การประยุกต์ใช้เทคโนโลยีปัญญาประดิษฐ์เชิงกำเนิดในการแปลงเสียง
The Application of Generative Artificial Intelligence Technology in Voice Conversion

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

[1]สำนักบัณฑิตศึกษา โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช

[2]ภาควิชาคอมพิวเตอร์ โรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราช

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

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

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

anon_b@rtaf.mi.th

Received 03 January 2025

Revised 07 April 2025

Accepted 17 April 2025

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์ 1) เพื่อศึกษาแนวทางการประยุกต์เทคโนโลยีปัญญาประดิษฐ์ที่มีความเหมาะสมในการปลอมแปลงเสียงบุคคล 2) เพื่อพัฒนาแบบจำลองการปลอมแปลงเสียงด้วยเทคโนโลยีปัญญาประดิษฐ์เชิงกำเนิดและศึกษาแนวทางการปรับแต่งแบบจำลองให้มีความเหมาะสมกับการใช้งานบนมิติไซเบอร์ 3) เพื่อประเมินประสิทธิภาพและศักยภาพในการหลอกลวงโดยใช้เสียงปลอมแปลงที่สร้างจากแบบจำลอง 4) เพื่อนำเสนอแนวทางการประยุกต์ใช้งานเทคโนโลยีปัญญาประดิษฐ์เชิงกำเนิดในการปฏิบัติการทางไซเบอร์เชิงรุก

ผลการศึกษาพบว่า MaskCycleGAN-VC เป็นแบบจำลองปัญญาประดิษฐ์เชิงกำเนิดที่มีประสิทธิภาพสูงและเหมาะสมสำหรับการปลอมแปลงเสียงภาษาไทย แบบจำลองนี้สามารถสร้างเสียงปลอมที่มีความใกล้เคียงกับต้นฉบับได้อย่างเป็นธรรมชาติ ทั้งในด้านจังหวะการเว้นวรรค น้ำหนักเสียงและการถ่ายทอดอารมณ์ จุดเด่นสำคัญคือใช้เวลาเพียง 1 วันในการฝึกฝนแบบจำลองด้วยทรัพยากรคอมพิวเตอร์ระดับปานกลาง เสียงปลอมสามารถหลอกลวงผู้ฟังได้ 56% ขณะที่เสียงจริงถูกเข้าใจผิดว่าเป็นเสียงปลอมสูงสุด 59% สะท้อนถึงความท้าทายในการแยกแยะเสียงในสภาพแวดล้อมที่มีสัญญาณรบกวน โดยมีตัวชี้วัดดังนี้ 1) ค่าคะแนนความคิดเห็นเฉลี่ย (Mean Opinion Score: MOS) ความเป็นธรรมชาติสูงสุด 3.9 และความคล้ายคลึงสูงสุด 4.2 2) ค่าความผิดเพี้ยนของเมลเซปสตรัม (Mel Cepstral Distortion: MCD) ต่ำสุด 5 dB 3) ค่าระยะห่างเสียงเชิงลึกแบบเคอร์เนล (Kernel Deep Speech Distance: KDSD) ต่ำสุด 15.9 mKDSD แบบจำลองนี้มีศักยภาพสูงสำหรับการประยุกต์ใช้งานด้านความมั่นคงและการปฏิบัติการทางไซเบอร์เชิงรุก อย่างไรก็ตามการใช้งานควรดำเนินการอย่างระมัดระวัง เพื่อป้องกันการนำไปใช้ในทางที่ไม่เหมาะสม

คำสำคัญ: ปัญญาประดิษฐ์เชิงกำเนิด, โครงข่ายปฏิปักษ์เชิงกำเนิด, การแปลงเสียง, สงครามไซเบอร์

Abstract

This research aims to 1) explore the appropriate application of artificial intelligence (AI) technology for voice spoofing, 2) develop a generative AI-based voice spoofing model and investigate optimization strategies to enhance its suitability for cyber domain applications, 3) evaluate the performance and deception potential of synthetic voices generated by the model, and 4) propose practical applications of generative AI technology in offensive cyber operations.

The findings indicated that MaskCycleGAN-VC was a highly effective generative artificial intelligence model suitable for voice spoofing in the Thai language. This model could generate synthetic voices that closely resembled the original in terms of naturalness, including rhythm, intonation, and emotional expression. A key feature of the model was its ability to be developed and trained within just one day, using only moderate computational resources. The synthetic voices generated by the model could deceive listeners into believing they were genuine voices with an accuracy of up to 56%, while genuine voices were misclassified as synthetic in up to 59% of cases. This highlighted the challenges of distinguishing between genuine and synthetic voices in noisy environments. Performance metrics included a Mean Opinion Score (MOS) score for naturalness of up to 3.9 and similarity of up to 4.2, with a minimum Mel Cepstral Distortion (MCD) of 5 dB and Kernel Deep Speech Distance (KDSD) of 15.9 mKDSD. This model demonstrated significant potential for applications in security and offensive cyber operations, including support for intelligence activities, confusion in emergency scenarios, and simulated training exercises. However, its usage should be approached with caution to prevent misuse in unethical contexts.

Keywords: Generative Artificial Intelligence, Generative Adversarial Network, Voice Conversion, Cyber Warfare

1. บทนำ

ในยุคดิจิทัลที่ทุกอย่างสามารถเข้าถึงได้ง่ายเพียงแค่มีอินเทอร์เน็ต นำมาซึ่งภัยคุกคามทางไซเบอร์หลากหลายรูปแบบ ซึ่งมีแนวโน้มส่งผลกระทบรุนแรงขึ้นทั้งต่อบุคคลและองค์กรตามการพัฒนาของเทคโนโลยีที่เปลี่ยนแปลงไป เมื่อพิจารณาแนวโน้มด้านความปลอดภัยทางไซเบอร์ ในปี 2024 ซึ่งจัดอันดับโดย Gartner [1] พบว่า หนึ่งในเทคโนโลยีที่พัฒนาอย่างก้าวกระโดดและน่าจับตามองที่สุดคือ ปัญญาประดิษฐ์เชิงกำเนิด โดยมีความสามารถสร้างสรรค์ผลงานที่มีความเสมือนจริงออกมาได้หลากหลาย [2] ทั้งข้อความ ภาพ เสียง และวิดีโอ บริษัท Gogolook ผู้พัฒนาแอปพลิเคชัน Whoscall ได้ออกมาเตือนถึงภัยคุกคามจากการหลอกลวงด้วยเทคโนโลยีปัญญาประดิษฐ์ [3] โดยเฉพาะการปลอมแปลงเสียงที่เพิ่มมากขึ้นในปัจจุบัน มิจฉาชีพนำเทคโนโลยีนี้ มาสร้างเสียงเลียนแบบบุคคลใกล้ชิด เช่น สมาชิกครอบครัวหรือเพื่อนสนิท เพื่อหลอกลวงเหยื่อให้โอนเงินหรือให้ข้อมูลส่วนตัวที่สำคัญ กระทรวงดิจิทัลเพื่อ

เศรษฐกิจและสังคมและสำนักงานตำรวจแห่งชาติ [4] ได้ออกมาเตือนประชาชนให้ระมัดระวังต่อภัยลักษณะนี้ โดยแนะนำให้ตรวจสอบตัวตนของผู้โทรทุกครั้งก่อนทำธุรกรรมหรือให้ข้อมูลส่วนตัว Gogolook ยังเน้นย้ำถึงความสำคัญของการไม่ไว้วางใจสายโทรเข้าที่ไม่คุ้นเคยและแนะนำให้ใช้เครื่องมือป้องกัน เช่น Whoscall ในการกรองและตรวจสอบเบอร์โทรที่อาจเป็นมิจฉาชีพ เพื่อป้องกันการตกเป็นเหยื่อของการโจมตีทางไซเบอร์ที่พัฒนาซับซ้อนมากขึ้นในยุคดิจิทัลปัจจุบัน

ด้วยเหตุนี้ตามแผนพัฒนาวิทยาศาสตร์และเทคโนโลยีป้องกันประเทศ [5] ยุทธศาสตร์กองทัพอากาศ [6] และนโยบายผู้บัญชาการทหารอากาศ [7] มีความสอดคล้องภายใต้ยุทธศาสตร์ชาติ [8] ในประเด็นด้านความมั่นคง เสริมสร้างศักยภาพและความพร้อมในทุกด้าน ทั้งคน เครื่องมือ ยุทโธปกรณ์ เทคโนโลยีข้อมูลขนาดใหญ่ และการแก้ไขปัญหาสำคัญ อาทิ ภัยคุกคามทางไซเบอร์ รวมถึงยุทธศาสตร์ชาติในประเด็นด้านการสร้างความสามารถในการแข่งขัน โดยให้การสนับสนุนการวิจัยและพัฒนาประยุกต์ใช้เทคโนโลยีวิทยาศาสตร์ข้อมูลและปัญญาประดิษฐ์กับการปฏิบัติภารกิจบนมิติไซเบอร์ทั้งเชิงรุกและเชิงรับ [9] และเพื่อให้กองทัพอากาศสามารถปฏิบัติภารกิจหลักด้านความมั่นคงได้อย่างสมบูรณ์และเกิดประโยชน์สูงสุดต่อประเทศ จึงจำเป็นต้องพัฒนาขีดความสามารถและเตรียมพร้อมรับมือกับภัยคุกคามรูปแบบใหม่ที่อาจเกิดขึ้นได้ซึ่งเป็นทั้งความท้าทายและโอกาสอันดีในการนำเทคโนโลยีนี้มาใช้อย่างมีจริยธรรมและปลอดภัย

อย่างไรก็ตามปัจจุบันงานวิจัยในประเทศไทยยังไม่มีการศึกษาเชิงลึกเกี่ยวกับเทคโนโลยีการแปลงเสียงที่มีความทันสมัย เพื่อใช้เป็นเครื่องมือสำหรับการปฏิบัติการทางไซเบอร์เชิงรุก งานวิจัยนี้มุ่งเน้นศึกษาและพัฒนาปัญญาประดิษฐ์เชิงกำเนิดในการแปลงเสียง [10-14] คือ กระบวนการเปลี่ยนแปลงเสียงผู้พูด ให้เป็นเสียงของอีกคนที่มีเสียงเป็นเอกลักษณ์เฉพาะบุคคล โดยยังคงรักษาเนื้อหาทางภาษา จังหวะ และน้ำเสียงของผู้พูดดั้งเดิมไว้ นอกจากนี้การประเมินคุณภาพเสียงที่แปลงเป็นปัจจัยสำคัญในการพัฒนาแบบจำลองให้มีความแม่นยำและใกล้เคียงเสียงต้นฉบับที่สุด งานวิจัยนี้ใช้ Kernel Deep Speech Distance (KDSD) เป็นตัวชี้วัดความคล้ายคลึงเชิงลึก โดยอาศัยฟีเจอร์จากแบบจำลองรู้จำเสียงพูดที่เหมาะกับภาษาไทย เช่น Facebook AI’s wav2vec 2.0 ซึ่งผ่านการฝึกด้วยข้อมูลเสียงขนาดใหญ่และสามารถดึงคุณลักษณะเชิงลึกได้อย่างแม่นยำ

2. ขอบเขตงานวิจัย

1. งานวิจัยนี้มุ่งเน้นศึกษาแบบจำลองที่พัฒนาจากเทคโนโลยีปัญญาประดิษฐ์เชิงกำเนิดที่ได้รับการเผยแพร่ผ่านข้อมูลวิจัยเท่านั้น
2. เสียงที่ใช้ในการทดสอบเป็นเสียงของบุคคลที่ให้ความยินยอมในการบันทึกเสียงเท่านั้น
3. การพัฒนาแบบจำลองอยู่บนข้อจำกัดของอุปกรณ์การประมวลผลและทรัพยากรการวิจัยที่มีอยู่

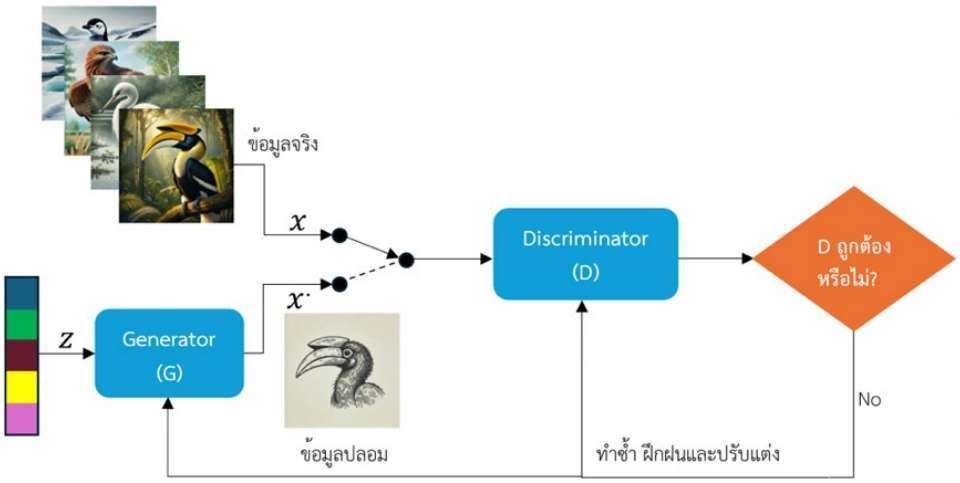
3. ทฤษฎีที่เกี่ยวข้อง

3.1 ปัญญาประดิษฐ์เชิงกำเนิด (Generative AI: GenAI)

