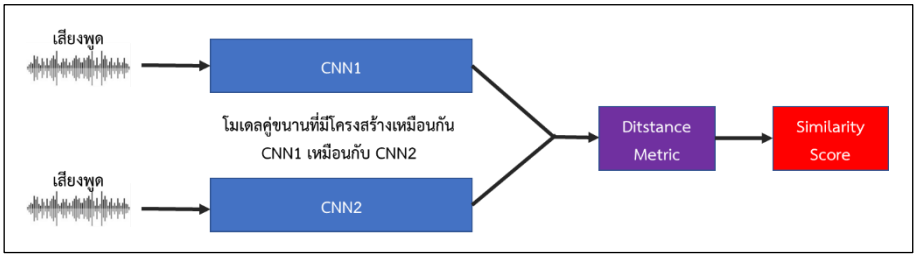
4.1 การศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องเพื่อทดสอบหาโมเดลที่เหมาะสม

คณะผู้วิจัยได้ทำการศึกษาและทบทวนวรรณกรรมที่เกี่ยวข้องกับเทคโนโลยีการรู้จำตัวบุคคลด้วยเสียงที่ได้รับการตีพิมพ์ในที่ประชุมทางวิชาการที่มีชื่อเสียงและเป็นที่ยอมรับว่าเป็นโมเดลที่มีศักยภาพสูง (State-of-the-Art Models) รวมถึงแนวทางการประยุกต์ใช้งานภายใต้ข้อจำกัดต่าง ๆ โดยผลของการศึกษาพบว่า แนวทางการสร้างโมเดล เพื่อรู้จำตัวบุคคลจากเสียงของบุคคล โดยทั่วไปจะใช้โมเดลจำแนกประเภท (Classification Model) และใช้การเรียนรู้เชิงลึก (Deep Learning) ผ่านโมเดลประเภทโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network: CNN)

4.1.1 ข้อจำกัดของโมเดลปัจจุบันและคุณลักษณะโมเดลที่มีความเหมาะสม

โมเดลโครงข่ายประสาทแบบคอนโวลูชันมีข้อจำกัดที่สำคัญของการนำไปประยุกต์ใช้งานจริง โดยเฉพาะในระบบราชการที่มีการโยกย้ายทุก ๆ 6 เดือน โมเดลในลักษณะนี้ไม่สามารถรองรับการเปลี่ยนจำนวนกำลังพล โดยหากมีการเพิ่มจำนวนกำลังพล โมเดลจะต้องถูกสร้างและฝึกฝนขึ้นมาใหม่ เพื่อรองรับการเปลี่ยนแปลงดังกล่าว รวมถึงการเรียนรู้ของโมเดลยังต้องการต้นแบบเสียงของแต่ละคนจำนวนมาก ซึ่งในการประยุกต์ใช้งานจริง ผู้ดูแลระบบมีความจำเป็นต้องให้กำลังพลใหม่บันทึกเสียงของตนเองเป็นจำนวนมาก เพื่อใช้ในการฝึกฝนโมเดล ซึ่งอาจสร้างความยุ่งยากในการประยุกต์ใช้ได้ ดังนั้นโมเดลที่มีความเหมาะสม จึงต้องเป็นโมเดลที่มีคุณสมบัติที่สามารถรองรับการเปลี่ยนแปลงจำนวนกำลังพลได้ เช่น โมเดลที่ไม่จำเป็นต้องฝึกฝนโมเดลใหม่ รวมถึงการฝึกฝนโมเดลไม่จำเป็นต้องใช้เสียงของบุคคลที่มากเกินไป โดยในปัจจุบันมีองค์ความรู้ที่สามารถสร้างโมเดลเพื่อรองรับความต้องการดังกล่าว โดยองค์ความรู้นี้เรียกว่า One-Shot Learning (OSL) ซึ่ง OSL เป็นกระบวนการเรียนรู้เพื่อสร้างโมเดล โดยใช้ข้อมูลจำนวนน้อย หรือใช้เพียงแค่ 1 ตัวอย่างเสียงต่อการเรียนรู้เพื่อสร้างโมเดลในการแยกแยะว่าเป็นคนเดียวกันหรือคนละคนกัน

ในปัจจุบันมีโมเดลที่มีคุณลักษณะดังกล่าวและถูกนำไปประยุกต์ใช้งานในงานด้าน การยืนยันตัวบุคคลด้วยเทคโนโลยีชีวมิติ (Biometric Authentication) เช่น การรู้จำใบหน้าของบุคคล นั่นคือ โมเดล Siamese Networks ซึ่งได้รับแรงบันดาลในจากแฝดสยามหรือ ฝาแฝดอินจันทร์ เพราะว่าโมเดลนี้เป็นการสร้างโมเดลคู่ขนาน หรือ 2 โมเดลที่เหมือนกันทุกประการ เพื่อเรียนรู้ในการจำแนกว่าวัตถุทั้งสองว่าเป็นสิ่งเดียวกันหรือแตกต่างกัน ดังแสดงในรูปที่ 3



รูปที่ 3 โครงสร้างของโมเดล Siamese Networks

โมเดลดังกล่าวเป็นโมเดลการเรียนรู้เชิงลึก เช่น โมเดล CNN สองโมเดล ดังแสดงในรูปที่ 3 กล่าว คือ โมเดล CNN1 และ CNN2 ซึ่งเป็นโมเดลคู่ขนานจะมีโครงสร้างเหมือนกันทุกประการ โดยในการฝึกฝนโมเดลจะทำการส่งคู่เสียงหรือเสียงพูด 2 เสียง ที่ประกอบด้วยคู่เสียงที่เป็นเสียงพูดคนเดียวกัน และคู่เสียงที่เป็นเสียงพูดคนละคนกัน เพื่อให้โมเดลเรียนรู้การเปรียบเทียบผ่าน Distance Metric ซึ่งใช้วัดค่าความคล้ายของข้อมูล 2 ชุด โดยผลลัพธ์คือคะแนนความคล้าย (Similarity Score) ซึ่งคะแนนยิ่งมากแสดงว่าข้อมูลดังกล่าวยิ่งมีความคล้ายกันมาก ในทางตรงกันข้ามหากข้อมูลที่เป็นข้อมูลชนิดต่างกันหรือเสียง 2 เสียงนี้เป็นคนละคนกันก็จะมีคะแนนความคล้ายน้อย ในขั้นตอนของการฝึกฝนโมเดล โมเดลจะทำการเรียนรู้จากคู่เสียงซ้ำ ๆ กัน ทั้งคู่เสียงจากคนเดียวกันและคู่เสียงจากคนละคนกัน ก็จะสามารถนำโมเดลดังกล่าวไปใช้แยกแยะได้ว่าเสียงที่ต้องการเปรียบเทียบ เช่น เสียงต้นแบบหรือเสียงของบุคคลที่ต้องการยืนยัน กับเสียงของบุคคลปริศนาเป็นคน ๆ เดียวกันหรือไม่

