KKU SCIENCE JOURNAL
Journal Home Page : https://ph01.tci-thaijo.org/index.php/KKUSciJ
Published by the Faculty of Science, Khon Kaen University, Thailand

การพัฒนาแบบจำลองการตรวจจับเสียงปลอมแปลงของบุคคลสาธารณะเพื่อบรรเทาภัยคุกคามจากข้อมูลบิดเบือนบนมิติไซเบอร์

Development of a Fake Voice Detection Model of Public Figures to Mitigate Disinformation in a Cyber Domain

อานนท์ บางเสน[1*] และ พายัพ ศิรินาม[2]

Anon Bangsan[1*] and Payap Sirinam[2]

[1]สำนักบัณฑิตศึกษา โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช จังหวัดสระบุรี 18180

[2]ภาควิชาคอมพิวเตอร์ โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช จังหวัดสระบุรี 18180

[1]The Graduate School of Navaminda Kasatriyadhiraj Royal Air Force Academy, Navaminda Kasatriyadhiraj Royal Air Force Academy, Saraburi, 18180, Thailand

[2]Department of Computer Science, Navaminda Kasatriyadhiraj Royal Air Force Academy, Saraburi, 18180, Thailand

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อ 1) ศึกษาแนวทางการตรวจจับเสียงปลอมแปลงด้วยเทคโนโลยีปัญญาประดิษฐ์ 2) พัฒนาแบบจำลองการตรวจจับเสียงปลอมแปลงของบุคคลสาธารณะที่มีประสิทธิภาพ และ 3) นำเสนอแนวทางการประยุกต์ใช้งานในการปฏิบัติการทางไซเบอร์เชิงรับ กระบวนการวิจัยประกอบด้วยการเตรียมข้อมูลเสียงที่ครอบคลุมเทคโนโลยีการปลอมแปลงหลายประเภท เช่น การแปลงเสียง การโคลนเสียง และการแปลงข้อความเป็นเสียง พร้อมทั้งการสกัดคุณลักษณะเสียงด้วยเมลสเปกโตรแกรม (Mel-spectrogram) จากนั้นพัฒนาและต่อยอดสถาปัตยกรรมเชิงลึก 3 แบบ ได้แก่ Mask CNN-4 Mask SE-ResNet-34 และ Mask SE-ResNeXt-152 เพื่อสร้างแบบจำลองเฉพาะด้านสำหรับการแยกแยะเสียงจริงและเสียงปลอมของบุคคลสาธารณะ ผลการประเมินด้วยชุดทดสอบและชุดข้อมูลใหม่ที่แบบจำลองไม่เคยเห็นมาก่อน พบว่าแบบจำลองทั้งหมดสามารถจำแนกได้อย่างแม่นยำ โดยเฉพาะ Mask SE-ResNeXt-152 ซึ่งให้ประสิทธิภาพสูงสุดที่ความแม่นยำ (accuracy) 98.33% นอกจากนี้ยังได้พัฒนาเว็บแอปพลิเคชันเพื่อประยุกต์ใช้งานจริง โดยสามารถแสดงผลการพยากรณ์ ค่าความเชื่อมั่น ค่าระยะห่างเสียงเชิงลึกแบบเคอร์เนล และการถอดเสียงพูดเป็นข้อความ เพื่อสนับสนุนการตัดสินใจเชิงปฏิบัติการทางไซเบอร์เชิงรับ การทดสอบด้วยชุดข้อมูล NKRAFA Thai ยังยืนยันว่าแบบจำลองมีประสิทธิภาพสูงกว่าการฟังของมนุษย์ โดยอัตราการตัดสินผิดสองกรณีหลัก (เสียงปลอมถูกตัดสินว่าเป็นเสียงจริงและเสียงจริงถูกตัดสินว่าเป็นเสียงปลอม) ลดลงจาก 56% และ 41% เหลือศูนย์ สะท้อนศักยภาพของแบบจำลองในการบรรเทาภัยคุกคามจากข้อมูลบิดเบือนในมิติไซเบอร์

*Corresponding Author, E-mail: boy_anonbangsan@hotmail.com

Received date: 6 June 2025 | Revised date: 19 September 2025 | Accepted date: 26 September 2025

doi: 10.14456/kkuscij.2026.5

ABSTRACT

This research aims to 1) investigate approaches for detecting voice spoofing using artificial intelligence (AI), 2) develop an effective model for identifying synthetic voices of public figures, and 3) propose an application framework for defensive cyber operations. The research process included preparing a speech dataset covering multiple spoofing technologies such as Voice Conversion, Voice Cloning, and Text-to-Speech synthesis, with Mel-spectrograms employed for feature extraction. Three deep architectures, Mask CNN-4, Mask SE-ResNet-34, and Mask SE-ResNeXt-152, originally derived from image recognition models, were further extended and customized to build specialized models for distinguishing between genuine and synthetic voices of public figures. Experimental results with testing and evaluation sets, including previously unseen data, showed that all models achieved accurate classification, with Mask SE-ResNeXt-152 yielding the highest performance at 98.33% accuracy under optimized parameters. Furthermore, the model was deployed in a web-based application that presents predictions, confidence scores, Kernel Deep Speech Distance (KDSD), and transcription outputs to support decision-making in cyber defense operations. Real-world testing using the NKRAFA Thai dataset confirmed that the model outperformed human listeners in identifying voice spoofing, reducing the misclassification rates of false acceptance and false rejection from 56% and 41% respectively to zero. These findings demonstrate the strong potential of the proposed model to mitigate disinformation threats in the cyber domain.

คำสำคัญ: ปัญญาประดิษฐ์เชิงกำเนิด โครงข่ายประสาทเทียมแบบคอนโวลูชัน การตรวจจับเสียงปลอมแปลง เสียงสังเคราะห์ ความมั่นคงปลอดภัยไซเบอร์

Keywords: Generative Artificial Intelligence, Convolutional Neural Networks, Fake Voice Detection, Voice Synthesis, Cybersecurity

บทนำ

ปัญญาประดิษฐ์เชิงกำเนิด (generative artificial intelligence) ได้แสดงศักยภาพในการสร้างเนื้อหาที่เสมือนจริงอย่างน่าทึ่ง ไม่ว่าจะเป็นภาพ เสียง หรือวิดีโอ (Jovanović and Campbell, 2022) ความสามารถนี้ได้นำไปสู่การปฏิรูปกระบวนการผลิตข้อมูลในหลายภาคส่วน อาทิ อุตสาหกรรมบันเทิงที่สามารถสร้างนักแสดงเสมือน การพัฒนาระบบสนทนาเชิงลึกในแชตบอต ตลอดจนการจำลองสถานการณ์ขั้นสูงเพื่อใช้ในการฝึกอบรมและการตัดสินใจเชิงยุทธศาสตร์ อย่างไรก็ตาม ความก้าวหน้านี้ยังมาพร้อมกับความท้าทายด้านจริยธรรมและความเสี่ยงต่อการนำไปใช้ในทางที่ผิด โดยเฉพาะการปลอมแปลงข้อมูลที่มีความสมจริงจนยากต่อการแยกแยะระหว่างของจริงกับของปลอม ซึ่งส่งผลกระทบต่อความน่าเชื่อถือในสังคมดิจิทัลอย่างมีนัยสำคัญ ทั้งนี้ Gartner (2025) คาดการณ์ว่า ภายในปี พ.ศ. 2569 ร้อยละ 80 ของความเสี่ยงจากการใช้ปัญญาประดิษฐ์จะเกิดจากการละเมิดภายในองค์กร เช่น การเปิดเผยข้อมูลที่ไม่เหมาะสม การสร้างสื่อปลอมที่ยากต่อการตรวจจับ ซึ่งรวมถึงภาพ เสียง และวิดีโอปลอมที่สร้างขึ้นโดยเทคโนโลยีปัญญาประดิษฐ์เชิงกำเนิด

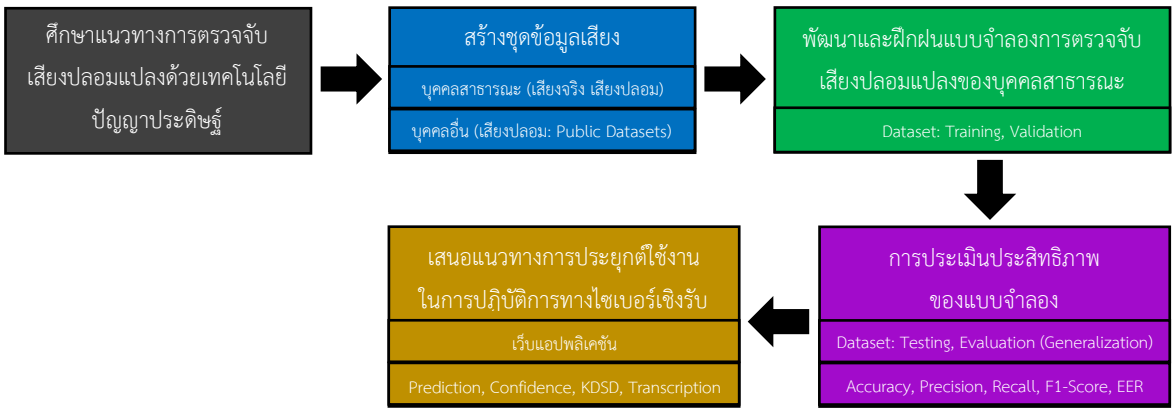
หนึ่งในภัยคุกคามที่น่ากังวลคือการปลอมแปลงเสียงบุคคลด้วยปัญญาประดิษฐ์ ซึ่งถูกนำไปใช้ในการก่ออาชญากรรมไซเบอร์ เช่น การแอบอ้างเป็นเสียงของบุคคลใกล้ชิดเพื่อหลอกลวงเหยื่อให้โอนเงินหรือเปิดเผยข้อมูลส่วนบุคคล รายงานจากสำนักงานตำรวจแห่งชาติในปี พ.ศ. 2567 (สำนักงานตำรวจแห่งชาติ, 2567) ระบุว่าปัญญาประดิษฐ์ กำลังถูกนำมาใช้ในทางที่ผิดเพิ่มขึ้นเรื่อย ๆ โดยเฉพาะในรูปแบบที่ซับซ้อนและตรวจจับได้ยาก กรณีตัวอย่างที่ตอกย้ำความรุนแรงของปัญหานี้ คือ เหตุการณ์เมื่อวันที่ 15 มกราคม พ.ศ.2568 ที่นางสาวแพทองธาร ชินวัตร นายกรัฐมนตรี (ไทยรัฐออนไลน์, 2568) เปิดเผยว่า

ตนเองตกเป็นเป้าหมายของแก๊งคอลเซนเตอร์ที่ใช้ปัญญาประดิษฐ์ปลอมเสียงผู้นำประเทศต่างชาติ ส่งคลิปเสียงและข้อความชักชวนให้โอนเงินบริจาคไปยังต่างประเทศ แม้เหตุการณ์จะไม่ก่อให้เกิดความเสียหาย แต่ก็สะท้อนว่าแม้แต่บุคคลระดับสูงก็ยังตกเป็นเป้าหมายของเทคโนโลยีหลอกลวงได้

