การพัฒนาต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความสำหรับการประชุมอิเล็กทรอนิกส์ สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม

พิมลพรรณ แซ่เหลี่ยว[1*] , จิรันดร บู๊ฮวดใช้[1] และเดช ธรรมศิริ[2]

[1]สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม
[2]สาขาวิชาธุรกิจดิจิทัล คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏนครปฐม
*pimonpan_sa@webmail.npru.ac.th

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์ 1) เพื่อพัฒนาต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความสำหรับการประชุมอิเล็กทรอนิกส์ สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม 2) เพื่อเปรียบเทียบความแม่นยำของโมเดลสำหรับต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความสำหรับการประชุมอิเล็กทรอนิกส์ สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม เนื่องจากปัญหาที่ผ่านมาผู้วิจัยต้องใช้ระยะเวลานานในการถอดไฟล์เสียงสรุปการประชุม และไม่ต้องการนำไฟล์เสียงไปประมวลผลแบบกลุ่มเมฆ เพื่อรักษาความลับภายในองค์กร ผู้วิจัยจึงพัฒนาต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความ โดยใช้โมเดลภาษาขนาดใหญ่ Whisper ด้วยภาษา PHP และภาษา Python โดยใช้ข้อมูลเสียงที่สร้างขึ้นจำนวน 3 ข้อความ โดยมีความยาว 10 ตัวอักษร 42 ตัวอักษร และ 121 ตัวอักษร และใช้เครื่องคอมพิวเตอร์ส่วนบุคคล ทดลองซ้ำจำนวน 3 รอบ ด้วยโมเดล Whisper จำนวน 6 ขนาด ได้แก่ 1) Tiny 2) Base 3) Small 4) Medium 5) Large และ 6) Turbo

ผลการวิจัยพบว่าโมเดล Whisper ขนาด Turbo มีค่าความแม่นยำเฉลี่ย 67.81% และใช้ระยะเวลาประมวลผลเฉลี่ย 38.54 วินาที เหมาะสมที่สุดจากโมเดลทั้งหมด โดยความยาวตัวอักษร 10 ตัวอักษร มีค่าความแม่นยำเฉลี่ย 100% ระยะเวลาประมวลผลเฉลี่ย 28.72 วินาที ความยาวตัวอักษร 42 ตัวอักษร มีค่าความแม่นยำเฉลี่ย 57.14% ระยะเวลาประมวลผลเฉลี่ย 30.41 วินาที และความยาวตัวอักษร 121 ตัวอักษร มีค่าความแม่นยำเฉลี่ย 46.28% ระยะเวลาประมวลผลเฉลี่ย 56.50 วินาที จากนั้นผู้วิจัยนำโมเดล Whisper ขนาด Turbo ไปใช้งาน และปรับปรุงผลลัพธ์หลังการประมวลผลการแปลงเสียงเป็นข้อความภาษาไทยให้มีความแม่นยำดียิ่งขึ้น

คำสำคัญ: เว็บแอปพลิเคชัน แปลงเสียงเป็นข้อความ การประชุมอิเล็กทรอนิกส์ โมเดลภาษาขนาดใหญ่ Whisper

Development of a Prototype of the Speech-to-Text Web Application for Electronic Meetings for the Research and Development Institute of Nakhon Pathom Rajabhat University

Pimonpan Saliew[1*] , Jirundon Buhuatchai[1] and Dech Thammasiri[2]

[1]Research and Development Institute, Nakhon Pathom Rajabhat University
[2]Department of Digital Bussiness, Faculty of management science, Nakorn Pathom Rajabhat University
*pimonpan_sa@webmail.npru.ac.th

Abstract

The objectives of this research are to 1) develop a prototype of the speech-to-text web application for electronic meetings for the Research and Development Institute of Nakhon Pathom Rajabhat University. 2) To compare the accuracy of models for the prototype of the speech-to-text web application for electronic meetings for the Research and Development Institute of Nakhon Pathom Rajabhat University. Due to the previous problem, the researcher had to spend a long time transcribing the meeting summary audio files and did not want to process the audio files in the cloud to maintain confidentiality within the organization. The researchers therefore developed a prototype of a speech-to-text web application using the large language model Whisper using PHP and Python to use the generated voice data of 3 messages with lengths of 10 characters, 42 characters, and 121 characters. And using a personal computer, the experiment was repeated 3 times with 6 sizes of Whisper models: 1) Tiny 2) Base 3) Small 4) Medium 5) Large, and 6) Turbo.