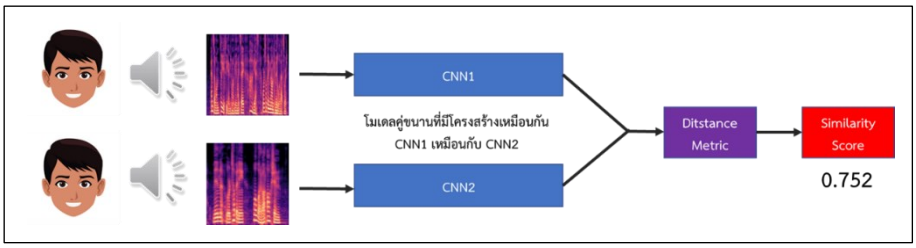
4.1.2 ชุดข้อมูลที่มีความเหมาะสม

คณะผู้วิจัยได้ทำการคัดเลือกข้อมูลไฟล์เสียงซึ่งรวบรวมมาจากชุดข้อมูลเสียงที่ใช้ในงานวิจัยในระดับนานาชาติ ได้แก่ ชุดข้อมูล LibriSpeech [9] ซึ่งประกอบไปด้วยไฟล์เสียงพูดประโยคภาษาอังกฤษหลาย ๆ ประโยคที่มีความยาวรวมกันประมาณ 1,000 ชั่วโมง โดยแต่ละไฟล์เสียงมีความยาวตั้งแต่ 5 ถึง 30 วินาทีต่อไฟล์เสียง ชุดข้อมูลดังกล่าวได้ทำการบันทึกเสียงจากผู้พูด 2,514 คนจากทั้งชายและหญิง โดยในการประยุกต์ใช้งานในงานวิจัยนี้ คณะผู้วิจัยใช้เพียงบางส่วนของเสียงในชุดข้อมูลนี้ เนื่องด้วยข้อจำกัดด้านพลังการประมวลผล สำหรับขั้นตอนการฝึกฝน (Training) ใช้ชุดข้อมูล *Train-Clean-100* ซึ่งมีความยาว 100 ชั่วโมง จากเสียงผู้พูดจำนวน 251 คน สำหรับการทดสอบ (Testing) ใช้ชุดข้อมูล *Test-Clean* ซึ่งมีความยาว 5.4 ชั่วโมงจากเสียงผู้พูดจำนวน 40 คน

4.1.3 การทดสอบโมเดลในเบื้องต้น

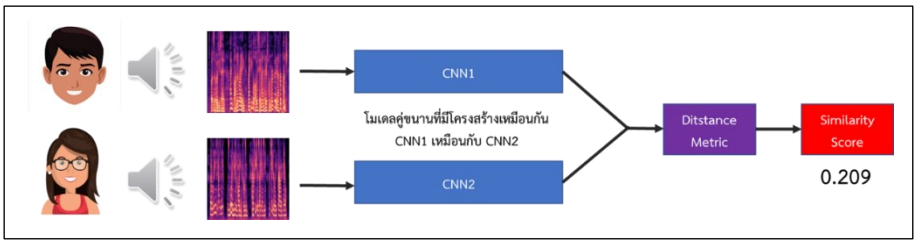
คณะผู้วิจัยประสบความสำเร็จในการฝึกฝนโมเดลในขั้นต้นโดยใช้โมเดล Siamese Networks กับชุดข้อมูล LibriSpeech หลังจากได้โมเดลที่ฝึกฝนในขั้นต้นแล้ว ขั้นตอนต่อไป คณะผู้วิจัยได้ทำการทดสอบโมเดลด้วยการทดสอบประสิทธิภาพของการแยกแยะผ่านการทดลองคู่ไฟล์เสียง ทั้งคู่เสียงของบุคคลคนเดียวกัน รวมถึงคู่เสียงของบุคคลคนละคน โดยการส่งคู่เสียงผ่านโมเดลคู่ขนาน และวัดค่าความคล้ายของเสียงทั้งสองเสียงดังกล่าว โดยในกรณีแรก เป็นการทดสอบกรณีที่ไฟล์เสียงทั้ง 2 เป็นเสียงจากบุคคลเดียวกัน ดังแสดงในรูปที่ 4 ซึ่งเป็นการทดลองของคู่เสียงที่เป็นเสียงของบุคคลเดียวกัน

โดยก่อนทำการส่งไฟล์เสียงเข้าสู่โมเดล ไฟล์เสียงจะถูกแปลงเป็น Spectrogram ซึ่งเป็นการแปลงค่าเสียงเป็นแผนภาพที่แสดงความถี่และความเข้มข้องเสียงก่อน และทำการส่งแผนภาพนี้ต่อเข้าไปยังโมเดลเพื่อวัดค่าความคล้าย จะเห็นว่าคะแนนค่าความคล้ายอยู่ที่ 0.752 ซึ่งถือว่าสูง (คะแนน 1.0 แสดงว่าเหมือนกันทุกประการ) ซึ่งผลคะแนนดังกล่าวสอดคล้องกับข้อเท็จจริงที่ว่า คู่เสียงดังกล่าวเป็นเสียงจากบุคคลเดียวกัน ดังนั้น โมเดลจึงประสบความสำเร็จในการแยกแยะเพื่อยืนยันว่าเป็นบุคคลคนเดียวกัน



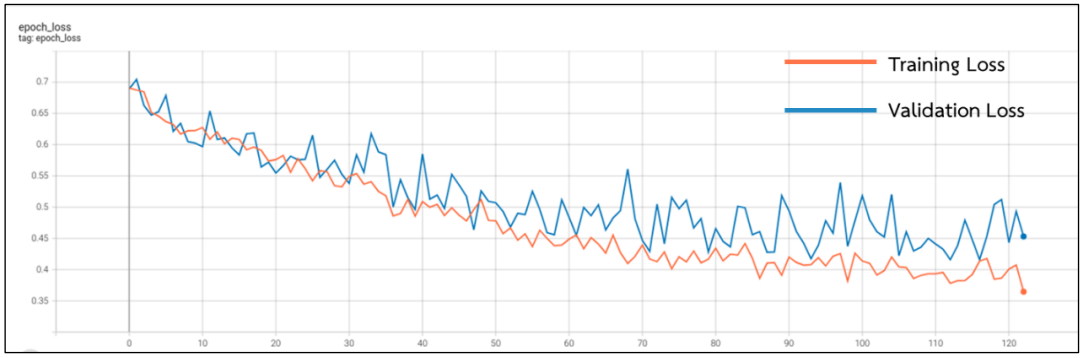
รูปที่ 4 การทดสอบโมเดลขั้นต้น (กรณีเป็นเสียงจากบุคคลเดียวกัน)