อย่างไรก็ตาม แม้ว่าจะมีงานวิจัยในประเทศไทยที่เกี่ยวข้องบางส่วน เช่น การพัฒนาโมเดลเลียนเสียงเชิงลึกเพื่อการประยุกต์ด้านสงครามไซเบอร์ (พายัพและประสงค์, 2566) และการศึกษาเทคโนโลยีการแปลงเสียงด้วยปัญญาประดิษฐ์เชิงกำเนิด (อานนท์และพายัพ, 2568) แต่งานเหล่านี้มุ่งเน้นที่การสร้างเสียงสังเคราะห์ อีกทั้งงานตรวจจับเสียงที่มีอยู่โดยทั่วไป (กฤติภูมิและคณะ, 2566; Fathima *et al*., 2024) มักมุ่งแยกแยะเพียง “เสียงจริงของมนุษย์” กับ “เสียงสังเคราะห์ของมนุษย์” ในภาพรวม และใช้ชุดข้อมูลที่เป็นภาษาอังกฤษเท่านั้น โดยไม่ได้ศึกษาในระดับเสียงเฉพาะบุคคล ซึ่งต้องการความแม่นยำสูงสุดเพื่อยืนยันเอกลักษณ์ของบุคคลนั้น ๆ การมุ่งตรวจจับเสียงปลอมของบุคคลสาธารณะอย่างจำเพาะจึงมีความสำคัญต่อการประยุกต์ใช้เป็นเครื่องมือดิจิทัลเพื่อการพิสูจน์หลักฐาน ที่สามารถลดอคติจากการเรียนรู้เสียงของมนุษย์คนอื่น และสร้างความน่าเชื่อถือในการใช้งานจริง โดยเฉพาะในบริบทที่ต้องการความถูกต้องสูง เช่น การสืบสวนสอบสวนหรือการพิจารณาคดีในชั้นศาล ดังนั้น งานวิจัยนี้จึงมุ่งพัฒนาแบบจำลองเฉพาะด้านสำหรับการแยกแยะเสียงจริงและเสียงปลอมของบุคคลสาธารณะ โดยเฉพาะเสียงปลอมที่สร้างขึ้นจากเทคโนโลยีปัญญาประดิษฐ์เชิงกำเนิดซึ่งมีความสามารถในการเลียนแบบเอกลักษณ์ของเสียงมนุษย์ได้อย่างแนบเนียน ทั้งในด้านจังหวะ น้ำเสียง และการถ่ายทอดอารมณ์ เพื่อสนับสนุนการตรวจสอบและยับยั้งการใช้เทคโนโลยีในทางที่ผิด ตลอดจนเสริมสร้างความมั่นคงปลอดภัยไซเบอร์และความเป็นธรรมทางกระบวนการยุติธรรมอย่างยั่งยืน

ในงานวิจัยนี้ คำว่า “บุคคลสาธารณะ” หมายถึง บุคคลที่มีบทบาทในที่สาธารณะซึ่งมีโอกาสที่เสียงจะถูกนำไปใช้สร้างข้อมูลปลอมและส่งผลกระทบเชิงกว้างต่อสังคม เช่น นักการเมือง ผู้ดำรงตำแหน่งระดับสูง ดารา ศิลปิน และผู้มีอิทธิพลบนสื่อสังคมออนไลน์ โดยมีเกณฑ์การคัดเลือกคือเป็นบุคคลที่ปรากฏในสื่อสาธารณะอย่างต่อเนื่องและคนทั่วไปมีความคุ้นเคยกับเสียงดังกล่าว ซึ่งทำให้เสียงของบุคคลเหล่านี้มีความเสี่ยงต่อการถูกนำไปเลียนแบบหรือบิดเบือน ทั้งนี้เพื่อหลีกเลี่ยงการละเมิดสิทธิส่วนบุคคล งานวิจัยจึงได้จำลองสภาวะเสมือนจริง โดยบุคคลสาธารณะสมมติในกรณีนี้คือหัวหน้านักเรียนนายเรืออากาศ ซึ่งเป็นบุคคลที่พบปะพูดคุยกับนักเรียนนายเรืออากาศทุกวัน ทำให้เสียงของบุคคลนั้นเป็นที่รู้จักและคุ้นเคยในหมู่นักเรียนนายเรืออากาศ ส่งผลให้การทดลองมีความสมจริง ขณะเดียวกันก็ยังรักษาจริยธรรมการวิจัยและหลักการคุ้มครองข้อมูลส่วนบุคคลอย่างเคร่งครัด โดยบุคคลสาธารณะสมมติดังกล่าวได้ให้ความยินยอมอย่างเป็นลายลักษณ์อักษรเป็นที่เรียบร้อย

วิธีการดำเนินการวิจัย



รูปที่ 1 กรอบแนวคิดงานวิจัย

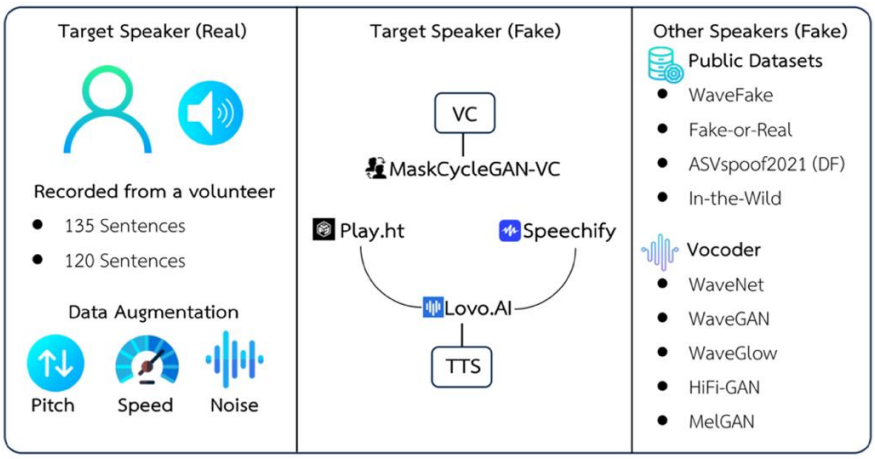
ผู้วิจัยได้กำหนดกรอบแนวคิดงานวิจัย แสดงดังรูปที่ 1 ประกอบด้วยการศึกษาแนวทางการตรวจจับเสียงปลอมแปลงด้วยเทคโนโลยีปัญญาประดิษฐ์โดยใช้ชุดข้อมูลเสียงที่สร้างขึ้นใหม่ ได้แก่ ข้อมูลเสียงจริงและเสียงปลอม ผู้วิจัยได้พัฒนาต่อยอดจากแบบจำลองในงานรู้จำภาพ ได้แก่ CNN SE-ResNet-34 และ SE-ResNeXt-152 และปรับแต่งค่าพารามิเตอร์สำหรับการฝึกฝนในแต่ละแบบจำลองให้มีความเหมาะสม รวมถึงประเมินประสิทธิภาพของแบบจำลองด้วยตัวชี้วัดต่าง ๆ สุดท้ายนำเสนอแนวทางการประยุกต์ใช้ในการปฏิบัติการทางไซเบอร์เชิงรับด้วยเว็บแอปพลิเคชัน แสดงผลการพยากรณ์พร้อมข้อมูลสนับสนุน

1. ศึกษาแนวทางการตรวจจับเสียงปลอมแปลงด้วยเทคโนโลยีปัญญาประดิษฐ์

ผู้วิจัยได้ศึกษาแนวทางการประยุกต์ใช้เทคโนโลยีปัญญาประดิษฐ์ในการสร้างเสียงปลอมแปลงซึ่งส่งผลกระทบต่อความมั่นคงปลอดภัยไซเบอร์ กรณีศึกษาที่เกี่ยวข้องกับการสร้างเสียงปลอมด้วยภาษาไทย ได้แก่ การแปลงเสียง (Voice Conversion: VC) หมายถึง กระบวนการเปลี่ยนแปลงเสียงของผู้พูดคนหนึ่งให้เป็นเสียงของอีกคน โดยยังคงรักษาเนื้อหาทางภาษาและสไตล์การพูดดั้งเดิมไว้ (อานนท์และพายัพ, 2568) การโคลนเสียง (voice cloning) หมายถึง กระบวนการเลียนแบบเอกลักษณ์เสียงบุคคลและสร้างเสียงปลอมของบุคคลนั้นผ่านการแปลงข้อความเป็นเสียง (Text-To-Speech: TTS) (พายัพและประสงค์, 2566) พร้อมทั้งสำรวจชุดข้อมูลเสียงมาตรฐานและวิเคราะห์แบบจำลองต่าง ๆ จากงานวิจัยที่เกี่ยวข้อง เพื่อนำมาเป็นแนวทางในการพัฒนาแบบจำลองการตรวจจับเสียงปลอมแปลงของบุคคลสาธารณะ

2. สร้างชุดข้อมูลเสียง

ชุดข้อมูลเสียงที่ใช้ในงานวิจัยนี้ แสดงดังรูปที่ 2 ผู้วิจัยแบ่งชุดข้อมูลเสียงออกเป็น 4 ส่วน ดังนี้ ชุดฝึกฝน (training set) ชุดตรวจสอบ (validation set) ชุดทดสอบ (testing set) ในอัตราส่วน 74:18:8 และชุดประเมินผล (evaluation set) เพื่อค้นหาความสามารถการเรียนรู้ทั่วไปของแบบจำลอง (model generalization performance) ซึ่งเป็นตัวชี้วัดสำคัญว่าแบบจำลองสามารถนำความรู้จากชุดฝึกฝนไปประยุกต์ใช้กับข้อมูลใหม่ที่ไม่เคยพบมาก่อนได้อย่างมีประสิทธิภาพเพียงใด หากแบบจำลองขาดความสามารถนี้ แม้จะให้ผลลัพธ์แม่นยำในชุดฝึกฝน แต่อาจล้มเหลวเมื่อนำไปใช้ในสถานการณ์จริง ส่งผลให้ระบบขาดความน่าเชื่อถือ (Barbiero *et al.*, 2020; Gohil *et al.*, 2024) ตามตารางที่ 1 - 2 ความหลากหลายของข้อมูลครอบคลุมทั้งเสียงจริงและเสียงปลอมที่สร้างจากเทคโนโลยีหลายรูปแบบ ไม่ว่าจะเป็นการแปลงเสียง การโคลนเสียง การแปลงข้อความเป็นเสียง รวมถึงชุดข้อมูลมาตรฐานสากลและแหล่งข้อมูลออนไลน์ โดยมีความแตกต่างในเชิงคุณลักษณะของผู้พูด ทั้งเพศชาย - หญิง หลากหลายช่วงอายุ รวมถึงหลายภาษา (ภาษาอังกฤษเป็นส่วนใหญ่) และสภาพแวดล้อมที่หลากหลาย ทั้งการบันทึกในสตูดิโอที่ควบคุมเสียงรบกวน ไปจนถึงสภาพแวดล้อมจริงที่มีเสียงแทรกและบริบทแตกต่างกัน เพื่อให้แบบจำลองสามารถเผชิญกับความแตกต่างของเสียงที่ใกล้เคียงกับการใช้งานจริงมากที่สุด



รูปที่ 2 แผนภาพสรุปชุดข้อมูลเสียงที่ใช้ในงานวิจัยนี้

ตารางที่ 1 การฝึกฝนแบบจำลองด้วยชุดฝึกฝนและชุดตรวจสอบ พร้อมการประเมินประสิทธิภาพเบื้องต้นด้วยชุดทดสอบ

ชุดข้อมูล	เสียงจริง[1]	เสียงปลอม[2]				
	บุคคลสาธารณะ Data Augmentation	บุคคลสาธารณะ MaskCycleGAN-VC	บุคคลอื่น WaveFake	บุคคลอื่น Fake-or-Real	บุคคลอื่น ASVspoof2021 (DF)	บุคคลอื่น In-the-Wild
Training	1800	100	400	400	400	500
Validation	450	25	100	100	100	125
Testing	180	10	40	40	40	50

เสียงจริง[1]: เสียงจริงของหัวหน้านักเรียนนายเรืออากาศ (TM03) โดยจำลองเป็นบุคคลสาธารณะ เพื่อเพิ่มความแม่นยำในการรู้จำและลดความลำเอียงจากเสียงบุคคลอื่น ดำเนินการบันทึกในสภาพแวดล้อมที่ปราศจากเสียงรบกวน ด้วยอุปกรณ์โทรศัพท์มือถือ รวมทั้งสิ้น 135 ประโยค ในการฝึกฝนและทดสอบแบบจำลอง ผู้วิจัยได้ดำเนินการตามขั้นตอนการแบ่งข้อมูลอย่างเคร่งครัด โดยเริ่มจากการแบ่งข้อมูลต้นฉบับออกเป็นชุดฝึกฝน ชุดตรวจสอบ และชุดทดสอบก่อนเป็นอันดับแรก จากนั้นจึงนำเทคนิคการเพิ่มข้อมูล (data augmentation) มาใช้ เพื่อประเมินประสิทธิภาพของแบบจำลองในการทำงานกับข้อมูลใหม่ได้อย่างเป็นกลาง โดยเทคนิคที่ใช้ประกอบด้วย การปรับคีย์เสียง (pitch shift ± 2 semitones) การปรับความเร็ว (speed factor 0.9 - 1.1 เท่า) และการเพิ่มสัญญาณรบกวน (gaussian noise SNR 15-30 dB) ส่งผลให้จำนวนข้อมูลเสียงเพิ่มขึ้นจากเดิมเป็น 2,430 ประโยค