ปัญญาประดิษฐ์เชิงกำเนิดจัดเป็นหนึ่งในเทคโนโลยีปัญญาประดิษฐ์ในปัจจุบันที่มาแรงและล้ำหน้าที่สุด [2] ถูกพัฒนาความสามารถจากการเรียนรู้เชิงลึกให้ชาญฉลาดและมีความอัจฉริยะมากยิ่งขึ้น เปรียบเสมือนสมองซีกขวาโดยเน้นจินตนาการ ความคิดสร้างสรรค์ สร้างข้อมูลใหม่แบบอัตโนมัติที่มีความเสมือนจริงคล้ายกับข้อมูลจริงที่มีอยู่ เช่น ภาพ เสียง วิดีโอ ข้อความ ถือเป็นโครงข่ายประสาทเทียมแบบกลับหลังซึ่งต่างจากการเรียนรู้เชิงลึกแบบเดิมที่เน้นการวิเคราะห์ข้อมูลเป็นหลัก โดยมีหลักการเริ่มต้นจากการระบุสิ่งที่ต้องการและแบบจำลองจะสร้างผลลัพธ์มาให้หลากหลายรูปแบบ

3.2 โครงข่ายปฏิปักษ์เชิงกำเนิด (Generative Adversarial Network: GAN)

โครงข่ายปฏิปักษ์เชิงกำเนิดหรือแบบจำลองแกนเป็นเทคนิคที่ได้รับความนิยมและใช้กันอย่างแพร่หลายที่สุดในปัจจุบันด้านปัญญาประดิษฐ์เชิงกำเนิด [2, 15-16] จัดเป็นประเภทการเรียนรู้แบบไม่มีผู้สอน อัลกอริทึมประกอบด้วยสองโครงข่ายประสาทเทียมที่ทำงานช่วยเหลือกันด้วยการแข่งขันลองสร้างและจับผิด ดังรูปที่ 1 ได้แก่ โครงข่ายผู้สร้าง (Generator) ทำหน้าที่วิเคราะห์ เรียนรู้ สกัดคุณลักษณะของสิ่งที่ต้องการสร้าง เช่น รูปร่าง ลายเส้น การตัดกันของสี เป็นต้น โดยส่วนใหญ่เริ่มต้นมักจะใช้จากข้อมูลสุ่ม โดยมีเป้าหมายเพื่อสร้างข้อมูลใหม่ที่เสมือนจริงใกล้เคียงกับข้อมูลจริงมากที่สุด ความโดดเด่นของโครงข่ายผู้สร้าง คือ ได้ผลลัพธ์ที่หลากหลายรูปแบบ ในขณะเดียวกันโครงข่ายผู้แยกแยะ (Discriminator) ทำหน้าที่วิเคราะห์ ฝึกฝน เรียนรู้ สกัดคุณลักษณะจากข้อมูลจริงและข้อมูลปลอมที่ถูกสร้างขึ้นจากโครงข่ายผู้สร้าง โดยมีเป้าหมายเพื่อให้สามารถแยกแยะข้อมูลทั้งสองได้อย่างแม่นยำ อีกทั้งคอยช่วยเหลือ ปรับปรุงและเพิ่มประสิทธิภาพของโครงข่ายผู้สร้างให้ดีขึ้น โดยใช้ผลการประเมินที่ได้รับจากโครงข่ายผู้แยกแยะในการตรวจสอบแต่ละครั้ง กระบวนการนี้จะทำซ้ำไปเรื่อยๆ จนกว่าโครงข่ายผู้สร้างจะสร้างข้อมูลใหม่ที่เสมือนจริงใกล้เคียงกับข้อมูลจริงมากที่สุด และโครงข่ายผู้แยกแยะไม่สามารถแยกแยะได้ว่าข้อมูลนั้นเป็นของจริงหรือของปลอม



รูปที่ 1 โครงข่ายปฏิปักษ์เชิงกำเนิด

3.3 การแปลงข้อความเป็นเสียง (Text-to-Speech: TTS)

การแปลงข้อความเป็นเสียงเป็นเทคนิคการสร้างเสียงสังเคราะห์ด้วยระบบคอมพิวเตอร์ที่ได้รับความนิยมมากที่สุด เปรียบเสมือนผู้ช่วยเสียง เช่น Google Assistant และ Siri ช่วยอำนวยความสะดวกในการอ่านออกเสียงแบบอัตโนมัติจากข้อความ [17] เป็นเทคโนโลยีที่มีประโยชน์และมีบทบาทในชีวิตประจำวันเป็นอย่างมาก ช่วยให้ผู้พิการเข้าถึงข้อมูลในรูปแบบดิจิทัลได้ง่ายขึ้น สนับสนุนการพัฒนาทักษะการฟังและการพูดของนักเรียนและนักศึกษา รวมถึงเพิ่มความสะดวกในการฟังข้อความระหว่างทำกิจกรรมอื่น เช่น ขับรถ ออกกำลังกาย หรือเล่นเกม นอกจากนี้ยังช่วยครีเอเตอร์ปกปิดตัวตนด้วยเสียง AI เพื่อสร้างความปลอดภัยบนโลกออนไลน์ พร้อมพากย์เสียงและสร้างเนื้อหาได้อย่างรวดเร็วและน่าสนใจสำหรับเผยแพร่บนแพลตฟอร์มต่าง ๆ เช่น YouTube, Facebook, Instagram และ TikTok

ผลการศึกษาและทดสอบการแปลงข้อความเป็นเสียงด้วยแบบจำลอง Voice Clone และ Fast Speech 2 พบว่าไม่เหมาะสำหรับการใช้งานในภาษาไทย ด้วยข้อจำกัดในการเตรียมข้อมูลเพื่อใช้สำหรับฝึกฝนแบบจำลอง เนื่องจากในปัจจุบันยังไม่มีชุดข้อมูลเสียงภาษาไทยแบบคู่ขนานและเป็นมาตรฐานที่พร้อมใช้งาน ได้แก่ ไฟล์เสียงและ

ไฟล์ข้อความ ผู้วิจัยจึงต้องรวบรวมชุดข้อมูลเสียงเองทั้งหมดซึ่งใช้เวลานานและได้ชุดข้อมูลจำนวนน้อย ส่งผลโดยตรงต่อความแม่นยำของแบบจำลอง ทำให้คุณภาพเสียงที่ได้มีความไม่เป็นธรรมชาติ [16]

3.4 การแปลงเสียง (Voice Conversion: VC)

การแปลงเสียง [10-14] คือ กระบวนการเปลี่ยนแปลงเสียงผู้พูด (Source) ให้เป็นเสียงผู้พูดของอีกคน (Target) ที่มีเสียงเป็นเอกลักษณ์เฉพาะบุคคล โดยยังคงรักษา เนื้อหาทางภาษา จังหวะ และน้ำเสียงของผู้พูดดั้งเดิมไว้ การทำงานพื้นฐานในการแปลงเสียง ดังรูปที่ 2 และมีรายละเอียดการทำงานในแต่ละส่วนดังนี้

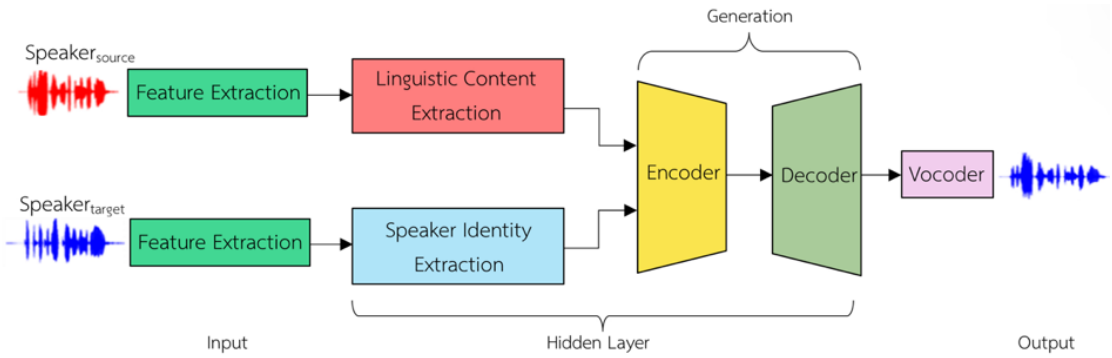
• Feature Extraction แปลงสัญญาณเสียง (Sound Wave) ให้อยู่ในรูปแบบ Mel-frequency Cepstrum Coefficients (MFCC) [11, 12] หรือ Mel-Spectrogram [13-14, 18]

• Linguistic Content Extraction สกัดคุณลักษณะเนื้อหาทางภาษาและสไตล์การพูด เช่น ข้อความหรือประโยค ตลอดจนจังหวะและน้ำเสียง

• Speaker Identity Extraction สกัดคุณลักษณะเอกลักษณ์เสียงบุคคล ซึ่งในทางฟิสิกส์เรียกว่า คุณภาพของเสียง (Timbre) [19] หมายถึงคุณสมบัติที่ช่วยให้แยกแยะเสียงจากแหล่งกำเนิดต่างๆได้ แม้จะมีระดับเสียงและความถี่เท่ากัน แหล่งกำเนิดเสียงเมื่อสั่นจะสร้างความถี่หลายค่าออกมาพร้อมกัน เรียกว่า ฮาร์มอนิก (Harmonics) ซึ่งประกอบด้วยความถี่พื้นฐาน (Fundamental Frequency: F0) และความถี่อื่นที่เป็นทวีคูณของความถี่พื้นฐาน เสียงที่เราได้ยินเกิดจากการรวมกันของฮาร์มอนิกหลายค่า โดยฮาร์มอนิกและแอมพลิจูดที่แตกต่างกันในแต่ละแหล่งกำเนิดเสียง ส่งผลให้เสียงมีเอกลักษณ์เฉพาะตัว คุณลักษณะนี้จึงถูกนำมาใช้ในการวิเคราะห์และสกัดเอกลักษณ์เสียงพูดของมนุษย์

• Generation รวมข้อมูลที่สกัดและประมวลผลเพื่อสร้างเสียงที่ปรับแต่ง โดยผลลัพธ์คือ สเปกโตรแกรมของเสียงพูดที่ผ่านการแปลงแล้ว

• Vocoder แปลงสเปกโตรแกรมของเสียงพูดให้เป็นสัญญาณเสียง



รูปที่ 2 การทำงานพื้นฐานในการแปลงเสียง

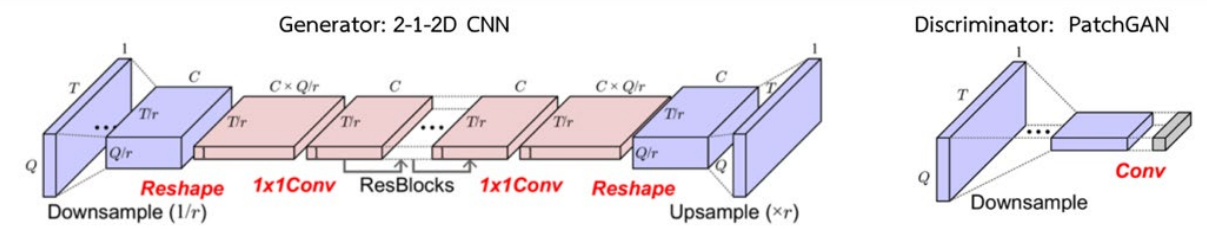
3.5 การแปลงเสียงด้วยโครงข่ายประสาทเทียมแบบแกนสำหรับการใช้งานในภาษาไทย

ผู้วิจัย [16] ได้ศึกษาและทดสอบแบบจำลอง StarGAN-VC โดยมุ่งเน้นการแปลงเสียงจากกลุ่มบุคคลทั่วไปไปยังกลุ่มบุคคลเป้าหมาย (Many-to-Many Voice Conversion) ซึ่งช่วยให้สามารถแปลงเสียงในกลุ่มเป้าหมายได้จำนวนมากในคราวเดียว โดยไม่ถูกจำกัดให้ใช้เฉพาะบุคคลใดบุคคลหนึ่งเหมือนแบบจำลอง CycleGAN-VC ที่เป็นรูปแบบหนึ่งต่อหนึ่ง (One-to-One Voice Conversion) การประเมินคุณภาพเสียงของทั้งสองแบบจำลองโดยใช้คะแนนเฉลี่ยความคิดเห็น (Mean Opinion Score: MOS) [20] ซึ่งเป็นมาตรวัดเชิงคุณภาพตามมาตรฐานสากล พบว่าไม่มีความแตกต่างอย่างมีนัยสำคัญ และเมื่อประเมินประสิทธิผลในการหลอกผู้ฟังพบว่าเสียงปลอมของบุคคลเป้าหมายสามารถหลอกลวงผู้ฟังได้ในอัตราร้อยละ 20-40 ผลการวิจัยชี้ให้เห็นว่าการแปลงเสียงบุคคลให้เป็นเสียงบุคคล

