การพัฒนาโมเดลการเลียนเสียงเชิงลึกในการประยุกต์ใช้งานด้านสงครามไซเบอร์
The Development of a Deep Voice Model Implemented for Cyber Warfare

บทความวิจัย

พายัพ ศิรินาม[1] และ ประสงค์ ปราณีตพลกรัง[2]
Payap Sirinam and Prasong Praneetpolgrang
โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช กรุงเทพฯ ประเทศไทย 10220
Navaminda Kasatriyadhiraj Royal Air Force Academy, Bangkok, Thailand 10220
E-mail: payap_siri@rtaf.mi.th[1] and E-mail: prasong_pra@rtaf.mi.th[2]

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์ 1) เพื่อศึกษาแนวทางการประยุกต์ใช้งานเทคโนโลยีการเลียนแบบเสียงเชิงลึกของมนุษย์ในงานทั่วไป 2) เพื่อศึกษาโมเดลในการพัฒนาตัวแบบการเลียนแบบเสียงเชิงลึกของมนุษย์ที่มีความเหมาะสมสำหรับการใช้งานภาษาไทย 3) เพื่อวิเคราะห์และประเมินประสิทธิภาพของเสียงที่ถูกเลียนแบบ เมื่อถูกนำมาใช้จริงกับผู้ใช้งานในระบบอินเทอร์เน็ต และ 4) เพื่อนำเสนอแนวทางการใช้งานเทคโนโลยีการเลียนแบบเสียงเชิงลึกของมนุษย์ในงานด้านสงครามไซเบอร์ (Cyber Warfare)

ผลการศึกษาพบว่า การประยุกต์ใช้งานเทคโนโลยีการเลียนแบบเสียงเชิงลึก (การเลียนแบบเสียงโดยใช้การเรียนรู้เชิงลึก) ของมนุษย์ในงานทั่วไปโดยเฉพาะกรณีเสียงพูดเป็นภาษาอังกฤษ สามารถใช้โมเดลการสังเคราะห์เสียงของมนุษย์จากประโยคข้อความต่าง ๆ (Text-to-speech Synthesis) ได้ ถึงอย่างไรก็ตาม โมเดลดังกล่าวมีข้อจำกัดสำคัญ คือ ความต้องการข้อมูลคู่ขนานรวมถึงระยะเวลาในการเตรียมการฝึกฝนโมเดลที่ใช้ภาษาไทย โดยงานวิจัยนี้ ผู้วิจัยทำการศึกษาและเสนอแนะแนวทางการพัฒนาและประยุกต์ใช้โมเดลการเลียนเสียงเชิงลึกสำหรับงานด้านสงครามไซเบอร์ผ่านการใช้โมเดลแกน (GAN Model) เพื่อก้าวข้ามข้อจำกัดของโมเดลการสังเคราะห์เสียงของมนุษย์จากประโยคข้อความ

---

[1] นาวาอากาศตรี ผศ. ดร., อาจารย์กองการศึกษา โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช
Squadron Leader Assist.Prof. Dr., Lecture of Academic Faculty, Navaminda Kasatriyadhiraj Royal Air Force Academy

[2] พลอากาศตรี ศ. ดร., ผู้อำนวยการกองการศึกษา โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช
Air Vice Marshal Prof. Dr., Dean of Academic Faculty, Navaminda Kasatriyadhiraj Royal Air Force Academy

วันที่รับบทความ: 20 พ.ย.65 วันที่แก้ไขบทความ: 16 พ.ค.66 วันที่ตอบรับบทความ: 29 พ.ค.66

ผลการศึกษาพบว่า โมเดล StarGAN-VC และ CycleGAN-VC สามารถใช้ในการแปลงเสียงของบุคคลทั่วไปให้กลายเป็นบุคคลเป้าหมาย เช่น นักการเมือง ผู้บริหารประเทศ เพื่อสร้างข่าวปลอมในสงครามไซเบอร์ได้และมีค่าคุณภาพเสียงสำหรับการรับฟัง (Mean Opinion Score: MOS) สูงสุดที่ 3.59 และมีศักยภาพในการหลอกลวงผู้ใช้งานอินเทอร์เน็ตให้หลงเชื่อว่าเสียงที่ถูกปลอมแปลงเป็นเสียงของบุคคลเป้าหมายจริง โดยในกรณีที่ร้ายแรงที่สุด กลุ่มตัวอย่าง ร้อยละ 40 ถูกหลอกลวงด้วยเสียงปลอมแปลง ผลการศึกษาดังกล่าวเน้นย้ำและสร้างความตระหนักต่อภัยคุกคามรูปแบบใหม่รวมถึงการแสวงหาแนวทางการตรวจจับและป้องกันเสียงปลอมแปลงที่อาจกระทบต่อความมั่นคงของชาติในอนาคต

คำสำคัญ: การเรียนรู้เชิงลึก, การเลียนแบบเสียงเชิงลึก, โมเดลแกน, สงครามไซเบอร์

Abstract

The objectives of this research were 1) to study the implementations of a deep voice technology, 2) to investigate an appropriate model for the development of a deep voice model for Thai language use, 3) to analyze and evaluate the efficiency of the deep voice model after being used by the Internet users, and 4) to present guidelines for using the deep voice technology in the field of cyber warfare.

This study showed that the model for text-to-speech synthesis could be employed in the application of a deep voice technology (Voice Cloning using Deep Learning) in general work, especially in the case of English speech. However, the model had significant limitations including the requirements of the parallel data and the preparation time required for a Thai-language model training. In this work, the researcher had studied and presented guidelines to develop and apply deep voice models for cyber warfare using a GAN (Generative Adversarial Networks) model to overcome the limitations of human speech synthesis models from text sentences. The results revealed that the StarGAN-VC (StarGAN Voice Conversion) and CycleGAN-VC (CycleGAN Voice Conversion) models could be used to transform the voices of ordinary people into those of the targeted people such as politicians and national executives to create fake news in cyber warfare. Moreover, spoofed voices generated from this model could achieve a highest mean opinion score (MOS) of 3.59 and had the potential to deceive Internet users into believing that the spoofed voice was the voice of a real target. In most severe cases, 40% of the samples were duped by fake voices. These findings highlight and raise awareness of new threats that might affect national security in the future, and the search for ways to detect and prevent them.

Keywords: Deep Learning, Deep Voice, GAN Model, Cyber Warfare

ความเป็นมาและความสำคัญของปัญหา

มิติด้านไซเบอร์[3] ถือเป็นมิติทางการรบที่มีความสำคัญในฐานะสื่อกลางในการสื่อสารข้อมูลที่ต้องดำเนินงานด้านข้อมูลข่าวสารบนมิติไซเบอร์ เพื่อสนับสนุนภารกิจในการป้องกันประเทศและรักษาไว้ซึ่งความมั่นคงแห่งชาติในปัจจุบัน ด้วยเทคโนโลยีการติดต่อสื่อสารที่ทันสมัยขึ้น ทำให้ธรรมชาติของการแลกเปลี่ยนข้อมูลข่าวสารของมนุษย์มีแนวโน้มที่จะเกิดขึ้นบนมิติทางไซเบอร์ เช่น การผลิตและส่งผ่านข้อมูลข่าวสารผ่านเครือข่ายอินเทอร์เน็ต การแสดงความคิดเห็น/แนวคิดต่าง ๆ ที่อาจส่งผลกระทบต่อความมั่นคงของชาติผ่านช่องทางสื่อสังคมออนไลน์ เป็นต้น

ในปัจจุบันเทคโนโลยีปัญญาประดิษฐ์[4] (Artificial Intelligence) ได้มีการพัฒนารุดหน้าไปอย่างรวดเร็ว โดยเฉพาะอย่างยิ่งการเรียนรู้เชิงลึก[5] (Deep Learning) จนส่งผลให้คอมพิวเตอร์มีความสามารถในการเรียนรู้และทำงานได้เทียบเท่าหรือเหนือกว่ามนุษย์ (LeCun, Bengio, and Hinton, 2015, p.1-7) โดยความก้าวหน้าดังกล่าวได้ถูกนำมาใช้เพื่องานอรรถประโยชน์ และช่วยให้มนุษย์มีคุณภาพชีวิตและความสะดวกสบายขึ้น อย่างไรก็ตาม จากข่าวสารและเหตุการณ์ในปัจจุบัน พบว่ามีการใช้เทคโนโลยีปัญญาประดิษฐ์ในทางที่ไม่ถูกต้อง (Misuse) ส่งผลกระทบในแง่ลบต่อการใช้ชีวิตของมนุษย์ รวมถึงอาจส่งผลกระทบโดยตรงต่อความมั่นคงของชาติ (National Security) (Cooke, 2017, p.211-221; Allcott and Gentzkow, 2017, p.21-36)

หนึ่งในเทคโนโลยีปัญญาประดิษฐ์ที่เกิดจากการเรียนรู้เชิงลึก ที่ได้รับการพัฒนาและมีแนวโน้มการใช้งานในทางลบ คือ เทคโนโลยีการเลียนแบบเสียงของมนุษย์ด้วยการเรียนรู้เชิงลึก[6] (Deep Voice) ซึ่งเป็นหนึ่งในการประยุกต์ใช้เทคโนโลยีการเรียนรู้เชิงลึกเพื่อการสร้างสิ่งลวง (Deep Fakes) โดยเทคโนโลยีดังกล่าวมีความสามารถในการเลียนแบบและสังเคราะห์เสียงของมนุษย์ (Li, 2018, p.1462-1474) โดยใช้โมเดลที่ถูกสร้างขึ้นมาจากการเรียนรู้เสียงของบุคคลเป้าหมาย (Target) ซึ่งเทคโนโลยีดังกล่าวทำให้ขอบเขตของภัยคุกคามทางมิติไซเบอร์ได้ถูกขยายบริเวณออกไป ไม่ถูกจำกัดแต่เพียงการโจมตีผ่านระบบเครือข่ายคอมพิวเตอร์ แต่ยังหมายรวมถึงการทำวิศวกรรมทางสังคม[7] (Social Engineering) เพื่อการล่อลวงหรือสร้างความเข้าใจผิด โดยในช่วงที่ผ่านมามีรายงานของการใช้เทคโนโลยีดังกล่าวเพื่อการทุจริต (Frauds) ทั้งในด้านการเงิน การธนาคาร ผ่านการล่อหลอกด้วยวิธีวิศวกรรมทางสังคม (Kietzmann, Lee, McCarthy, & Kietzmann, 2020, p.135-146)