โดยในกรณีถัดไป เป็นการทดสอบกรณีที่ไฟล์เสียงทั้ง 2 เป็นเสียงจากบุคคลคนละคนกันดังแสดงในรูปที่ 5 จากนั้นทำการส่งแผนภาพ Spectrograms ของแต่ละเสียงต่อเข้าไปยังโมเดล เพื่อวัดค่าความคล้าย จะเห็นว่าคะแนนค่าความคล้ายอยู่ที่ 0.209 ซึ่งถือว่าต่ำ ซึ่งผลคะแนนดังกล่าวสอดคล้องกับข้อเท็จจริงที่ว่า คู่เสียงนี้เป็นเสียงจากบุคคลคนละกัน ดังนั้น โมเดลจึงประสบความสำเร็จในการแยกแยะเพื่อยืนยันว่าเป็นบุคคลคนละคนกัน



รูปที่ 5 การทดสอบโมเดลขั้นต้น (กรณีเป็นเสียงจากบุคคลคนละคนกัน)

จากผลในขั้นต้นที่บ่งชี้ว่า โมเดลมีศักยภาพในการแยกแยะเสียงของบุคคลดังนั้น คณะผู้วิจัยจึงได้ทำการฝึกฝนโมเดลเดิม ให้มีจำนวนรอบการฝึกฝนที่มากขึ้น โดยในขั้นต้นทำการฝึกฝนไปเพียง 50 รอบของการฝึกฝน แต่ในครั้งนี้ได้ดำเนินการฝึกฝนโมเดลจำนวน 125 รอบของการฝึกฝนและใช้ไฟล์คู่เสียงจำนวน 150,000 คู่เสียง โดยสาเหตุที่หยุดการฝึกฝนไว้ที่ 125 รอบ เนื่องจากกลไก Early Stopping หรือกลไกที่ตรวจสอบว่า โมเดลยังสามารถพัฒนาประสิทธิภาพได้อีกหรือไม่ โดยกลไกได้บ่งชี้ว่าโมเดลถึงจุดที่ประสิทธิภาพโมเดลทำได้ดีที่สุดแล้ว ดังจะเห็นได้จาก ณ หลังจากรอบการฝึกฝนที่ 90 อัตราการลดลงของค่า Loss Value เริ่มคงที่ จนกระทั่งที่รอบที่ 125 กลไกที่ค่าคงที่ไม่มีการลดลงแล้วดังแสดงในรูปที่ 6 ดังนั้น โมเดลจึงหยุดการฝึกฝน โดยใช้เวลาในการฝึกฝนโมเดลทั้งสิ้นประมาณ 18 ชั่วโมง และใช้โมเดลดังกล่าวเป็นโมเดลหลักในการทำการทดลองตลอดงานวิจัยนี้



รูปที่ 6 กราฟ Loss Value ระหว่างการฝึกฝนโมเดลในขั้นสุดท้าย

4.2 การเก็บข้อมูลเสียงเพื่อทำการทดสอบ

ในการศึกษาปัจจัยที่มีผลกระทบต่อประสิทธิภาพของโมเดลเมื่อนำไปใช้งานจริง คณะผู้วิจัยได้ทำเก็บรวบรวมไฟล์เสียงภาษาไทยจากบุคคล 10 คน โดยใช้เสียงจากกลุ่มตัวอย่างจากนักเรียนนายเรืออากาศ ตามสถานการณ์และรูปแบบการทดสอบในแต่ละประเภท คนละ 50 เสียง รวมทั้งสิ้น 500 เสียง

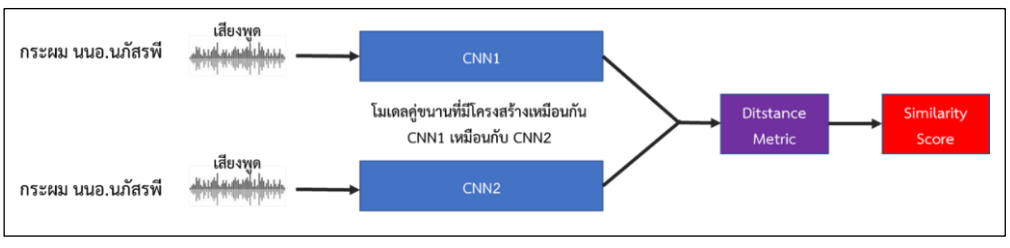
4.3 การทดสอบประสิทธิภาพของโมเดลกับภาษาไทย

การเก็บรวบรวมไฟล์เสียงกว่า 500 ไฟล์เสียง ก็เพื่อนำมาใช้ในการทดสอบประสิทธิภาพของโมเดลบนพื้นฐานสถานการณ์ที่แตกต่างกันไป โดยคณะผู้วิจัยได้กำหนดสถานการณ์ไว้ทั้งหมดจำนวน 6 การทดลอง โดยมีรายละเอียดดังนี้

4.3.1 การทดสอบในภาพรวม (ไม่มีสิ่งรบกวน)

ในการทดสอบแรกเป็นการทดสอบประสิทธิภาพของโมเดลในภาพรวม โดยค่าที่ใช้ในการวัดประสิทธิภาพโมเดลคือ ค่าความถูกต้อง (Accuracy) ซึ่งเป็นการวัดความสามารถของโมเดลในการจำแนกเสียงว่าสามารถจำแนกคู่เสียงที่เป็นต้นแบบ และเสียงของบุคคลที่เข้าทำการยืนยันตัวบุคคลได้ถูกต้องมากน้อยเพียงใด โดยในการวัดประสิทธิภาพดังกล่าว จะใช้การวัดค่าร้อยละหรือเปอร์เซ็นต์ความถูกต้องของจำนวนเสียงที่แยกแยะได้ถูกต้อง จากจำนวนคู่เสียงที่ใช้ในการทดลองทั้งหมด

โดยในการทดสอบ คณะผู้วิจัยได้ทำการสุ่มคู่เสียงทั้งคู่เสียงที่เป็นคนคนเดียวกัน และคู่เสียงของคนละคนกัน เพื่อวัดว่าโมเดลสามารถทำนายได้ถูกต้องจำนวนกี่ครั้ง โดยเสียงที่ใช้ในการบันทึกนั้นไม่มีสัญญาณรบกวน และคู่เสียงใช้ประโยคเดียวกันในกรณีที่เป็นคนคนเดียวกัน ยกตัวอย่าง เช่น หาก นักเรียนนายเรืออากาศ นภัสรพี อู่วิเชียร ทำการบันทึกเสียง เพื่อลงทะเบียนในระบบยืนยันตัวบุคคลว่า *"กระผม นักเรียนนายเรืออากาศ นภัสรพี อู่วิเชียร ครับ"* เพื่อเป็นเสียงต้นแบบในการยืนยันตนเอง เมื่อ นักเรียนนายเรืออากาศ นภัสรพี ฯ ต้องการยืนยันตัวเองก็ต้องพูดว่า *"กระผม นักเรียนนายเรืออากาศ นภัสรพี อู่วิเชียร ครับ"* เหมือนกัน ดังนั้นในกรณีนี้ ผู้ใช้งานต้องจำประโยคที่เป็นเหมือนรหัสผ่านให้ได้ ดังแสดงในรูปที่ 7