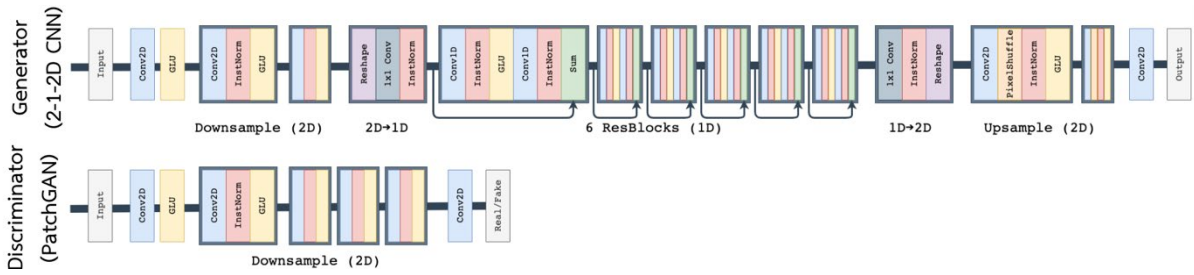
เป้าหมายมีศักยภาพเหมาะสมสำหรับปฏิบัติการไซเบอร์ในแง่การสร้างข่าวปลอมเพื่อหลอกลวงและสร้างความเข้าใจผิดว่าเสียงที่ได้ยินเป็นเสียงของบุคคลนั้นจริง หากนำเทคโนโลยีไปใช้ในเชิงลบ อาจกลายเป็นภัยคุกคามทางไซเบอร์ เช่น การทำวิศวกรรมทางสังคมเพื่อการล่อลวง ทุจริตด้านการเงิน และเผยแพร่ข่าวสารเท็จที่ส่งผลกระทบต่อความมั่นคงของประเทศและประชาชน

3.6 MaskCycleGAN-VC

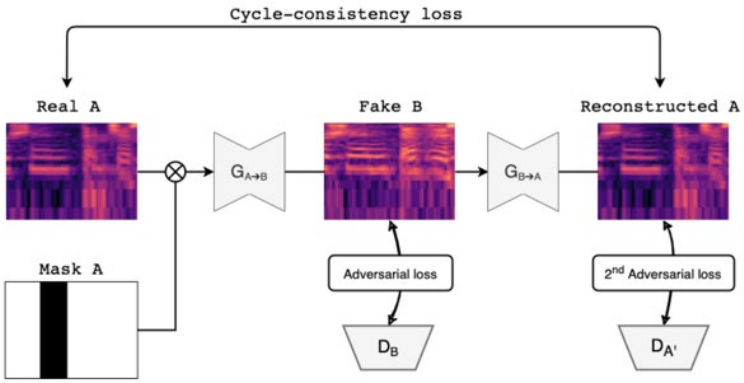
MaskCycleGAN-VC เป็นแบบจำลองแปลงเสียงรูปแบบหนึ่งต่อหนึ่ง [14] ที่ได้รับการยอมรับว่าเป็นวิธีมาตรฐานในปัจจุบัน (Current Benchmark Method) [21] สถาปัตยกรรมประกอบด้วย Generator และ Discriminator ดังรูปที่ 3 โดยใช้โครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neural Network: CNN) ดังรูปที่ 4 ข้อมูลนำเข้าอยู่ในรูปแบบ Mel-Spectrogram ซึ่งแสดงการกระจายพลังงานของเสียงในแกนเวลาและความถี่ที่สอดคล้องกับการได้ยินของมนุษย์ แบบจำลองนี้มีความโดดเด่นในการวิเคราะห์และเรียนรู้คุณลักษณะเสียง สามารถสร้างเสียงปลอมได้เสมือนจริงด้วยเทคนิคคงสภาพโครงสร้างฮาร์โมนิกหลังการแปลงเสียง โดยใช้วิธีเติมเต็มเฟรม (Filling in Frame) ที่ฝึกให้เรียนรู้เฟรมข้อมูลที่เสียหายและเติมเต็มแบบอัตโนมัติผ่านหน้ากาก (Mask) ดังรูปที่ 5 วิธีการนี้พัฒนาต่อยอดจาก CycleGAN-VC3 [13] ดังรูปที่ 6 และได้รับการปรับปรุงให้มีประสิทธิภาพยิ่งขึ้น



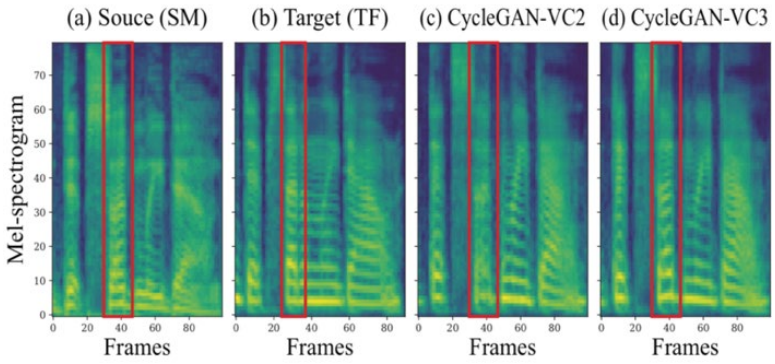
รูปที่ 3 สถาปัตยกรรม MaskCycleGAN-VC ประกอบด้วย Generator และ Discriminator [14]



รูปที่ 4 สถาปัตยกรรมภายใน Generator และ Discriminator [14]



รูปที่ 5 การฝึกฝนแบบจำลอง [14]



รูปที่ 6 เทคนิคการคงสภาพของโครงสร้างฮาร์โมนิก [13]

3.7 การประเมินประสิทธิภาพของเสียงที่สร้างจากแบบจำลอง MaskCycleGAN-VC

3.7.1 Mean Opinion Score (MOS)

MOS เป็นตัวชี้วัดการประเมินเชิงคุณภาพที่ใช้วัดคุณภาพเสียงในระดับคะแนน 1-5 [20] โดยผู้ประเมินพิจารณาจากความเป็นธรรมชาติ (Naturalness) ซึ่งสะท้อนถึงความสมจริงและลักษณะเสียงที่ฟังดูเหมือนมนุษย์ และความคล้ายคลึง (Similarity) ซึ่งแสดงถึงความใกล้เคียงระหว่างเสียงที่แปลงกับเสียงต้นฉบับ ค่า MOS ที่สูงบ่งชี้ว่าเสียงที่แปลงมีคุณภาพดีและสมจริง

3.7.2 Mel Cepstral Distortion (MCD)

MCD เป็นตัวชี้วัดการประเมินเชิงปริมาณที่ใช้วัดความแตกต่างเชิงพลังงานระหว่างเสียงที่แปลง (Converted speech) กับเสียงต้นฉบับ (Original speech) [22] งานวิจัยนี้ใช้ WORLD Analyzer [23] ในการสกัดสเปกตรัมของเสียงและแปลงเป็น Mel-Cepstral Coefficients (MCC) ที่ความละเอียด 35 มิติ ซึ่งเป็นข้อมูลสำคัญในการคำนวณค่า MCD ตามสมการที่ 1 [24] ค่าที่ต่ำบ่งชี้ว่าเสียงที่แปลงมีความใกล้เคียงกับเสียงต้นฉบับ ตัวชี้วัด MCD ยังสัมพันธ์กับการรับรู้ของมนุษย์ในด้านความสมจริงและความคล้ายคลึง โดยแสดงค่าในหน่วย dB (Decibel)

$$MCD = \frac{10}{\ln 10}\sqrt{2\sum_{n=1}^{N}\left(c_n^{(1)}(t) - c_n^{(2)}(t)\right)^2} \quad (1)$$

โดยที่ N คือ จำนวนมิติของ MCC ที่ใช้

$c_n^{(1)}(\mathrm{t})$ คือ ค่าสัมประสิทธิ์ MCC ของเสียงต้นฉบับในมิติที่ n และเฟรม t

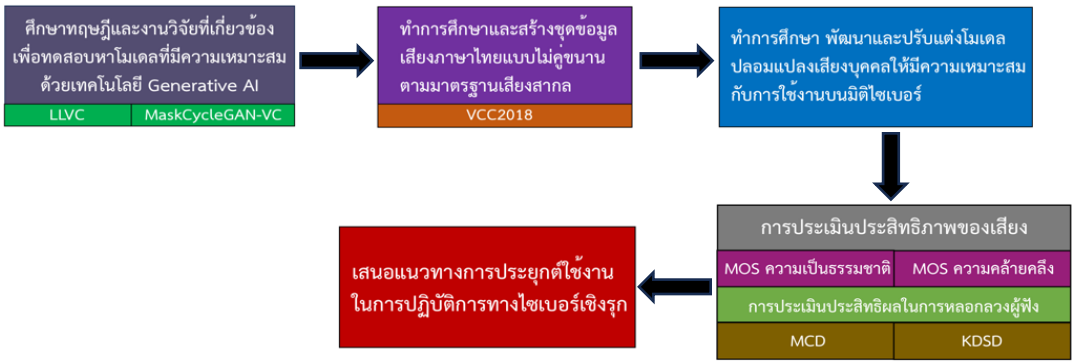
$c_n^{(2)}(\mathrm{t})$ คือ ค่าสัมประสิทธิ์ MCC ของเสียงที่แปลงในมิติที่ n และเฟรม t

3.7.3 Kernel Deep Speech Distance (KDSD)

KDSD เป็นตัวชี้วัดการประเมินเชิงปริมาณที่ใช้วัดความคล้ายคลึงของคุณลักษณะเสียงระหว่างเสียงที่แปลงกับเสียงต้นฉบับ [25] โดยอาศัยข้อมูลเชิงลึกของเสียง (Embedding) ที่สกัดจากแบบจำลองการรู้จำเสียงพูด DeepSpeech2 [26] ซึ่งพัฒนาในปี 2016 รองรับภาษาอังกฤษและภาษาจีนกลาง ค่าที่ต่ำบ่งชี้ว่าเสียงที่แปลงมีคุณลักษณะเชิงลึกที่ใกล้เคียงกับเสียงต้นฉบับ โดยแสดงค่าในหน่วย KDSD

4. การดำเนินการวิจัย

จากรูปที่ 7 แสดงกรอบแนวคิดงานวิจัย โดยผู้วิจัยเริ่มต้นการวิจัยโดยศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง ได้แก่ กระบวนการศึกษาและพัฒนาแบบจำลองการแปลงเสียงโดยใช้ LLVC และ MaskCycleGAN-VC ซึ่งเป็นแบบจำลองที่เป็นต้นแบบการใช้งานแบบจำลองปัญญาประดิษฐ์เชิงกำเนิด ร่วมกับชุดข้อมูลที่สร้างขึ้นใหม่ตามมาตรฐานเสียงสากล VCC2018 โดยในขั้นตอนดังกล่าว ผู้วิจัยได้ทำการปรับแต่งแบบจำลองปลอมแปลงเสียงให้มีความเหมาะสมกับการใช้งานบนมิติไซเบอร์ เช่น การปรับแต่งค่าพารามิเตอร์ของแบบจำลอง รวมถึงการทดสอบประสิทธิภาพและการทำงานของแบบจำลอง จากนั้นทำการประเมินประสิทธิภาพของเสียงโดยใช้ตัวชี้วัดได้แก่ 1) MOS ด้านความเป็นธรรมชาติและด้านความคล้ายคลึง 2) MCD 3) KDSD รวมทั้งประเมินประสิทธิผลในการหลอกลวงผู้ฟัง สุดท้ายนำเสนอแนวทางการประยุกต์ใช้ในการปฏิบัติการทางไซเบอร์เชิงรุก



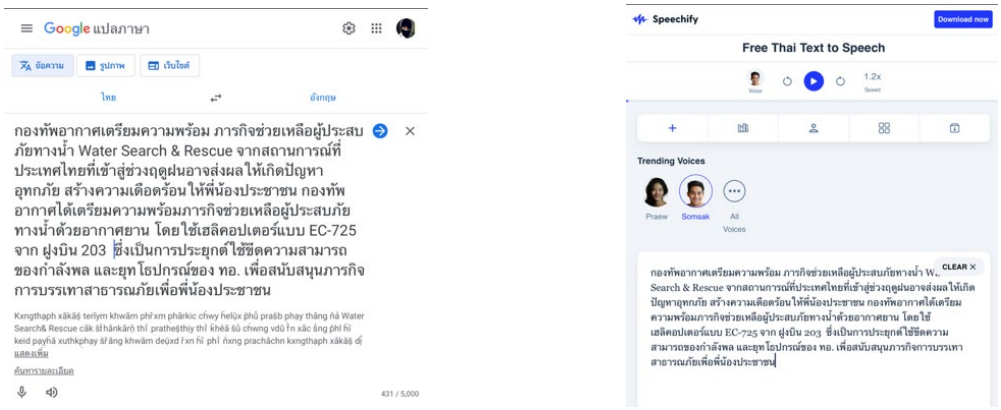
รูปที่ 7 กรอบแนวคิดงานวิจัย

4.1 ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องเพื่อทดสอบหาแบบจำลองที่มีความเหมาะสม