The research results found that the Turbo Whisper model has an average accuracy of 67.81% and an average processing time of 38.54 seconds, making it the most suitable among all models. For a character length of 10 characters, the average accuracy is 100%, with an average processing time of 28.72 seconds. For a character length of 42 characters, the average accuracy is 57.14%, with an average processing time of 30.41 seconds. For a character length of 121 characters, the average accuracy is 46.28%, with an average processing time of 56.50 seconds. The researchers then applied the Turbo Whisper model and improved the post-processing results of Thai speech-to-text conversion to achieve better accuracy.

Keywords: Web Application, Speech-to-Text, Electronic Meetings, Large Language Models: LLMs, Whisper

1. บทนำ

ปัจจุบันสถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม เป็นหน่วยงานที่มีหน้าที่รับผิดชอบภารกิจด้านการวิจัยและพัฒนา การจัดสรรทุนอุดหนุนการวิจัยงบประมาณรายได้ เช่น ทุนพัฒนานักวิจัยรุ่นใหม่ ทุนพัฒนานักวิจัยรุ่นกลาง

ทุนอุดหนุนการวิจัย โครงการวิจัยบูรณาการนักศึกษาและอาจารย์เพื่อการพัฒนาท้องถิ่น ทุนอุดหนุนการวิจัย โครงการวิจัยสถาบันบูรณาการงานวิจัยกับงานประจำ R to R เป็นต้น ตลอดจนการให้รางวัลเพื่อสร้างแรงจูงใจแก่อาจารย์ เช่น ค่าตอบแทนตีพิมพ์บทความวิจัยระดับชาติและนานาชาติ ค่าตอบแทนอนุสิทธิบัตรหรือสิทธิบัตร เป็นต้น ซึ่งการพิจารณาวาระดังกล่าวจะต้องผ่านความเห็นชอบจากการประชุมคณะอนุกรรมการบริหารจัดการทุนอุดหนุนการวิจัย หรือคณะอนุกรรมการกลั่นกรองข้อเสนอโครงการวิจัยเรียบร้อยแล้ว จึงเสนอให้คณะกรรมการกองทุนเพื่อการวิจัยเป็นผู้อนุมัติให้ทุนอุดหนุนการวิจัยหรือค่าตอบแทนดังกล่าว เนื่องจากปัญหาที่ผ่านมา เมื่อเสร็จสิ้นการประชุมแล้ว เจ้าหน้าที่ผู้ปฏิบัติงานจะต้องใช้เวลานานอย่างมากในการสรุปมติที่ประชุมหรือการถอดเทปเสียงการประชุมของคณะกรรมการ โดยเฉพาะการฟังเสียงของผู้เข้าร่วมประชุมแต่ละท่าน เพื่อนำมาสรุปการอภิปรายและมติที่ประชุมในแต่ละวาระ

อย่างไรก็ตามด้วยความก้าวหน้าทางเทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence: AI) ในปัจจุบันทำให้เกิดการพัฒนาโมเดลภาษาขนาดใหญ่ (Large Language Models: LLMs) เช่น GPT, Llama หรือ Gemini ซึ่งเป็นโมเดลที่ได้รับการฝึกฝนด้วยข้อมูลจำนวนมหาศาล จนมีความสามารถในการทำความเข้าใจและสร้างภาษามนุษย์ได้อย่างเป็นธรรมชาติ โดยสามารถนำไปประยุกต์ใช้กับงานได้อย่างหลากหลาย รวมถึงการพัฒนาระบบรู้จำเสียงพูดอัตโนมัติ (Automatic Speech Recognition: ASR) ที่สามารถวิเคราะห์เสียง ถอดข้อความ หรือสรุปสาระสำคัญ หรือแม้กระทั่งการตอบคำถามได้อย่างมีประสิทธิภาพดียิ่งขึ้น แต่การประยุกต์ใช้โมเดลภาษาขนาดใหญ่ภายในองค์กร ยังคงเผชิญกับข้อจำกัดหลายประการ ทั้งในด้านความปลอดภัยของข้อมูล และข้อจำกัดด้านโครงสร้างพื้นฐานของเครื่องแม่ข่าย (Server) ที่ต้องมีสมรรถนะในการประมวลผลขั้นสูง เพื่อรองรับโมเดลที่มีความซับซ้อน โดยเฉพาะอย่างยิ่งในกรณีที่ต้องประมวลผลไฟล์เสียงเป็นข้อความ นอกจากนี้อาจมีเนื้อหาที่เป็นความลับหรืออ่อนไหวต่อองค์กร ด้วยเหตุนี้ ผู้วิจัยจึงมีแนวคิดในการประยุกต์ใช้โมเดลภาษาขนาดใหญ่ เพื่อเพิ่มประสิทธิภาพในการปฏิบัติงาน โดยหลีกเลี่ยงการพึ่งพาระบบประมวลผลบนกลุ่มเมฆ (Cloud Computing) และมุ่งเน้นการประมวลผลข้อมูลภายในขอบเขตขององค์กร หรือบนคอมพิวเตอร์ส่วนบุคคลเท่านั้น เพื่อรักษาความปลอดภัยของข้อมูล และลดความเสี่ยงที่อาจเกิดจากการเปิดเผยข้อมูลต่อบุคคลภายนอกได้