รูปที่ 7 โครงสร้างโมเดลสำหรับการทดสอบในภาพรวม (ไม่มีสิ่งรบกวน)

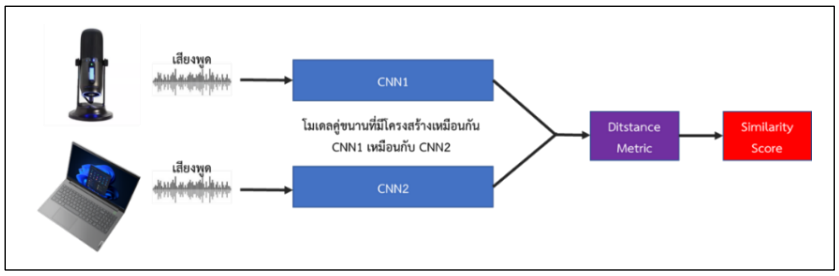
4.3.2 การทดสอบเมื่อคู่เสียงมีประโยคไม่เหมือนกัน

การทดสอบมีรูปแบบการทดลองที่คล้ายกับการทดลองแรก แต่มีข้อแตกต่างที่สำคัญก็คือ ทำการทดสอบโดยมีการเพิ่มเงื่อนไข ข้อจำกัดและความยากให้กับโมเดล โดยในการทดสอบนี้ กำหนดให้คู่เสียงจะใช้ประโยคต้นแบบและประโยคที่ใช้ในการยืนยันตัวตนไม่เหมือนกัน ดังนั้นในกรณีนี้ ผู้ใช้งานไม่จำเป็นต้องจำประโยคที่เป็นเหมือนรหัสผ่าน ซึ่งจะมีความง่ายมากขึ้นสำหรับผู้ใช้งาน ดังแสดงในรูปที่ 10

รูปที่ 8 โครงสร้างโมเดลสำหรับการทดสอบเมื่อคู่เสียงมีประโยคไม่เหมือนกัน

4.3.3 การทดสอบเมื่อมีสิ่งรบกวน (ไมโครโฟน)

ในการทดสอบนี้เป็นการศึกษาปัจจัยที่เกี่ยวข้องกับการประยุกต์ใช้งานจริง ในแง่ของอุปกรณ์ที่ใช้ในการบันทึกเสียง โดยในการทดสอบในข้อ 4.3.1 และ 4.3.2 เสียงทั้งหมด ทั้งเสียงที่ใช้ในการบันทึกเสียงไฟล์ต้นแบบและบันทึกเสียงสำหรับการยืนยันตัวบุคคลถูกบันทึกจากไมโครโฟนรุ่นเดียวกัน จึงเกิดคำถามทางการวิจัยขึ้นว่า ในการประยุกต์ใช้งานระบบการยืนยันตัวบุคคลด้วยเสียงในสถานการณ์จริง หากผู้ใช้งานบันทึกเสียงโดยใช้ไมโครโฟนคนละชนิดกัน จะส่งผลต่อประสิทธิภาพของโมเดลอย่างไร โดยในการทดลองนี้ ไฟล์คู่เสียงจะถูกบันทึกจากไมโครโฟนคนละชนิดกัน ดังแสดงในรูปที่ 9



รูปที่ 9 การทดสอบเมื่อมีสิ่งรบกวน (ไมโครโฟน)

4.3.4 การทดสอบเมื่อมีสิ่งรบกวน (เสียงจากสิ่งแวดล้อม)

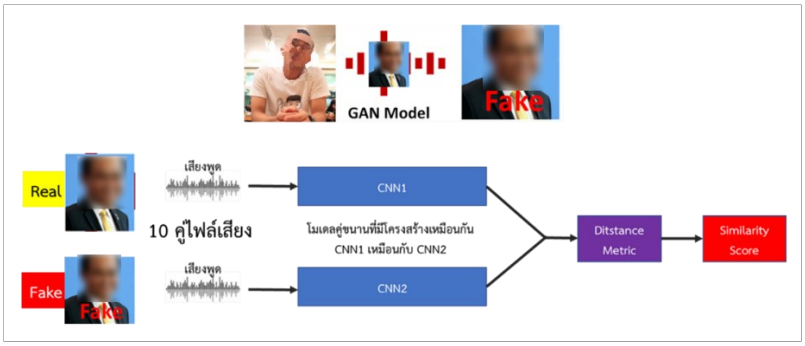
ในการทดสอบนี้ เป็นการทดสอบเพื่อศึกษาปัจจัยที่เกี่ยวข้องกับเสียงรบกวนที่มาจากสภาพแวดล้อม เช่น เสียงจากการก่อสร้างที่อาจเข้ามารบกวนในระหว่างที่มีการยืนยันตัวบุคคลด้วยเสียง โดยในการทดลองนี้ ไฟล์คู่เสียงนั้นเสียงหนึ่งซึ่งเป็นเสียงที่ผู้ใช้งานได้ลงทะเบียนไว้ สำหรับใช้เป็นไฟล์เสียงต้นแบบ จะถูกบันทึกในสภาพแวดล้อมที่ถูกควบคุม คือ ไม่มีเสียงรบกวน และอีกเสียงหนึ่งที่ใช้ในการเปรียบเทียบ จะเป็นการบันทึกเสียงพร้อมกับเปิดไฟล์เสียงที่ได้รับการรบกวนจากสิ่งแวดล้อม โดยในกรณีนี้ เป็นเสียงจากการก่อสร้างที่คณะผู้วิจัยได้จำลองสถานการณ์โดยเปิดไฟล์เสียงก่อสร้างจากแอปพลิเคชัน YouTube [10] คลอไปด้วยระหว่างบันทึกเสียงซึ่งถือเป็นการจำลองสถานการณ์ว่า ในขณะที่ทำการยืนยันตัวมีเสียงรบกวนจากการก่อสร้าง

4.3.5 การทดสอบเมื่อมีสิ่งรบกวน (การใส่หน้ากากอนามัย)