เสียงปลอม[2]: เสียงปลอมของบุคคลสาธารณะ ผู้วิจัยได้สร้างขึ้นจากการแปลงเสียงด้วย MaskCycleGAN-VC (Kaneko *et al.*, 2021) โดยใช้เสียงต้นทางจากผู้วิจัย (เนื้อหาภาษาไทย) แปลงเป็นเสียงบุคคลสาธารณะ สำหรับเสียงปลอมของบุคคลอื่น ผู้วิจัยใช้ชุดข้อมูลเสียงมาตรฐาน ได้แก่ WaveFake (Frank and Schönherr, 2021) Fake-or-Real (Abdeldayem, 2021) ASVspoof2021 Deepfake (ASVspoof Consortium, 2021; Liu *et al.*, 2023) และ In-the-Wild (Mohamed, 2022) ด้วยการสุ่มตัวอย่างตามจำนวนที่กำหนด เพื่อเพิ่มความแม่นยำของแบบจำลองในการตรวจจับเสียงปลอมแปลงจากแหล่งข้อมูลที่ไม่เคยเห็นมาก่อนและลดความเสี่ยงจากการทำนายผิดพลาดซึ่งครอบคลุมเทคนิคการสร้างเสียงสังเคราะห์ (vocoder) หลากหลายประเภท (Oord *et al.*, 2016; Donahue *et al.*, 2019; Kumar *et al.*, 2019; Kong *et al.*, 2020) เช่น WaveNet WaveGAN WaveGlow HiFi-GAN และ MelGAN เป็นต้น

ตารางที่ 2 การประเมินประสิทธิภาพแบบจำลอง เพื่อค้นหาความสามารถการเรียนรู้ทั่วไปของแบบจำลองด้วยชุดประเมินผล

ชุดข้อมูล	เสียงจริง[1]	เสียงปลอม[2]			
	บุคคลสาธารณะ New Environment	บุคคลสาธารณะ MaskCycleGAN-VC	บุคคลสาธารณะ Play.ht	บุคคลสาธารณะ Speechify	บุคคลสาธารณะ Lovo.AI
Evaluation	120	30	30	30	30

เสียงจริง[1]: เสียงจริงของบุคคลสาธารณะ บันทึกเสียงในสภาพแวดล้อมใหม่ เช่น ในสถานที่ที่มีเสียงรบกวนแตกต่างจากเดิม รวมถึงลักษณะของบทพูดในรูปแบบใหม่ กล่าวคือ เป็นการเล่าเรื่องจากบทความ โดยใช้สไตล์การพูดของผู้บันทึกเสียง รวม 120 ประโยค

เสียงปลอม[2]: เสียงปลอมของบุคคลสาธารณะ ผู้วิจัยได้สร้างขึ้นในบริบทสภาพแวดล้อมใหม่ จากการแปลงเสียงด้วย MaskCycleGAN-VC โดยใช้เสียงต้นทางจากผู้วิจัย (เนื้อหาภาษาไทย) แปลงเป็นเสียงบุคคลสาธารณะ นอกจากนี้ยังสร้างขึ้นจากการโคลนเสียงจากเว็บไซต์ ดังนี้ Play.ht (Play.ht, 2025), Speechify (Speechify, 2025) และ Lovo.AI (Lovo.AI, 2025) โดยอัปโหลดเสียงของบุคคลสาธารณะ (เนื้อหาภาษาไทย) และใช้ข้อความภาษาอังกฤษสร้างเสียงปลอมผ่านการแปลงข้อความเป็นเสียง

3. พัฒนาและฝึกฝนแบบจำลองการตรวจจับเสียงปลอมแปลงของบุคคลสาธารณะ

3.1 การพัฒนาแบบจำลอง

ผู้วิจัยได้พัฒนาต่อยอดแบบจำลองจากงานรู้จำภาพด้วยโครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neural Networks: CNN) ตามตารางที่ 3 ซึ่งมีการประยุกต์ใช้งานจริงสำหรับการตรวจจับเสียงปลอมแปลง โดยอ้างอิงจากการสำรวจของ Yi *et al.* (2023) ที่ชี้ให้เห็นว่ากลุ่มแบบจำลอง CNN และ ResNet เป็นโครงสร้าง backbone พื้นฐานที่ใช้กันอย่างกว้างขวางในการตรวจจับเสียงปลอมและสถาปัตยกรรมที่ผสาน Squeeze-and-Excitation (SE) เช่น SE-ResNet และ SE-ResNeXt ได้รับการยืนยันว่าช่วยเพิ่มประสิทธิภาพในการเน้นคุณลักษณะที่สำคัญระหว่างช่องสัญญาณ (channel)

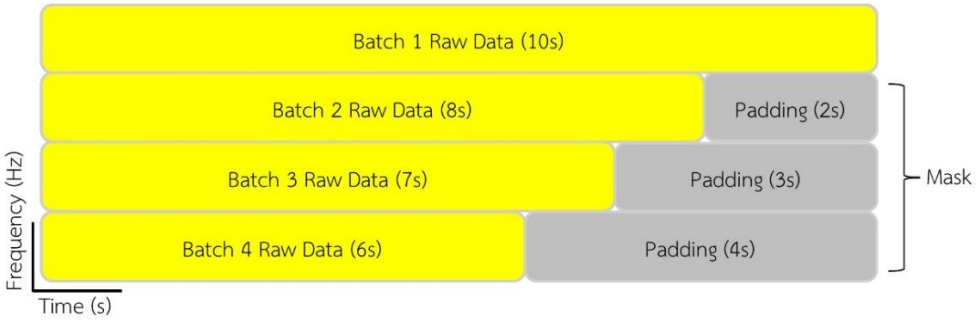
ทำให้การจำแนกเสียงจริงและเสียงปลอมมีความแม่นยำยิ่งขึ้น ดังนั้นการศึกษานี้จึงเลือกใช้ CNN (กฤติภูมิและคณะ, 2566; Fathima *et al.*, 2024) SE-ResNet-34 และ SE-ResNeXt-152 (He *et al.*, 2016; Hu *et al.*, 2018; StickCui, 2019) เป็นแบบจำลองหลัก

ตารางที่ 3 เทคนิคการปรับปรุงสถาปัตยกรรมโดยย่อ

เทคนิค	รายละเอียด	Reference
สถาปัตยกรรม	- Mask CNN-4: ใช้ convolutional layers จำนวน 4 ชั้น พร้อม receptive field ซึ่งเป็นขอบเขตรับรู้ของนิวรอนในโครงข่าย โดยมีขนาดเล็ก ทำให้เหมาะกับข้อมูลที่มีขนาดจำกัด - Mask SE-ResNet-34: ใช้ residual blocks จำนวน 34 ชั้น ร่วมกับ SE blocks ซึ่งช่วยปรับน้ำหนักช่องสัญญาณตามบริบท ทำให้การเรียนรู้คุณลักษณะมีความแม่นยำมากขึ้น - Mask SE-ResNeXt-152: ใช้โครงสร้างลึก 152 ชั้น ร่วมกับกลไก split–transform–merge และ SE Blocks ซึ่งช่วยดึงคุณลักษณะเสียงที่ซับซ้อนได้ละเอียดกว่า	Source Code[1]
Flexible Input: Mel-Spectrogram	ภาพขาวดำแบบช่องสัญญาณเดียว จัดเก็บเป็นอาร์เรย์สองมิติ แสดงการกระจายพลังงานของสัญญาณเสียงในแกนเวลาและความถี่ โดยสอดคล้องกับลักษณะการรับรู้ของระบบการได้ยินของมนุษย์	อานนท์และพายัพ (2568)
Mask[2]	เพื่อจดจำเฉพาะค่าจริงและเพิกเฉยกับค่าที่เติมเข้ามาในขณะฝึกฝนแบบจำลอง เพื่อเพิ่มเสถียรภาพและความแม่นยำในเรียนรู้	Shashank *et al.* (2024)
LeakyReLU Activation	เพื่อลดปัญหาการตายของนิวรอน (dead neuron problem) ส่งผลให้แบบจำลองเรียนรู้ได้ต่อเนื่องและมีเสถียรภาพมากขึ้น	Maas *et al.* (2013)
Embedding	ข้อมูลเชิงลึก (embedding) จัดเก็บเป็นข้อมูลหนึ่งมิติ งานวิจัยนี้ได้วิเคราะห์ความคล้ายคลึงของคุณลักษณะเสียง ด้วยวิธีการ KDSD จากค่าข้อมูลเชิงลึก	อานนท์และพายัพ (2568)
Dropout	เพื่อสุ่มปิดการทำงานของนิวรอนบางส่วน ลดความเสี่ยงจากการเรียนรู้เกินพอดี (overfitting) ทำให้สามารถใช้งานกับข้อมูลใหม่ได้ดียิ่งขึ้น	กฤติภูมิและคณะ (2566) Fathima *et al.* (2024)
Output: Binary Classification	จำแนกเสียงจริง (real) หรือเสียงปลอม (fake) ของบุคคลสาธารณะ จากค่าอาร์เรย์สองมิติของ fully connected ในชั้นสุดท้าย	

Source Code[1]: https://github.com/NKRAFAVoiceAI/FakeVoiceDetection

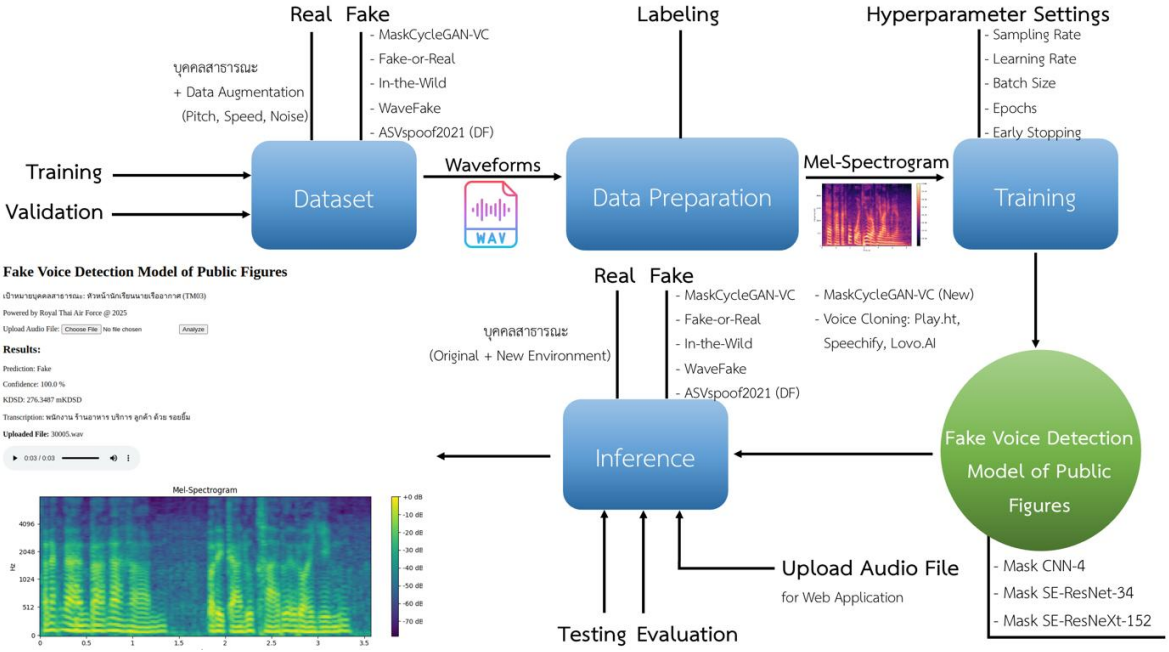
Mask[2]: ตัวอย่างการใช้งานแสดงดังรูปที่ 3 โดยผู้วิจัยกำหนด batch size = 4 เพื่อใช้เป็นตัวอย่างประกอบการอธิบายขั้นตอนการจัดการข้อมูล ได้แก่ การจัดเรียงตามความยาว การทำ Padding ให้มีความยาวเท่ากัน และการใช้ Masking เพื่อเพิกเฉยส่วนที่ไม่ใช่ข้อมูลจริง (Raw Data) วิธีการนี้สอดคล้องกับผลการศึกษาของ Shashank *et al*. (2024) ที่แสดงว่าเทคนิค Masking มีส่วนช่วยเพิ่มเสถียรภาพและความแม่นยำในการฝึกแบบจำลองโดยละเว้นข้อมูลที่ไม่จำเป็น ในการฝึกฝนจริงขนาดของ batch ไม่ได้จำกัดอยู่ที่ 4 แต่สามารถปรับให้เหมาะสมตามข้อจำกัดและประสิทธิภาพของหน่วยประมวลผลกราฟิก (GPU) ทั้งนี้ การใช้ batch ที่มีขนาดใหญ่ขึ้นจะช่วยให้การประมวลผลรวดเร็วขึ้น เนื่องจากการคำนวณ Gradient ซึ่งเป็นค่าความชันของฟังก์ชันสูญเสีย (loss function) และชี้ทิศทางการปรับน้ำหนักของแบบจำลอง อาศัยการเฉลี่ยจากหลายตัวอย่างในรอบเดียว อย่างไรก็ตาม หากมีขนาดใหญ่เกินไป อาจส่งผลต่อความสามารถในการเรียนรู้ทั่วไปของแบบจำลอง (Keskar *et al.*, 2017; Masters and Luschi, 2018)