ผู้วิจัยได้ศึกษาค้นคว้าและทบทวนวรรณกรรมในการสังเคราะห์เสียงด้วยระบบคอมพิวเตอร์ มี 2 วิธี ได้แก่

4.1.1 การแปลงข้อความเป็นเสียง

ผู้วิจัยได้ทดลองใช้การแปลงข้อความเป็นเสียงที่มีให้ใช้งานฟรีและรองรับภาษาไทย จากเว็บไซต์ระดับโลก เช่น Google Translate และ Speechify [27] โดยพิมพ์ข้อความภาษาไทยและกดฟังเสียงเพื่อรับฟังเสียงสังเคราะห์ที่ถูกสร้างจากข้อความ ดังรูปที่ 8



(ก) เว็บไซต์ Google Translate (ข) เว็บไซต์ Speechify

รูปที่ 8 ทดลองใช้การแปลงข้อความเป็นเสียง

4.1.2 การแปลงเสียง

ผู้วิจัยได้ทดสอบแบบจำลองการแปลงเสียงด้วยเทคโนโลยีปัญญาประดิษฐ์เชิงกำเนิด ประกอบด้วย LLVC [28] และ MaskCycleGAN-VC [14] โดยมีขั้นตอนดังนี้ 1) เตรียมทรัพยากรที่จำเป็นรวมถึงระบบปฏิบัติการและซอฟต์แวร์สำหรับแบบจำลอง โดยสามารถศึกษารายละเอียดเพิ่มเติมตามจากซอร์สโค้ดที่กำหนด 2) ทดสอบการฝึกฝนแบบจำลองโดยใช้ชุดข้อมูลเสียงเริ่มต้น 3) สร้างชุดข้อมูลเสียง NKRAFA Thai 4) ทดลองปรับแต่งการตั้งค่าพารามิเตอร์ในการฝึกฝนแบบจำลองเพื่อให้เหมาะสมกับการใช้งานบนมิติไซเบอร์ 5) วิเคราะห์ผลการทดลองเพื่อประเมินประสิทธิภาพของแบบจำลองทั้งสอง

4.2 ศึกษาและสร้างชุดข้อมูลเสียงภาษาไทย

ข้อมูลเสียงที่ใช้ในการฝึกฝนแบบจำลองถูกสร้างขึ้นใหม่ โดยใช้ชื่อ NKRAFA Thai Dataset ตามตารางที่ 1 ประกอบด้วยเสียงต้นแบบจากผู้วิจัย (SM01) และเสียงบุคคลเป้าหมายจากอาจารย์ (TM02, TF04, TF05) รวมถึงหัวหน้านักเรียนนายเรืออากาศ (TM03) ทั้งหมดรวม 5 คน ซึ่งได้ให้ความยินยอมเป็นลายลักษณ์อักษรในการบันทึกเสียงเพื่อใช้ในการวิจัย โดยก่อนดำเนินการวิจัยในขั้นตอนการทดสอบ ผู้วิจัยได้ชี้แจงถึงวัตถุประสงค์ของการวิจัย รวมถึงแนวทางการเก็บข้อมูลเสียงของกลุ่มตัวอย่างบนพื้นฐานของการรักษาข้อมูลส่วนบุคคล รวมถึงให้คำยืนยันว่าแบบจำลองดังกล่าวจะถูกนำไปใช้เพื่อการวิจัยเท่านั้น ชุดข้อมูลนี้อ้างอิงตามมาตรฐาน VCC2018 Dataset [29] ซึ่งออกแบบมาสำหรับงานแปลงเสียงโดยเฉพาะและถูกใช้ในงานวิจัยหลายฉบับ [12-14] โดยมีเสียงจากผู้พูดที่หลากหลายทั้งชายและหญิง

ตารางที่ 1 การใช้งาน VCC2018 Dataset ในงานวิจัย [14] และ NKRAFA Thai Dataset ในงานวิจัยนี้

Characteristic	VCC2018 Dataset	NKRAFA Thai Dataset
Record	High-quality microphones	(iPad) Built-in microphones
Bit Depth / Sampling Rate	16-bit / 22050 Hz	16-bit / 22050 Hz
Audio Type / Format	MONO / WAV	MONO / WAV
Content	Sentences spoken by Speakers	Sentences spoken by Speakers
Duration	Varies	Varies
Language	English	Thai
Source / Target Speakers	2 / 2	1 / 4
Parallel / Non-Parallel	Both	Non-Parallel
Training / Evaluation / Testing	81 / 35 files	100 / 25 / 10 files

4.3 ทำการศึกษาพัฒนาและปรับแต่งแบบจำลองปลอมแปลงเสียงบุคคลให้มีความเหมาะสม

เครื่องมือและอุปกรณ์งานวิจัย HP-Notebook CPU: Intel® CoreTM i7-9750H @ 2.60GHz x 12, GPU: NVIDIA GeForce GTX 1650 Mobile 4GB และ Primary/Secondary Storage: RAM 8 GB / SSD 512 GB

เนื่องจากข้อจำกัดด้านฮาร์ดแวร์ตามขอบเขตงานวิจัยนี้ การใช้ Google Colab อาจเป็นทางเลือกที่ช่วยเพิ่มประสิทธิภาพ โดยสามารถเลือกใช้ GPU ของ Google เพื่อลดภาระการประมวลผลของเครื่อง ลดระยะเวลาในการฝึกฝนแบบจำลอง และจัดเก็บข้อมูลผ่าน Google Drive อย่างไรก็ตาม เวอร์ชันฟรีมีข้อจำกัดด้านระยะเวลาการใช้งาน หากต้องการใช้งานต่อเนื่องหรือเพิ่มทรัพยากร อาจจำเป็นต้องสมัครบริการแบบชำระเงินซึ่งมาพร้อมกับค่าใช้จ่ายที่เพิ่มขึ้น

ผู้วิจัยเลือกใช้แบบจำลอง MaskCycleGAN-VC โดยมีขั้นตอนการฝึกฝนแบบจำลอง ดังรูปที่ 9 และมีรายละเอียดการตั้งค่าพารามิเตอร์ดังนี้

• Epochs = 7000 รอบ เพื่อสังเกตแนวโน้มของค่าความผิดพลาดและประสิทธิภาพของแบบจำลองโดยรวม อ้างอิงจากค่าเริ่มต้นที่กำหนดในซอร์สโค้ด

• Sampling Rate = 22050 Hz เพื่อให้สอดคล้องกันกับชุดข้อมูลเสียงที่จัดเตรียมไว้

• Batch Size = 4 เพื่อเพิ่มประสิทธิภาพและลดระยะเวลาการฝึกฝนแบบจำลอง โดยเลือกค่าสูงที่สุดและเหมาะสมกับหน่วยความจำของ GPU ที่ใช้งาน

• Generator Learning Rate = 0.0002

• Discriminator Learning Rate = 0.0005 เพื่อช่วยให้โครงข่ายผู้แยกแยะแข็งแกร่งในช่วงแรกและเพื่อให้โครงข่ายผู้สร้างสามารถเรียนรู้ได้เร็วขึ้นและป้องกันปัญหาการฝึกที่ไม่สมดุลในระยะยาว

โดยที่

• Epochs คือ จำนวนรอบที่แบบจำลองฝึกฝนบนชุดข้อมูลทั้งหมดตั้งแต่ต้นจนจบหนึ่งครั้ง

• Iterations คือ หนึ่งรอบของการประมวลผลข้อมูลหนึ่งชุด (Batch) ผ่านแบบจำลองในการฝึกฝน

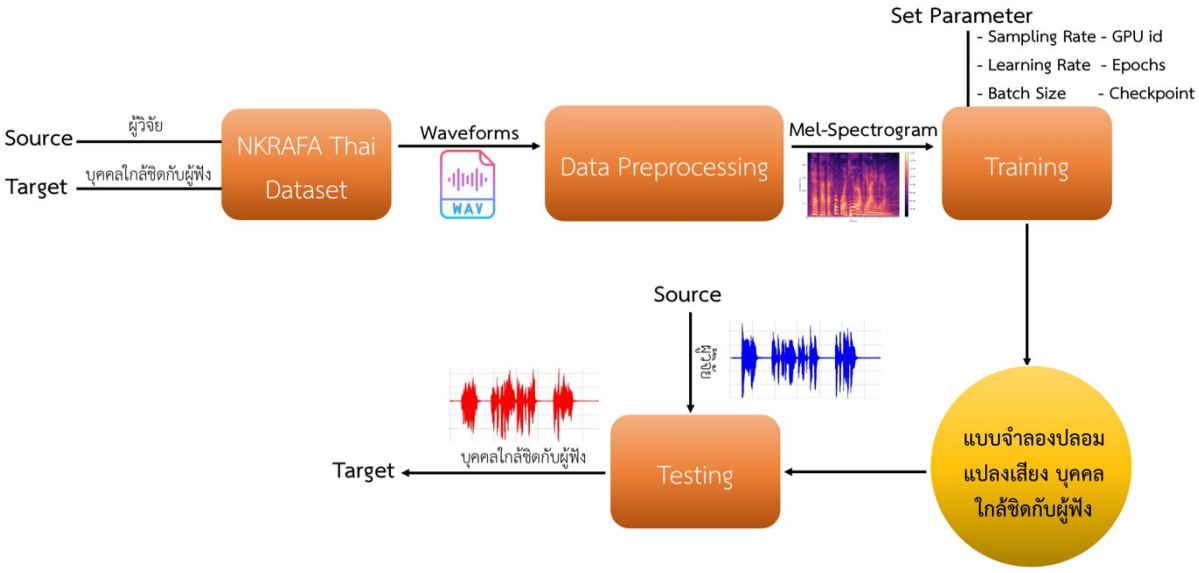
• Batch Size คือ จำนวนตัวอย่างข้อมูลที่แบบจำลองใช้ในการประมวลผล

• Generator Learning Rate คือ อัตราการเรียนรู้ของโครงข่ายผู้สร้างในการฝึกแต่ละรอบ (Iteration) โดยอิงตามค่าความผิดพลาด (Loss) ที่ได้รับจากการคำนวณระหว่างเสียงปลอมที่สร้างขึ้นและเสียงจริง ซึ่งโครงข่ายผู้สร้างจะพยายามลดค่าความผิดพลาดนี้เพื่อสร้างเสียงปลอมที่เหมือนจริงมากขึ้น

• Discriminator Learning Rate คือ อัตราการเรียนรู้ของโครงข่ายผู้แยกแยะ ซึ่งทำหน้าที่แยกแยะว่าเสียงที่ได้รับเป็นเสียงจริงหรือเสียงปลอม โดยอิงตามค่าความผิดพลาดของโครงข่ายผู้แยกแยะที่เกิดจากการทำนายผิดพลาดว่าเสียงปลอมเป็นเสียงจริง หรือเสียงจริงเป็นเสียงปลอม ซึ่งโครงข่ายผู้แยกแยะจะพยายามลดค่าความผิดพลาดนี้เพื่อเพิ่มประสิทธิภาพในการแยกแยะข้อมูล

• GPU ID คือ ค่าที่กำหนดเพื่อระบุว่าจะใช้ GPU ตัวใดในการฝึกฝนแบบจำลอง

• Checkpoint คือ ไฟล์ที่บันทึกสถานะของแบบจำลองระหว่างการฝึกฝน รวมถึงน้ำหนักของแบบจำลอง ค่า Optimizer และจำนวน Epochs เพื่อให้สามารถหยุดการฝึกแล้วกลับมาฝึกต่อหรือการนำแบบจำลองไปใช้งานหรือปรับแต่งเพิ่มเติมได้ในภายหลัง



รูปที่ 9 ขั้นตอนการฝึกฝนแบบจำลอง

4.4 การประเมินประสิทธิภาพของเสียง

ผู้วิจัยได้ฝึกฝนแบบจำลองจนกระทั่งพร้อมใช้งาน จำนวน 4 แบบจำลอง สำหรับการนำไปใช้งานจริง จะพิจารณาคัดเลือกเวอร์ชันที่มีค่า Generator Loss น้อยที่สุด เนื่องจากเป็นตัวชี้วัดว่าค่ายิ่งต่ำ ยิ่งสามารถสร้างเสียงปลอมที่ใกล้เคียงกับเสียงจริงได้มากที่สุด โดยใช้ 2 ตัวชี้วัดในการประเมินประสิทธิภาพดังนี้