---

[3] มิติด้านไซเบอร์ (Cyber Domain) เป็นมิติที่เกี่ยวข้องกับการสื่อสารและส่งต่อข้อมูลในระบบคอมพิวเตอร์ และถือเป็นโดเมนที่ 5 ของการรบนอกเหนือจาก มิติภาคพื้นดิน (Land Domain) มิติภาคพื้นน้ำ (Sea Domain) มิติภาคอากาศ (Air Domain) และมิติด้านอวกาศ (Space Domain)

[4] ปัญญาประดิษฐ์ (Artificial Intelligence) หมายถึง ระบบประมวลผลของคอมพิวเตอร์ หุ่นยนต์ เครื่องจักร หรืออุปกรณ์อิเล็กทรอนิกส์ต่าง ๆ ที่มีการวิเคราะห์เชิงลึกคล้ายความฉลาดของมนุษย์ และสามารถก่อให้เกิดผลลัพธ์ที่เป็นการกระทำได้

[5] การเรียนรู้เชิงลึก (Deep Learning) เป็นวิธีการการเรียนรู้ของเครื่องบนพื้นฐานของโครงข่ายประสาทเทียม (Neural Networks) โดยการเรียนรู้รูปแบบดังกล่าวสามารถเป็นได้ทั้งแบบการเรียนรู้แบบมีผู้สอน (Supervised Learning) การเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised Learning) และการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning)

[6] เทคโนโลยีการเรียนรู้เชิงลึกเพื่อการสร้างสิ่งลวง (Deep Fakes) เป็นการใช้ประโยชน์จากเทคโนโลยีปัญญาประดิษฐ์ที่ถูกพัฒนาด้วยเทคโนโลยีการเรียนรู้แบบการเรียนรู้เชิงลึก เพื่อเลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ซึ่งทำให้เทคโนโลยีดังกล่าวมีความสามารถในการปลอมแปลงคุณลักษณะบุคคลต่าง ๆ ได้อย่างอัตโนมัติและมีความแนบเนียน

[7] วิศวกรรมทางสังคม (Social Engineering) เป็นศิลปะการหลอกลวงผู้คนเพื่อผลประโยชน์ตามที่ผู้ไม่ประสงค์ดีต้องการ โดยอาศัยจุดอ่อนและความไม่รู้รวมถึงความประมาทเลินเล่อ ทำให้เป็นการโจมตีที่ได้ผลดีมากเมื่อเทียบกับการโจมตีไซเบอร์รูปแบบอื่น ๆ โดยเฉพาะกับคนที่ไม่มีความรู้ทางด้านความมั่นคงปลอดภัย

ในแง่ของความมั่นคงของชาติ เทคโนโลยีการเลียนแบบเสียงอาจเป็นหนึ่งในเครื่องมือที่ทรงพลังในงานด้านสงครามไซเบอร์ผ่านการสร้างข่าวเท็จ (Fake News) หรือการทำให้เกิดความเข้าใจผิดว่าข่าวสารนั้นมาจากบุคคลดังกล่าวจริง ตัวอย่างภัยคุกคามที่อาจเกิดขึ้น คือ ถ้ามีผู้ไม่ประสงค์ดีนำเอาเสียงของผู้บริหารภาครัฐ/ผู้บังคับบัญชาระดับสูง ที่อาจได้มาจากการแอบบันทึกเสียงหรือจากสื่อวิทยุโทรทัศน์และนำเสียงบุคคลเหล่านั้นมาลอกเลียนแบบ และส่งสารผ่านประโยคหรือให้ข่าวที่อาจกระทบต่อภาพลักษณ์และความมั่นคงของกองทัพและประเทศชาติ (Sindermann, Cooper, & Montag, 2020, p.44-48; Karlsen & Aalberg, 2021, p.1-17) จากนั้นก็เผยแพร่เสียงดังกล่าวไปยังสื่อสังคมออนไลน์ เช่น *“คลิปลับเสียงของ...”* หรือ หากมีการเลียนแบบเสียงของผู้บังคับบัญชาระดับสูง และมีอำนาจสั่งการทางยุทธการและใช้คลิปลับเสียงดังกล่าวโทรไปสั่งการให้ผู้ใต้บังคับบัญชากระทำการใด ๆ ก็อาจก่อให้เกิดความเสียหายอย่างร้ายแรงได้ ดังนั้น จึงจำเป็นอย่างยิ่งที่ต้องมีการดำเนินการวิจัยและพัฒนาเพื่อศึกษาและทำความเข้าใจเทคโนโลยีดังกล่าว ทั้งในแง่ความเข้าใจพื้นฐานของการสร้าง พัฒนา และกลไกของเทคโนโลยีดังกล่าว เพื่อพัฒนาองค์ความรู้ของหน่วยงานด้านความมั่นคงในการใช้ประโยชน์และต่อต้านการใช้ประโยชน์จากผู้ไม่หวังดีรวมถึงกระบวนการและวิธีการตรวจจับการเลียนแบบเสียง พร้อมทั้งคำอธิบายทางวิทยาศาสตร์ที่เชื่อถือได้เพื่อรองรับต่อการปฏิบัติการในสงครามไซเบอร์ต่อไป

วัตถุประสงค์การวิจัย

1. เพื่อศึกษาแนวทางการประยุกต์ใช้งานเทคโนโลยีการเลียนแบบเสียงเชิงลึกของมนุษย์ในงานทั่วไป
2. เพื่อศึกษาโมเดลในการพัฒนาตัวแบบการเลียนแบบเสียงเชิงลึกของมนุษย์ที่มีความเหมาะสมสำหรับการใช้งานภาษาไทย
3. เพื่อวิเคราะห์และประเมินประสิทธิภาพของเสียงที่ถูกเลียนแบบ เมื่อถูกนำมาใช้จริงกับผู้ใช้งานในระบบอินเทอร์เน็ต
4. เพื่อนำเสนอแนวทางการใช้งานเทคโนโลยีการเลียนแบบเสียงเชิงลึกของมนุษย์ ในงานด้านสงครามไซเบอร์ (Cyber Warfare)

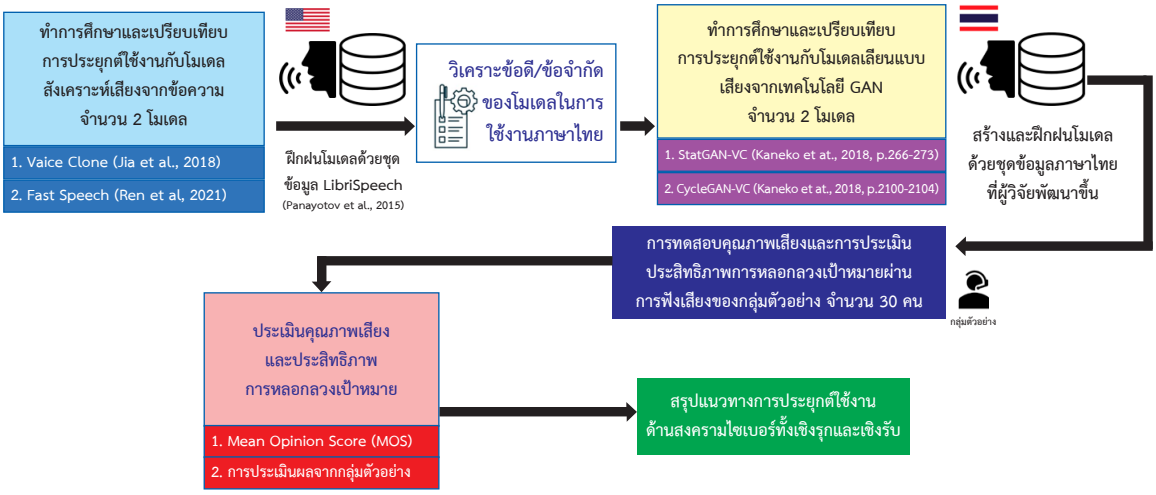
สมมติฐาน

1. ได้แนวทางในการประยุกต์ใช้โมเดลเลียนแบบเสียงเชิงลึกของมนุษย์ที่มีความเหมาะสม รวมถึงแนวทางการนำโมเดลดังกล่าวไปประยุกต์ใช้งานด้านสงครามไซเบอร์
2. องค์ความรู้ที่ได้รับจากการดำเนินการวิจัยสามารถนำมาใช้ในการประยุกต์ใช้ในงานด้านข่าวกรองไซเบอร์เชิงรุก (การเลียนแบบเสียงเพื่อปฏิบัติการข้อมูลข่าวสาร) และเชิงรับ (การตรวจจับและพิสูจน์ทราบ) ซึ่งเป็นหนึ่งในการปฏิบัติการด้านสงครามไซเบอร์ทั้งในปัจจุบันและอนาคต

ขอบเขตงานวิจัย

1. การวิจัยมุ่งเน้นศึกษาแนวทางการพัฒนาและประยุกต์ใช้โมเดลให้เหมาะสมอย่างน้อย 2 โมเดล และศึกษาแนวทางการประยุกต์ใช้งานในด้านสงครามไซเบอร์ พร้อมทำการวัดประสิทธิภาพและประเมินผลโดยใช้กรอบการศึกษาอยู่ในบริบทของการปฏิบัติงานของกองทัพอากาศ
2. การวิจัยในการประเมินผลกับกลุ่มตัวอย่างดำเนินการภายใต้ข้อจำกัดในสถานการณ์การแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 (โควิด-19)