รูปที่ 3 Mel-spectrogram input for 4-batch processing with masked padding ignored during training

3.2 ขั้นตอนการฝึกฝนแบบจำลอง

ภาพรวมการฝึกฝนแบบจำลอง แสดงดังรูปที่ 4 โดยมีรายละเอียดในแต่ละส่วนดังนี้



รูปที่ 4 ภาพรวมการฝึกฝนแบบจำลอง

1) การรวบรวมข้อมูล (data collection)

งานวิจัยนี้ใช้ชุดข้อมูลเสียง ตามที่ระบุไว้ในตารางที่ 1 - 2

2) การเตรียมข้อมูล (data preparation)

ข้อมูลนำเข้าของแบบจำลอง คือ สัญญาณเสียงที่ถูกแปลงให้อยู่ในรูปแบบ Mel-spectrogram (McFee *et al.*, 2015) และได้ใช้ตัวกรองแบบสามเหลี่ยมบนสเกลเมลด้วย Mel-filter banks จำนวน 128 ตัว เพื่อแปลงข้อมูลความถี่ให้เป็นค่าบนแกนแนวตั้ง ส่วนความยาวของแกนนอนแปรผันตามระยะเวลาของแต่ละไฟล์เสียง จากนั้นทำการระบุป้ายกำกับ (label) ของข้อมูลว่าเป็นเสียงจริง (real) หรือเสียงปลอม (fake) และจัดเก็บข้อมูลเพื่อใช้ในการประมวลผลในขั้นตอนถัดไป

3) การตั้งค่าพารามิเตอร์ (hyperparameter settings)

เนื่องจากการกำหนดค่าพารามิเตอร์เริ่มต้นมีลักษณะสุ่ม (random seed) ซึ่งอาจทำให้ผลลัพธ์ของแต่ละการฝึกแตกต่างกัน เพื่อเพิ่มความน่าเชื่อถือและสะท้อนความสามารถในการเรียนรู้ทั่วไปของแบบจำลอง (Henderson *et al.*, 2018) ผู้วิจัยจึงทำการฝึกซ้ำจำนวน 10 ครั้ง ระหว่างการฝึกฝนแบบจำลองได้ใช้เทคนิค early stopping เพื่อหยุดการฝึก

อัตโนมัติเมื่อค่าความผิดพลาดในชุดตรวจสอบ (validation loss) ไม่ลดลงต่อเนื่องหลังจากรอบที่กำหนด ซึ่งช่วยป้องกันการเรียนรู้เกินพอดีและลดเวลาที่ไม่จำเป็น จากนั้นผู้วิจัยได้ประเมินประสิทธิภาพโดยรวมของแบบจำลอง และใช้ตัวชี้วัด Equal Error Rate (EER) ในการเลือกเวอร์ชันที่ดีที่สุดสำหรับการนำไปใช้งาน สำหรับการตั้งค่าพารามิเตอร์หลัก ได้อ้างอิงแนวปฏิบัติที่นิยมในงานวิจัยที่เกี่ยวข้องและปรับเพิ่มเติมจากการทดลองเบื้องต้น โดย weight decay ถูกกำหนดเพื่อควบคุมความซับซ้อนของแบบจำลอง เพื่อลดความเสี่ยงของการเรียนรู้เกินพอดี โดยการลงโทษค่าน้ำหนักที่มีค่ามากเกินไป ส่วน step size และ gamma เป็นพารามิเตอร์ของ learning rate scheduler แบบ StepLR ที่ใช้ควบคุมการลดค่าอัตราการเรียนรู้ (learning rate) ตามรอบการฝึก (epochs) เพื่อช่วยให้การเรียนรู้มีความเสถียรและไม่หยุดนิ่งเร็วเกินไป ตามตารางที่ 4

ตารางที่ 4 การตั้งค่าพารามิเตอร์สำหรับการฝึกฝนแบบจำลอง

Model	Batch Size	Epochs	Early Stopping	Learning Rate	Weight Decay	Step Size	Gamma	Random Seed[1]	Device
Mask CNN-4	8	100	10	1e-4	1e-4	10	0.1	✓	HP-Notebook[2]
Mask SE-ResNet-34	8	100	10	1e-5	1e-4	10	0.1	✓	HP-Notebook[2]
Mask SE-ResNeXt-152	8	100	10	1e-4	1e-4	10	0.1	✓	Google Colab[3]

Random Seed[1]: ใช้ค่าพารามิเตอร์เริ่มต้นแบบสุ่มในการฝึกแต่ละครั้ง เพื่อสะท้อนสภาวะการทำงานจริงของแบบจำลอง

HP-Notebook[2] CPU: Intel® CoreTM i7-9750H @ 2.60GHz x 12 GPU: NVIDIA GeForce GTX 1650 Mobile 4GB Primary/Secondary Storage: RAM 8 GB / SSD 512 GB + HD 1 TB

Google Colab[3] GPU: T4 15GB Primary / Secondary Storage: RAM 12.7 GB / 112.6 GB (Free ~ 4 hours)

4. การประเมินประสิทธิภาพแบบจำลอง

โครงสร้างของ confusion matrix (Vanacore *et al.*, 2024) สำหรับ binary classification จำแนกเสียงจริงหรือเสียงปลอม ตามตารางที่ 5 โดยใช้ตัวชี้วัดดังสมการที่ 1 - 7 (Pedregosa *et al.*, 2011; Yi *et al.*, 2023; Liu *et al.*, 2023)

ตารางที่ 5 Confusion Matrix

นิยาม	ข้อมูลจริง	ทำนายผล	
TP (True Positive)	เสียงปลอม	เสียงปลอม	(ถูกต้อง)
TN (True Negative)	เสียงจริง	เสียงจริง	(ถูกต้อง)
FP (False Positive)	เสียงจริง	เสียงปลอม	(ผิดพลาด)
FN (False Negative)	เสียงปลอม	เสียงจริง	(ผิดพลาด)

Accuracy ใช้วัดความแม่นยำของการทำนายที่ถูกต้องทั้งหมดในภาพรวม

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision ใช้วัดความเที่ยงของการทำนายว่าเป็นเสียงปลอม

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall ใช้วัดความถูกต้องในการตรวจจับเสียงปลอม

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-Score ใช้วัดสมดุลระหว่าง Precision และ Recall

$$F1-Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (4)$$

False Positive Rate (FPR) ใช้วัดอัตราความผิดพลาดจากการทำนายผลว่าเป็นเสียงปลอม

$$FPR = \frac{FP}{FP+TN} \tag{5}$$

False Negative Rate (FNR) ใช้วัดอัตราความผิดพลาดจากการทำนายผลว่าเป็นเสียงจริง

$$FNR = \frac{FN}{FN+TP} \tag{6}$$

Equal Error Rate (EER) ใช้วัดอัตราความผิดพลาดที่เท่ากัน ซึ่งได้จากการทดลองปรับค่า threshold ตัวชี้วัดนี้นิยมใช้ในการประเมินระบบ biometric verification และ speaker verification หากค่า EER ต่ำ หมายความว่าแบบจำลองมีความแม่นยำและมีความน่าเชื่อถือสูง แต่หากค่า EER สูง แสดงว่าแบบจำลองยังไม่สามารถแยกแยะเสียงจริงและเสียงปลอมได้อย่างมีประสิทธิภาพ

$$EER = FPR = FNR \tag{7}$$

5. เสนอแนวทางการประยุกต์ใช้งานในการปฏิบัติการทางไซเบอร์เชิงรับ

5.1 ค่าระยะห่างเสียงเชิงลึกแบบเคอร์เนล (Kernel Deep Speech Distortion: KDSD)

รูปที่ 5 ขั้นตอนการคำนวณ KDSD

ผู้วิจัยได้พัฒนา KDSD (Binkowski *et al.*, 2020; อานนท์และพายัพ, 2568) แสดงดังรูปที่ 5 ซึ่งเป็นตัวชี้วัดที่ใช้ในการวัดความคล้ายคลึงของคุณลักษณะเชิงลึกของเสียง ระหว่างเสียงอ้างอิง (reference speech) กับ เสียงที่ต้องการทดสอบ (test speech) โดยอาศัยข้อมูลเชิงลึกของเสียงที่สกัดได้จากแบบจำลอง ก่อนนำมาคำนวณด้วยสูตร KDSD ดังสมการที่ 8 ค่า KDSD จะสะท้อนถึงความแตกต่างเชิงการกระจายของคุณลักษณะเสียง หากเสียงทั้งสองมีคุณลักษณะใกล้เคียงกัน ค่าที่ได้จะมีค่าใกล้ศูนย์และหากมีความแตกต่างสูงค่าจะมีค่ามากขึ้น ดังนั้นค่ายิ่งต่ำหมายถึงความคล้ายคลึงกันยิ่งสูง ทั้งนี้เพื่อความสะดวกในการรายงานผลลัพธ์ งานวิจัยนี้นำเสนอค่าในหน่วย milli-KDSD (mKDSD) นอกจากนี้ผู้วิจัยได้พัฒนาซอร์สโค้ดสำหรับคำนวณ KDSD โดยใช้การสกัดคุณลักษณะจากแบบจำลองและนำผลลัพธ์ข้อมูลเชิงลึกที่ได้มาเปรียบเทียบเชิงสถิติจริงภายในระบบ แนวคิดของ KDSD อ้างอิงจาก Maximum Mean Discrepancy (MMD) ร่วมกับ Gaussian RBF Kernel โดยให้ X และ Y เป็นชุดข้อมูลเชิงลึกของเสียงอ้างอิงและเสียงที่ต้องการตรวจสอบตามลำดับ ค่าของ KDSD สามารถนิยามได้ดังนี้

$$KDSD(X,Y) = \frac{1}{n^2}\sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{m^2}\sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{nm}\sum_{i,j} k(x_i, y_j) \tag{8}$$

เมื่อ $k(u,v) = exp\left(-\gamma||u-v||^2\right)$ เป็น Gaussian RBF Kernel โดยที่ค่าพารามิเตอร์ γ มีบทบาทสำคัญในการกำหนดความไวต่อความแตกต่างของเวกเตอร์ หาก γ มีค่ามาก Kernel จะตอบสนองต่อความแตกต่างเล็กน้อยได้ไวขึ้น ขณะที่ค่าที่ต่ำจะทำให้การเปรียบเทียบมีความหยาบมากขึ้น

5.2 การถอดเสียงพูดเป็นข้อความ (Transcription)

ผู้วิจัยได้พัฒนาการถอดเสียงพูดเป็นข้อความ โดยใช้แบบจำลอง wav2vec2-large-xlsr-53-th (HuggingFace, 2021) ซึ่งรองรับภาษาไทยและพัฒนามาจาก Wav2Vec2 ของ Facebook AI

5.3 การใช้งานแบบจำลองผ่านเว็บแอปพลิเคชัน

ผู้วิจัยได้พัฒนาเว็บแอปพลิเคชันต้นแบบสำหรับใช้งานแบบจำลอง โดยผู้ใช้งานสามารถอัปโหลดไฟล์เสียงที่ต้องการตรวจสอบ ซึ่งระบบจะแปลงสัญญาณเสียงให้อยู่ในรูปแบบ Mel-spectrogram และนำเข้าสู่แบบจำลองในโหมดการใช้งานเพื่อดำเนินการประมวลผล หลังจากสิ้นสุดการพยากรณ์ แบบจำลองจะแสดงผลลัพธ์ว่าข้อมูลเสียงนั้นเป็นเสียงจริงหรือเสียงปลอม โดยอ้างอิงจากการปรับค่า threshold ที่เหมาะสมกับแบบจำลองที่ถูกเลือกใช้งาน นอกจากนี้ ระบบยังแสดงข้อมูลเพื่อสนับสนุนการตัดสินใจเพิ่มเติม ได้แก่ ค่าความเชื่อมั่น ค่าระยะห่างเสียงเชิงลึกแบบเคอร์เนลและการถอดเสียงพูดเป็นข้อความ