4.4.1 ตัวชี้วัดการประเมินเชิงคุณภาพ

วัตถุประสงค์:

1) เพื่อประเมินคุณภาพเสียงปลอมที่ถูกสร้างจากแบบจำลองว่ามีคะแนนเฉลี่ยความคิดเห็นด้านความเป็นธรรมชาติและมีความคล้ายคลึงกันมากเพียงใด โดยมีระดับคะแนนดังนี้ 5: ดีเยี่ยม, 4: ดี, 3: พอใช้, 2: แย่, 1: แย่มาก

2) เพื่อประเมินประสิทธิผลในการหลอกลวงผู้ฟังสำหรับการใช้วิจารณญาณของตนเองในการตัดสินใจว่าเสียงของบุคคลใกล้ชิดดังกล่าว คือ เสียงจริงหรือเสียงปลอม

กลุ่มตัวอย่าง: นักเรียนนายเรืออากาศ จำนวน 50 คน

รูปแบบการวัดผล: ผู้วิจัยออกแบบการประเมินจำนวน 50 ชุด (1 ชุด ต่อ 1 คน) โดยแต่ละชุดประกอบด้วยเสียงจริงและเสียงปลอมรวม 12 ไฟล์เสียง ทั้งนี้เสียงทั้งหมดถูกสุ่มจากชุดเสียงจริงและเสียงปลอมที่คัดเลือกไว้ล่วงหน้าและผ่านการตรวจสอบความสมบูรณ์แล้ว

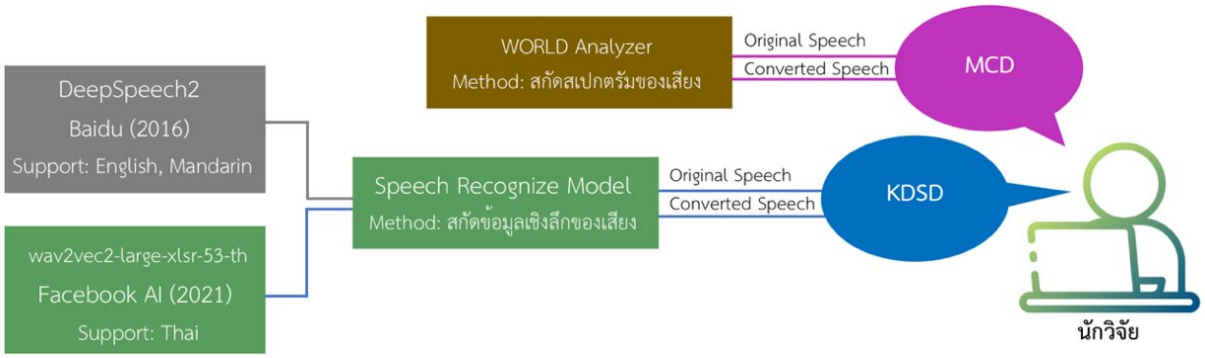
ขั้นตอนการวัดผล: ผู้วิจัยเริ่มต้นจากอธิบายวัตถุประสงค์ของการดำเนินการวิจัยให้กับกลุ่มตัวอย่างฟัง หลังจากนั้นเปิดตัวอย่างเสียงครั้งละ 1 เสียง เมื่อกลุ่มตัวอย่างฟังเสร็จในแต่ละเสียง กลุ่มตัวอย่างจะให้คะแนนคุณภาพเสียงทั้งในด้านความเป็นธรรมชาติและความคล้ายคลึง รวมทั้งตัดสินใจว่าเสียงที่ได้ยินว่าเป็นเสียงจริงหรือเสียงปลอม ซึ่งแบบประเมินจะถูกบันทึกโดยผู้วิจัย ดังรูปที่ 10

รูปที่ 10 ตัวชี้วัดการประเมินเชิงคุณภาพ

4.4.2 ตัวชี้วัดการประเมินเชิงปริมาณ

ผู้วิจัยได้คำนวณค่าความแตกต่างเชิงพลังงานระหว่างเสียงที่แปลงกับเสียงต้นฉบับ (พูดภาษาไทยประโยคเดียวกัน) โดยใช้ MCD เนื่องจากเอกสารวิจัยต้นฉบับมิได้เปิดเผยซอร์สโค้ดในส่วนนี้ ผู้วิจัยจึงเขียนซอร์สโค้ดใหม่โดยอ้างอิงและปรับปรุงการคำนวณ MCD ในหน่วยเดซิเบลจากแบบจำลอง StarGAN-VC [30] ให้เหมาะสมกับงานวิจัยนี้และเป็นไปตามมาตรฐานเดียวกัน เพื่อความสอดคล้องในการเปรียบเทียบค่า MCD

สำหรับการวัดความคล้ายคลึงของคุณลักษณะเสียงระหว่างเสียงที่แปลงกับเสียงต้นฉบับ ผู้วิจัยได้ใช้ KDSD และเขียนซอร์สโค้ดใหม่ทั้งหมด โดยปรับเปลี่ยนแบบจำลองรู้จำเสียงพูดเป็น wav2vec2-large-xlsr-53-th [31] ปี 2021 ซึ่งรองรับภาษาไทยและพัฒนามาจาก Wav2Vec2 ของ Facebook AI [32] นอกจากนี้ งานวิจัยนี้ยังนำเสนอค่า KDSD ในเกณฑ์ใหม่ที่เหมาะสมกับประโยคภาษาไทย ดังรูปที่ 11



รูปที่ 11 ตัวชี้วัดการประเมินเชิงปริมาณ

5. ผลการวิจัย

ผู้วิจัยได้ดำเนินการตามขั้นตอนวิจัย โดยมีรายละเอียดดังนี้

5.1 ผลการศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องเพื่อทดสอบหาแบบจำลองที่มีความเหมาะสม

5.1.1 การแปลงข้อความเป็นเสียง

ผลการศึกษาและทดลองใช้การแปลงข้อความเป็นเสียง จากเว็บไซต์ Google Translate และ Speechify ตามตารางที่ 2

ตารางที่ 2 Google Translate VS Speechify

เว็บไซต์การแปลงข้อความเป็นเสียง	Google Translate	Speechify
การรองรับภาษาไทย	✓	✓
การปรับความเร็วเสียง	ไม่มี	✓
ความเร็วในการแปลงเสียง	เร็วมาก	ช้า
เสียงผู้พูด	ไม่สามารถเลือกเสียงได้	เลือกเสียงผู้ชายและผู้หญิงได้
ความเป็นธรรมชาติ	มีความเป็นธรรมชาติระดับปานกลาง	มีความเป็นธรรมชาติมากกว่า
การถ่ายทอดอารมณ์ใกล้เคียงเสียงมนุษย์	ไม่มี	✓
จังหวะการพูดและการเว้นวรรค	อาจไม่ลื่นไหลในบางกรณี	มีการเว้นวรรคที่สมจริงมากขึ้น
วัตถุประสงค์หลัก	ใช้เพื่อฟังข้อความจากการแปลภาษา	ใช้เพื่อฟังข้อความทั่วไป
ข้อจำกัดของเวอร์ชันฟรี	ใช้งานได้ฟรี 100%	จำกัดจำนวนข้อความต่อวัน

5.1.2 การแปลงเสียง

ผลการศึกษาและทดลองใช้แบบจำลอง MaskCycleGAN-VC และ LLVC ตามตารางที่ 3

ตารางที่ 3 MaskCycleGAN-VC VS LLVC

แบบจำลอง	**MaskCycleGAN-VC**	**LLVC**
สถาปัตยกรรม	Generative Adversarial Network	Generative Adversarial Network
ประเภทของการแปลงเสียง	One-to-One	Any-to-One
การสกัดคุณลักษณะ (Feature Extraction)	Mel-Spectrogram	Mel-Spectrogram
ตัวสังเคราะห์เสียง (Vocoder)	MelGAN	HiFi-GAN
ชุดข้อมูลเสียง	VCC2018	LibriSpeech360
ลักษณะของชุดข้อมูลเสียง	แบบไม่คู่ขนาน (ไฟล์เสียง)	แบบคู่ขนาน (ไฟล์เสียง+ไฟล์บทพูด)
การเรียนรู้ของแบบจำลอง	ใช้เทคนิค Cycle-Consistency Loss แปลงเสียงต้นฉบับพร้อมคงเนื้อหาเดิมไว้ เป็นเสียงของผู้พูดเป้าหมาย และใช้เทคนิค Filling in Frames (FIF) ช่วยให้แบบจำลองเติมเฟรมเสียงที่ขาดหาย ส่งผลให้การเรียนรู้โครงสร้าง เชิงความถี่และเวลา (Mel-Spectrogram) มีประสิทธิภาพมากขึ้น	ใช้เทคนิค Waveformer Encoder-Decoder แยกเนื้อหาคำพูดออกจากเสียงต้นฉบับแล้วสร้างเสียงใหม่ให้เหมือนผู้พูดเป้าหมาย
การยอมรับว่าเป็นวิธีมาตรฐานในปัจจุบัน	✓	ไม่มี
การพัฒนาเวอร์ชันแบบจำลองอย่างต่อเนื่อง	✓ [11-14]	ไม่มี
การเผยแพร่ผ่านวารสารทางวิชาการ	✓	ไม่มี
ความสมบูรณ์ของทรัพยากร (Library)	✓	ไม่มี
โค้ดโอเพนซอร์สบนแพลตฟอร์ม GitHub	✓	✓
การตั้งค่าพารามิเตอร์สำหรับฝึกฝน	ง่าย	ยาก
การประเมินเชิงคุณภาพด้วยค่า MOS ความเป็นธรรมชาติ / ความคล้ายคลึง	4.43 / 4.08	3.78 / 3.83
การใช้งานจริง	เหมาะสำหรับการแปลงเสียงที่ต้องการคุณภาพสูง	เหมาะสำหรับการแปลงเสียงที่เป็นแบบเรียลไทม์

5.1.3 ความแตกต่างระหว่างการแปลงข้อความเป็นเสียงและการแปลงเสียง

ความแตกต่างระหว่างการแปลงข้อความเป็นเสียงและการแปลงเสียง ตามตารางที่ 4

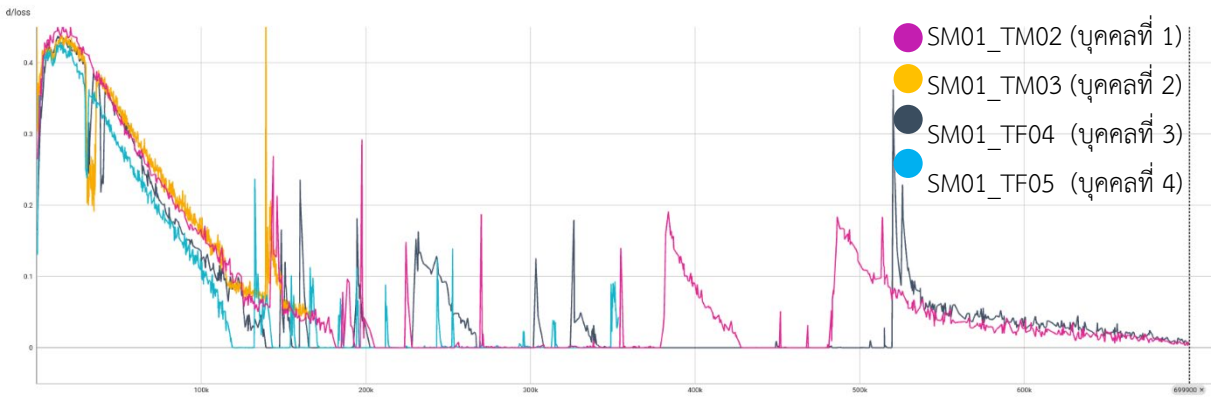
ตารางที่ 4 Text-to-Speech VS Voice Conversion

Speech Synthesis	Text-to-Speech	Voice Conversion
ความหมาย	เปลี่ยนข้อความเป็นเสียงพูด	เปลี่ยนเสียงพูดของบุคคลหนึ่งให้เหมือนกับเสียงของบุคคลอื่น
Input / Output	Text / Speech	Speech / Speech
ลักษณะของชุดข้อมูลเสียง	แบบคู่ขนาน (ไฟล์เสียง+ไฟล์บทพูด)	แบบไม่คู่ขนาน (ไฟล์เสียง)
ภาษาที่ใช้ในการฝึกฝน	ขึ้นกับภาษา	ไม่ขึ้นกับภาษา
วิธีการปกปิดตัวตนด้วยเสียง	✓	✓
ความเป็นธรรมชาติ	เสียงอาจฟังดูไม่เป็นธรรมชาติ	เสียงมีความเป็นธรรมชาติสูง
การใช้วิจารณญาณในการตัดสินใจ	แยกแยะได้โดยง่าย	แยกแยะได้ยากเพราะเสียงมีความคล้ายคลึงกับต้นฉบับและเป็นธรรมชาติสูง
เทคโนโลยีที่เลือกใช้	- Convolutional Neural Network - Recurrent Neural Network - Generative Adversarial Network	- Convolutional Neural Network - Generative Adversarial Network
ตัวอย่างการใช้งาน	- เน้นอำนวยความสะดวก - ระบบผู้ช่วยเสียง / อ่านออกเสียง - ระบบนำทาง GPS - พากย์เสียง - ช่วยเหลือผู้พิการทางสายตา	- เน้นเลียนแบบเสียงบุคคล - สร้างเนื้อหาที่น่าดึงดูดใจ - ความบันเทิง - พากย์เสียง - ช่วยเหลือผู้ที่มีความผิดปกติในการพูด [33]