กรอบแนวคิดงานวิจัย



ภาพที่ 1 กรอบแนวคิดงานวิจัย

ทฤษฎีที่เกี่ยวข้อง

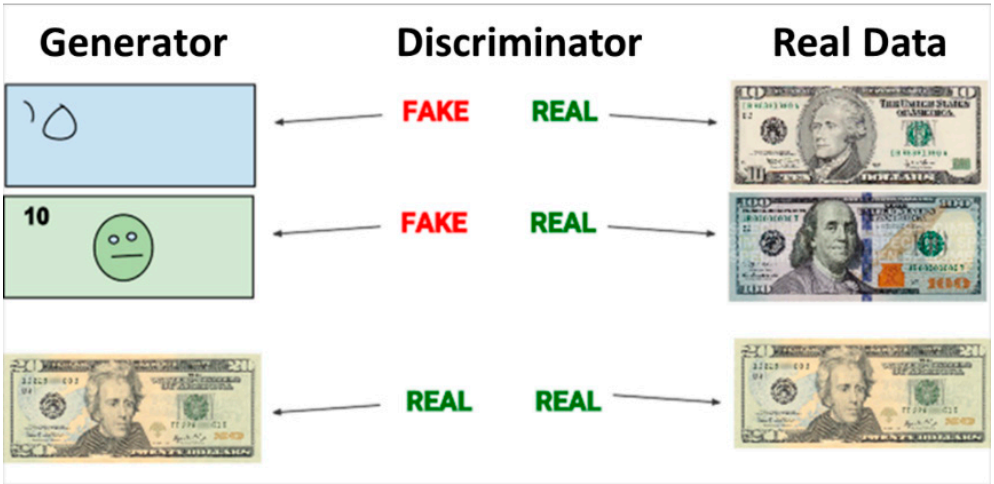
1. การคัดลอกเสียงด้วยเทคโนโลยีปัญญาประดิษฐ์

1.1 การสังเคราะห์เสียงจากข้อความ

การสังเคราะห์เสียงจากข้อความ (Text-to-Speech Synthesis: TTS) คือ เทคโนโลยีที่สามารถสร้างเสียงคำพูดใด ๆ จากข้อความที่ระบุร่วมกับเทคโนโลยีด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ทำให้ได้เทคโนโลยีการสังเคราะห์เสียงจากข้อความ (Text-to-Speech Synthesis: TTS) (Jia et al., 2018, p.1-11) โดยโมเดลที่สร้างขึ้นอยู่บนพื้นฐานของโมเดลการเรียนรู้เชิงลึกในรูปแบบโครงข่ายประสาทเทียมเชิงลึก (Deep Neural Networks: DNN) ที่จะเรียนรู้จากไฟล์คู่ขนาน (เสียงพูดและไฟล์ประโยคข้อความของเสียงพูด) เพื่อสร้างโมเดลที่สามารถสังเคราะห์เสียงจากข้อความที่ผู้ใช้พิมพ์ขึ้น

1.2 การโคลนเสียงพูด

การโคลนเสียงพูด (Voice Cloning) คือ เทคโนโลยีที่สามารถทำการเลียนแบบ หรือแปลงเสียงคำพูดของบุคคลเป้าหมาย (Ren et al., 2021, p.1-15) โดยสามารถดำเนินการผ่านเทคโนโลยีการสังเคราะห์เสียงจากข้อความที่ต้องมีข้อมูลแบบคู่ขนานเช่นเดียวกับกระบวนการสังเคราะห์เสียงจากข้อความที่กล่าวมาข้างต้น หรืออาจใช้ข้อมูลแบบไม่คู่ขนาน (Non-Parallel) กล่าวคือ ใช้แต่เพียงเสียงอย่างเดียว โดยโมเดลที่ไม่ใช้ข้อมูลแบบคู่ขนานมีลักษณะเด่น คือ มีความสะดวกรวดเร็ว ผู้ใช้งานใช้เพียงเสียงต้นแบบที่ต้องการเลียนแบบและเสียงของผู้ใช้ที่ต้องการพูดออกไป โดยโมเดลจะทำการแปลงเสียงของผู้ใช้ให้มีความคล้ายคลึงกับเสียงต้นแบบให้มากที่สุด



ภาพที่ 2 ตัวอย่างการใช้งาน GAN เพื่อสร้างธนบัตรปลอม

1.3 การแปลงเสียงพูดผ่านโมเดลแกน

แกน (Generative Adversarial Networks: GAN) เป็นเทคโนโลยีการเรียนรู้เชิงลึกที่มีความก้าวหน้าและได้รับความนิยมเป็นอย่างมากในปัจจุบัน ด้วยความสามารถในการสร้างและเลียนแบบข้อมูลได้อย่างมีประสิทธิภาพ โดยหลักคิดพื้นฐานของโมเดลแกน คือ การใช้เครือข่ายประสาทเทียมจำนวน 2 เครือข่าย ในการฝึกฝนให้คอมพิวเตอร์ได้เรียนรู้ลักษณะของชุดข้อมูลที่ดีพอ เพื่อให้สามารถสร้างภาพปลอมได้เหมือนจริงมากที่สุด (Goodfellow et al., 2020, p.139-144) โดยเครือข่ายเพื่อใช้ในการฝึกฝนดังกล่าวประกอบไปด้วย

1.3.1 เครือข่ายสำหรับสร้างข้อมูล (Generator) มีหน้าที่ในการเรียนรู้ว่าสิ่งที่ต้องการจะสร้างมีรูปแบบวิธีการสร้างอย่างไร ยกตัวอย่างเช่น หากเราต้องการศึกษาวิธีการสร้างธนบัตรปลอมขึ้นมา เครือข่ายสำหรับสร้างข้อมูลก็จะทำการเรียนรู้ในการพิมพ์ลายเส้น สี และรูปแบบของธนบัตรให้มีความแนบเนียนและเหมือนกับธนบัตรจริงมากที่สุด

1.3.2 เครือข่ายสำหรับแยกแยะ (Discriminator) มีหน้าที่ในการตรวจสอบว่าสิ่งที่ถูกสร้างจากเครือข่ายสำหรับสร้างข้อมูลมีความเหมือนกับสิ่งที่ต้องการเลียนแบบมากเพียงใด โดยทุก ๆ ครั้งของการตรวจสอบ เครือข่ายสำหรับแยกแยะก็จะทำการส่งผลการประเมินกลับไปยังเครือข่ายสำหรับสร้างข้อมูลเพื่อทำการปรับปรุงให้สามารถเลียนแบบได้เร็วขึ้น โดยหากเปรียบเทียบกับกระบวนการสร้างธนบัตรปลอม เครือข่ายสำหรับแยกแยะจะทำหน้าที่เหมือนผู้ตรวจสอบธนบัตรว่า ธนบัตรที่ถูกสร้างขึ้นมามีความเหมือนกับธนบัตรจริงมากน้อยเพียงใดดังแสดงในภาพที่ 2 โดยกระบวนการในการสร้างวัตถุขึ้นมาเพื่อเลียนแบบนั้นจะเกิดขึ้นซ้ำ ๆ กัน จนกระทั่งเครือข่ายสำหรับแยกแยะไม่สามารถแยกแยะสิ่งที่ถูกสร้างและวัตถุจริงออกจากกันได้ และเมื่อเครือข่ายสามารถเรียนรู้ได้ถึงจุดดังกล่าว เราก็จะสามารถใช้เครือข่ายดังกล่าวในการสร้างวัตถุที่เหมือนกับวัตถุจริงได้อย่างมีประสิทธิภาพ

ด้วยคุณลักษณะสำคัญของโมเดลแกนในการสร้างโมเดล เพื่อเรียนรู้การสร้างเพื่อเลียนแบบวัตถุได้อย่างมีประสิทธิภาพ ดังนั้น จึงมีการใช้งานโมเดลแกนในการสร้างโมเดลเพื่อเลียนแบบวัตถุชนิดอื่น ๆ โดยการเลียนแบบเสียงของมนุษย์ ก็เป็นอีกหนึ่งการประยุกต์การใช้งาน โดยใช้หลักการในการเรียนรู้วิธีสร้างเสียงจากเสียงทั่ว ๆ ไป เพื่อเลียนแบบเสียงของเป้าหมาย

2. การทดสอบคุณภาพของเสียง

ในการดำเนินการวิจัยเพื่อสร้างหรือเลียนแบบเสียงขึ้นมานั้น จำเป็นอย่างยิ่งที่ต้องมีการวัดและประเมินผลคุณภาพของเสียงที่ถูกสร้างขึ้นในแต่ละโมเดล โดยมาตรฐานสากลที่เป็นที่ยอมรับอย่างแพร่หลาย ซึ่งใช้ในการทดสอบคุณภาพของเสียงคือ คะแนนค่าเฉลี่ยความคิดเห็น (Mean Opinion Score: MOS) ผ่านการวัดค่าความพึงพอใจในคุณภาพของเสียงที่กลุ่มตัวอย่างได้รับฟัง (Streijl, Winkler, and Hands, 2016, p.213-227) โดยกลุ่มตัวอย่างจะให้คะแนนตามระดับของคุณภาพเสียง โดยมีค่าคะแนนในแต่ละระดับและการแปลความหมายดังแสดงในตารางที่ 1