โดยโมเดล Whisper เป็นโมเดล ASR ที่พัฒนาจากบริษัท OpenAI โดยรุ่นล่าสุดในปัจจุบันคือ Whisper large-v3 Turbo ซึ่งได้รับการฝึกฝนด้วยข้อมูลเสียงมากกว่า 680,000 ชั่วโมง จากหลากหลายภาษา ส่งผลทำให้โมเดลสามารถจดจำสำเนียงและศัพท์เฉพาะทางได้อย่างครอบคลุมมากยิ่งขึ้น นอกจากนี้โมเดล Whisper ยังสามารถแปลงเสียงเป็นข้อความในหลายภาษารวมถึงแปลเสียงเป็นภาษาไทยได้อีกด้วย ทั้งนี้โมเดลดังกล่าวให้ใช้งานในแบบโอเพนซอร์ส (Open Source) บนคอมพิวเตอร์ส่วนบุคคลได้ และสามารถนำมาใช้ประโยชน์ในการวิจัยได้โดยไม่เสียค่าใช้จ่าย

จากแนวคิดดังกล่าว ผู้วิจัยจึงเสนอแนวทางการพัฒนาต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความสำหรับการประชุมอิเล็กทรอนิกส์ สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม โดยใช้โมเดล Whisper เพื่อยกระดับกระบวนการจัดทำรายงานสรุปการประชุม ลดระยะเวลาการสรุปประชุมของเจ้าหน้าที่ผู้ปฏิบัติงาน และรักษาความปลอดภัยของข้อมูลภายในองค์กรอย่างมีประสิทธิภาพ

2. วัตถุประสงค์

2.1 เพื่อพัฒนาต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความสำหรับการประชุมอิเล็กทรอนิกส์ สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม

2.2 เพื่อเปรียบเทียบความแม่นยำของโมเดลสำหรับต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความ สำหรับการประชุมอิเล็กทรอนิกส์ สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม

3. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

3.1 ทฤษฎีที่เกี่ยวข้อง

โมเดลภาษาขนาดใหญ่ (Large Language Models: LLMs) คือ อัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning) ที่ได้รับการฝึกฝนจากชุดข้อมูลจำนวนมหาศาล ทำให้สามารถเข้าใจและสร้างภาษาธรรมชาติ (Natural Language Processing: NLP) ได้ใกล้เคียงกับภาษาของมนุษย์มากที่สุด ทั้งยังสามารถนำไปประยุกต์ใช้งานได้หลากหลาย อาทิ เช่น การสรุปข้อมูลสาระสำคัญ การตอบคำถามแบบอัจฉริยะ รวมถึงการสร้างรูปภาพ เสียง หรือวิดีโอ เป็นต้น โดยกระบวนการฝึกโมเดลภาษาขนาดใหญ่ประกอบด้วย 2 ขั้นตอน ได้แก่ 1) การฝึกเบื้องต้น (Pre-training) ซึ่งเป็นการเรียนรู้จากข้อมูลขนาดใหญ่ที่มีความหลากหลายเพื่อทำความเข้าใจโครงสร้างภาษาทั่วไป และ 2) การปรับแต่งเฉพาะงาน (Fine-tuning) ซึ่งเป็นการฝึกฝนเพิ่มเติมด้วยข้อมูลเฉพาะทาง เพื่อทำให้โมเดลสามารถใช้งานได้อย่างเฉพาะเจาะจง และมีประสิทธิภาพมากยิ่งขึ้น

Radford et al. [1] Whisper คือ โมเดลรู้จำเสียงพูดอัตโนมัติ (ASR: Automatic Speech Recognition) เปิดตัวในเดือนตุลาคม 2022 โดยรุ่นล่าสุดในปัจจุบันคือ Whisper large-v3 Turbo ซึ่งได้รับการฝึกข้อมูลเสียงและข้อความที่มีความหลากหลายกว่า 680,000 ชั่วโมง จากหลายภาษาและหลายงาน ด้วยโมเดลที่ใช้สถาปัตยกรรม Transformer แบบ sequence-to-sequence เพื่อถอดเสียง แปลภาษา และระบุภาษา โดยใช้เพียงโมเดลเดียวผ่านการเข้ารหัสเป็นลำดับ token ที่ระบุงานต่าง ๆ โมเดล Whisper มีทั้งหมด 6 ขนาด โดยแต่ละขนาดมีความแตกต่างด้านพารามิเตอร์ ขนาดพื้นที่หน่วยความจำโดยประมาณ ความเร็ว และความแม่นยำ ดังตารางที่ 1 ซึ่งวัดประสิทธิภาพจากการแปลงเสียงเป็นภาษาอังกฤษด้วย GPU A100 ทั้งนี้ความเร็วในการใช้งานจริงอาจแตกต่างไปตามภาษา และทรัพยากรของฮาร์ดแวร์ที่ใช้งาน