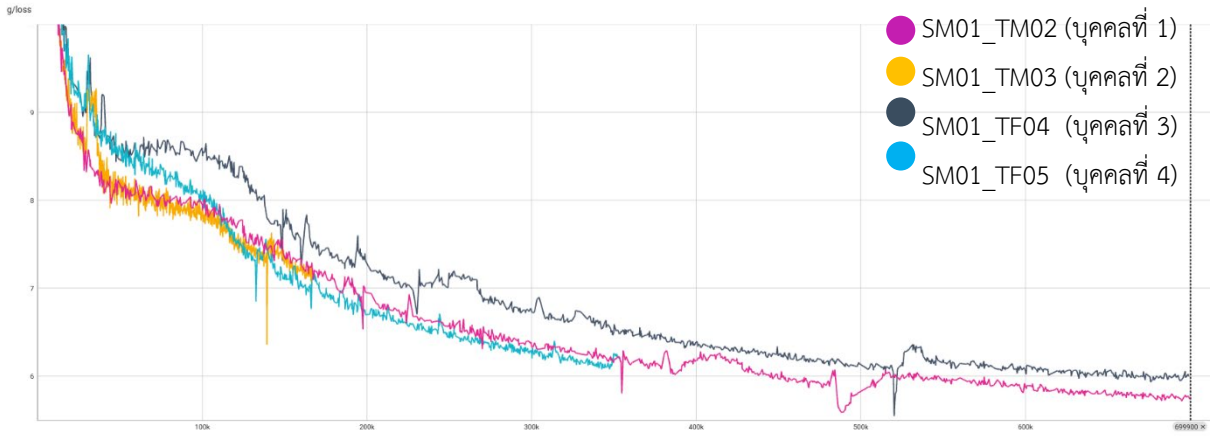
5.2 ผลการศึกษาพัฒนาและปรับแต่งแบบจำลอง MaskCycleGAN-VC ปลอมแปลงเสียงบุคคลให้มีความเหมาะสมกับการใช้งานบนมิติไซเบอร์

5.2.1 การฝึกฝนแบบจำลองที่จำนวน 7000 Epochs

ผลการทดลองดังรูปที่ 12, 13 และตารางที่ 5



รูปที่ 12 Discriminator Loss (7000 Epochs)



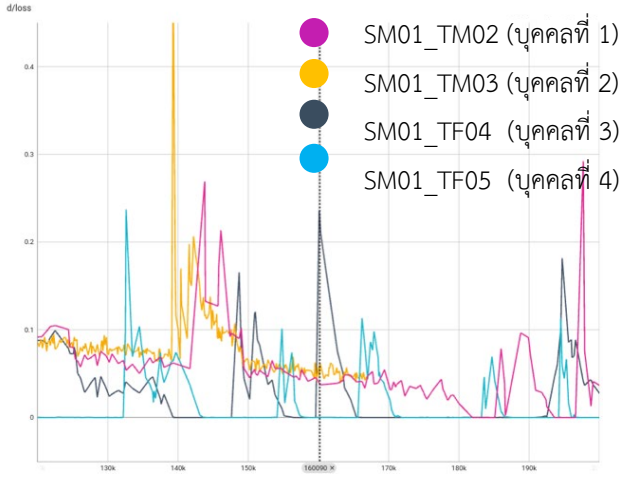
รูปที่ 13 Generator Loss (7000 Epochs)

ตารางที่ 5 การฝึกฝนแบบจำลอง (7000 Epochs)

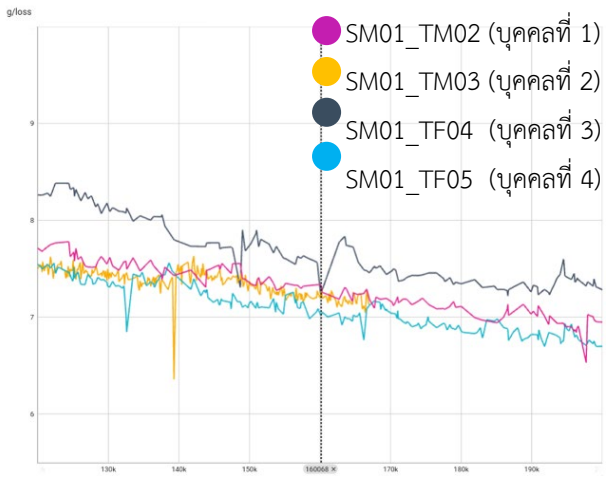
Model	Generator Loss	Discriminator Loss	Duration
SM01_TM02	5.7398	0.0060	106 ชั่วโมง
SM01_TF04	6.0054	0.0069	106 ชั่วโมง
Average	5.8726	0.0065	106 ชั่วโมง

5.2.2 การฝึกฝนแบบจำลองด้วยเทคนิค Early Stopping

ผลการทดลองดังรูปที่ 14, 15 และตารางที่ 6



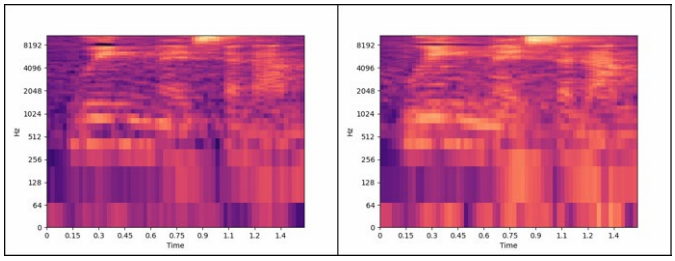
รูปที่ 14 Discriminator Loss (1600 Epochs)



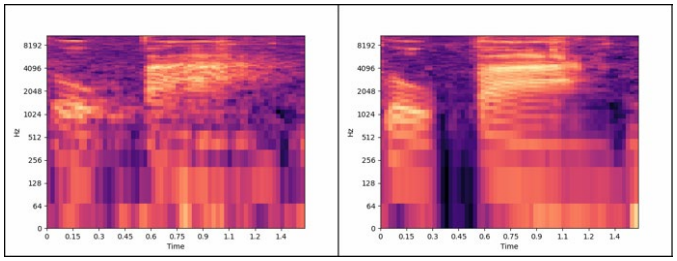
รูปที่ 15 Generator Loss (1600 Epochs)

ตารางที่ 6 การฝึกฝนแบบจำลอง (1600 Epochs)

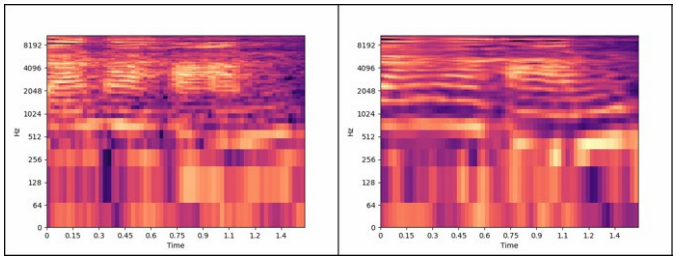
Model	Generator Loss	Discriminator Loss	Duration	Mel-Spectrogram
SM01_TM02	7.3355	0.0395	24 ชั่วโมง	ดังรูปที่ 16
SM01_TM03	7.2543	0.0475	24 ชั่วโมง	ดังรูปที่ 17
SM01_TF04	7.2170	0.2357	24 ชั่วโมง	ดังรูปที่ 18
SM01_TF05	7.0200	0.0002	24 ชั่วโมง	ดังรูปที่ 19
Average	7.2067	0.0807	24 ชั่วโมง	



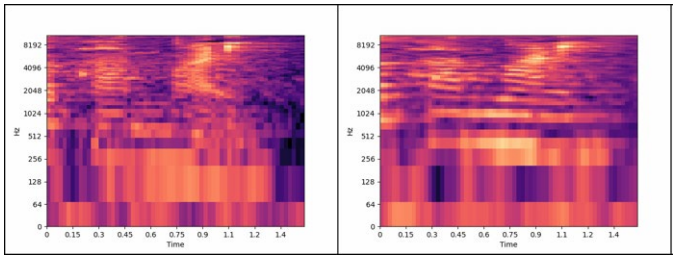
รูปที่ 16 Mel-Spectrogram เสียงจริง (ผู้วิจัย) แปลงเป็น เสียงปลอม (บุคคลเป้าหมายที่ 1)



รูปที่ 17 Mel-Spectrogram เสียงจริง (ผู้วิจัย) แปลงเป็น เสียงปลอม (บุคคลเป้าหมายที่ 2)



รูปที่ 18 Mel-Spectrogram เสียงจริง (ผู้วิจัย) แปลงเป็น เสียงปลอม (บุคคลเป้าหมายที่ 3)



รูปที่ 19 Mel-Spectrogram เสียงจริง (ผู้วิจัย) แปลงเป็น เสียงปลอม (บุคคลเป้าหมายที่ 4)

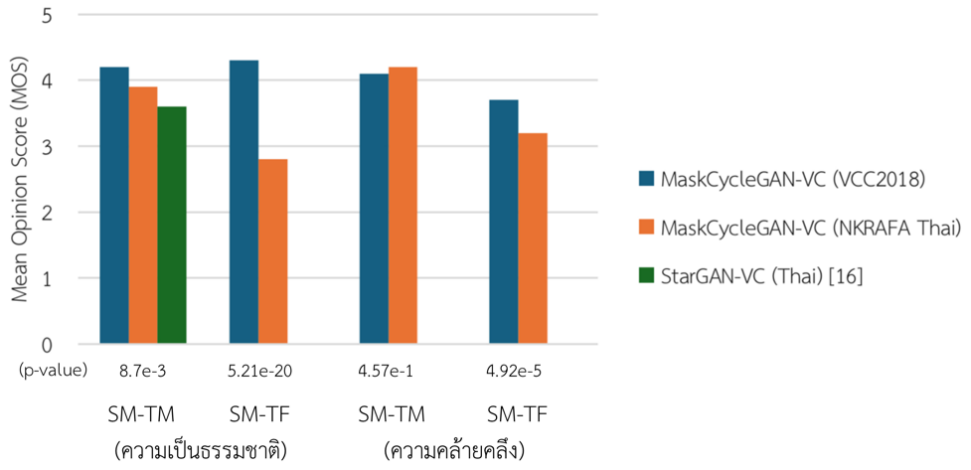
5.3 ผลการประเมินประสิทธิภาพของเสียง

5.3.1 คุณภาพของเสียงที่ได้จากการแปลงเสียง

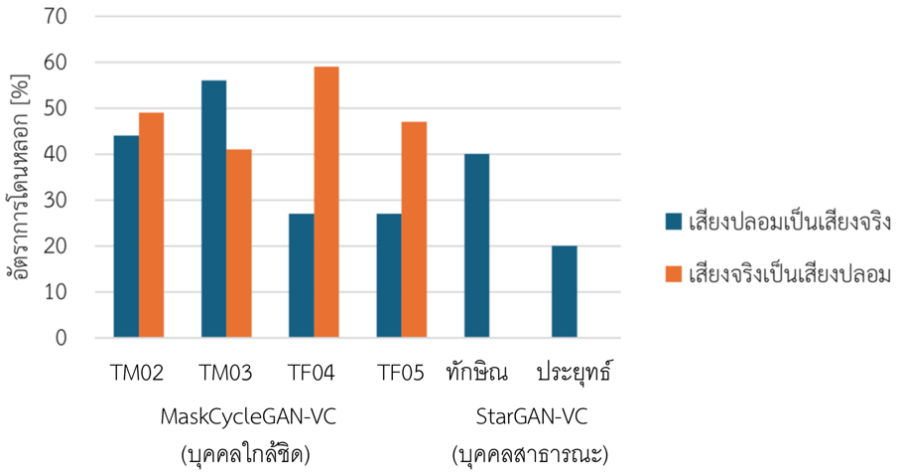
ผลการทดสอบแบบจำลอง พบว่ามีสัญญาณรบกวนในระดับน้อยถึงปานกลาง ทำให้คุณภาพเสียงต่ำและไม่เหมาะสมสำหรับการใช้งานจริง เนื่องจากเสียงปลอมที่สร้างขึ้นสามารถแยกแยะได้ง่าย เช่น เสียงพูดไม่ชัด เสียงเพี้ยน หรือเสียงบิดเบี้ยวที่คล้ายเสียงเอเลี่ยน เพื่อแก้ไขปัญหานี้ ผู้วิจัยจึงคัดเลือกเฉพาะเสียงปลอมที่มีคุณภาพดี และเพิ่ม สัญญาณรบกวนให้เสียงจริงอยู่ในระดับใกล้เคียงกับเสียงปลอม เพื่อให้การประเมินประสิทธิภาพมีความยุติธรรมมากขึ้น