ตารางที่ 1 การแปลความหมายของ MOS

MOS	คุณภาพ	ลักษณะ
5	ยอดเยี่ยม (Excellent)	ดีเยี่ยมจนแยกไม่ออก (Imperceptible)
4	ดี (Good)	ใช้ได้ดี (Perceptible)
3	พอใช้ (Fair)	น่าหงุดหงิดเล็กน้อย (Slightly Annoying)
2	ไม่ดี (Poor)	ไม่น่าพอใจ (Annoying)
1	แย่ (Bad)	ไม่พอใจ (Very Annoying)

ที่มา: Streijl, Winkler, and Hands, 2016

วิธีการดำเนินการวิจัย

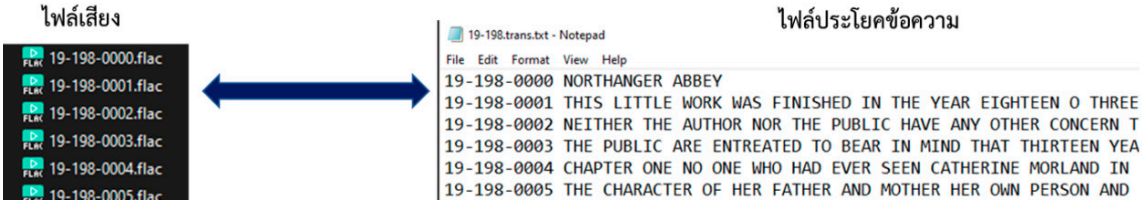
การดำเนินการวิจัยเพื่อศึกษารูปแบบการประยุกต์ใช้งานโมเดลในการสังเคราะห์และเลียนแบบเสียงเพื่อใช้ในงานด้านสงครามไซเบอร์ ใช้รูปแบบการประเมินผลจากการทดลอง (Experimental Evaluations) และการศึกษาจากกลุ่มตัวอย่างผู้ใช้งาน (User Study) โดยวิธีการดำเนินการวิจัยมีรายละเอียดดังนี้

1. การศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องเพื่อทดสอบหาโมเดลที่เหมาะสม

ผู้วิจัยได้ทำการศึกษาและทบทวนวรรณกรรมที่เกี่ยวข้องกับการสังเคราะห์และแปลงเสียงมนุษย์ที่ได้รับการตีพิมพ์ในที่ประชุมทางวิชาการที่มีชื่อเสียงและเป็นที่ยอมรับว่าเป็นโมเดลที่มีศักยภาพสูง (State of the Art Models) โดยพบว่าในการสร้างเสียงสังเคราะห์โดยคอมพิวเตอร์เพื่อเลียนแบบเสียงมนุษย์มีรูปแบบการดำเนินการ 2 รูปแบบประกอบไปด้วย

1.1 การใช้โมเดลในการแปลงข้อความเป็นเสียงมนุษย์

การแปลงข้อความเป็นเสียงมนุษย์ (Text to Speech) เป็นเทคโนโลยีการใช้การเรียนรู้ของเครื่อง (Machine Learning) รวมถึงการเรียนรู้เชิงลึกเพื่อเรียนรู้การออกเสียงของแต่ละคำ การเชื่อมประโยคข้อความจากข้อมูลคู่ขนานที่ประกอบไปด้วย เสียงการอ่านประโยคและประโยคข้อความ ดังแสดงในภาพที่ 3



ภาพที่ 3 ข้อมูลคู่ขนานที่ประกอบไปด้วยไฟล์เสียงและไฟล์ประโยคข้อความที่เกี่ยวข้อง

โดยในการศึกษานี้ คณะผู้วิจัยมุ่งเน้นศึกษาใน 2 โมเดล ประกอบไปด้วย

1.1.1 โมเดล Voice Clone ซึ่งใช้รูปแบบการเรียนรู้เชิงลึกผ่านโมเดล โดยการใช้ข้อมูลคู่ขนาน 2 ส่วน คือไฟล์เสียงและไฟล์ข้อความเป็นข้อมูลสำหรับการฝึกฝนโมเดล จากนั้นโมเดลจะทำการแปลงเป็นไฟล์ข้อความเป็นวิธีการออกเสียงหรือโฟนีม (Phoneme) เพื่อให้โมเดลเปรียบเทียบและเรียนรู้จากไฟล์เสียงต้นแบบจากมนุษย์ (Jia et al., 2018, p.1-11) ดังแสดงในภาพที่ 4 โดยกระบวนการเรียนรู้ดังกล่าวจะถูกดำเนินการซ้ำ ๆ กันด้วยไฟล์คู่ขนานจำนวนมหาศาลจากไฟล์เสียงพูดของบุคคลและประโยคการพูดที่แตกต่างกันออกไป จนโมเดลเรียนรู้กระบวนการสังเคราะห์เสียงให้สอดคล้องกับเสียงของบุคคลต้นแบบและประโยคข้อความ โดยผู้วิจัยได้ใช้ชุดข้อมูลคู่ขนานของ LibriSpeech[8] ในการฝึกฝนโมเดล โดยชุดข้อมูลดังกล่าวเป็นชุดข้อมูลขนาดใหญ่ที่ประกอบด้วยเสียงพูดของมนุษย์และประโยคข้อความที่เกี่ยวข้อง มีความยาวรวมทั้งสิ้นประมาณ 1,000 ชั่วโมง

Raw Text Sequence: "Thailand is the land of smile"
Phoneme Sequence: {T AY1 L AE2 N D IH0 Z DH AH0 L AE1 N D AH0 V S M AY1 L}

ภาพที่ 4 รูปแบบการแปลงประโยคข้อความจากไฟล์เสียงให้อยู่ในมาตรฐานโฟนีม

1.1.2 โมเดล Fast Speech 2 เป็นอีกหนึ่งโมเดลใช้รูปแบบการเรียนรู้เชิงลึกเช่นเดียวกับโมเดล Voice Clone รวมถึงขั้นตอนการดำเนินการฝึกฝนโมเดลที่ต้องการข้อมูลคู่ขนานและรูปแบบโฟนีมเพื่อเป็นข้อมูลนำเข้า (Ren et al., 2021, p.1-15) โดยโมเดล Fast Speech 2 ใช้โครงข่ายใยประสาทเทียมแบบคอนโวลูชัน (Convolution Neural Network: CNN)

1.2 การใช้โมเดลแกนในการแปลงเสียงมนุษย์

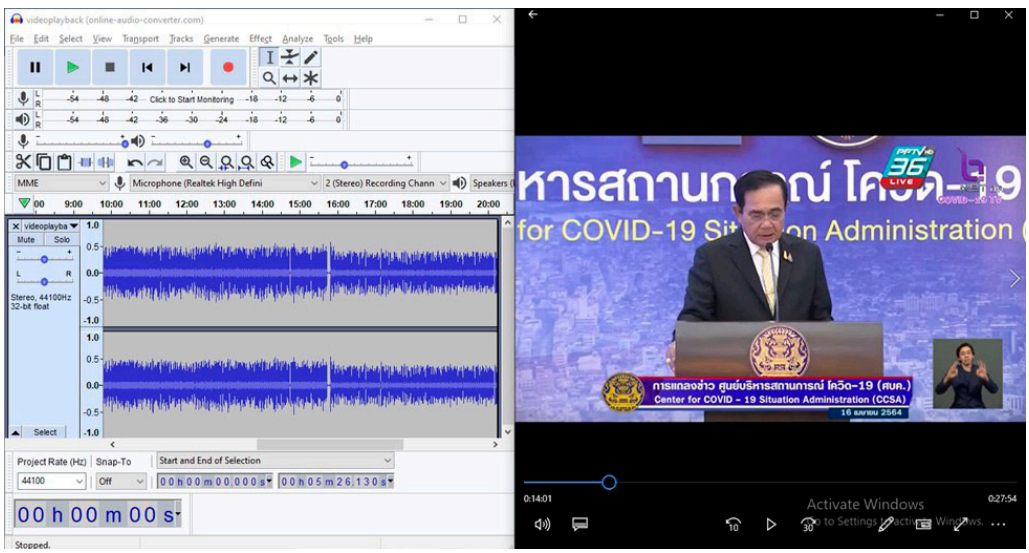
ผู้วิจัยได้เลือกใช้โมเดลแกน ในการเลียนแบบเสียงของมนุษย์ (Kameoka, Kaneko, Tanaka, & Hojo 2018, p.266-273) (Keneko & Kameoka, 2018, p.2100-2104) เพื่อแก้ไขข้อจำกัดของการใช้โมเดลการแปลงข้อความเป็นเสียงมนุษย์ โดยโมเดลที่ถูกนำมาใช้ในการทดสอบและประเมินผลในงานวิจัยครั้งนี้

[8] ชุดข้อมูลคู่ขนานของ LibriSpeech จาก https://www.openslr.org/12

ประกอบไปด้วย StarGAN และ CycleGAN สำหรับหลักการพื้นฐานของโมเดล สถาปัตยกรรมของโมเดล รวมถึงขั้นตอนในการฝึกฝนโมเดลทั้งสอง ได้ถูกอธิบายไว้ในส่วนของทฤษฎีที่เกี่ยวข้อง