ในการทดสอบนี้ เป็นการทดสอบเพื่อศึกษาปัจจัยที่เกี่ยวข้องกับเสียงรบกวนที่มาจากสภาพแวดล้อมเช่นเดียวกัน โดยจำลองสถานการณ์ในการยืนยันตัวบุคคลในห้วงการแพร่ระบาดของโรค COVID - 19 ที่ผู้ยืนยันตัวบุคคลต้องใส่หน้ากากอนามัย โดยในการทดลองนี้ กำหนดให้ไฟล์คู่เสียง ประกอบไปด้วยเสียงที่ผู้ใช้งานได้ลงทะเบียนไว้ก่อนแล้วซึ่งยังไม่มีการแพร่ระบาดของโรค COVID - 19 เพื่อเก็บเอาไว้เป็นไฟล์เสียงต้นแบบ โดยถูกบันทึกในสภาพแวดล้อมที่มีการควบคุม คือ ไม่มีการใส่หน้ากากอนามัย และอีกเสียงหนึ่งที่ใช้ในการเปรียบเทียบ ซึ่งจะเป็นเสียงที่ผู้ใช้ต้องบันทึกเพื่อยืนยันตัวบุคคล โดยเสียงนี้จะเป็นการบันทึกเสียงจากผู้ที่ใส่หน้ากากอนามัย

4.3.6 การทดสอบต่อเสียงปลอมแปลง

ในการทดสอบนี้ เป็นการทดสอบ เพื่อจำลองสถานการณ์ว่า หากมีคนใช้เสียงปลอมแปลง เพื่อพยายามหลอกโมเดล โดยใช้การปลอมแปลงเสียงพูดของบุคคลด้วยเทคโนโลยี Deep Voice ผลการทดลองบ่งชี้ว่า มนุษย์เมื่อฟังเสียงแปลงแล้ว มีผู้ถูกหลอก 3 ใน 10 คน เข้าใจผิดว่าเสียงปลอมแปลงดังกล่าวนั้น คือเสียงจริง จึงเกิดคำถามงานวิจัยว่า เสียงของมนุษย์ที่ถูกปลอมแปลงด้วยเทคโนโลยี Deep Voice จะมีศักยภาพในการหลอกโมเดลคอมพิวเตอร์สำหรับการยืนยันตัวบุคคลได้มากน้อยเพียงใด โดยในการทดลองนี้ คณะผู้วิจัยได้ใช้ไฟล์เสียงที่ถูกปลอมแปลงและไฟล์จริงของบุคคลจริงจากชุดข้อมูลของงานวิจัย Deep Voice ในปีที่ผ่านมา โดยไฟล์คู่เสียงนั้นประกอบไปด้วย เสียงหนึ่งซึ่งเป็นเสียงจริงที่ผู้ใช้งาน (ในกรณีนี้คือ บุคคลสำคัญ) ได้ลงทะเบียนไว้ล่วงหน้า และอีกเสียงหนึ่งที่ใช้ในการเปรียบเทียบซึ่งจะเป็นเสียงที่ถูกปลอมแปลงขึ้นเป็นเสียงของบุคคลสำคัญ ดังแสดงในรูปที่ 10



รูปที่ 10 การทดสอบต่อเสียงปลอมแปลง

4.4 การนำเสนอแนวทางการพัฒนาระบบสำหรับกองทัพอากาศไทย

หลังจากดำเนินการทดลองเสร็จสิ้นทั้ง 6 การทดลองข้างต้น คณะผู้วิจัยได้ดำเนินการสรุปผลการศึกษา เพื่อจัดทำ General Implementation Framework ซึ่งเป็นการระบุขั้นตอน ข้อเสนอแนะ ข้อจำกัด รวมถึงข้อพิจารณาในการพัฒนาแพลตฟอร์มการยืนยันตัวบุคคลด้วยเทคโนโลยีชีวมิติที่มีความมั่นคงปลอดภัยสำหรับกองทัพอากาศไทยต่อไป

5. ผลการวิจัย

5.1 การทดสอบในภาพรวม (กรณีไม่มีสิ่งรบกวน)

ผลการทดสอบจากคู่เสียงที่ทำการสุ่มตัวอย่างจำนวนกว่า 1,000 ตัวอย่างคู่เสียง ในค่าขีดแบ่ง (Threshold) ที่แตกต่างกัน โดยค่าดังกล่าวเป็นตัวเลขที่กำหนดว่า หากค่า Similarity Score ที่ได้จากการคำนวนความคล้ายของคู่เสียงตัวอย่างมีค่ามากกว่าหรือเท่ากับค่าขีดแบ่ง ตัวโมเดลจะทำนายว่า คู่เสียงดังกล่าวเป็นคนเดียวกัน แต่หากต่ำกว่าค่าขีดแบ่ง

โมเดลจะทำนายว่าคู่เสียงดังกล่าวเป็นคนละคนกัน เช่น สมมติว่า ใช้ค่าขีดแบ่งที่ 0.9 นั่น หมายความว่า โมเดลดังกล่าวจะทำนายคู่เสียงดังกล่าวจะเป็นบุคคลคนเดียวกันก็ต่อเมื่อค่าความคล้ายมีค่ามากกว่าหรือเท่ากับ 0.9 เท่านั้น จากนั้นจะใช้ผลการทำนายตัวอย่างทั้ง 1,000 คู่เสียงมาเทียบกับข้อเท็จจริง เพื่อคำนวณหาค่าความถูกต้อง โดยจากผลการทดลองพบว่า ประสิทธิภาพของโมเดลในภาพรวม มีค่าความถูกต้องหรือ Accuracy อยู่ที่ 75.88 % ณ ค่าขีดแบ่งที่ 0.9 ดังแสดงในตารางที่ 1

ตารางที่ 1 ค่าความถูกต้องของโมเดลในการทดสอบในภาพรวม (ไม่มีสิ่งรบกวน)

ค่าขีดแบ่ง (Threshold)	**0.7**	**0.8**	**0.9**	**0.95**
ค่าความถูกต้อง (Accuracy)	56.11	64.22	75.88	62.00

จากผลการทดสอบในแง่ของประสิทธิภาพพบว่า โมเดลมีค่าความแม่นยำอยู่ในระดับค่อนข้างดี โดยมีค่า Accuracy อยู่ที่ประมาณ 75% โดยค่า Threshold ที่มีความเหมาะสมในกรณีนี้คือ 0.9 และสิ่งหนึ่งที่น่าสนใจคือ การที่โมเดลฝึกฝนโดยใช้ภาษาอังกฤษ แต่ทดสอบกับเสียงที่เป็นภาษาไทย ซึ่งจากค่าความถูกต้องดังกล่าวพบว่า ภาษาไทยมีผลต่อประสิทธิภาพการจำแนก ดังนั้น ในการประยุกต์ใช้ในอนาคต ควรมีการใช้ไฟล์เสียงภาษาไทยในการฝึกฝนโมเดล นอกจากนี้ระดับค่า Accuracy ที่โมเดลนี้ทำได้ที่ประมาณ 75% ยังอยู่ในระดับที่ไม่สามารถใช้งานเพื่อการยืนยันตัวบุคคลด้วยเสียงได้อย่างสมบูรณ์ ซึ่งควรต้องมีการพัฒนาต่อไป