ตารางที่ 1 แสดงโมเดล Whisper จำนวน 6 ขนาด

Size	Parameters	Required VRAM	Relative speed
tiny	39 M	~1 GB	~10x
base	74 M	~1 GB	~7x
small	244 M	~2 GB	~4x
medium	769 M	~5 GB	~2x
large	1550 M	~10 GB	1x
turbo	809 M	~6 GB	~8x

การติดตั้งและใช้งานโมเดล Whisper

โมเดล Whisper สามารถติดตั้งและใช้งานผ่านภาษา Python เวอร์ชัน 3.8 ขึ้นไป และต้องใช้ PyTorch เวอร์ชัน 1.10 หรือใหม่กว่า นอกจากนี้ระบบที่ใช้งานยังจำเป็นต้องติดตั้งโปรแกรม ffmpeg ซึ่งเป็นเครื่องมือสำคัญในการจัดการไฟล์เสียงและวิดีโอ โดยสามารถ Download ได้จากเว็บไซต์ https://ffmpeg.org สำหรับขั้นตอนการติดตั้งและใช้งานโมเดล Whisper สามารถศึกษารายละเอียดเพิ่มเติมได้ที่ https://github.com/openai/whisper

เว็บแอปพลิเคชัน (Web Application) คือ ซอฟต์แวร์ที่พัฒนาขึ้น สำหรับใช้งานผ่านเว็บเบราว์เซอร์ โดยไม่จำเป็นต้องติดตั้งโปรแกรมเพิ่มเติมบนอุปกรณ์ของผู้ใช้งาน สามารถเข้าใช้งานได้ง่าย รองรับการทำงานร่วมกัน ทำให้เว็บแอปพลิเคชันเป็นทางเลือกที่เหมาะสมสำหรับองค์กรในยุคดิจิทัล ซึ่งสอดคล้องกับแนวทางของ Ayon et al. [2] ที่นำเสนอวิธีการใหม่ในสร้างเว็บแอปพลิเคชันระดับองค์กร โดยใช้โมเดลภาษาขนาดใหญ่ (Large Language Models: LLMs) เพื่อเพิ่มคุณภาพวิศวกรรมแบบอัจฉริยะ (Intelligent Quality Engineering) ช่วยลดเวลา ความซับซ้อนในการพัฒนา การบำรุงรักษา และการทดสอบเว็บแอปพลิเคชันแบบอัตโนมัติ

สถิติที่ใช้ในการวิจัย

ค่าความแม่นยำ (Accuracy) คือ ค่าความถูกต้องของตัวอักษรที่ทำนายถูกต้อง คำนวณได้ตามสมการที่ 1 ดังนี้

$$\text{Accuracy} = \frac{\textit{Correct Predictions}}{\textit{Total Characters}} \times 100 \quad (1)$$

Correct Predictions คือ จำนวนตัวอักษรที่ทำนายถูกต้อง
Total Characters คือ จำนวนตัวอักษรทั้งหมด

ค่าระยะเวลาประมวลผล (Processing Time) คือ ค่าระยะเวลาประมวลผลในการแปลงเสียงเป็นข้อความ โดยเริ่มจับเวลาตั้งแต่เริ่มต้นที่โมเดลประมวลผลด้วย Python จนสิ้นสุดการประมวลผล คำนวณได้ตามสมการที่ 2 ดังนี้

$$\text{Processing Time} = T_{end} - T_{start} \quad (2)$$

T_{start} คือ เวลาที่เริ่มต้นกระบวนการประมวลผล (Start Time)
T_{end} คือ เวลาที่สิ้นสุดกระบวนการประมวลผล (End Time)