2. การเตรียมข้อมูลในการฝึกฝนโมเดล

ข้อมูลที่ใช้ในการฝึกฝนโมเดลแกนเพื่อใช้ในการสร้างเสียงเลียนแบบเพื่อใช้ในสงครามไซเบอร์ ได้ถูกสร้างขึ้นมาใหม่โดยยึดมาตรฐานตามที่ระบุไว้ในบทความวิจัยที่ได้รับการตีพิมพ์ของโมเดลทั้งสอง โดยผู้วิจัยได้ทำการสร้างชุดข้อมูลที่มีความเหมาะสมเพื่อใช้ในการทดสอบคุณภาพของเสียงที่ถูกสร้างขึ้น รวมถึงศักยภาพในการใช้เป็นเครื่องมือในงานด้านสงครามไซเบอร์ ในแง่ของความสามารถในการหลอกลวงเพื่อสร้างข่าวปลอมในบริบทการใช้งานของประเทศไทย ดังนั้น ข้อมูลเสียงที่ใช้ในการฝึกฝนและสร้างโมเดลจึงเป็นข้อความเสียงภาษาไทย โดยผู้วิจัยได้ใช้ไฟล์เสียงต้นแบบเป็นบุคคลทั่วไป และไฟล์เสียงเป้าหมายเป็นบุคคลที่สังคมรู้จัก รวมถึงพิจารณาในบริบทที่ว่า หากไฟล์เสียงของบุคคลเป้าหมายถูกบิดเบือนอาจก่อให้เกิดความเสียหายต่อความมั่นคงของชาติ โดยเสียงของบุคคลเป้าหมายที่ผู้วิจัยเลือกจำนวน 2 คน ประกอบไปด้วย พลเอก ประยุทธ์ จันทร์โอชา นายกรัฐมนตรี และ ดร.ทักษิณ ชินวัตร อดีตนายกรัฐมนตรี เนื่องจากบุคคลทั้งสองเป็นบุคคลที่มีอิทธิพลในทางความคิดทางการเมือง รวมถึงสถานการณ์ทางการเมืองในปัจจุบันที่มีปัจจัยเสี่ยง ที่หากเสียงของบุคคลทั้งสองถูกบิดเบือนย่อมส่งผลโดยตรงต่อความมั่นคงของชาติ



ภาพที่ 5 กระบวนการสร้างไฟล์เสียงเป้าหมาย[9]

[9] ผู้วิจัยได้คำนึงถึงจริยธรรมในการทำวิจัย รวมถึงการใช้ข้อมูลส่วนบุคคล ดังนั้นการบันทึกเสียงเป้าหมายของบุคคลสำคัญทั้ง 2 ท่าน จึงมาจากเสียงที่ถูกเผยแพร่จากสื่อสาธารณะเท่านั้น สำหรับการบันทึกเสียงต้นแบบจากนักเรียนนายเรืออากาศก็ได้รับการยินยอมจากนักเรียนนายเรืออากาศผู้ทำการบันทึกเสียงเป็นที่เรียบร้อย

การเตรียมชุดข้อมูลไฟล์เสียงภาษาไทย มีขั้นตอน ดังนี้

2.1 ไฟล์เสียงต้นแบบ (Source) เป็นไฟล์เสียงที่เป็นตัวแทนของบุคคลทั่วไป ซึ่งในกรณีนี้ผู้วิจัยได้ใช้ไฟล์เสียงของนักเรียนนายเรืออากาศ ผ่านการบันทึกเสียงการพูดในประโยคที่แตกต่างตั้งแต่ประโยคที่สั้นไปจนถึงประโยคที่ยาว จำนวน 200 ไฟล์เสียง ซึ่งแต่ละประโยคมีความยาวของประโยคไม่เกิน 30 วินาที โดย 100 ไฟล์เสียงแรก ใช้เพื่อแปลงเป็นเสียงของ พลเอก ประยุทธ์ จันทร์โอชา และ 100 ไฟล์เสียงถัดมา ใช้เพื่อแปลงเป็นเสียงของ ดร.ทักษิณ ชินวัตร

2.2 ไฟล์เสียงเป้าหมาย (Target) หรือเสียงที่บุคคลทั่วไปต้องการจะปลอมแปลง ผู้วิจัยได้พิจารณาหาแหล่งข้อมูลไฟล์เสียงการให้สัมภาษณ์ แถลงการณ์ของบุคคลทั้งสองผ่านแพลตฟอร์ม YouTube ดังแสดงในภาพที่ 5 จากนั้นจึงทำการถอดไฟล์เสียงออกมาและสร้างไฟล์เสียงของคำพูดจากประโยคที่แตกต่างกันออกมาเป็นไฟล์จำนวน 100 ไฟล์เสียงของ พลเอก ประยุทธ จันทร์โอชา (เป้าหมายที่ 1) และ 100 ไฟล์เสียงของ ดร.ทักษิณ ชินวัตร (เป้าหมายที่ 2) ซึ่งไฟล์เสียงแต่ละประโยคมีความยาวของประโยคไม่เกิน 10 วินาที

3. การวิเคราะห์และประเมินผล

เมื่อดำเนินการสร้างเสียงแปลงเพื่อเลียนแบบเสียงเป้าหมายเป็นที่เรียบร้อยและผู้วิจัยได้ทำการวัดและประเมินผลผ่านกลุ่มตัวอย่างใน 2 ตัวชี้วัด ได้แก่ คุณภาพของเสียงสังเคราะห์ (กลุ่มตัวอย่าง A) และประสิทธิผลในการหลอกลวงผู้ฟัง โดยมีกลุ่มตัวอย่าง จำนวน 2 กลุ่มในแต่ละตัวชี้วัด กลุ่มตัวอย่าง A แทนกลุ่มตัวอย่างสำหรับการประเมินคุณภาพของเสียงสังเคราะห์จำนวน 30 คน และกลุ่มตัวอย่าง B แทนกลุ่มตัวอย่างสำหรับการประเมินประสิทธิผลในการหลอกลวงผู้ฟัง และกลุ่มตัวอย่าง A และ B ต้องไม่มีประชากรที่ซ้ำกัน (Mutually Exclusive) เพื่อหลีกเลี่ยงปัญหาที่อาจเกิดจากอคติของการได้ยิน (Bias)

สำหรับเหตุผลในการใช้กลุ่มตัวอย่างเป็นนักเรียนนายเรืออากาศเป็นหลัก เนื่องจากการดำเนินการวิจัยดำเนินการภายใต้สถานการณ์การแพร่ระบาดของโรคโควิด-19 ที่ค่อนข้างรุนแรง จึงจำเป็นต้องคัดเลือกกลุ่มตัวอย่างที่สามารถดำเนินการได้ในระบบปิด รวมถึงขั้นตอนการดำเนินการเก็บข้อมูลที่ผู้วิจัยจำเป็นต้องดำเนินการอย่างรัดกุม มีการสังเกตและควบคุมการทดสอบอย่างใกล้ชิด เพื่อให้มั่นใจว่า ผู้เข้ารับการทดสอบอยู่ภายใต้สภาพแวดล้อมที่ใกล้เคียงกับการรับรู้ข่าวสาร (รับฟังเสียงจริงและเสียงปลอมแปลง) ผ่านการใช้งานอินเทอร์เน็ตอย่างแท้จริง สำหรับการวัดและประเมินผลเสียงสังเคราะห์ในแต่ละตัวชี้วัดมีรายละเอียดดังนี้

3.1 การประเมินคุณภาพของเสียงสังเคราะห์

วัตถุประสงค์: เพื่อประเมินว่า เสียงที่สังเคราะห์จากการแปลงเสียงของโมเดล มีความเป็นธรรมชาติ รวมถึงสามารถใช้ในการสื่อสารได้ดีเพียงใด

กลุ่มตัวอย่าง: นักเรียนนายเรืออากาศ จำนวน 30 คน

รูปแบบการวัดผล: กลุ่มตัวอย่างแต่ละคนจะทำแบบประเมินจำนวน 2 ชุด แต่ละชุดแทนเสียงต้นแบบและเสียงสังเคราะห์ของบุคคลที่แตกต่างกันโดยชุดที่ 1 ประกอบด้วยเสียงต้นแบบของบุคคลที่ 1 จำนวน 1 เสียง และเสียงสังเคราะห์ (เสียงที่ถูกปลอมแปลงขึ้น) ของบุคคลที่ 1 จำนวน 5 เสียง และชุดที่ 2 ประกอบด้วยเสียงต้นแบบของบุคคลที่ 2 จำนวน 1 เสียง และเสียงสังเคราะห์ของบุคคลที่ 2 จำนวน 5 เสียง โดยเสียงสังเคราะห์จำนวน 5 เสียงมาจากการสุ่มโดยคอมพิวเตอร์จากเสียงสังเคราะห์จำนวน 10 เสียง สำหรับเกณฑ์การประเมินใช้มาตรวัดของลิเคิร์ท (Likert Scale) โดยระดับคะแนนใช้มาตรฐานการวัดคุณภาพเสียง (Mean Opinion Score: MOS) ที่ถูกปรับปรุงคำอธิบายให้เข้าใจได้ง่ายขึ้นดังแสดงในตารางที่ 2

ตารางที่ 2 เกณฑ์การวัดและประเมินคุณภาพของเสียงสังเคราะห์

ระดับคะแนน	การแปลความหมาย
5 - ดีเยี่ยม	เสียงมีความใกล้เคียงกับการสนทนาแบบตัวต่อตัวหรือการรับฟังจากเครื่องขยายเสียง
4 - ดี	เสียงยังมีความไม่สมบูรณ์บ้างแต่ยังคงชัดเจนใกล้เคียงกับเสียงสนทนาผ่านโทรศัพท์มือถือ
3 - พอใช้	เสียงมีความไม่สมบูรณ์ มีเสียงรบกวน ทำให้อาจก่อให้เกิดปัญหาในการสื่อสาร
2 - แย่	เสียงมีความไม่สมบูรณ์ จนเกือบไม่สามารถสื่อสารได้
1 - ไม่ดี	เสียงที่ได้ยินฟังไม่ออก ไม่สามารถสื่อสารได้