5.2 การทดสอบเมื่อคู่เสียงมีประโยคไม่เหมือนกัน

ผลการทดสอบพบค่าความถูกต้องของโมเดลลดลงอย่างมีนัยยะสำคัญเมื่อเทียบกับการทดสอบในข้อ 5.1 โดยเฉพาะกับโมเดลที่ใช้ค่า Threshold ที่ 0.9 โดยเมื่อเปรียบเทียบผลของการหาค่าความถูกต้องจะพบว่าค่าความถูกต้องลดลงจาก 75.88 % เหลือเพียง 59.67 % นอกจากนี้ ค่าความถูกต้องของค่าขีดแบ่งอื่น ๆ ก็ลดลงเช่นเดียวกัน ดังผลในตารางที่ 2

ตารางที่ 2 ค่าความถูกต้องของโมเดลในการทดสอบเมื่อคู่เสียงมีประโยคไม่เหมือนกัน

ค่าขีดแบ่ง (Threshold)	**0.7**	**0.8**	**0.9**	**0.95**
ค่าความถูกต้อง (Accuracy)	56.00	61.44	59.67	51.56

จากผลการทดสอบแสดงให้เห็นว่า การใช้รูปประโยคที่แตกต่างกันในการยืนยันตัวบุคคลมีผลโดยตรงต่อประสิทธิภาพโมเดล ดังนั้น คณะผู้วิจัยจึงเสนอแนะว่า หากต้องประยุกต์ใช้งานจริงในระบบการยืนยันตัวบุคคลด้วยเสียงควรใช้ไฟล์คู่เสียงที่เป็นรูปประโยคเดียวกัน ทั้งนี้ ผู้ใช้จำเป็นต้องจำว่าตนประโยคยืนยันตัวบุคคลของตนเองเป็นอะไร ซึ่งถือเป็น Trade Off หรือการแลกเปลี่ยนที่ผู้พัฒนาระบบต้องเลือกระหว่างประสิทธิภาพของโมเดล หรือความง่ายต่อการใช้งาน

5.3 การทดสอบเมื่อมีสิ่งรบกวน (ไมโครโฟน)

ผลการทดสอบพบว่า ค่าความถูกต้องของโมเดลนั้น ลดลงอย่างมีนัยยะสำคัญ โดยเฉพาะกับโมเดลที่ใช้ค่า Threshold ที่ 0.9 โดยเมื่อเปรียบเทียบผลของการหาค่าความถูกต้องในการทดสอบแรกในข้อ 5.1 จะพบว่า ค่าความถูกต้องที่ลดลงจาก 75.88 % เหลือเพียง 69.67 % นอกจากนี้ ค่าความถูกต้องของค่าขีดแบ่งอื่น ๆ ก็ลดลงเช่นเดียวกัน ในอัตรา 1-5 % อย่างไรก็ตาม อัตราการลดลงของประสิทธิภาพโมเดลยังน้อยกว่าการทดลองในข้อ 5.2 ซึ่งใช้รูปประโยคข้อความที่แตกต่างกัน ดังแสดงในตารางที่ 3

ตารางที่ 3 การทดสอบเมื่อมีสิ่งรบกวน (ไมโครโฟน)

ค่าขีดแบ่ง (Threshold)	0.7	0.8	0.9	0.95
ค่าความถูกต้อง (Accuracy)	50.77	60.62	69.23	61.54

จากผลการทดสอบแสดงให้เห็นว่า การใช้ประเภทของไมโครโฟนในการบันทึกเสียงส่งผลโดยตรงต่อประสิทธิภาพ ดังนั้น ในการประยุกต์ใช้งานจริงของระบบการยืนยันตัวบุคคลด้วยเสียง ควรใช้ไฟล์คู่เสียงที่ถูกบันทึกจากไมโครโฟนชนิดเดียวกัน

5.4 การทดสอบเมื่อมีสิ่งรบกวน (เสียงจากสิ่งแวดล้อม)

ผลการทดสอบพบค่าความถูกต้องของโมเดลลดลงอย่างมีนัยยะสำคัญเช่นเดียวกับการทดลองก่อนหน้านี้ โดยเฉพาะกับโมเดลที่ใช้ค่า Threshold ที่ 0.7 โดยเมื่อเปรียบเทียบผลของการหาค่าความถูกต้องในการทดสอบแรก จะพบว่า ค่าความถูกต้องที่ลดลงจาก 56.11 % เหลือเพียง 49.85 % อยู่ในระดับที่โมเดลใช้งานไม่ได้ แต่หากพิจารณา ค่า Threshold 0.9 ซึ่งเป็นค่า Threshold ที่มีความเหมาะสมจากการทดสอบแรกพบว่า ลดลงเพียงเล็กน้อยในระดับประมาณ 2% เท่านั้น ดังตารางที่ 4

ตารางที่ 4 การทดสอบเมื่อมีสิ่งรบกวน (เสียงจากสิ่งแวดล้อม)

ค่าขีดแบ่ง (Threshold)	0.7	0.8	0.9	0.95
ค่าความถูกต้อง (Accuracy)	49.85	61.54	72.62	58.15

จากผลการทดสอบในแง่ของประสิทธิภาพจะเห็นว่า ค่าความถูกต้องส่วนใหญ่ลดลงไม่มาก เมื่อเทียบกับการทดสอบแรก ดังนั้น จึงสรุปได้ว่า เสียงรบกวนจากสภาพแวดล้อม เช่น เสียงจากการก่อสร้างที่ไม่ดังจนเกินไป ไม่ได้มีผลกระทบต่อประสิทธิภาพของโมเดล หากเลือกค่า Threshold ที่มีความเหมาะสม ดังนั้น ในการประยุกต์ใช้งานจริงในระบบการยืนยันตัวบุคคลด้วยเสียง จึงอาจไม่จำเป็นต้องติดตั้งอุปกรณ์บันทึกเสียงในพื้นที่ปิดเงียบสงัด แต่อย่างไรก็ตาม ยังจำเป็นต้องศึกษาเสียงรบกวนประเภทอื่น ๆ ประกอบกันในอนาคตต่อไป

5.5 การทดสอบเมื่อมีสิ่งรบกวน (การใส่หน้ากากอนามัย)

ผลการทดสอบพบว่า ค่าความถูกต้องของโมเดลลดลงอย่างมีนัยยะสำคัญเฉพาะกับโมเดลที่ใช้ค่า Threshold ที่ 0.7 และ 0.8 แต่ในค่า Threshold ที่ 0.9 ซึ่งเป็นค่ามาตรฐานที่ได้จากการทดลองครั้งที่ 1 และค่า Threshold ที่ 0.95 ค่าความถูกต้องไม่เปลี่ยนแปลงมากนัก แสดงดังตารางที่ 5

ตารางที่ 5 การทดสอบเมื่อมีสิ่งรบกวน (การใส่หน้ากากอนามัย)