3.2 งานวิจัยที่เกี่ยวข้อง

Labied et al. [3] ได้ทำการศึกษาวิจัยเรื่อง Speech Translation From Darija to Modern Standard Arabic: Fine-Tuning Whisper on the Darija-C Corpus Across Six Model Versions โดยมีการปรับแต่งโมเดล Whisper จำนวน 6 ขนาด ได้แก่ 1) small 2) medium 3) large 4) large-v2 5) large-v3 และ 6) turbo เพื่อแปลคำพูดจากภาษาโมร็อกโกดาริจา (Darija) เป็นภาษาอาหรับมาตรฐาน (MSA) โดยใช้ชุดข้อมูล Darija-C ที่มีเสียงและข้อความแปลตรงกัน ซึ่งมีลักษณะทางภาษาดาริจาที่ครอบคลุม โดยการทดลองวัดประสิทธิภาพของการแปล คุณภาพของผลลัพธ์ ระยะเวลาในการฝึกโมเดล และการใช้หน่วยความจำ ผลการทดลองพบว่า โมเดลทุกขนาดสามารถทำงานได้ตามค่าคะแนน BLEU ที่ใกล้เคียงกัน (0.744) แต่แตกต่างกันในด้านความเร็วและประสิทธิภาพ โดยโมเดลขนาดเล็ก เช่น small และ medium สามารถเรียนรู้ได้เร็ว และใช้ทรัพยากรน้อย เหมาะสำหรับการใช้งานในสภาพแวดล้อมที่มีข้อจำกัดด้านทรัพยากรของคอมพิวเตอร์ ส่วนโมเดลขนาดใหญ่มีความสามารถในการจัดการความซับซ้อนของภาษาได้ดีกว่าแต่ต้องใช้ GPU และเวลาในการเรียนรู้มากขึ้น ในขณะที่โมเดลขนาด turbo เป็นโมเดลที่มีประสิทธิภาพสูง และใช้ทรัพยากรน้อย เหมาะสำหรับการใช้งานแบบเรียลไทม์

Tripathi et al. [4] ได้ทำการศึกษาวิจัยเรื่อง Enhancing Whisper’s Accuracy and Speed for Indian Languages through Prompt-Tuning and Tokenization งานวิจัยนี้นำเสนอแนวทางปรับปรุงค่าความแม่นยำและความเร็วของโมเดล Whisper สำหรับการรู้จำเสียงพูดในภาษากลุ่มอินเดียที่มีทรัพยากรจำกัด โดยใช้ 2 เทคนิคหลัก ได้แก่ 1) การปรับจูนด้วย prompt ที่ระบุข้อมูลกลุ่มภาษา เพื่อเพิ่มความแม่นยำในการเรียนรู้ภาษาศาสตร์ให้มีความใกล้เคียงกัน และ 2) การออกแบบ tokenizer รูปแบบใหม่ที่ลดจำนวน token ที่สร้าง ส่งผลทำให้ประมวลผลได้เร็วขึ้น ผลการทดลองพบว่าทั้ง 2 แนวทางช่วยลด Word Error Rate (WER) และเพิ่มประสิทธิภาพในหลายขนาดของโมเดล Whisper ได้อย่างมีนัยสำคัญ

Zezario et al. [5] ได้ทำการศึกษาวิจัยเรื่อง A study on zero-shot non-intrusive speech assessment using large language models งานวิจัยนี้เสนอแนวทางการประเมินเสียงพูดด้วยใช้โมเดลภาษาใหญ่ (LLMs) 2 รูปแบบ ได้แก่ GPT-4o ที่วิเคราะห์เสียงโดยตรง และ GPT-Whisper ซึ่งใช้โมเดล Whisper แปลงเสียงเป็นข้อความและประเมินความถูกต้องของข้อความ ผลการทดลองพบว่า GPT-Whisper มีความแม่นยำสูงกว่า GPT-4o และมีความสัมพันธ์ระดับปานกลางกับการประเมินโดยมนุษย์ และความสัมพันธ์ระดับสูงกับ Character Error Rate (CER) ของ ASR และยังแสดงประสิทธิภาพเหนือกว่าโมเดลประเมินเสียงอื่นในด้าน intelligibility โดยไม่ต้องใช้ข้อมูลฝึกเพิ่มเติม

Andreyev [6] ได้ทำการศึกษางานวิจัยเรื่อง Quantization for OpenAI's Whisper Models: A Comparative Analysis บทความนี้วิเคราะห์เปรียบเทียบประสิทธิภาพของโมเดล Whisper ของ OpenAI และรูปแบบที่ปรับปรุงสำหรับการถอดเสียงสดและการถอดเสียงแบบออฟไลน์ โดยชี้ให้เห็นปัญหาเรื่องการสร้างข้อความที่ไม่ตรงกับเสียงจริง (hallucination)

และความล่าช้าของโมเดลขนาดใหญ่ และทดลองการใช้ quantization (INT4, INT5, INT8) ต่ออัตราความผิดพลาด (WER) และเวลาหน่วง (latency) โดยใช้ชุดข้อมูล LibriSpeech ผลการทดลองพบว่า quantization ลดเวลาหน่วงได้ 19% และลดขนาดโมเดลได้ 45% โดยไม่กระทบต่อความแม่นยำของการถอดเสียง ทั้งนี้เหมาะกับการใช้งาน edge devices