5.4 การประยุกต์ใช้งานในการปฏิบัติการทางไซเบอร์เชิงรับเพื่อบรรเทาภัยคุกคามจากข้อมูลบิดเบือนบนมิติไซเบอร์

จากงานวิจัยของ อานนท์และพายัพ (2568) พบว่ามีการประเมินประสิทธิผลในการหลอกลวงผู้ฟังซึ่งเป็นบุคคลใกล้ชิดของหัวหน้านักเรียนนายเรืออากาศ โดยจำลองให้บุคคลดังกล่าวเป็นบุคคลสาธารณะในบริบทการวิจัยนี้ กลุ่มตัวอย่างคือ นักเรียนนายเรืออากาศจำนวน 50 คน (1 คนต่อ 1 ชุดการประเมิน) โดยแต่ละชุดประกอบด้วยไฟล์เสียงจริงและไฟล์เสียงปลอม รวมทั้งหมด 12 ไฟล์ ซึ่งสุ่มเลือกจากชุดข้อมูล NKRAFA Thai ที่คัดเลือกตัวอย่างเสียงคุณภาพดีที่สุดตามที่แสดงในตารางที่ 6 ขั้นตอนการประเมิน ผู้วิจัยเปิดไฟล์เสียงครั้งละหนึ่งไฟล์ และให้ผู้เข้าร่วมใช้วิจารณญาณส่วนบุคคลตัดสินว่าเสียงที่ได้ยินเป็น “เสียงจริง” หรือ “เสียงปลอม” ผลการตัดสินใจทั้งหมดถูกบันทึกลงในแบบประเมิน สำหรับงานวิจัยฉบับนี้ ได้นำผลดังกล่าวมาประยุกต์ใช้ในการวิเคราะห์เพิ่มเติม โดยเปรียบเทียบกับผลการจำแนกเสียงของปัญญาประดิษฐ์ (แบบจำลองที่มีประสิทธิภาพโดยรวมสูงที่สุดจากการประเมินด้วยชุดข้อมูลประเมินผล)

ตารางที่ 6 NKRAFA Thai Dataset

ชุดข้อมูล	เสียงจริง	เสียงปลอม
	หัวหน้านักเรียนนายเรืออากาศ บันทึกเสียง	หัวหน้านักเรียนนายเรืออากาศ MaskCycleGAN-VC
Evaluation	10	10

ผลการวิจัยและวิจารณ์ผล

1. ผลการศึกษาแนวทางการตรวจจับเสียงปลอมแปลงด้วยเทคโนโลยีปัญญาประดิษฐ์

ผลการศึกษาพบว่าเสียงปลอมที่สร้างขึ้นจากเทคโนโลยีปัญญาประดิษฐ์เชิงกำเนิด มีความสามารถในการเลียนแบบเอกลักษณ์ของเสียงมนุษย์ได้อย่างแนบเนียน ทั้งในด้านจังหวะ น้ำเสียงและการถ่ายทอดอารมณ์ ซึ่งยากต่อการตรวจจับด้วยการฟังของมนุษย์เพียงอย่างเดียว โดยเทคโนโลยีที่นิยมใช้ในการสร้างเสียงปลอม ได้แก่ การแปลงเสียง การโคลนเสียง และการแปลงข้อความเป็นเสียง จากการศึกษาสถาปัตยกรรมภายในของปัญญาประดิษฐ์เชิงกำเนิดพบว่า ส่วนใหญ่ใช้โครงข่ายปฏิปักษ์เชิงกำเนิด (Generative Adversarial Network: GAN) ซึ่งประกอบด้วยโครงข่ายประสาทเทียมแบบคอนโวลูชัน โดยมีโครงข่ายผู้สร้าง (generator) สำหรับสร้างเสียงปลอมและโครงข่ายผู้แยกแยะ (discriminator) สำหรับตรวจสอบความแตกต่างระหว่างเสียงจริงและเสียงปลอม ทั้งนี้ผู้วิจัยได้เสนอแนวทางการพัฒนาและปรับปรุงโครงข่ายผู้แยกแยะเพื่อใช้ในการตรวจจับเสียงปลอมแปลงของบุคคลสาธารณะ พร้อมเปรียบเทียบผลการทดลองด้วยงานวิจัยนี้กับงานวิจัยที่เกี่ยวข้อง ตามตารางที่ 7

ตารางที่ 7 เปรียบเทียบผลการทดลองด้วยงานวิจัยนี้กับงานวิจัยที่เกี่ยวข้อง

Reference	Method	Features Extraction	Dataset	Accuracy	Model Purpose
กฤติภูมิและคณะ (2566)	CNN-4	MFCC (Fixed Input)	YouTube, Fakeyou	0.970[1]	General[6]
Fathima *et al.* (2024)	CNN-2	Mel-Spectrogram (Fixed Input)	ASVspoof2019 (LA)	0.950[2]	General[6]
งานวิจัยนี้	Mask CNN-4 Mask SE-ResNet-34 Mask SE-ResNeXt-152[1st]	Mel-Spectrogram (Flexible Input)	บันทึกเสียง (หัวหน้านักเรียนนายเรืออากาศ: TM03) NKRAFA Thai Play.ht Speechify Lovo.AI WaveFakeFake-or-Real ASVspoof2021 (DF) In-the-Wild MaskCycleGAN-VC	1.000[3] 0.983[4] 1.000[5]	Specific[7]

0.970[1]: ได้จากการคำนวณด้วยชุดทดสอบ

0.950[2]: ได้จากการคำนวณด้วยชุดทดสอบ

1.000[3]: ได้จากการคำนวณด้วยชุดทดสอบ สำหรับการประเมินประสิทธิภาพเบื้องต้น

0.983[4]: ได้จากการคำนวณด้วยชุดประเมินผลที่ผ่านการปรับแต่งแบบจำลองแล้ว สำหรับการประเมินประสิทธิภาพแบบจำลองในการประยุกต์ใช้งานจริงกับข้อมูลใหม่ที่ไม่เคยเห็นมาก่อน

1.000[5]: ได้จากการคำนวณด้วยชุดข้อมูล NKRAFA Thai สำหรับการทดสอบในสถานการณ์จริง

General Purpose[6]: แบบจำลองที่ออกแบบมาเพื่อใช้งานทั่วไป สำหรับการแยกแยะเสียงมนุษย์ (real) และเสียงสังเคราะห์ (fake) ของบุคคลทั่วไป ซึ่งเหมาะกับการใช้งานในวงกว้าง แต่ความแม่นยำอาจไม่สูงพอสำหรับบริบทที่ต้องยืนยันตัวบุคคลอย่างเข้มงวด

Specific Purpose[7]: แบบจำลองที่ออกแบบมาเพื่อใช้งานเฉพาะด้าน สำหรับการแยกแยะเสียงจริง (real) และเสียงปลอม (fake) ของบุคคลสาธารณะ งานวิจัยนี้มุ่งเน้นแนวทางนี้ เนื่องจากบุคคลสาธารณะเป็นกลุ่มเป้าหมายที่เสียงของพวกเขามักถูกนำไปสร้างข้อมูลปลอมเพื่อบิดเบือนหรือนำไปใช้ในทางที่ผิด และอาจส่งผลกระทบต่อความมั่นคงของชาติ ดังนั้น การตรวจจับเสียงปลอมของบุคคลเหล่านี้จึงต้องการความแม่นยำสูงสุดเพื่อใช้เป็นเครื่องมือดิจิ ทัลในการพิสูจน์หลักฐานและสร้างความน่าเชื่อถือในเชิงกฎหมาย อีกทั้งการพัฒนาแบบจำลองเฉพาะบุคคลยังช่วยลดอคติจากการเรียนรู้เสียงของบุคคลอื่น ทำให้ผลลัพธ์สะท้อนเอกลักษณ์ของบุคคลนั้นได้ตรงกว่า ความท้าทายสำคัญคือการเก็บข้อมูลเสียงของบุคคลเดียวกันในหลายสภาพแวดล้อมซึ่งทำได้ยากและใช้เวลามาก รวมถึงเสียงปลอมที่สร้างจากเทคโนโลยีสังเคราะห์รุ่นใหม่ซึ่งเลียนแบบเอกลักษณ์ของเสียงมนุษย์ได้แนบเนียนมากขึ้น แบบจำลองลักษณะนี้จึงมีความซับ ซ้อนและต้องใช้ทรัพยากรมากกว่าการสร้างแบบจำลองทั่วไป แต่ก็ให้ความแม่นยำและความน่าเชื่อถือสูงกว่า

2. ผลการประเมินประสิทธิภาพแบบจำลองด้วยชุดทดสอบ

ผลการประเมินประสิทธิภาพแบบจำลองด้วยชุดทดสอบ ตามตารางที่ 8

ตารางที่ 8 ผลการประเมินประสิทธิภาพแบบจำลองด้วยชุดข้อมูลทดสอบ

Model	Average ± S.D.			
	Accuracy	Precision	Recall	F1-Score
Mask CNN-4	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
Mask SE-ResNet-34	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
Mask SE-ResNeXt-152	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0

จากผลการทดลองพบว่า ทุกแบบจำลองสามารถทำงานได้อย่างสมบูรณ์แบบในทุกตัวชี้วัด โดย accuracy precision recall และ F1-score มีค่าเฉลี่ยเท่ากับ 1.0 และมีค่าเบี่ยงเบนมาตรฐานเป็น 0.0 แสดงถึงความแม่นยำเสถียรภาพ และมีความสามารถในการแยกแยะเสียงจริงและเสียงปลอมได้อย่างครบถ้วน ทั้งนี้เพื่อยืนยันว่าผลลัพธ์ไม่ได้เกิดจากการแบ่งข้อมูลเพียงครั้งเดียวและเพื่อป้องกันการเกิด overfitting ผู้วิจัยได้ประเมินเพิ่มเติมสองวิธีได้แก่ 1) การใช้ 5-fold cross-validation โดยแบ่งข้อมูลออกเป็นห้าส่วนและสลับเลือกหนึ่งส่วนเป็นชุดตรวจสอบในการฝึกฝน ผลลัพธ์ที่ได้สอดคล้องกับการประเมินประสิทธิภาพเบื้องต้น และ 2) การใช้ชุดประเมินผลเพื่อค้นหาความสามารถการเรียนรู้ทั่วไปของแบบจำลอง

3. ผลการประเมินประสิทธิภาพแบบจำลองด้วยชุดประเมินผล

ผลการประเมินประสิทธิภาพแบบจำลองด้วยชุดประเมินผล ตามตารางที่ 9

ตารางที่ 9 ผลการประเมินประสิทธิภาพแบบจำลองด้วยชุดประเมินผล

Model	Average ± S.D.					
	Accuracy	Precision	Recall	F1-Score	EER[1st]	EER[1st]
Mask CNN-4	0.829 ± 0.034	0.987 ± 0.036	0.673 ± 0.103	0.794 ± 0.050	0.008	7.680×10^{-6}
Mask SE-ResNet-34	0.855 ± 0.075	0.924 ± 0.049	0.773 ± 0.141	0.836 ± 0.093	0.050	0.314
Mask SE-ResNeXt-152	0.863 ± 0.064	0.980 ± 0.041	0.746 ± 0.152	0.838 ± 0.086	0.016	0.029

EER[1st]: ผู้วิจัยคัดเลือกค่าที่มี EER ต่ำที่สุด (อันดับ 1) จากการฝึกซ้ำ 10 ครั้งของแต่ละแบบจำลอง มาใช้เป็นตัวแทนในการรายงานผล เพื่อสะท้อนประสิทธิภาพสูงสุดที่แบบจำลองสามารถทำได้ภายใต้เงื่อนไขการฝึกที่แตกต่างกัน

Threshold[2]: ค่าขอบเขตการตัดสินใจที่ปรับเพื่อหาจุดสมดุลระหว่าง FPR และ FNR โดยจุดที่ทั้งสองค่าเท่ากันจะถูกใช้ในการคำนวณ EER ของแต่ละแบบจำลอง ดังนั้น ค่า Threshold จึงสะท้อนเกณฑ์ที่แบบจำลองใช้ในการจำแนกเสียงจริงกับเสียงปลอม ซึ่งอาจแตกต่างกันไปตามคุณลักษณะที่แบบจำลองเรียนรู้ได้