5.3.2 ผลการประเมินเชิงคุณภาพ

ผลการประเมินคุณภาพเสียงด้วยค่า MOS และการทดสอบทางสถิติด้วย One-sample T-test ดังรูปที่ 20 และผลการประเมินประสิทธิผลในการหลอกลวงผู้ฟัง ดังรูปที่ 21



รูปที่ 20 การประเมินคุณภาพเสียงด้วยค่า MOS (ค่า Confidence Intervals ที่ 95%)



รูปที่ 21 การประเมินประสิทธิผลในการหลอกลวงผู้ฟัง

5.3.3 ผลการประเมินเชิงปริมาณ

ผลการประเมินเชิงปริมาณด้วยค่า MCD และ KDSD ซึ่งเป็นตัวชี้วัดความแตกต่างระหว่างเสียงที่แปลงกับเสียงต้นฉบับ ตามตารางที่ 7

ตารางที่ 7 การประเมินเชิงปริมาณด้วยค่า MCD และ KDSD

ลำดับ	ชุดข้อมูล	แบบจำลองรู้จำเสียงพูด	MCD [dB] / KDSD [KDSD]	
			SM-TM (ชาย-ชาย)	SM-TF (ชาย-หญิง)
1	NKRAFA Thai	wav2vec2-large-xlsr-53-th (2021)	6.01 / 15.90 x 10^{-3}	5.00 / 16.23 x 10^{-3}
2	VCC2018	DeepSpeech2 (2016)	6.77 / 83.8 x 10^{5}	7.64 / 502 x 10^{5}

5.4 การอภิปรายผล

5.4.1 การแปลงข้อความเป็นเสียง

ตามตารางที่ 2 พบว่า การเลือกใช้งานขึ้นอยู่กับวัตถุประสงค์ หากต้องการระบบอ่านออกเสียงที่มีความรวดเร็วและใช้งานฟรี Google Translate จะเหมาะสมกว่า เนื่องจากเป็นแบบจำลองที่เน้นการใช้งานแปลภาษาเป็นหลัก โดยไม่เน้นว่าจะต้องแปลงเป็นเสียงบุคคลใด ทำให้มีความซับซ้อนน้อยกว่า แต่หากต้องการคุณภาพเสียงที่สมจริง Speechify จะเป็นตัวเลือกที่ดีกว่า

5.4.2 การแปลงเสียง

ตามตารางที่ 3 พบว่า การเลือกใช้แบบจำลองขึ้นอยู่กับวัตถุประสงค์ของงาน หากต้องการการแปลงเสียงที่มีคุณภาพสูง MaskCycleGAN-VC จะเหมาะสมกว่า เนื่องจากแบบจำลองนี้เป็นการแปลงเสียงรูปแบบหนึ่งต่อหนึ่ง ทำให้สามารถเรียนรู้เสียงบุคคลเป้าหมายได้อย่างแม่นยำ และใช้เทคนิค Cycle-Consistency และ FIF เป็นการเพิ่มคุณภาพเสียงที่แปลงให้มีความสมจริงมากยิ่งขึ้นโดยที่ยังคงเนื้อหาเดิมไว้ รวมถึงใช้ชุดข้อมูลเสียงแบบไม่คู่ขนานซึ่งสามารถเตรียมได้ง่ายกว่า ขณะที่ LLVC เหมาะสำหรับการแปลงเสียงที่ต้องการความรวดเร็วแบบเรียลไทม์ แต่มีข้อจำกัดด้านคุณภาพเสียงที่ต่ำกว่า

5.4.3 ความแตกต่างระหว่างการแปลงข้อความเป็นเสียงและการแปลงเสียง

ตามตารางที่ 4 พบว่า การแปลงข้อความเป็นเสียงและการแปลงเสียง มีจุดประสงค์และแนวทางที่แตกต่างกัน โดยการแปลงข้อความเป็นเสียงเหมาะสำหรับงานที่ต้องการระบบอ่านออกเสียง ขณะที่การแปลงเสียงเป็นการเปลี่ยนเสียงของบุคคลหนึ่งให้คล้ายกับอีกบุคคลเหมาะสำหรับงานที่มุ่งเน้นการสร้างเสียงที่สมจริงและเป็นธรรมชาติ

5.4.4 การฝึกฝนแบบจำลองที่จำนวน 7000 Epochs

ดังรูป 12, 13 และ ตารางที่ 5 พบว่า ค่า Generator Loss ลดลงอย่างรวดเร็วในช่วงต้นและเริ่มเข้าสู่จุดอิ่มตัวที่ประมาณ 4000 - 5000 Epochs ขณะที่ Discriminator Loss มีค่าต่ำมาก แสดงว่าโครงข่ายผู้แยกแยะมีประสิทธิภาพสูงในการจำแนกเสียงจริงกับเสียงปลอม อย่างไรก็ตาม การลดลงของ Generator Loss ที่ช้าลงบ่งบอกถึงข้อจำกัดของแบบจำลองในการปรับปรุงคุณภาพเสียงเพิ่มเติม ซึ่งอาจใช้เทคนิค Early Stopping หรือปรับ Learning Rate เพื่อลดเวลาฝึกควบคู่กับการฟังตัวอย่างเสียงเพื่อประเมินคุณภาพก่อนนำไปใช้งานจริง

5.4.5 การฝึกฝนแบบจำลองด้วยเทคนิค Early Stopping

ดังรูปที่ 14, 15 และ ตารางที่ 6 พบว่า การหยุดการฝึกที่จำนวน 1600 Epochs เป็นแนวทางที่เหมาะสมเพื่อลดเวลาฝึก โดยผู้วิจัยพิจารณาจำนวนดังกล่าวนี้จากจุดตัดของกราฟที่ได้จากการสังเกตแนวโน้มของค่า

ความผิดพลาดและประสิทธิภาพของแบบจำลองโดยรวม ขณะเดียวกันยังสามารถรักษาประสิทธิภาพของแบบจำลองได้โดยไม่ต้องใช้ทรัพยากรเกินจำเป็น

5.4.6 คุณภาพของเสียงที่ได้จากการแปลงเสียง

ปัญหาคุณภาพเสียงในงานวิจัยนี้เกิดจาก NKRAFA Thai Dataset ที่ใช้ยังไม่เทียบเท่ากับมาตรฐานอย่าง VCC2018 Dataset ซึ่งได้รับการออกแบบมาโดยเฉพาะสำหรับงานแปลงเสียง ปัญหาหลักมาจากกระบวนการเก็บตัวอย่างเสียงที่อาจไม่มีการควบคุมปัจจัยสำคัญ เช่น ระยะไมค์จากแหล่งกำเนิดเสียง ชนิดของไมโครโฟนที่ใช้ในการบันทึก ความแรงของสัญญาณรบกวนภายนอก และสภาพแวดล้อมของห้องอัดเสียง หากไม่มีการจัดการที่ดี ปัจจัยเหล่านี้อาจทำให้เกิดความคลาดเคลื่อนในคุณลักษณะของเสียงที่ถูกบันทึก เช่น ความผิดเพี้ยนของโทนเสียง ระดับเสียงที่ไม่สม่ำเสมอ หรือความผิดพลาดในโครงสร้างทางสเปกตรัมของเสียง ซึ่งส่งผลต่อการเรียนรู้ของแบบจำลองโดยตรงทำให้ไม่สามารถสร้างเสียงที่มีคุณภาพสูงหรือใกล้เคียงกับต้นฉบับได้อย่างแม่นยำ คุณภาพของชุดข้อมูลจึงเป็นปัจจัยสำคัญที่ต้องได้รับการพัฒนาเพื่อลดข้อผิดพลาดและปรับปรุงผลลัพธ์ของแบบจำลองให้ดียิ่งขึ้น

5.4.7 การประเมินเชิงคุณภาพ

ดังรูปที่ 20 พบว่า MaskCycleGAN-VC มีประสิทธิภาพสูงกว่า StarGAN-VC โดยเฉพาะในกรณีภาษาไทยที่มีค่าความเป็นธรรมชาติสูงกว่าอย่างมีนัยสำคัญ อย่างไรก็ตามค่าความคล้ายคลึงของเสียงในกรณีแปลงเพศเดียวกันไม่มีความแตกต่างทางสถิติอย่างชัดเจน ยกเว้นกรณีแปลงเสียงข้ามเพศที่แตกต่างอย่างมีนัยสำคัญ ($p < 0.05$) ซึ่งสะท้อนถึงข้อจำกัดของชุดข้อมูล NKRAFA Thai เมื่อเทียบกับชุดข้อมูลมาตรฐาน VCC2018

ดังรูปที่ 21 พบว่า ผู้ฟังในกลุ่มบุคคลใกล้ชิดมีการตัดสินใจผิดพลาดใน 2 กรณีหลัก ได้แก่ 1) เสียงปลอมถูกตัดสินว่าเป็นเสียงจริง โดยมีอัตราการโดนหลอกสูงสุดถึง 56% ซึ่งมากกว่ากลุ่มบุคคลสาธารณะที่ผู้ฟังไม่คุ้นเคยกับเสียง แสดงให้เห็นถึงศักยภาพของการสร้างเสียงปลอมที่สามารถสร้างได้ใกล้เคียงจนยากต่อการตรวจจับ 2) เสียงจริงถูกตัดสินว่าเป็นเสียงปลอม โดยมีอัตราการโดนหลอกสูงสุดถึง 59% ชี้ให้เห็นว่าสัญญาณรบกวนที่ใส่ในเสียงจริงมีผลกระทบต่อการตัดสินใจของผู้ฟังอย่างมีนัยสำคัญ ความคล้ายคลึงระหว่างเสียงจริงและเสียงปลอมในสถานการณ์ที่มีเสียงรบกวนทำให้เกิดความสับสนและความยากในการแยกแยะ สะท้อนถึงความท้าทายในการวิเคราะห์เสียงในสภาพแวดล้อมจริงและตอกย้ำว่าคุณภาพเสียงรวมถึงระดับสัญญาณรบกวนเป็นปัจจัยสำคัญที่ต้องพิจารณาในการวิจัยด้านเสียงปลอมแปลง

5.4.8 การประเมินเชิงปริมาณ

ตามตารางที่ 7 พบว่า MCD และ KDSD ในภาษาไทยมีค่าต่ำกว่าในภาษาอังกฤษ สะท้อนถึงความคล้ายคลึงของเสียงในภาษาไทยที่มากกว่า สำหรับ KDSD ที่มีค่าแตกต่างกันมากนั้น สาเหตุอาจมาจากการใช้เลือกใช้แบบจำลองรู้จำเสียงที่ทันสมัยในปัจจุบันและมีความเหมาะสมกับภาษาไทย นอกจากนี้วิธีการนี้ยังมีศักยภาพในการนำไปใช้ตรวจจับเสียงปลอมเบื้องต้น ซึ่งอาจเป็นประโยชน์ต่อการวิจัยที่เกี่ยวข้องในอนาคต

5.5 การเผยแพร่ซอร์สโค้ดและชุดข้อมูลเสียง

ผู้วิจัยได้เผยแพร่ซอร์สโค้ด รวมถึง NKRAFA Thai Dataset ที่ใช้ในการวิจัยนี้ ผ่านแพลตฟอร์ม GitHub เว็บไซต์ https://github.com/NKRAFAVoiceAI/MaskCycleGAN-VC เพื่อประโยชน์ในแวดวงงานวิจัยและการพัฒนาต่อยอดในอนาคต

6. สรุป

6.1 ผลการวิจัยการประยุกต์ใช้เทคโนโลยีปัญญาประดิษฐ์ในการปลอมแปลงเสียงบุคคลที่มีความเหมาะสม งานวิจัยนี้ได้ศึกษาประเภทของปัญญาประดิษฐ์ที่สามารถนำมาใช้ในการปลอมแปลงเสียง พบว่า เทคโนโลยีปัญญาประดิษฐ์เชิงกำเนิดที่ทันสมัยที่สุดและได้รับการยอมรับว่าเป็นวิธีมาตรฐานในปัจจุบัน คือ แบบจำลอง

MaskCycleGAN-VC มีศักยภาพสูงในการสร้างเสียงปลอมที่สมจริง เลียนแบบจังหวะ น้ำเสียง และอารมณ์ของเสียงต้นฉบับได้แม่นยำ โดยสามารถเรียนรู้จากข้อมูลเสียงที่ไม่คู่ขนาน เหมาะสำหรับการใช้งานในมิติไซเบอร์ แม้ยังมีข้อจำกัดในกรณีเสียงมีสัญญาณรบกวนสูงหรือแปลงเสียงระหว่างเพศ เพื่อแก้ไขปัญหาดังกล่าว ผู้วิจัยเสนอทั้งการปรับแต่งแบบจำลองและการเตรียมชุดข้อมูลให้มีคุณภาพตรงตามมาตรฐาน