Guo et al. [7] ได้ทำการศึกษางานวิจัยเรื่อง SQ-Whisper: Speaker-Querying Based Whisper Model for Target-Speaker ASR โดยเสนอโมเดลทางเสียงที่มักจะมีข้อจำกัดในการทำงานกับเสียงของผู้พูดคนเดียว ทำให้ไม่สามารถจัดการกับเสียงที่มีผู้พูดซ้อนทับกันในสถานการณ์จริงได้ งานวิจัยนี้มุ่งศึกษาการปรับแต่งโมเดลพื้นฐานเสียง เพื่อแยกเสียงของผู้พูดเป้าหมายออกจากเสียงซ้อนทับ โดยใช้โมเดล Whisper เป็นโมเดลพื้นฐานและเปรียบเทียบเทคนิคการปรับแต่งของผู้พูดเป้าหมาย จากนั้นเสนอโมเดลใหม่ที่มีชื่อว่า Speaker-Querying Whisper (SQ-Whisper) โดยมีการใช้ชุดข้อมูลเรียนรู้ เพื่อดึงคุณลักษณะเฉพาะผู้พูดเป้าหมาย ผลการทดลองพบว่า SQ-Whisper สามารถลดอัตราความผิดพลาดในการรู้จำเสียง (WER) ได้มากกว่าระบบ TS-HuBERT สูงสุด 15% และ 10% ด้วยชุดข้อมูล Libri2Mix และ WSJ0-2Mix ตามลำดับ

Shah et al. [8] ได้ทำการศึกษางานวิจัยเรื่อง Enhancing Multimedia Accessibility: Automated Video Captioning and Translation System Using OpenAI Whisper งานวิจัยนี้นำเสนอระบบสร้างคำบรรยายอัตโนมัติแบบเรียลไทม์เพื่อเพิ่มความสามารถในการเข้าถึงสื่อวิดีโอดิจิทัล โดยผสานการทำงานของ Flask, MoviePy, OpenCV และโมเดล Whisper จากบริษัท OpenAI ระบบสามารถวิเคราะห์วิดีโอเพื่อสร้างคำบรรยายที่ถูกต้องและแสดงพร้อมวิดีโอได้แบบเรียลไทม์ อีกทั้งยังรองรับการแปลหลายภาษาอย่างแม่นยำถึง 98.4% และสามารถใช้งานผ่านอินเทอร์เฟซที่เป็นมิตรกับผู้ใช้ เหมาะสำหรับการนำไปใช้งานจริงในสื่อดิจิทัลเพื่อความครอบคลุมทางภาษาและการมีส่วนร่วมของผู้ชม

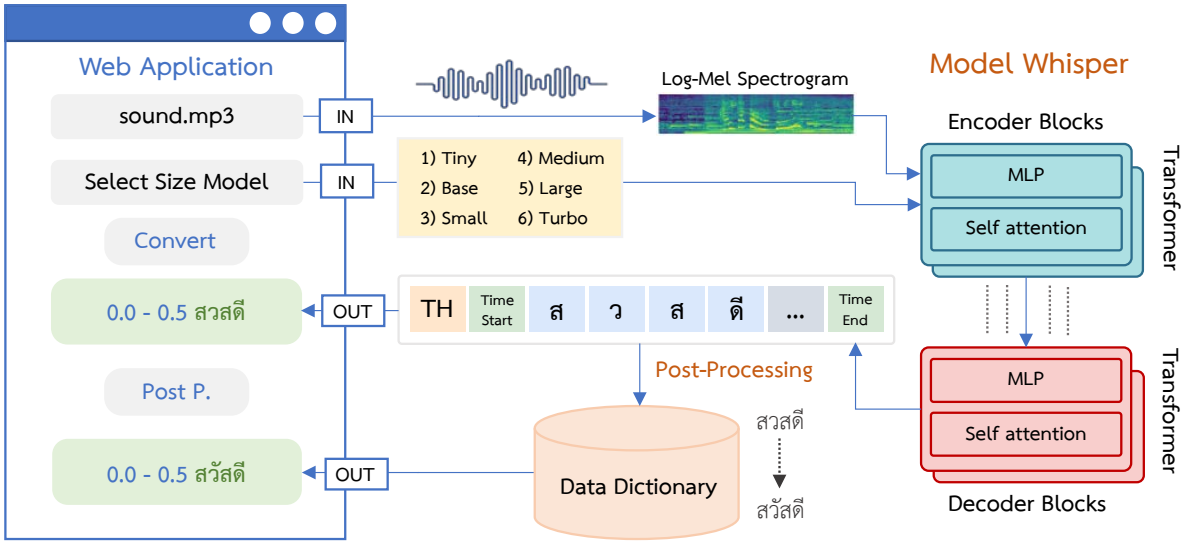
Aung et al. [9] ได้ทำการศึกษางานวิจัยเรื่อง Thonburian Whisper: Robust Fine-tuned and Distilled Whisper for Thai โดยนำเสนอ Thonburian Whisper ซึ่งเป็นโมเดลที่ผ่านการปรับแต่ง (fine-tuning) จากโมเดล Whisper ดั้งเดิม เพื่อเพิ่มความสามารถในการจำเสียงพูดภาษาไทยให้ดียิ่งขึ้น โดยใช้ชุดข้อมูลเสียงภาษาไทยกว่า 1,300 ชั่วโมง และใช้เทคนิคเพิ่มข้อมูล (audio augmentation) การแบ่งส่วนเสียงระหว่างการฝึก และใช้เทคนิค distillation เพื่อทำให้โมเดลมีขนาดเล็กลง ผลการทดลองพบว่า Thonburian Whisper ลดอัตราความผิดพลาดของคำ (WER) อย่างมีนัยสำคัญเมื่อเทียบกับโมเดลต้นแบบและโมเดล ASR ภาษาไทยอื่น ๆ ขณะที่โมเดลยังคงรักษาความแม่นยำใกล้เคียงเดิมแม้ใช้พารามิเตอร์และข้อมูลฝึกน้อยกว่า