จากผลการทดลองพบว่าแบบจำลอง Mask SE-ResNeXt-152 มีประสิทธิภาพโดยรวมสูงที่สุดเมื่อเทียบกับ Mask CNN-4 และ Mask SE-ResNet-34 โดยมีค่า accuracy เฉลี่ย 0.863 และ F1-score เฉลี่ย 0.838 ซึ่งสูงที่สุดในบรรดาทั้งสามแบบจำลอง แสดงถึงความสามารถในการจำแนกเสียงจริงและเสียงปลอมได้อย่างมีประสิทธิภาพ แม้ว่า Mask CNN-4 จะมี precision เฉลี่ยสูงที่สุดที่ 0.987 แต่มี recall ต่ำที่ 0.673 สะท้อนว่าแบบจำลองสามารถทำนายเสียงปลอมได้แม่นยำแต่ยังมีการตรวจจับตกหล่น ขณะที่ Mask SE-ResNeXt-152 ให้สมดุลระหว่าง precision และ recall ได้ดีกว่า และยังมีค่า EER 0.016 ที่ต่ำมาก แสดงถึงอัตราความผิดพลาดที่มีค่าน้อยมาก โดยรวมแล้วผลการทดลองยืนยันว่า Mask SE-ResNeXt-152 เป็นแบบจำลองที่เหมาะสมที่สุดในการนำไปประยุกต์ใช้การตรวจจับเสียงปลอมแปลงของบุคคลสาธารณะ ภายใต้ชุดประเมินผลซึ่งเป็นข้อมูลใหม่ที่แบบจำลองไม่เคยเห็นมาก่อน

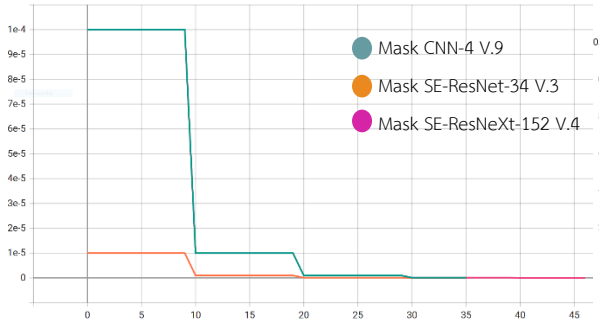
4. ผลการฝึกฝนและการทดสอบแบบจำลอง

ผู้วิจัยได้คัดเลือกแบบจำลองที่ดีที่สุดในแต่ละแบบจากผลการประเมินประสิทธิภาพแบบจำลองด้วยชุดประเมินผล (ตารางที่ 9) โดยมีผลการฝึกฝนและการทดสอบแบบจำลอง ตามตารางที่ 10 พร้อมทั้งแสดงกราฟ ดังรูปที่ 6 - 7

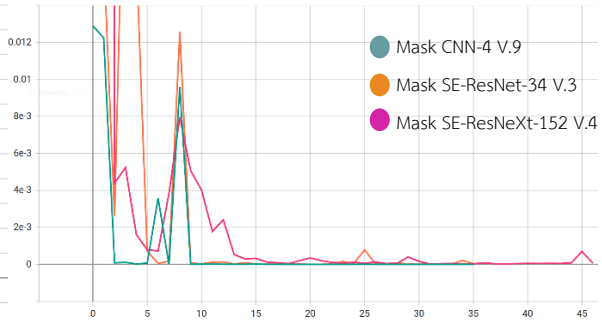
ตารางที่ 10 ผลการฝึกฝนและการทดสอบแบบจำลอง

Model	Mask CNN-4	Mask SE-ResNet-34	Mask SE-ResNeXt-152
Version[1]	9	3	4
Epochs (Early Stopping)	35	35	46
Training Duration (Minutes)	12.19	31.28	85.38
Number of Parameters (Million)	4.58	21.43	36.04
Model Size (MB)	18.40	85.90	145.20
GPU Memory Usage (MB)	3486	2904	15000
CPU Latency (ms)	8.86	15.00	26.22

Version[1]: แต่ละแบบจำลองจะถูกฝึกฝนซ้ำจำนวน 10 ครั้ง ภายใต้พารามิเตอร์และสถาปัตยกรรมเดียวกัน แต่ใช้การสุ่มค่าเริ่มต้นที่ต่างกันและระบบจะกำหนดหมายเลขลำดับ (Version 1 - 10) ตามลำดับการฝึกที่เสร็จสิ้น เพื่อให้ง่ายต่อการติดตามและประเมินผลลัพธ์ที่ได้จากแต่ละรอบการฝึก



รูปที่ 6 Learning Rate

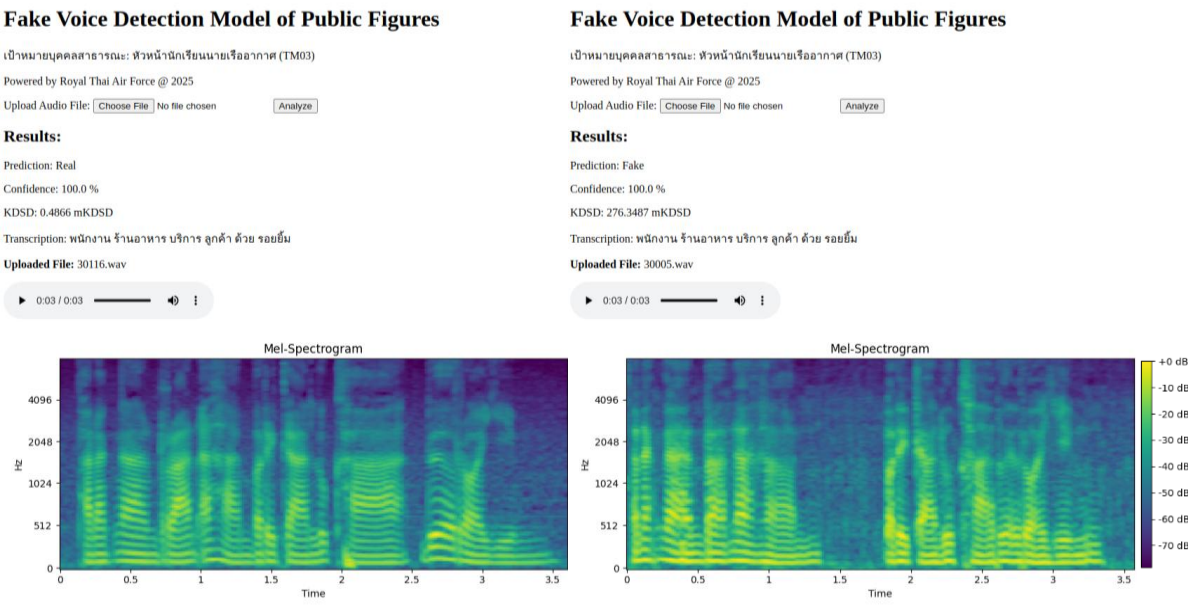


รูปที่ 7 Validation Loss

เมื่อเปรียบเทียบแบบจำลองทั้งสาม ตามตารางที่ 10 พบว่า Mask CNN-4 มีข้อดีที่เด่นคือใช้เวลาฝึกสั้นที่สุดเพียง 12.19 นาที มีขนาดเล็ก ใช้หน่วยความจำ GPU ต่ำ และจากทดสอบโดยใช้ความยาวเสียง 1 วินาที ประมวลผลบน CPU มี latency ต่ำที่สุด เหมาะสำหรับการใช้งานในระบบที่ต้องการความรวดเร็วและมีข้อจำกัดด้านทรัพยากร จากรูปที่ 6 และ 7 ยังพบว่า Mask CNN-4 สามารถปรับค่า learning rate และลดค่า validation loss ลงได้อย่างรวดเร็วและคงที่ แต่ข้อเสียคือประสิทธิภาพการจำแนกเสียงยังต่ำกว่าแบบจำลองอื่น สำหรับ Mask SE-ResNet-34 ใช้เวลาฝึก 31.28 นาที มีขนาดและการใช้ทรัพยากรเพิ่มขึ้น แต่ผลจากกราฟ learning rate และ validation loss แสดงให้เห็นว่าแบบจำลองมีการปรับตัวที่ค่อยเป็นค่อยไป ทำให้เกิดความสมดุลระหว่างความเร็วในการฝึกและความแม่นยำของผลลัพธ์ จึงถือว่ามีความสมดุลที่ดีกว่า Mask CNN-4 ส่วน Mask SE-ResNeXt-152 แม้จะใช้เวลาฝึกนานที่สุดถึง 85.38 นาที และใช้หน่วยความจำรวมถึง latency สูงสุด แต่จากกราฟทั้งสองแสดงให้เห็นว่าแบบจำลองสามารถรักษา validation loss ให้อยู่ในระดับต่ำอย่างต่อเนื่อง และปรับลด learning rate ได้สอดคล้องกับความซับซ้อนของแบบจำลอง ส่งผลให้ได้ประสิทธิภาพในการจำแนกเสียงจริงและเสียงปลอมดีที่สุด เหมาะสำหรับงานที่ให้ความสำคัญกับความแม่นยำมากกว่าความเร็วในการประมวลผล

5. ผลการใช้งานแบบจำลอง Mask SE-ResNeXt-152 ผ่านเว็บแอปพลิเคชัน

ผู้วิจัยได้ประยุกต์การตรวจจับเสียงปลอมแปลงของบุคคลสาธารณะผ่านเว็บแอปพลิเคชัน โดยเลือกใช้แบบจำลอง Mask SE-ResNeXt-152 เวอร์ชัน 4 ตามตารางที่ 10 ซึ่งให้ผลลัพธ์ดีที่สุดจากการประเมินด้วยชุดข้อมูลประเมินผล หลังจากปรับค่า threshold แล้ว แบบจำลองสามารถทำ accuracy ได้สูงถึง 0.983 โดยมีค่า FPR และ FNR ต่ำเพียง 0.016 และ 0.008 ตามลำดับ ตัวอย่างการทดสอบแสดงในรูปที่ 8 แม้เสียงจริงและเสียงปลอมจะพูดข้อความเดียวกันและหูมนุษย์แทบจะแยกไม่ออก แต่เมื่อพิจารณา Mel-spectrogram จะเห็นความแตกต่างของฮาร์มอนิก (harmonics) ซึ่งหมายถึงคลื่นความถี่ที่เกิดขึ้นร่วมกับความถี่พื้นฐาน (Fundamental frequency: F0) โดยเป็นทวีคูณของ F0 การผสมกันของฮาร์มอนิกเหล่านี้สร้างคุณลักษณะของเสียงที่เรียกว่า คุณภาพเสียง (timbre) ทำให้เสียงของแต่ละบุคคลมีเอกลักษณ์เฉพาะ แม้จะพูดด้วยระดับเสียงและความดังเท่ากันก็ตาม (สถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี, 2563; อานนท์และพายัพ, 2568) ดังนั้น แบบจำลองสามารถใช้คุณลักษณะเชิงฟิสิกส์เหล่านี้ซึ่งมนุษย์ฟังแทบไม่ออก เป็นเกณฑ์ตัดสินผลได้อย่างแม่นยำจากค่าความเชื่อมั่นและผลการพยากรณ์ อีกทั้งค่า KDSD ยังแสดงความแตกต่างได้ชัดเจน โดยเสียงจริงมีค่าต่ำมาก ซึ่งสะท้อนว่าคุณลักษณะเชิงสเปกตรัมมีความใกล้เคียงและคงเอกลักษณ์ของผู้พูดได้สม่ำเสมอ ในขณะที่เสียงปลอมมีค่าสูงกว่าอย่างมีนัยสำคัญ บ่งชี้ว่าการกระจายของคุณลักษณะเสียงแตกต่างไปจากเสียงจริงอย่างชัดเจน แม้หูมนุษย์อาจรับรู้ไม่ออกก็ตาม ทั้งนี้ งานวิจัยนำเสนอผลในหน่วย mKDSD เพื่อให้การรายงานและเปรียบเทียบค่ามีความสะดวกและละเอียดมากขึ้น นอกจากนี้การถอดเสียงพูดด้วยแบบจำลองรู้จำเสียงพูดของ Facebook AI พบว่า เสียงปลอมที่สร้างด้วยเทคโนโลยีปัญญาประดิษฐ์เชิงกำเนิดยังสามารถถอดเป็นข้อความได้ถูกต้องทุกประการ ซึ่งสะท้อนถึงความแนบเนียนและความสมจริงของเสียงปลอมและตอกย้ำความท้าทายในการตรวจจับในสถานการณ์จริง