ค่าขีดแบ่ง (Threshold)	0.7	0.8	0.9	0.95
ค่าความถูกต้อง (Accuracy)	42.46	56.61	75.69	61.85

จากผลการทดสอบในแง่ของประสิทธิภาพจะเห็นว่า ค่าความถูกต้องที่เป็นค่ามาตรฐานที่ค่า Threshold ที่ 0.9 ลดลงไม่มาก เมื่อเทียบกับการทดสอบแรก ดังนั้น จึงสรุปได้ว่าการใส่หน้ากากอนามัย (Mask) ไม่มีผลกระทบโดยตรงต่อประสิทธิภาพของโมเดล

5.6 การทดสอบต่อเสียงปลอมแปลง (Deep Voice)

ผลการทดสอบพบว่า ค่าความถูกต้องของโมเดลนั้น ลดลงอย่างมีนัยยะสำคัญ ดังจะเห็นได้จากค่า Threshold ที่ 0.7 และ 0.8 ค่าความถูกต้องอยู่ที่ 0 นั่นหมายความว่า โมเดลถูกหลอกลวงอย่างสมบูรณ์ และสำหรับค่า Threshold ระดับ 0.9 ซึ่งเป็นค่ามาตรฐาน ค่าความถูกต้องอยู่ที่ 30 % ซึ่งแปลความได้ว่า โมเดลจำแนกประเภทได้ถูกต้อง 3 ใน 10 เสียง และโมเดลถูกหลอกลวง 7 ใน 10 เสียง กล่าวคือ โมเดลเข้าใจว่า เสียงที่ปลอมแปลงของบุคคลสำคัญ คือ เสียงที่แท้จริงของบุคคลสำคัญนั้น แต่เป็นที่น่าสังเกตว่า ณ ค่า Threshold ที่ 0.95 ค่าความถูกต้องอยู่ที่ 80 % ซึ่งหมายความว่าโมเดลทำนายถูก 8 ใน 10 เสียง ซึ่งถือว่าอยู่ในระดับที่สูง แต่หากเลือกระดับค่า Threshold ดังกล่าวในการประยุกต์ใช้งานจริง เพื่อป้องกันเสียงปลอมแปลง ต้องแลกมาด้วยค่าความถูกต้องที่ลดลง เมื่อโมเดลต้องทำนายกับเสียงปกติ แสดงผลดังตารางที่ 6

ตารางที่ 6 การทดสอบต่อเสียงปลอมแปลง (Deep Voice)

ค่าขีดแบ่ง (Threshold)	**0.7**	**0.8**	**0.9**	**0.95**
ค่าความถูกต้อง (Accuracy)	42.46	56.61	75.69	61.85

จากผลการทดสอบจะเห็นว่า โมเดลการยืนยันบุคคลด้วยเสียงยังมีความเปราะบางอย่างมากกับเสียงแปลงด้วยเทคโนโลยี Deep Voice ดังนั้น การศึกษาแนวทางในการป้องกันการหลอกลวงโมเดล เป็นสิ่งที่มีความจำเป็น ก่อนที่จะนำโมเดลดังกล่าวไปใช้งานจริง โดยเฉพาะการใช้งานในหน่วยงานด้านความมั่นคง

6. สรุป

คณะผู้วิจัยได้ทำการวิเคราะห์ผลการวิจัย และสรุปผลการวิจัย เพื่อจัดทำแนวทางการประยุกต์ใช้งานทั่วไป (General Implementation Framework) ของเทคโนโลยีการรู้จำตัวบุคคลด้วยเสียงเพื่อใช้ในการยืนยันตัวบุคคลด้วยเทคโนโลยีชีวมิติที่มีความมั่นคงปลอดภัยสำหรับกองทัพอากาศไทย และมีรายละเอียดของการประยุกต์ใช้งานดังนี้

6.1 รูปแบบโมเดลที่มีความเหมาะสม

จากผลการวิจัย สรุปได้ว่า โมเดลที่มีความเหมาะสมต่อการใช้งานเพื่อยืนยันตัวบุคคลที่มีประสิทธิภาพสูง และรองรับต่อการเปลี่ยนแปลงกำลังพล ซึ่งต้องมีการเพิ่มหรือลดอยู่บ่อยครั้ง โดยไม่จำเป็นต้องทำการฝึกฝนโมเดลใหม่ คือ โมเดลประเภทการเรียนรู้เชิงลึกที่ใช้รูปแบบการเรียนรู้โมเดลแบบ One-Shot-Learning (OSL) ประเภท Siamese Model

6.2 ชุดข้อมูลเพื่อฝึกฝนโมเดล

ในการฝึกฝนโมเดลให้มีประสิทธิภาพมีความจำเป็นต้องใช้ชุดข้อมูลที่มีมาตรฐานในการฝึกฝนโมเดลสำหรับการเริ่มต้นชุดข้อมูลที่มีความเหมาะสมได้แก่ชุดข้อมูล LibriSpeech อย่างไรก็ตาม จากผลการวิจัยพบว่า เสียงภาษาไทยส่งผลต่อประสิทธิภาพของการยืนยันตัวบุคคลของโมเดล โดยเฉพาะอย่างยิ่งโมเดลที่ทำการฝึกฝน และทดสอบโดยใช้ภาษาคนละภาษากัน ดังนั้น การประยุกต์ใช้งานในอนาคตการฝึกฝนโมเดลให้มีประสิทธิภาพ ควรมีชุดข้อมูลภาษาไทยเพื่อการฝึกฝน โดยชุดข้อมูลดังกล่าวต้องมีขนาดที่ใหญ่เพียงพอในการฝึกฝนโมเดล

6.3 รูปแบบไฟล์เสียงเพื่อยืนยันตัวบุคคล

การใช้รูปแบบของไฟล์เสียงเพื่อยืนยันตัวบุคคลส่งผลกระทบต่อประสิทธิภาพของโมเดลอย่างมีนัยยะสำคัญ โดยพบว่า การใช้รูปประโยคที่แตกต่างกันระหว่างประโยคที่ใช้ในการลงทะเบียน และประโยคที่ใช้ในการยืนยันตัวบุคคลส่งผลโดยตรงทำให้ประสิทธิภาพของโมเดลในแง่ค่าความถูกต้อง (Accuracy) ลดลงอย่างมีนัยยะสำคัญ ดังนั้น ในการประยุกต์ใช้ในสถานการณ์จริง หากต้องการเพิ่มประสิทธิภาพของโมเดล จึงควรใช้ประโยคที่ในการยืนยันตัวบุคคลที่เป็นประโยคเดียวกัน