Chaovarit Janpirom et al. [10] ได้ทำการศึกษางานวิจัยเรื่อง การพัฒนาระบบแปลงไฟล์ภาพและเสียงเป็นข้อความจากหน้าซองจดหมายด้วยการประมวลผลภาพ โดยมีวัตถุประสงค์เพื่อออกแบบและพัฒนาระบบ ประเมินประสิทธิภาพของระบบแปลงไฟล์ภาพและเสียงเป็นข้อความจากหน้าซองจดหมายด้วยการประมวลผลภาพ และศึกษาความพึงพอใจต่อระบบดังกล่าว โดยทดลองกับกลุ่มตัวอย่างจำนวน 30 คน วิเคราะห์ข้อมูลด้วยสถิติ ได้แก่ จำนวน ร้อยละ ค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน ผลการวิจัยพบว่าระบบการแปลงภาพเป็นข้อความภาษาไทยมีค่าความแม่นยำอยู่ที่ 35% และภาษาอังกฤษอยู่ที่ 91.95% ระบบการแปลงเสียงภาษาไทยและอังกฤษมีค่าความแม่นยำอยู่ที่ 91.06% และระดับความพึงพอใจของผู้ใช้งานต่อระบบอยู่ในระดับมาก

4. ขั้นตอนและวิธีดำเนินการวิจัย

4.1 กรอบแนวคิดการทำวิจัย

ผู้วิจัยดำเนินการตามกรอบแนวคิดการทำวิจัย ดังภาพที่ 1



ภาพที่ 1 กรอบแนวคิดการทำวิจัย

4.2 ขั้นตอนการดำเนินงานวิจัย

4.2.1 การศึกษาข้อมูลเบื้องต้น

4.2.1.1 ผู้วิจัยศึกษาปัญหา ทฤษฎี และงานวิจัยที่เกี่ยวข้องในการแปลงเสียงเป็นข้อความสำหรับภาษาไทย รวมถึงแนวทางการประยุกต์ร่วมกับเทคโนโลยีเว็บแอปพลิเคชัน (Web Application) ภายในองค์กร

4.2.1.2 ผู้วิจัยศึกษาขั้นตอนการติดตั้งและการใช้งานโมเดล Whisper ซึ่งเป็นโมเดลแปลงเสียงเป็นข้อความ

4.2.2 การออกและพัฒนาต้นแบบเว็บแอปพลิเคชัน

4.2.2.1 ผู้วิจัยออกแบบส่วนติดต่อผู้ใช้งาน (User Interface: UI) โดยมุ่งเน้นความเรียบง่าย และสะดวกในการใช้งานสามารถอัปโหลดไฟล์เสียงที่มีนามสกุล .mp3 ผ่านหน้าเว็บแอปพลิเคชันได้โดยตรง จากนั้นระบบจะดำเนินการประมวลผลแปลงเสียงเป็นข้อความด้วยโมเดล Whisper โดยใช้ภาษา PHP ในการจัดการฝั่งเซิร์ฟเวอร์ (Server) และเชื่อมต่อกับภาษา Python ในการประมวลผลแปลงเสียงเป็นข้อความ และส่งผลลัพธ์มายังหน้าเว็บแอปพลิเคชัน