ที่มา: Streijl, Winkler, and Hands, 2016

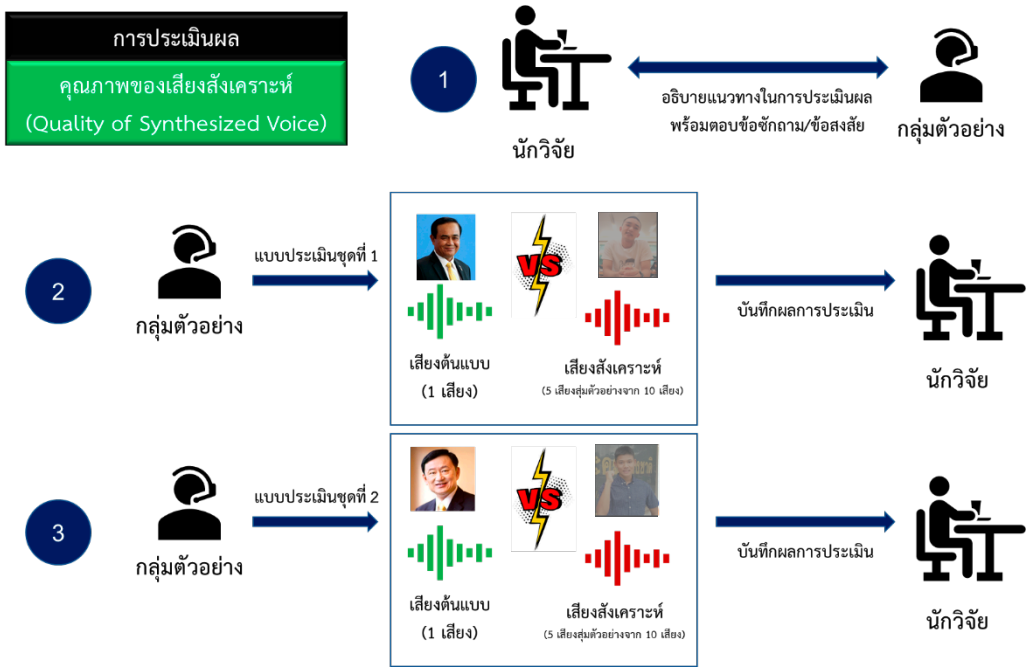
ภาพที่ 6 แสดงภาพรวมในการวัดผลและประเมินผลโดยผู้วิจัยเริ่มต้นจากการอธิบายวัตถุประสงค์ของการดำเนินการวิจัยให้แก่กลุ่มตัวอย่าง จากนั้นทำการเปิดตัวอย่างเสียงสังเคราะห์ทีละ 1 เสียงจากแบบประเมินชุดที่ 1 (เป้าหมายที่ 1) และเมื่อกลุ่มตัวอย่างฟังเสร็จในแต่ละเสียง กลุ่มตัวอย่างให้คะแนนคุณภาพเสียงโดยคณะผู้วิจัยจะเป็นผู้บันทึกคะแนนเองผ่านแบบประเมินคุณภาพของเสียงสังเคราะห์ จากนั้นผู้วิจัยดำเนินกระบวนการประเมินอีกครั้งกับแบบประเมินชุดที่ 2 (เป้าหมายที่ 2)

3.2 การประเมินประสิทธิผลในการหลอกลวงผู้ฟัง

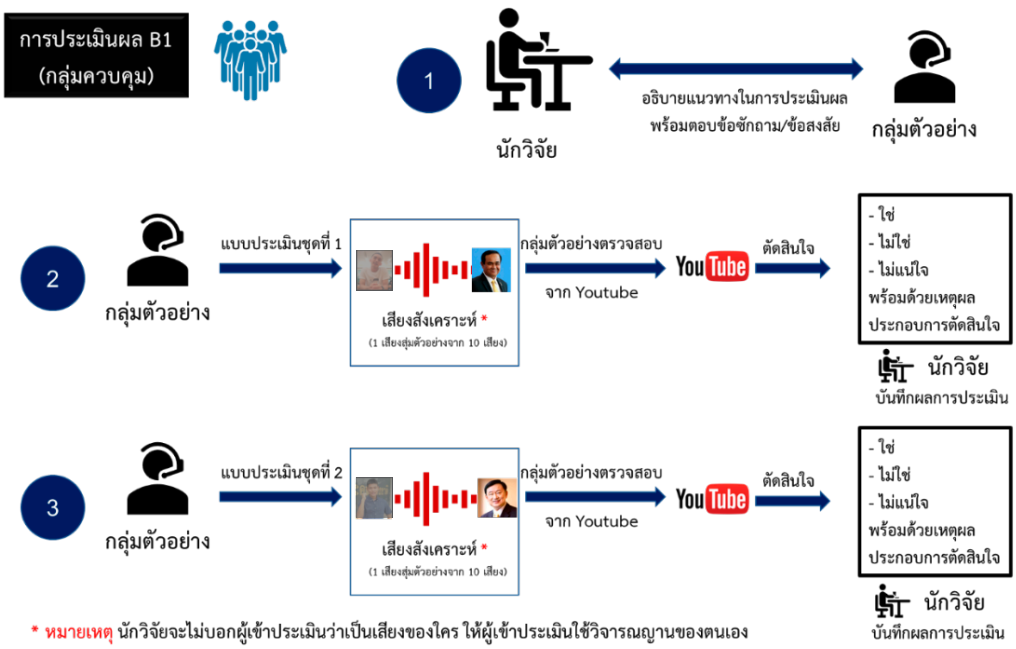
วัตถุประสงค์: เพื่อประเมินว่า เสียงที่สังเคราะห์ขึ้นมีศักยภาพในการหลอกลวงผู้ฟังให้เข้าใจผิดว่าเป็นเสียงต้นแบบมากน้อยเพียงใด

กลุ่มตัวอย่าง: นักเรียนนายเรืออากาศ จำนวน 30 คน โดยกลุ่มตัวอย่างถูกแบ่งเป็น 2 กลุ่ม ประกอบด้วย กลุ่มควบคุม (Controlled Group) จำนวน 15 คน และกลุ่มทั่วไป (General Group) จำนวน 15 คน โดยกลุ่มย่อยทั้ง 2 กลุ่มต้องไม่มีประชากรที่ซ้ำกัน

รูปแบบการวัดผล: เพื่อให้การดำเนินการวัดประสิทธิผลของเสียงสังเคราะห์มีความสอดคล้องกับสภาวะแวดล้อมที่เกิดขึ้นจริงในการรับรู้ข่าวสารของผู้ใช้งานอินเทอร์เน็ต ผู้วิจัยจึงได้ออกแบบการประเมินบนสมมติฐานที่ว่า กลุ่มตัวอย่างต้องใช้วิจารณญาณของตนเองในการตัดสินใจว่า เสียงดังกล่าวคือเสียงของบุคคลใดและตัดสินใจว่า ใช่/ไม่ใช่/ไม่แน่ใจ ดังนั้น ผู้วิจัยจะไม่แจ้งกับกลุ่มตัวอย่างว่าเสียงดังกล่าวเป็นเสียงของบุคคลใดมากไปกว่านั้น การแบ่งกลุ่มตัวอย่างออกเป็น 2 กลุ่มย่อย มุ่งเน้นศึกษาความเป็นไปได้ของพฤติกรรมของผู้ใช้งานอินเทอร์เน็ตที่แตกต่างกันออกไป ว่ามีผลต่อประสิทธิผลของการหลอกลวงโดยเสียงสังเคราะห์หรือไม่ โดยกลุ่มควบคุมแทนกลุ่มประชากรผู้ใช้งานอินเทอร์เน็ต ที่เมื่อได้ฟังเสียงสังเคราะห์แล้วมีโอกาสในการตรวจสอบข้อมูลกับเสียงจริงจากแหล่งข้อมูลอินเทอร์เน็ต เช่น YouTube ผ่าน Keyword ชื่อของเสียงต้นแบบ แล้วจึงตัดสินใจว่าเสียงสังเคราะห์ดังกล่าว ใช่/ไม่ใช่/ไม่แน่ใจ บุคคลเป้าหมายหรือไม่ ดังแสดงในภาพที่ 7 ในทางกลับกัน กลุ่มทั่วไปจะแทนกลุ่มประชากรผู้ใช้งานอินเทอร์เน็ตที่เมื่อได้ฟังเสียงสังเคราะห์แล้ว ต้องตัดสินใจในทันทีว่าเสียงสังเคราะห์ดังกล่าว ใช่/ไม่ใช่/ไม่แน่ใจ บุคคลเป้าหมายหรือไม่



ภาพที่ 6 ขั้นตอนในการการวัดและประเมินคุณภาพของเสียงสังเคราะห์



ภาพที่ 7 ขั้นตอนการวัดและประเมินประสิทธิผลการหลอกลวง (กลุ่มควบคุม)

ผลการวิจัย

ผู้วิจัยได้ดำเนินการตามขั้นตอนการดำเนินการวิจัย เพื่อศึกษาแนวทางการประยุกต์ใช้งานเทคโนโลยีการแปลงเสียงเชิงลึกโดยในแง่ของข้อดี/ข้อจำกัด ของโมเดลที่มีอยู่ในปัจจุบัน รวมถึงโมเดลที่มีความเหมาะสมต่อการประยุกต์ใช้ในงานด้านสงครามไซเบอร์และการวัดผลประเมินผลโมเดล โดยผลการวิจัยมีรายละเอียดดังนี้

1. ผลการศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องเพื่อทดสอบหาโมเดลที่เหมาะสม