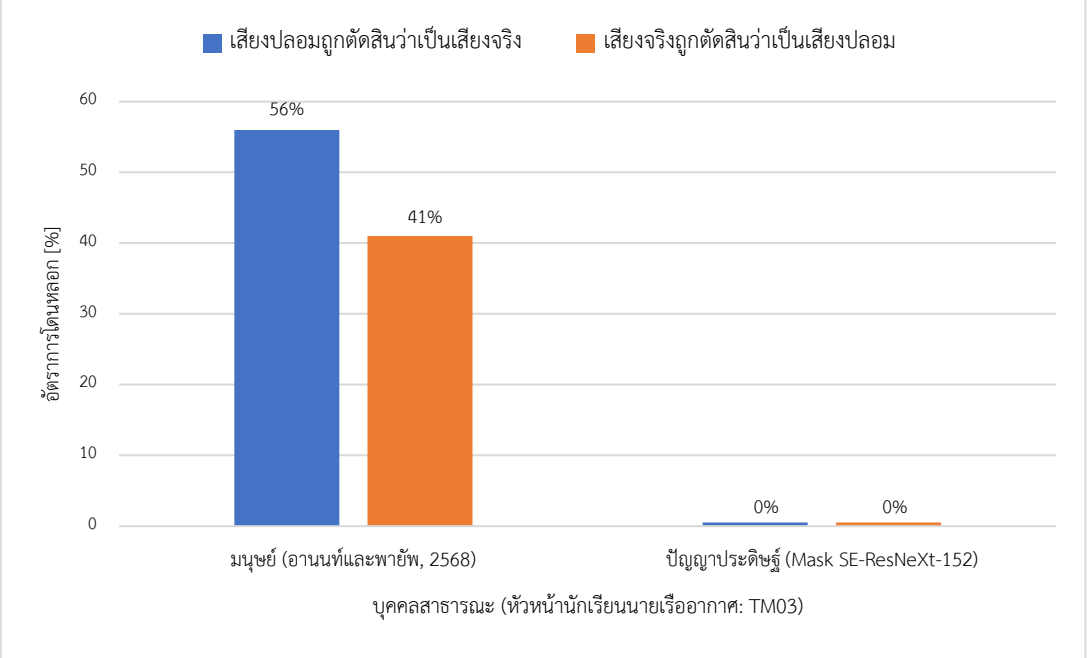
(ก) การทดสอบด้วยเสียงจริง (ข) การทดสอบด้วยเสียงปลอม

รูปที่ 8 การประยุกต์ใช้งานการตรวจจับเสียงปลอมแปลงของบุคคลสาธารณะผ่านเว็บแอปพลิเคชัน

6. ผลการประยุกต์ใช้งานในการปฏิบัติการทางไซเบอร์เชิงรับเพื่อบรรเทาภัยคุกคามจากข้อมูลบิดเบือนบนมิติไซเบอร์

จากการประยุกต์ใช้งานปัญญาประดิษฐ์ด้วยแบบจำลอง Mask SE-ResNeXt-152 ผ่านเว็บแอปพลิเคชัน โดยใช้ชุดข้อมูล NKRAFA Thai ตามตารางที่ 6 สำหรับการตรวจจับเสียงปลอมแปลงของบุคคลสาธารณะ แสดงดังรูปที่ 9 ผลการทดลองพบว่า แบบจำลองสามารถลดความผิดพลาดในการตัดสินใจได้อย่างชัดเจน โดยในกรณีที่มนุษย์ฟังเองยังมีอัตราการโดนหลอกสูงถึง 56% และ 41% (เสียงปลอมถูกตัดสินว่าเป็นเสียงจริงและเสียงจริงถูกตัดสินว่าเป็นเสียงปลอม) แต่เมื่อใช้แบบจำลองตรวจจับ อัตราความผิดพลาดเหล่านี้ได้ลดลงเหลือศูนย์ หรือถูกต้องครบถ้วนทุกไฟล์ สะท้อนถึงจุดแข็งสำคัญของแบบจำลองที่สามารถป้องกันการถูกหลอกจากเสียงปลอมได้ดีกว่ามนุษย์โดยสิ้นเชิง

นัยยะสำคัญ คือ เทคโนโลยีนี้สามารถนำไปประยุกต์ใช้เป็นเครื่องมือเชิงปฏิบัติการทางไซเบอร์เชิงรับ เช่น การยับยั้งการโจมตีหรือการบิดเบือนข้อมูลผ่านเสียงปลอม ซึ่งมีตัวอย่างการใช้งานในเครือข่ายสังคมออนไลน์เพื่อสร้างข่าวสารเท็จ (fake news) อย่างแพร่หลายในปัจจุบัน นอกจากนี้ ยังช่วยเพิ่มความน่าเชื่อถือในการตรวจสอบข้อเท็จจริง สนับสนุนการตัดสินใจของหน่วยงานด้านความมั่นคง และสามารถนำไปใช้ได้จริงในหลายบริบท เช่น นักวิเคราะห์ด้านความมั่นคง ผู้ตรวจสอบข้อเท็จจริง หน่วยงานภาครัฐ การพิสูจน์หลักฐาน การสืบสวนสอบสวน หรือการพิจารณาคดีในชั้นศาล



รูปที่ 9 การเปรียบเทียบการประเมินประสิทธิผลในการหลอกลวงผู้ฟังระหว่างมนุษย์และปัญญาประดิษฐ์ โดยใช้ชุดข้อมูล NKRAFA Thai ของบุคคลสาธารณะ (หัวหน้านักเรียนนายเรืออากาศ: TM03) เสียงจริง 10 ไฟล์ และเสียงปลอม 10 ไฟล์ รวม 20 ไฟล์ กลุ่มตัวอย่างนักเรียนนายเรืออากาศ 50 คน โดยแต่ละคนประเมินเสียงที่สุ่มเลือกจำนวน 12 ไฟล์

สรุปผลการวิจัย

1. ผลการศึกษาแนวทางการตรวจจับเสียงปลอมแปลงด้วยเทคโนโลยีปัญญาประดิษฐ์

งานวิจัยนี้ได้ศึกษาการตรวจจับเสียงปลอมแปลงด้วยเทคโนโลยีปัญญาประดิษฐ์และศึกษาข้อมูลเสียงจากงานวิจัยที่เกี่ยวข้อง ซึ่งช่วยให้ผู้วิจัยสามารถกำหนดแนวทางการการรวบรวมและการจัดเตรียมชุดข้อมูลได้อย่างเหมาะสมกับแบบจำลองเฉพาะด้านสำหรับการแยกแยะเสียงจริงและเสียงปลอมของบุคคลสาธารณะซึ่งพัฒนาต่อยอดจากงานรู้จำภาพด้วยโครงข่ายประสาทเทียมแบบคอนโวลูชัน ชุดข้อมูลเสียงที่ใช้ในงานวิจัยนี้ประกอบด้วยเสียงจริงและเสียงปลอมของบุคลสาธารณะ รวมถึงเสียงปลอมของบุคคลอื่น เสียงจริงใช้จากอาสาสมัครหัวหน้านักเรียนนายเรืออากาศ โดยจำลองเป็นบุคคลสาธารณะ ใช้เทคนิคการเพิ่มข้อมูล ทำให้เพิ่มจำนวนข้อมูลเสียงได้มากขึ้น สำหรับเสียงปลอมได้จากหลายแหล่ง ทั้งสร้างขึ้นเองด้วยเทคนิคการแปลงเสียง จาก MaskCycleGAN-VC โดยใช้เสียงต้นแบบจากผู้วิจัยแปลงเป็นเสียงของบุคคลสาธารณะ และใช้เทคนิคการโคลนเสียงจากเว็บไซต์ดังนี้ Play.ht Speechify Lovo.AI โดยอัปโหลดเสียงบุคคลสาธารณะ จากนั้นระบบจะเรียนรู้เอกลักษณ์เสียงบุคคลและสร้างเสียงปลอมด้วยเทคนิคการแปลงข้อความเป็นเสียง เพื่อเพิ่มความแม่นยำในการเรียนรู้ลักษณะเฉพาะของเสียงบุคคลสาธารณะและลดอคติที่อาจเกิดจากเสียงของบุคคลอื่น นอกจากนี้ยังใช้ข้อมูลเสียงปลอมของบุคคลอื่นจากชุดข้อมูลเสียงมาตรฐาน ได้แก่ WaveFake Fake-or-Real ASVspoof2021 Deepfake และ In-the-Wild ซึ่งครอบคลุมเทคนิคการสร้างเสียงสังเคราะห์หลากหลายประเภท เช่น WaveNet WaveGAN WaveGlow HiFi-GAN และ MelGAN เป็นต้น เพื่อเพิ่มความแม่นยำในการตรวจจับเสียงปลอมแปลงจากแหล่งข้อมูลที่ไม่เคยเห็นมาก่อนและลดความเสี่ยงจากการทำนายผิดพลาด ทั้งนี้ผู้วิจัยได้แบ่งชุดข้อมูลเสียงดังนี้ ชุดฝึกฝน ชุดตรวจสอบ ชุดทดสอบ ในอัตราส่วน 74:18:8 และชุดประเมินผล ชุดข้อมูลเสียงดังกล่าวนี้จะถูกแปลงให้อยู่ในรูปแบบ Mel-Spectrogram สำหรับการฝึกในแบบจำลองต่อไป

2. การพัฒนาแบบจำลองการตรวจจับเสียงปลอมแปลงของบุคคลสาธารณะที่มีประสิทธิภาพ

ผู้วิจัยได้พัฒนาแบบจำลองการตรวจจับเสียงปลอมแปลงของบุคคลสาธารณะ ได้แก่ Mask CNN-4 Mask SE-ResNet-34 และ Mask SE-ResNeXt-152 โดยทำการฝึกฝนแบบจำลองตามค่าที่ตั้งไว้ จำนวน 10 ครั้งในแต่ละแบบจำลอง เพื่อเพิ่มความสามารถการเรียนรู้ทั่วไปของแบบจำลองจากการกำหนดค่าพารามิเตอร์เริ่มต้นเป็นแบบสุ่ม ใช้เทคนิค mask เพื่อจดจำเฉพาะค่าจริงและเพิกเฉยกับค่าที่เติมเข้ามาในระหว่างการฝึกฝนแบบจำลอง ทำให้สามารถประมวลผลได้หลาย batch พร้อมกัน ซึ่งช่วยให้การฝึกฝนรวดเร็วขึ้นอย่างมีประสิทธิภาพ พร้อมเพิ่มเสถียรภาพและความแม่นยำในเรียนรู้จาก ข้อมูลดิบแบบ 100% นอกจากนี้ยังใช้เทคนิค early stopping เพื่อหยุดการฝึกฝนแบบอัตโนมัติ โดยพิจารณาจากการลดลงอย่างต่อเนื่องของค่าความผิดพลาดในชุดตรวจสอบเป็นหลัก ซึ่งระบบจะยุติการฝึกเมื่อค่าดังกล่าวเพิ่มขึ้นตามจำนวนครั้งที่กำหนดไว้ล่วงหน้า ทำให้ได้เวอร์ชันที่ดีที่สุดสำหรับการนำไปทดสอบ ผลการประเมินประสิทธิภาพด้วยชุดทดสอบแสดงให้เห็นว่าแบบจำลองทั้งหมดสามารถจำแนกได้อย่างแม่นยำสมบูรณ์ (ค่า accuracy เฉลี่ย 100%) ขณะที่การทดสอบด้วยชุดประเมินผลพบว่า Mask SE-ResNeXt-152 มีประสิทธิภาพสูงสุด (ค่า accuracy เฉลี่ย 86.33 ± 6.48%) และสามารถปรับแต่งค่าพารามิเตอร์เพื่อเพิ่มค่า accuracy ได้สูงสุดถึง 98.33% และ EER ต่ำเพียง 0.016