โดยผู้ใช้งานมีความจำเป็นต้องจดจำประโยคที่ตนเองลงทะเบียนเพื่อใช้ในการยืนยันตัวเองให้ถูกต้อง จึงอาจสร้างความภาระเพิ่มเติมให้แก่ผู้ใช้งาน แต่ก็เป็นสิ่งที่ต้องแลกเปลี่ยน (Trade Off) ระหว่างประสิทธิภาพและความง่ายของการใช้งาน

6.4 ปัจจัยสภาพแวดล้อม

จากผลการวิจัยสรุปได้ว่า ประสิทธิภาพของโมเดลได้รับผลกระทบในระดับที่แตกต่างกันออกไป ขึ้นอยู่กับสภาวะแวดล้อมต่าง ๆ โดยพบว่า เพื่อให้เกิดประสิทธิภาพสูงสุด และลดปัจจัยที่อาจกระทบเชิงลบต่อประสิทธิภาพของโมเดล ควรใช้อุปกรณ์บันทึกเสียงประเภทเดียวกันสำหรับการบันทึกเสียงที่ใช้ในการลงทะเบียนและการยืนยันตัวบุคคล รวมถึงการบันทึกเสียงอาจมีเสียงรบกวนได้บ้าง แต่ต้องไม่อยู่ในระดับที่ดังจนเกินไป และผู้ใช้งานสามารถสวมใส่หน้ากากอนามัยในการยืนยันตัวบุคคลได้

6.5 ปัจจัยภัยคุกคามด้านไซเบอร์

ผลการวิจัยพบว่า โมเดลมีความเปราะบางต่อเสียงที่ปลอมแปลงจากเทคโนโลยี Deep Voice โดยประสิทธิภาพของโมเดลลดลงอย่างมีนัยยะสำคัญ ทั้งนี้ ค่าความถูกต้องลดลงเหลือเพียง 30 % ดังนั้น แนวทางการแก้ไขจึงเป็นการตั้งค่าขีดแบ่งให้มีความเข้มงวดมากขึ้น เช่น ค่า Threshold ที่ 0.95 แต่ก็ต้องแลกมาด้วยอัตราค่า False Negative ที่สูงมาก จึงจำเป็นอย่างยิ่งที่จะต้องมีการวิจัยและพัฒนาในการพัฒนาโมเดลที่มีความทนทานต่อเสียงปลอมแปลงดังกล่าว

7. ข้อเสนอแนะ

ถึงแม้ว่าโมเดลจะมีค่าความแม่นยำอยู่ในระดับค่อนข้างดี โดยมีค่า Accuracy อยู่ที่ประมาณ 75 % โดยค่า Threshold ที่มีความเหมาะสมในกรณีนี้คือ 0.9 ซึ่งจากค่าความถูกต้องในระดับดังกล่าวยังอยู่ในระดับที่ไม่สามารถใช้งานเพื่อการยืนยันตัวบุคคลด้วยเสียงได้อย่างสมบูรณ์ จึงยังต้องมีการพัฒนาต่อไป อย่างไรก็ตามสามารถใช้การยืนยันตัวบุคคลด้วยเสียงควบคู่กับการยืนยันตัวบุคคลแบบหลายปัจจัยอื่น ๆ เช่น การยืนยันตัวบุคคลด้วยรหัสผ่าน เพื่อเป็นการเพิ่มความเชื่อมั่นในการยืนยันตัวบุคคล ดังนั้น การวิจัยและพัฒนาในเรื่องดังกล่าวจึงมีความจำเป็นต่อไปในอนาคต

นอกจากนี้ ควรมีการศึกษาวิจัยต่อไปเกี่ยวกับการเปรียบเทียบยืนยันตัวตนระหว่างการยืนยันตัวตนด้วยลายนิ้วมือและการยืนยันตัวตนด้วยเสียง เนื่องจากปัจจุบันลายนิ้วมือถูกใช้อย่างแพร่หลายจนเป็นที่ยอมรับ หากมีการพัฒนาเชิงเปรียบเทียบกับลายนิ้วมือ และหากผลลัพธ์การยืนยันตัวตนด้วยเสียงใกล้เคียงกับลายนิ้วมือ จะทำให้การยืนยันด้วยเสียงได้รับความน่าเชื่อถือมากขึ้น

8. กิตติกรรมประกาศ

งานวิจัยนี้สำเร็จลุล่วงได้ด้วยดี คณะผู้วิจัยต้องขอขอบคุณ ภาควิชาคอมพิวเตอร์ กองวิชาคณิตศาสตร์และคอมพิวเตอร์ กองการศึกษา โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช กองทัพอากาศที่ให้การสนับสนุนในการดำเนินการวิจัยจนงานวิจัยนี้สำเร็จลุล่วงไปได้ด้วยดี

9. เอกสารอ้างอิง

[1] กองทัพอากาศ. (2563). ยุทธศาสตร์กองทัพอากาศ 20 ปี (ฉบับปรับปรุง พ.ศ.2563). สืบค้น 12 มีนาคม 2565, จาก www.rtaf.mi.th/th/Documents/Publication/RTAF%20Strategy_Final_04122563.pdf

[2] Manikandan, K., & Chandra, E. (2016). Speaker Identification using a Novel Prosody with Fuzzy based Hierarchical Decision Tree Approach. *Indian Journal of Science and Technology*. 9: 44.

[3] Dhakal, P., Damacharla, P., Javaid, A. Y., & Devabhaktuni, V. (2019). A Near Real-Time Automatic Speaker Recognition Architecture for Voice-Based User Interface. *Machine Learning and Knowledge Extraction*. 1: 504 - 520.

[4] The Math Works, Inc. (1994). Speaker Identification Using Pitch and MFCC. Retrieved on May 5, 2022, from https://www.mathworks.com/help/audio/ug/speaker-identification-using-pitch-and-mfcc.html

[5] Zaharchuk, G., Gong, E., Wintermark, M., et al. (2018). Deep learning in neuroradiology. *AJNR Am J Neuroradiol.* 39: 1776 - 1784.

[6] Tirumala, S.S., Shahamiri, S. R., Garhwal, A. S., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications*. 90: 250 - 271.

[7] พายัพ ศิรินาม. (2562). การพัฒนาโมเดลการเลียนเสียงเชิงลึกสำหรับงานด้านสงครามไซเบอร์ (รายงานผลการวิจัย). กรุงเทพฯ: โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช.

[8] Adrian Rosebrock. (2020). GANs with Keras. Retrieved on May 12, 2022, from https://pyimagesearch.com/2020/11/16/gans-with-keras-and-tensorflow/

[9] Vassil Panayotov. (2015). LibriSpeech. Retrieved on June 6, 2022, from https://paperswithcode.com/dataset/librispeech

[10] Ambient Sound TH. (2565). เสียงคนงานกำลังทำงานก่อสร้าง. Retrieved on June 6, 2022, from https://www.youtube.com/watch?v=3mpmmCqx9fo&ab_channel=AmbientSound%E2%80%8BTH