1.1 การใช้โมเดลการแปลงข้อความเป็นเสียงมนุษย์

ผลการศึกษาและทดสอบโมเดลการแปลงข้อความเป็นเสียงมนุษย์จำนวน 2 โมเดล ได้แก่ Voice Clone และ Fast Speech 2 พบว่า โมเดลทั้งสองมีความเหมาะสมที่จะใช้งานกับงานด้านการแปลงข้อความเป็นเสียงมนุษย์ที่เป็นภาษาอังกฤษ เนื่องจากมีชุดข้อมูลคู่ขนานที่พร้อมใช้งาน แต่ยังไม่เหมาะสมกับการใช้งานด้านสงครามไซเบอร์โดยเฉพาะอย่างยิ่งในบริบทของการใช้ภาษาไทย ที่ในบางครั้งต้องการความรวดเร็วในการฝึกฝนโมเดลเนื่องจากต้องใช้ข้อมูลคู่ขนานในการฝึกฝนโมเดลจำนวนมาก รวมถึงข้อจำกัดในเชิงเทคนิคในการเตรียมข้อมูล เช่น การแปลงเสียงพูดเป็นไฟล์ประโยคที่ต้องใช้มนุษย์ดำเนินการและใช้เวลาที่ค่อนข้างนานในการเตรียมข้อมูล ซึ่งการที่ข้อมูลคู่ขนานของภาษาไทยยังมีไม่เพียงพอส่งผลให้คุณภาพของการสร้างเสียงภาษาไทยที่ยังมีความไม่เป็นธรรมชาติ ดังนั้น การใช้โมเดลการแปลงข้อความเป็นเสียงมนุษย์จึงจำเป็นต้องมีการศึกษา วิจัย และพัฒนาเพิ่มเติมในอนาคตต่อไป

1.2 โมเดลที่มีความเหมาะสมกับการใช้งานด้านสงครามไซเบอร์

ผลการศึกษาพบว่า เพื่อเป็นการก้าวข้ามขีดจำกัดในด้านความต้องการข้อมูลคู่ขนาน ข้อจำกัดด้านการใช้งานภาษาไทย รวมถึงระยะเวลาในการเตรียมข้อมูลของการใช้งานโมเดลการแปลงข้อความเป็นเสียงมนุษย์ คณะผู้วิจัยได้เลือกการใช้โมเดลแกน ในการเลียนแบบเสียงของมนุษย์ ซึ่งประกอบไปด้วยโมเดล StarGAN-VC และโมเดล CycleGAN-VC เนื่องจากทั้ง 2 โมเดล เป็นโมเดลที่ได้รับการยอมรับว่าเป็นหนึ่งในโมเดลที่ดีที่สุด (State-of-the-Art Models) ของการแปลงเสียง (Voice Conversion) รวมถึงความต้องการในการใช้ไฟล์เสียงสำหรับการฝึกฝนเพียง 100 ไฟล์เสียง/เป้าหมาย และไม่จำเป็นต้องมีการถอดประโยคข้อความออกมา ทำให้มีความสะดวกและรวดเร็วในการสร้างเสียงปลอมแปลง โดยผู้วิจัยได้ทำการทดสอบทั้ง 2 โมเดล เพื่อทำการศึกษาวิธีการตั้งแต่การเตรียมชุดข้อมูล ที่ใช้สำหรับการฝึกฝนโมเดลรวมแนวทางการใช้โมเดลในสถานการณ์จริง และการทดสอบประสิทธิภาพและประสิทธิผลของโมเดล

2. ผลการศึกษาแนวทางการประยุกต์ใช้งานโมเดล

2.1 การเข้าถึงซอร์สโค้ด

เนื่องจากงานโมเดลการแปลงเสียงทั้ง 2 โมเดล เป็นผลผลิตจากการวิจัยและพัฒนา และเป็นโอเพนซอร์สและสามารถเข้าถึงได้/แก้ไข/เปลี่ยนแปลงได้ผ่านระบบฐานข้อมูลงานวิจัยโดยไม่เสียค่าใช้จ่าย

2.2 ความต้องการพื้นฐานของระบบ

ซอร์สโค้ดสำหรับการใช้งานทั้ง 2 โมเดล ถูกออกแบบมาเพื่อรองรับการใช้งานทั้งกับระบบคอมพิวเตอร์ที่ไม่มีหน่วยประมวลผลกราฟิกซึ่งในกรณีนี้ ระบบจะใช้หน่วยประมวลผลกลาง (CPU) ในการฝึกฝนโมเดล ในขณะเดียวกัน หากระบบคอมพิวเตอร์ใช้หน่วยประมวลผลกราฟิก (GPU) ข้อมูลที่ใช้ในการฝึกฝนโมเดลจะถูกถ่ายโอนและประมวลผลในหน่วยประมวลผลกราฟิก โดยปัจจัยดังกล่าวส่งผลโดยตรงต่อระยะเวลาที่ใช้ในการฝึกฝนโมเดล โดยในการฝึกฝนโมเดลของระบบคอมพิวเตอร์ที่ใช้หน่วยประมวลผลกลาง และหน่วยประมวลผลกราฟิกอยู่ในช่วงภายใน 3 วัน ดังนั้น การใช้ประโยชน์เทคโนโลยีแปลงเสียงด้วยโมเดลทั้งสอง มีความเป็นไปได้ในทางปฏิบัติ ดังแสดงในตารางที่ 3

ตารางที่ 3 ระยะเวลาในการฝึกฝนโมเดล

โมเดล	ระยะเวลาที่ใช้ในการฝึกฝนโมเดล (ชั่วโมง)	
	หน่วยประมวลผลกลาง (CPU)	หน่วยประมวลผลกราฟิก (GPU)
StarGAN-VC	~50	~11
CycleGAN-VC	~70	~14

3. การประเมินผลเบื้องต้นสำหรับเสียงแปลงที่ได้จากโมเดล

หลังจากที่ผู้วิจัยทำการสร้างและฝึกฝนโมเดล StarGAN-VC และ CycleGAN-VC เรียบร้อยแล้ว คณะผู้วิจัยได้ทำการตรวจสอบคุณภาพเสียงของเสียงสังเคราะห์ในเบื้องต้นพบว่า คุณภาพของเสียงไม่มีความแตกต่างกันอย่างมีนัยยะสำคัญ แต่เมื่อพิจารณาแล้วโมเดล StarGAN-VC ใช้เวลาในการฝึกฝนโมเดลน้อยกว่าเมื่อเทียบกับโมเดล CycleGAN-VC ดังนั้น คณะผู้วิจัยจึงเลือกใช้เสียงสังเคราะห์ของบุคคลเป้าหมายจากโมเดล StarGAN-VC เป็นหลัก คณะผู้วิจัยได้ดำเนินการจัดทำตัวอย่างเสียงที่ถูกสังเคราะห์ขึ้น และแสดงผลผ่านหน้าเว็บไซต์ https://sites.google.com/view/cybervoice

4. ผลการประเมินผลโมเดลการสังเคราะห์เสียง

4.1 ผลการประเมินผลคุณภาพของเสียงสังเคราะห์

เสียงสังเคราะห์ของบุคคลเป้าหมายที่ 1 (พลเอก ประยุทธ์ จันทร์โอชา) มีค่าเฉลี่ยของ MOS อยู่ที่ 3.59 และค่าเบี่ยงเบนมาตรฐานอยู่ที่ 0.97 ซึ่งแปลความได้ว่าเสียงดังกล่าว อยู่ในระดับค่อนข้างดี โดยอาจมีความไม่สมบูรณ์อยู่บ้าง ใกล้เคียงกับเสียงสนทนาผ่านโทรศัพท์เคลื่อนที่ และอาจก่อให้เกิดปัญหาในการสื่อสารอยู่บ้าง ในขณะเดียวกัน ผลการประเมินผลคุณภาพของเสียงสังเคราะห์ของบุคคลเป้าหมายที่ 2 (ดร.ทักษิณ ชินวัตร) มีค่าเฉลี่ยของ MOS อยู่ที่ 2.62 และค่าเบี่ยงเบนมาตรฐานอยู่ที่ 0.94 ซึ่งแปลความได้ว่าเสียงดังกล่าว อยู่ในระดับค่อนข้างพอใช้ เสียงมีความไม่สมบูรณ์ซึ่งอาจก่อให้เกิดปัญหาในการสื่อสารอยู่บ้าง

4.2 ประสิทธิผลในการหลอกลวงผู้ฟัง (กลุ่มควบคุม)

ผู้วิจัยพบว่ากลุ่มตัวอย่างจำนวน 3 คน จากกลุ่มตัวอย่างทั้งหมด จำนวน 15 คน หรือคิดเป็นร้อยละ 20 ของกลุ่มตัวอย่างถูกหลอกลวงด้วยเสียงแปลงสำหรับเสียงของบุคคลเป้าหมายที่ 1 ในขณะเดียวกันพบว่า กลุ่มตัวอย่าง จำนวน 4 คน จากกลุ่มตัวอย่างทั้งหมด 15 คน หรือคิดเป็น ร้อยละ 27 ของกลุ่มตัวอย่างถูกหลอกลวงด้วยเสียงแปลงสำหรับเสียงของบุคคลเป้าหมายที่ 2 ดังแสดงในตารางที่ 4

ตารางที่ 4 ผลการประเมินประสิทธิผลในการหลอกลวงผู้ฟัง

กลุ่มตัวอย่าง	เสียงแปลงของ	ใช่ (ถูกหลอกลวง)	ไม่ใช่	ไม่แน่ใจ
กลุ่มควบคุม	เป้าหมายที่ 1 (พลเอก ประยุทธ์ฯ)	3	11	1
	เป้าหมายที่ 2 (ดร.ทักษิณฯ)	4	9	2
กลุ่มทั่วไป	เป้าหมายที่ 1 (พลเอก ประยุทธ์ฯ)	3	11	1
	เป้าหมายที่ 2 (ดร.ทักษิณฯ)	6	7	2

4.3 ประสิทธิผลในการหลอกลวงผู้ฟัง (กลุ่มทั่วไป)

พบว่า กลุ่มตัวอย่าง จำนวน 3 คน จากกลุ่มตัวอย่างทั้งหมด จำนวน 15 คน หรือคิดเป็น ร้อยละ 20 ของกลุ่มตัวอย่างถูกหลอกลวงด้วยเสียงแปลง สำหรับเสียงของบุคคลเป้าหมายที่ 1 ที่น่าสนใจ คือ กลุ่มตัวอย่างจำนวน 6 คน จากกลุ่มตัวอย่างทั้งหมด 15 คน หรือคิดเป็นร้อยละ 40 ของกลุ่มตัวอย่างถูกหลอกลวงด้วยเสียงแปลง สำหรับเสียงของบุคคลเป้าหมายที่ 2 ดังแสดงในตารางที่ 4