3. การเสนอแนวทางการประยุกต์ใช้งานในการปฏิบัติการทางไซเบอร์เชิงรับ

ผู้วิจัยได้ประยุกต์ใช้งานแบบจำลอง SE-ResNeXt-152 ผ่านเว็บแอปพลิเคชัน ซึ่งแสดงผลการพยากรณ์ค่าความเชื่อมั่น ค่าระยะห่างเสียงเชิงลึกแบบเคอร์เนล และการถอดเสียงพูดเป็นข้อความ จากการทดสอบในสถานการณ์จริงด้วยชุดข้อมูล NKRAFA Thai (20 ตัวอย่างเสียงที่ดีที่สุด : 10 เสียงจริง 10 เสียงปลอม) พบว่าสามารถจำแนกเสียงปลอมแปลงได้อย่างแม่นยำและมีประสิทธิภาพสูงกว่าการฟังของมนุษย์ โดยอัตราความผิดพลาดจากการฟังของมนุษย์ที่สูงถึง 56% และ 41% (เสียงปลอมถูกตัดสินว่าเป็นเสียงจริงและเสียงจริงถูกตัดสินว่าเป็นเสียงปลอม) ถูกลดลงจนเหลือศูนย์ สะท้อนถึงศักยภาพของแบบจำลองในการบรรเทาภัยคุกคามจากข้อมูลบิดเบือนบนมิติไซเบอร์ แบบจำลองนี้สามารถนำไปใช้เป็นเครื่องมือเชิงปฏิบัติการทางไซเบอร์เชิงรับ ไม่เพียงเพื่อการตรวจจับเสียงปลอมเท่านั้น แต่ยังช่วยยับยั้งการโจมตีหรือการบิดเบือนข้อมูลผ่านเสียงปลอมในเครือข่ายสังคมออนไลน์ เพิ่มความน่าเชื่อถือในการตรวจสอบข้อเท็จจริง สนับสนุนการตัดสินใจของหน่วยงานด้านความมั่นคง และสามารถต่อยอดไปสู่การใช้งานจริงในหลายบริบท เช่น การพิสูจน์หลักฐาน การสืบสวนสอบสวน และการพิจารณาคดีในชั้นศาล

4. ข้อจำกัดของงานวิจัยและการเสนอแนวทางการวิจัยในอนาคต

งานวิจัยนี้แม้จะแสดงให้เห็นว่าแบบจำลองที่พัฒนาขึ้นสามารถตรวจจับเสียงปลอมแปลงได้อย่างมีประสิทธิภาพ แต่ยังมีข้อจำกัดบางประการ ได้แก่ การจัดเตรียมชุดข้อมูลสำหรับบุคคลเฉพาะ ซึ่งต้องอาศัยการเก็บเสียงของบุคคลเดียวกันในหลายสภาพแวดล้อม ทำให้ได้ข้อมูลจำนวนจำกัดและยังไม่ครอบคลุมสถานการณ์จริงทั้งหมด อีกทั้งเทคนิคการสร้างเสียงปลอมที่ใช้ยังจำกัดเพียงบางประเภท เช่น การแปลงเสียงหรือการโคลนเสียง จึงอาจไม่ครอบคลุมเทคโนโลยีสังเคราะห์เสียงที่ซับซ้อนกว่าในปัจจุบัน ดังนั้นงานวิจัยในอนาคตควรเพิ่มการเก็บข้อมูลให้ครอบคลุมหลายสภาพแวดล้อมและหลากหลายเทคนิคการสร้างเสียงปลอม ตลอดจนพัฒนาแบบจำลองให้ซับซ้อนขึ้นเพื่อยกระดับความแม่นยำ ความทนทาน และศักยภาพในการนำไปใช้งานจริง

กิตติกรรมประกาศ

ผู้วิจัยขอแสดงความขอบคุณอย่างสุดซึ้งต่อกองทัพอากาศที่สนับสนุนทุนการศึกษา อันเป็นกำลังสำคัญในการดำเนินการวิจัยครั้งนี้ พร้อมทั้งขอขอบพระคุณโรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราชที่เอื้อเฟื้อสถานที่และอำนวยความสะดวกในทุกขั้นตอน ตลอดจนคณะกรรมการติดตามงานวิจัยและอาจารย์ที่ปรึกษาซึ่งได้ให้คำแนะนำและข้อเสนอแนะอันทรงคุณค่าอย่างต่อเนื่อง นอกจากนี้ขอขอบคุณนักเรียนนายเรืออากาศที่ให้ความร่วมมือในการจัดเตรียมข้อมูลเสียงและบุคลากรสำนักบัณฑิตศึกษาที่ให้ความช่วยเหลือด้านเอกสารและกระบวนการทางวิชาการด้วยความเอื้ออาทร

เอกสารอ้างอิง

กฤติภูมิ ผลาจันทร์, สุพศิน วงศ์ลาภสุวรรณ, ธนพล รุ่มนุ่ม, สัจจาภรณ์ ไวจรรยา และ ณัฐโชติ พรหมฤทธิ์. (2566). การจำแนกเสียงคนจริงและเสียงสังเคราะห์ปัญญาประดิษฐ์ด้วยโครงข่ายประสาทเทียมแบบคอนโวลูชัน. วารสารวิทยาศาสตร์ มข. 51(2): 170 - 179. doi: 10.14456/kkuscij.2023.15.

ไทยรัฐออนไลน์. (2568). แพทองธาร เผยโดนแก๊งคอลฯ ปลอมเสียงผู้นำประเทศหลอกโอนเงิน สั่ง รมว.ดีอี ตรวจสอบ. แหล่งข้อมูล: https://www.thairath.co.th/news/politic/2836244. ค้นเมื่อวันที่ 10 มีนาคม 2568.

พายัพ ศิรินาม และ ประสงค์ ปราณีตพลกรัง. (2566). การพัฒนาโมเดลการเลียนเสียงเชิงลึกในการประยุกต์ใช้งานด้านสงครามไซเบอร์. วารสารสถาบันวิชาการป้องกันประเทศ 14(1): 162 - 178.

สถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี. (2563). รายวิชาเพิ่มเติมวิทยาศาสตร์และเทคโนโลยี ฟิสิกส์ ชั้นมัธยมศึกษาปีที่ 5 เล่มที่ 4 [ฉบับ e-book]. พิมพ์ครั้งที่ 1. กรุงเทพฯ: สถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี. หน้า 29 - 30.

สำนักงานตำรวจแห่งชาติ. (2567). 4 รูปแบบอาชญากรรมออนไลน์ที่ต้องจับตามองในปี 2567 เมื่อ AI ถูกใช้ในด้านมืด ปลอมได้สารพัด. แหล่งข้อมูล: https://www.facebook.com/photo.php?fbid=766631268843499. ค้นเมื่อวันที่ 10 มีนาคม 2568.

อานนท์ บางเสน และ พายัพ ศิรินาม. (2568). การประยุกต์ใช้เทคโนโลยีปัญญาประดิษฐ์เชิงกำเนิดในการแปลงเสียง. วารสารวิทยาศาสตร์และเทคโนโลยีนายเรืออากาศ 21(2): 135 - 157.

Abdeldayem, M. (2021). The Fake-or-Real (FoR) Dataset (deepfake audio). Source: https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset. Retrieved date 14 February 2025.

ASVspoof Consortium. (2021). ASVspoof 2021 Deepfake (DF) database. Source: https://zenodo.org/record/4835108. Retrieved date 14 February 2025.

Barbiero, P., Squillero, G. and Tonda, A. (2020). Modeling generalization in machine learning: A methodological and computational study. arXiv preprint arXiv:2006.15680.

Binkowski, M., Donahue, J., Dieleman, S., Clark, A., Elsen, E., Casagrande, N., Cobo, L. C. and Simonyan, K. (2020). High fidelity speech synthesis with adversarial networks. In: Proceedings of the International Conference on Learning Representations (ICLR 2020). ICLR, Online.

Donahue, C., McAuley, J. and Puckette, M. (2019). Adversarial audio synthesis. In: International Conference on Learning Representations (ICLR) 2019. 1-18.

Fathima, G., Kiruthika, S., Malar, M. and Nivethini, T. (2024). Deepfake Audio Detection Model Based on Mel Spectrogram Using Convolutional Neural Network. International Journal of Creative Research Thoughts (IJCRT) 12(4): 208 - 215.

Frank, J. and Schönherr, L. (2021). WaveFake: A Data Set to Facilitate Audio Deepfake Detection. In: NeurIPS Datasets and Benchmarks Track 2021. NeurIPS Foundation, Online. Source: https://openreview .net/forum?id=IO7jcf63iDI. Retrieved date 14 February 2025.

Gartner. (2025). 2025 Gartner® Market Guide for AI Trust, Risk and Security Management (AI TRiSM). Source: https://www.proofpoint.com/us/resources/analyst-reports/gartner-market-guide-ai-trism. Retrieved date 10 March 2025.

Gohil, M., Yang, H., Tumanov, A., Delimitrou, C. and Venkataraman, S. (2024). Can machine learning models generalize across cloud environments? IEEE Computer Architecture Letters 23(1): 12 - 15.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). IEEE, Las Vegas, USA. 770-778. doi:10.1109/CVPR.2016.90.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D. and Meger, D. (2018). Deep Reinforcement Learning That Matters. In: 32nd AAAI Conference on Artificial Intelligence (AAAI 2018). AAAI Press, New Orleans, USA. 3207-3214.

Hu, J., Shen, L. and Sun, G. (2018). Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). IEEE, Salt Lake City, USA. 7132-7141. doi:10.1109/CVPR.2018. 00745.

HuggingFace. (2024). wav2vec2-large-xlsr-53-th. Source: https://huggingface.co/airesearch/wav2vec2-large-xlsr-53-th. Retrieved date 14 February 2025.

Jovanović, M. and Campbell, M. (2022). Generative Artificial Intelligence: Trends and Prospects. IEEE Computer Society 55(10): 107 - 112.

Kaneko, T., Kameoka, H., Tanaka, K. and Hojo, N. (2021). MaskCycleGAN-VC: Learning Non-parallel Voice Conversion with Filling in Frames. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021). IEEE, Toronto, Canada. 5919 - 5923.

Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P.T.P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. In: International Conference on Learning Representations (ICLR).

Kong, J., Kim, J. and Bae, J. (2020). HiFi-GAN: Generative adversarial network for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems 33: 17022-17033.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W.Z., Sotelo, J., de Brebisson, A., Bengio, Y. and Courville, A. (2019). MelGAN: Generative adversarial networks for conditional waveform synthesis. Advances in Neural Information Processing Systems 32: 14881-14892.

Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A. and Lee, K.A. (2023). ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31: 1 - 17.

LOVO. (2025). AI Voice Cloning & Text to Speech Online Tool. Source: https://lovo.ai. Retrieved date 14 February 2025.

Maas, A.L., Hannun, A.Y. and Ng, A.Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In: 30th International Conference on Machine Learning (ICML 2013). International Machine Learning Society, Atlanta, USA. 1 - 6.

Masters, D., and Luschi, C. (2018). Revisiting small batch training for deep neural networks. arXiv preprint arXiv:1804.07612.

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E. and Nieto, O. (2015). librosa: Audio and music signal analysis in Python. In: 14th Python in Science Conference (SciPy 2015). SciPy, Austin, USA. 18 - 25. doi:10.25080/Majora-7b98e3ed-003.

Mohamed, A. (2022). In-the-Wild Dataset. Source: https://www.kaggle.com/datasets/abdallamohamed312/in-the-wild-dataset. Retrieved date 14 February 2025.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12: 2825 - 2830.

Play.ht. (2025). AI Voice Cloning & Text to Speech Online Tool. Source: https://play.ht. Retrieved date 14 February 2025.

Shashank, N., Sathvik, S., Tanmaya, R. and Nethravathi, B. (2024). Enhancing sequence learning with masking in recurrent neural networks. International Research Journal of Modernization in Engineering Technology and Science (IRJMETS) 6(1): 2238 - 2246. doi: 10.56726/IRJMETS48489.

Speechify. (2025). AI Voice Cloning & Text to Speech Online Tool. Source: https://speechify.com. Retrieved date 14 February 2025.

StickCui. (2019). PyTorch-SE-ResNet. Source: https://github.com/StickCui/PyTorch-SE-ResNet/blob/master/model/model.py. Retrieved date 14 February 2025.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. In: 9th ISCA Speech Synthesis Workshop (SSW9). International Speech Communication Association (ISCA), Sunnyvale, California, USA. 125.

Vanacore, A., Pellegrino, M.S. and Ciardiello, A. (2024). Fair evaluation of classifier predictive performance based on binary confusion matrix. Computational Statistics 39: 363 - 383. doi: 10.1007/s00180-022-01301-9.

Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C.Y. and Zhao, Y. (2023). Audio Deepfake Detection: A Survey. arXiv preprint arXiv: 2308.14970.

❑❑❑❑❑