6.2 การพัฒนาและปรับแต่งแบบจำลองการปลอมแปลงเสียงด้วยปัญญาประดิษฐ์เชิงกำเนิดสำหรับการใช้งานบนมิติไซเบอร์ แบบจำลอง MaskCycleGAN-VC ถูกเลือกเนื่องจากสามารถรักษาคุณลักษณะเสียงได้ดีกว่าแบบจำลองอื่น โดยมีการใช้เทคนิค Filling in Frames (FIF) เพื่อให้เสียงที่แปลงมีความต่อเนื่อง ปรับพารามิเตอร์ให้มีความเหมาะสม และใช้ Early Stopping ที่ 1600 Epochs เพื่อเพิ่มประสิทธิภาพและลดเวลา อีกทั้งใช้ Mel-Spectrogram แทน MFCC ซึ่งเป็นการสกัดคุณลักษณะแบบเดิมที่ใช้ในงานวิจัยอื่นๆ เพื่อให้เรียนรู้โครงสร้างเสียงได้แม่นยำยิ่งขึ้น

6.3 ผลการประเมินประสิทธิภาพและศักยภาพในการหลอกลวงโดยใช้เสียงปลอมแปลงที่สร้างจากแบบจำลอง ผู้วิจัยได้ดำเนินการประเมินผลทั้งในเชิงคุณภาพและเชิงปริมาณ พบว่า เสียงปลอมที่สร้างขึ้นมีค่า MOS สูงสุด 3.9 และความคล้ายคลึง 4.2 โดยสามารถหลอกผู้ฟังได้ถึง 56% และเสียงจริงถูกเข้าใจผิดว่าเป็นเสียงปลอมถึง 59% ในเชิงปริมาณ ค่า MCD ต่ำสุด 5.0 dB และ KDSD ต่ำสุด 15.9 mKDSD แสดงถึงคุณภาพเสียงปลอมที่ใกล้เคียงเสียงจริง

6.4 ผู้วิจัยได้เสนอแนวทางการประยุกต์ใช้งานเทคโนโลยีปัญญาประดิษฐ์เชิงกำเนิดในการปฏิบัติการทางไซเบอร์เชิงรุกซึ่งประกอบด้วย การใช้เสียงปลอมในงานข่าวกรอง การทดสอบระบบรักษาความปลอดภัย และการฝึกซ้อมรับมือภัยไซเบอร์ พร้อมเตือนถึงความเสี่ยงด้านจริยธรรมและความมั่นคงจากการใช้เทคโนโลยีในทางที่ผิด จึงเสนอให้พัฒนาเทคโนโลยีตรวจจับเสียงปลอมควบคู่กัน เพื่อป้องกันการใช้งานที่ไม่เหมาะสมและเสริมความมั่นคงทางไซเบอร์

7. ข้อเสนอแนะ

ผลการวิจัยชี้ให้เห็นว่าปัญญาประดิษฐ์เชิงกำเนิดสามารถสร้างเสียงปลอมแปลงที่มีความสมจริงสูงจนยากที่มนุษย์จะตรวจจับได้ และจากการทดสอบพบว่าเสียงที่มีความยาว 3 วินาที ใช้ระยะเวลาในการแปลงเสียงเพียง 1.63 วินาที ซึ่งแสดงให้เห็นถึงประสิทธิภาพและศักยภาพในการพัฒนาให้รองรับการใช้งานแบบเรียลไทม์

สำหรับแนวทางการพัฒนาเพิ่มเติมผู้วิจัยเห็นว่าแบบจำลอง MaskCycleGAN-VC สามารถปรับปรุงประสิทธิภาพการสร้างเสียงได้โดยใช้ HiFi-GAN เป็นตัวสังเคราะห์เสียง เนื่องจากให้มีคุณภาพเสียงที่สูงขึ้น ลดปัญหาเสียงผิดเพี้ยน และสามารถสร้างเสียงที่มีความใกล้เคียงกับเสียงมนุษย์มากขึ้น อีกทั้งยังมีความเร็วในการสังเคราะห์สูง รองรับการใช้งานแบบเรียลไทม์ ซึ่งช่วยให้ MaskCycleGAN-VC สามารถแปลงเสียงได้อย่างมีประสิทธิภาพมากขึ้นและตอบโจทย์การใช้งานที่ต้องการความสมจริงและความรวดเร็ว

8. กิตติกรรมประกาศ

การดำเนินการวิจัยนี้สามารถดำเนินการจนประสบความสำเร็จลุล่วงไปได้ด้วยดี ขอขอบคุณกองทัพอากาศที่สนับสนุนทุนการศึกษาแก่ผู้ดำเนินการวิจัย ขอขอบคุณโรงเรียนนายเรืออากาศนวมินทกษัตริยาธิราชที่อนุญาตให้ใช้สถานที่เพื่อการวิจัย ขอขอบคุณคณะกรรมการติดตามงานวิจัยและอาจารย์ที่ปรึกษาที่ให้คำปรึกษาและให้ข้อเสนอแนะ ขอขอบคุณอาจารย์ภาควิชาคณิตศาสตร์และคอมพิวเตอร์และนักเรียนนายเรืออากาศที่ได้ให้ความอนุเคราะห์และความยินยอมในการบันทึกและใช้เสียงเพื่อการวิจัย ขอขอบคุณบุคลากรสำนักบัณฑิตศึกษาที่ให้การช่วยเหลือและอำนวยความสะดวกในเรื่องต่างๆ

9. เอกสารอ้างอิง

[1] Gartner, *"Gartner Identifies Top Cybersecurity Trends for 2024"* [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2024-02-22-gartner-identifies-top-cybersecurity-trends-for-2024. (Accessed: July. 20, 2024).

[2] M. Jovanović and M. Campbell, "Generative Artificial Intelligence: Trends and Prospects," *IEEE Computer Society*, vol. 55, no. 10, pp. 107-112, 2022.

[3] ฐานเศรษฐกิจ, *"มิจฉาชีพใช้ AI Clone เสียงหลอกโอนเงิน"* [ออนไลน์]. Available: https://www.thansettakij.com/technology/technology/576101. (เข้าถึงเมื่อ: 20 กรกฎาคม 2567).

[4] สำนักงานตำรวจแห่งชาติ, *"4 รูปแบบอาชญากรรมออนไลน์ที่ต้องจับตามองในปี 2567 เมื่อ AI ถูกใช้ในด้านมืด ปลอมได้สารพัด"* [ออนไลน์]. Available: https://www.facebook.com/photo.php?fbid=766631268843499. (เข้าถึงเมื่อ: 20 กรกฎาคม 2567).

[5] กระทรวงกลาโหม, *"แผนการพัฒนาวิทยาศาสตร์และเทคโนโลยีป้องกันประเทศ พ.ศ. 2566-2570"* [ออนไลน์]. Available: https://dstd.mod.go.th/getdoc/32fad3cd-50b1-46a0-b8b8-5c187ca9815a/planresearch-66-70.aspx. (เข้าถึงเมื่อ: 23 กรกฎาคม 2567).

[6] กองทัพอากาศ, *"ยุทธศาสตร์กองทัพอากาศ 20 ปี (พ.ศ.2561-2580) (ฉบับปรับปรุงพ.ศ.2563)"* [ออนไลน์]. Available: www.rtaf.mi.th/th/Documents/Publication/RTAF%20Strategy_Final_04122563.pdf. (เข้าถึงเมื่อ: 23 กรกฎาคม 2567).

[7] กองทัพอากาศ, *"นโยบายผู้บัญชาการทหารอากาศประจำปีพุทธศักราช 2567-2568"* [ออนไลน์]. Available: https://heyzine.com/flip-book/e12bf07274.html#page/1. (เข้าถึงเมื่อ: 23 กรกฎาคม 2567).

[8] สำนักเลขาธิการคณะรัฐมนตรี, *"ยุทธศาสตร์ชาติ พ.ศ. 2561-2580"* [ออนไลน์]. Available: https://www.ratchakitcha.soc.go.th/DATA/PDF/2561/A/082/T_0001.PDF. (เข้าถึงเมื่อ: 23 กรกฎาคม 2567).

[9] ศูนย์ไซเบอร์กองทัพอากาศ, *"ความรู้พื้นฐานสำหรับปฏิบัติการทางไซเบอร์ พ.ศ.2566"* [ออนไลน์]. Available: https://cybercenter.rtaf.mi.th/wp-content/uploads/2024/02/01.วิชาความรู้พื้นฐานสำหรับปฏิบัติการทางไซเบอร์-1.pdf. (เข้าถึงเมื่อ: 25 กรกฎาคม 2567).

[10] T. Walczyna and Z. Piotrowski, "Overview of voice conversion methods based on deep learning," *Applied Sciences*, vol. 13, no. 5, pp. 3100, 2023.

[11] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100-2104.

[12] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6820–6824.

[13] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-spectrogram Conversion," in *Proc. Interspeech*, 2020.

[14] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “MaskCycleGAN-VC: Learning Non-parallel Voice Conversion with Filling in Frames,” in *Proc. ICASSP*, 2021, pp. 5919-5923.

[15] I. H. Sarker, “Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions,” *Springer Journal*, vol. 2, no. 6, pp. 420, 2021.

[16] พายัพ ศิรินาม และ ประสงค์ ปราณีตพลกรัง, “การพัฒนาโมเดลการเลียนเสียงเชิงลึกในการประยุกต์ใช้งานด้านสงครามไซเบอร์,” *วารสารสถาบันวิชาการป้องกันประเทศ*, ปีที่ 14, ฉบับที่ 1, หน้า 162–178, มกราคม-มิถุนายน, 2566.

[17] F. Khanam, F. A. Munmun, N. A. Ritu, A. K. Saha, and M. F. Mridha, “Text to Speech Synthesis: A Systematic Review, Deep Learning Based Architecture and Future Research Direction,” *Journal of Advances in Information Technology*, vol. 13, no. 5, pp. 398-412, October 2022.

[18] Y. A. Li, A. A. Zare, and N. Mesgarani, “StarGANv2-VC: A Diverse, Unsupervised, Non-parallel Framework for Natural-Sounding Voice Conversion,”in *Proc. Interspeech*, 2021.

[19] สถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี, *รายวิชาเพิ่มเติมวิทยาศาสตร์และเทคโนโลยี ฟิสิกส์ ชั้นมัธยมศึกษาปีที่ 5 เล่มที่ 4*. พิมพ์ครั้งที่ 1. กรุงเทพมหานคร: สถาบันส่งเสริมการสอนวิทยาศาสตร์และเทคโนโลยี, 2563.

[20] R. C. Streijl, S. Winkler, and D. S. Hands, “Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives,” *Multimedia Systems*, vol. 22, no. 2, pp. 213-227, March 2016.

[21] X. Liang, Z. Bie, and S. Ma, “Pyramid Attention CycleGAN for Non-Parallel Voice Conversion,” in *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*, 2022, pp. 139-143.

[22] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *Proc. IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing*, 1993, pp. 125-128.

[23] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE Trans. Inf. Syst*., vol. 99, no. 7, pp. 1877-1884, July 2016.

[24] M. Shannon, *“MCD: Mel-Cepstral Distortion”* [Online]. Available: https://github.com/MattShannon/mcd. (Accessed: November. 12, 2024).

[25] M. Binkowski, et al., “High fidelity speech synthesis with adversarial networks,” in *Proc. ICLR*, 2020.

[26] D. Amodei, et al., “Deep Speech 2: End-to-end speech recognition in English and Mandarin,” in *Proc. ICML*, 2016, pp. 173-182.

[27] Speechify, *“Text-to-Speech Online in Thai,”* [Online]. Available: https://speechify.com/text-to-speech-online/thai/. (Accessed: July. 25, 2024).

[28] K. Sadov, M. Hutter, and A. Near, *"Low-latency real-time voice conversion on CPU,"* [Online]. Available: https://github.com/KoeAI/LLVC. (Accessed: July. 25, 2024).

[29] J. Lorenzo-Trueba, et al., "The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, Les Sables d'Olonne, France, Jun. 26–29, 2018.

[30] S. Broughton, *"Mel-Cepstral Distortion,"* [Online]. Available: https://github.com/SamuelBroughton/Mel-Cepstral-Distortion. (Accessed: November. 12, 2024).

[31] HuggingFace, *"wav2vec2-large-xlsr-53-th,"* [Online]. Available: https://huggingface.co/airesearch/wav2vec2-large-xlsr-53-th. (Accessed: November. 12, 2024).

[32] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, 2020.

[33] H. Y. Lwin, W. Kumwilaisak, C. Hansakunbuntheung, and N. Thatphithakkul, "Alaryngeal Speech Generation Using MaskCycleGAN-VC and Timbre-Enhanced Loss," in *Proc. 13th International Conference on Advances in Information Technology (IAIT 2023)*, December 06-09, Bangkok, Thailand, 2023.