จากผลการประเมินพบว่า เสียงของบุคคลทั่วไปที่ถูกแปลงเป็นเสียงของบุคคลเป้าหมาย มีศักยภาพในการหลอกลวงกลุ่มตัวอย่างให้หลงเชื่อว่าเป็นเสียงของบุคคลเป้าหมายตั้งแต่ ร้อยละ 20-40 โดยอัตราดังกล่าวถือว่าเป็นอัตราที่สูงและบ่งบอกถึงความมีศักยภาพของเสียงแปลงในงานด้านสงครามไซเบอร์ในแง่ของการสร้างข่าวปลอม เนื่องจากกลุ่มบุคคลที่หลงเชื่ออาจมีแนวโน้มในการแชร์เสียงปลอมแปลงดังกล่าวในสื่อสังคมออนไลน์และอาจมีการแชร์ข่าวปลอมดังกล่าวต่อเนื่องเป็นวงกว้างได้ เป็นที่น่าสนใจว่า จากผลการทดลองพบว่า คุณภาพของเสียงสังเคราะห์มิได้มีผลกระทบโดยตรงต่อประสิทธิภาพในการหลอกลวง ดังจะเห็นได้จากเสียงสังเคราะห์ของบุคคลเป้าหมายที่ 2 ที่ถึงแม้จะมี MOS ต่ำกว่าเสียงบุคคลเป้าหมายที่ 1 แต่มีประสิทธิภาพในการหลอกลวงที่สูงกว่า ซึ่งเมื่อพิจารณาแล้วสาเหตุอาจมีปัจจัยมาจากประสบการณ์/ความคุ้นเคยของกลุ่มตัวอย่าง เนื่องจากบุคคลเป้าหมายที่ 1 เป็นนายกรัฐมนตรีคนปัจจุบันของประเทศ จึงมีแนวโน้มที่กลุ่มตัวอย่างจะได้รับฟังเสียงจริงของบุคคลเป้าหมายที่ 1 จนสามารถจดจำข้อมูลของลักษณะ จังหวะและเนื้อเสียงการพูดได้ดีกว่าบุคคลเป้าหมายที่ 2 จนส่งผลให้สามารถแยกแยะว่าเสียงดังกล่าวเป็นเสียงจริงหรือเสียงสังเคราะห์ที่ถูกปลอมแปลงขึ้น มากไปกว่านั้น ผลการวิจัยยังชี้ให้เห็นว่า การที่ผู้ใช้งานอินเทอร์เน็ตได้มีโอกาสตรวจสอบข้อเท็จจริง (กลุ่มควบคุม) ก่อนตัดสินใจ มีแนวโน้มที่จะถูกหลอกลวงจากเสียงแปลงที่น้อยกว่า ดังนั้นการส่งเสริมให้ผู้ใช้งานอินเทอร์เน็ตมีวิจารณญาณผ่านการตรวจสอบข้อเท็จจริงจากแหล่งข้อมูลอื่น ๆ ถือเป็นกลไกสำคัญในการปกป้องการโจมตีด้านข้อมูลข่าวสารผ่านสงครามไซเบอร์ได้ ทั้งนี้ การศึกษาวิจัยเพิ่มเติมในปัจจัยที่มีผลกระทบต่อวิจารณญาณ รวมถึงแนวทางการเสริมสร้างวิจารณญาณในการรับข้อมูลข่าวสารบนมิติไซเบอร์ จึงเป็นหัวข้องานวิจัยหนึ่งที่มีความสำคัญในอนาคต

สรุปและอภิปรายผล

ผลการศึกษาและทดสอบโมเดล พบว่า โมเดลที่มีความเหมาะสมสำหรับการประยุกต์ใช้เทคโนโลยีการเลียนแบบเสียงเชิงลึก คือ โมเดลที่ใช้เทคนิคแกน ได้แก่ StarGAN-VC และ CycleGAN-VC ที่มีจุดเด่นที่สามารถใช้เพียงไฟล์เสียงของบุคคลทั่วไปและบุคคลเป้าหมายที่เราต้องการเลียนแบบเท่านั้น ลดระยะเวลาและความซับซ้อนในการจัดทำชุดข้อมูล ความต้องการทรัพยากรคอมพิวเตอร์ที่ไม่สูงมากนัก รวมถึงการสร้างโมเดลที่ใช้ระยะเวลาเร็วสุดเพียงไม่กี่ชั่วโมงและช้าสุดไม่เกิน 3 วัน ทำให้โมเดลดังกล่าวมีความเป็นไปได้ในทางปฏิบัติในหน่วยงานด้านความมั่นคง

ผลการทดสอบคุณภาพของเสียงสังเคราะห์และประสิทธิผลในการหลอกลวงยังชี้ให้เห็นว่า โมเดลมีศักยภาพในการปฏิบัติการบนสงครามไซเบอร์ ในแง่การสร้างข่าวปลอมเพื่อหลอกลวงผู้ใช้งานอินเทอร์เน็ตเพื่อให้เกิดความเข้าใจผิดว่าเสียงที่ถูกปลอมแปลงคือเสียงของบุคคลเป้าหมายจริง โดยผลการศึกษาจากกลุ่มตัวอย่าง พบว่ากลุ่มตัวอย่าง ร้อยละ 20 ถึง ร้อยละ 40 ถูกหลอกลวงโดยเสียงแปลงซึ่งในแง่ของความมั่นคงปลอดภัยไซเบอร์ถือว่า จำนวนผู้ที่ถูกหลอกลวงดังกล่าว ถือเป็นภัยคุกคามที่มีนัยยะสำคัญ ซึ่งผลการค้นพบดังกล่าวได้เน้นย้ำและสร้างความตระหนักต่อภัยคุกคามรูปแบบใหม่ รวมถึงการแสวงหาแนวทางการตรวจจับและป้องกันเสียงปลอมแปลงที่อาจกระทบต่อความมั่นคงของชาติในอนาคต

ข้อเสนอแนะ

จากผลการศึกษาดังกล่าวชี้ให้เห็นถึงศักยภาพของการใช้เทคโนโลยีการเลียนแบบเสียงเชิงลึกในงานด้านความมั่นคงโดยเฉพาะในรูปแบบของสงครามไซเบอร์ทั้งในเชิงรุก (การโจมตี) และเชิงรับ (การป้องกัน)

สำหรับการใช้งานในเชิงรุก ฝ่ายโจมตีสามารถสร้างข่าวสารอันเป็นเท็จผ่านการปลอมแปลงเสียงของบุคคลสำคัญของชาติหรือผู้บริหารองค์กร และทำการเผยแพร่ผ่านสื่อสังคมออนไลน์ เช่น Facebook Twitter และ Tik-Tok เป็นต้น พร้อมทั้งสร้างเนื้อหาเพิ่มเติมประกอบเสียงดังกล่าว เช่น *"ไฟล์เสียงลับหลุดจากห้องประชุมด้านความมั่นคง"* เพื่อสร้างความเสียหายและลดทอนความน่าเชื่อถือของบุคคลดังกล่าว รวมถึงองค์กรหรือหน่วยงานที่เกี่ยวข้อง

สำหรับการใช้งานเชิงรับ ฝ่ายความมั่นคงที่ดูแลความมั่นคงปลอดภัยไซเบอร์ จำเป็นอย่างยิ่งที่ต้องมีโมเดลสำหรับการตรวจสอบและดักจับเสียงเลียนแบบดังกล่าว รวมถึงการเตรียมการในการชี้แจง อธิบาย ผ่านสื่อสารมวลชนพร้อมทั้งหลักฐานการพิสูจน์ โดยในปัจจุบันเริ่มมีงานวิจัยและเทคโนโลยีการตรวจจับเสียงเลียนแบบ ดังนั้นการให้ความสำคัญในการศึกษาและพัฒนาโมเดลการตรวจจับเสียงเลียนแบบจึงเป็นสิ่งที่มีความจำเป็นเร่งด่วน

เอกสารอ้างอิง

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives, 31*(2), 21-36.

Cooke, N. A. (2017). Posttruth, truthiness, and alternative facts: Information behavior and critical information consumption for a new age. *The library quarterly, 87*(3), 211-221.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139-144.

Jia, Y., Zhang, Y., Weiss, R.J., Wang, Q., Shen, J., Ren, F., ...Wu, Y. (2018). Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. *NeurIPS,* 1-11.

Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2018). Stargan-vc: Non- parallel many-to-many voice conversion using star generative adversarial networks. In The Institute of Electrical and Electronics Engineers (Ed.), *2018 IEEE Workshop Spoken Language Technology Workshop (SLT)* (p.266-273). IEEE

Kaneko, T., & Kameoka, H. (2018). Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *26th European Signal Processing Conference (EUSIPCO)* (p.2100-2104).

Karlsen, R., & Aalberg, T. (2021). Social Media and Trust in News: An Experimental Study of the Effect of Facebook on News Story Credibility. *Digital Journalism,* 1-17.

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat?. *Business Horizons. 63*(2), 135-146

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature, 521*(7553), 436-444.

Li, J. (2018). Cyber security meets A.I.: A survey. *Frontiers of Information Technology & Electronic Engineering, 19*(12), 1462-1474.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech : an asr corpus based on public domain audio books. In The Institute of Electrical and Electronics Engineers (Ed.). *2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (p.5206-5210).

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., ...Liu, T. (2021). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech.

Sindermann, C., Cooper, A., & Montag, C. (2020). A short review on susceptibility to falling for fake political news. *Current opinion in psychology, 36,* 44-48.

Streijl, R. C., Winkler, S., & Hands, D. S. (2016). Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems, 22*(2), 213-227.