4.2.2.2 ผู้วิจัยพัฒนาเว็บแอปพลิเคชันต้นแบบ โดยใช้ภาษา JavaScript, PHP และ Python ร่วมกับไลบรารีที่เชื่อมต่อกับโมเดล Whisper เพื่อให้ระบบสามารถประมวลผลแปลงเสียงและแสดงข้อความผ่านเว็บแอปพลิเคชันได้ทันที

4.2.3 การเปรียบเทียบความแม่นยำของโมเดลสำหรับต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความ

4.2.3.1 เครื่องมือที่ใช้ในการทดลอง

ผู้วิจัยใช้เครื่องคอมพิวเตอร์ส่วนบุคคลที่มีทรัพยากรจำกัด ดังนี้

1. หน่วยประมวลผลกลาง (CPU)	Intel Core i5-7400 @ 3.00 GHz
2. หน่วยความจำหลัก (RAM)	DDR4 8 GB
3. หน่วยจัดเก็บข้อมูล (SSD)	250 GB
4. การ์ดแสดงผล (GPU)	NVIDIA GeForce GT 730
5. ระบบปฏิบัติการ	Windows 10

4.2.3.2 ข้อมูลเสียงที่ใช้ในการทดลอง

ผู้วิจัยใช้ชุดข้อมูลเสียงจำนวน 3 ข้อความ ซึ่งเป็นเสียงพูดภาษาไทยจากผู้พูดคนเดียวกัน โดยข้อความทั้งหมดถูกบันทึกด้วยเครื่องบันทึกเสียงชนิดเดียวกัน ภายใต้สภาพแวดล้อมในห้องประชุมที่ใช้งานจริง โดยรายละเอียดของข้อความเสียงแต่ละชุด แสดงดังตารางที่ 2

ตารางที่ 2 แสดงข้อความเสียง จำนวน 3 ข้อความ

ลำดับ	ข้อความเสียง	จำนวนอักษร
1	สวัสดีครับ	10 ตัวอักษร
2	สถาบันวิจัยและพัฒนามหาวิทยาลัยราชภัฏนครปฐม	42 ตัวอักษร
3	การพัฒนาต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความสำหรับการประชุมอิเล็กทรอนิกส์ สถาบันวิจัยและพัฒนามหาวิทยาลัยราชภัฏนครปฐม	121 ตัวอักษร

4.2.3.3 วิธีการทดสอบและการประเมินผล

ผู้วิจัยได้ทำการทดสอบและบันทึกข้อมูลผลลัพธ์ของโมเดล Whisper ทั้งหมด 6 ขนาด ได้แก่ 1) Tiny 2) Base 3) Small 4) Medium 5) Large และ 6) Turbo โดยทำการทดสอบโมเดลแต่ละขนาดจำนวน 3 ข้อความ (ไม่นับช่องว่าง) และทำซ้ำการทดสอบข้อความละ 3 ครั้ง เพื่อความแม่นยำและความน่าเชื่อถือของผลการทดลอง โดยใช้ชุดข้อมูลเสียงเดียวกันในการประเมินผล เพื่อเปรียบเทียบว่าโมเดลขนาดใดมีความเหมาะสมที่สุดสำหรับการนำไปใช้งานเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความ โดยพิจารณา 2 ตัวชี้วัดหลัก ได้แก่

1) ค่าความแม่นยำ (Accuracy)

2) ค่าระยะเวลาในการประมวลผล (Processing Time)

4.2.4 วิธีการ Post-Processing สำหรับต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความ

หลังจากที่โมเดล Whisper ทำการแปลงเสียงเป็นข้อความแล้ว ผู้วิจัยได้ดำเนินการปรับปรุงผลลัพธ์ที่ได้ด้วยวิธีการ Post-Processing เพื่อเพิ่มความถูกต้องของข้อความให้มีความแม่นยำยิ่งขึ้น โดยการแทนที่ข้อความที่ผิดให้เป็นข้อความที่ถูก ในขั้นตอนนี้จะต้องอาศัยฐานข้อมูลคำศัพท์ (Data Dictionary) ที่ผู้วิจัยสร้างขึ้น ซึ่งเป็นฐานข้อมูลที่เก็บคู่ของข้อความของระหว่าง “ข้อความที่โมเดลแปลงผิด” กับ “ข้อความที่ถูกต้อง” เมื่อระบบตรวจพบข้อความที่ผิด ระบบจะดำเนินการแทนที่ข้อความถูกต้องให้ทันที

การดำเนินการในขั้นตอนนี้ช่วยเพิ่มความแม่นยำของข้อความที่ได้จากการแปลงเสียงให้มีความถูกต้องมากยิ่งขึ้น และสามารถนำไปประยุกต์ใช้งานจริงในการถอดเทปเสียงจากการประชุมได้อย่างมีประสิทธิภาพและน่าเชื่อถือมากยิ่งขึ้น

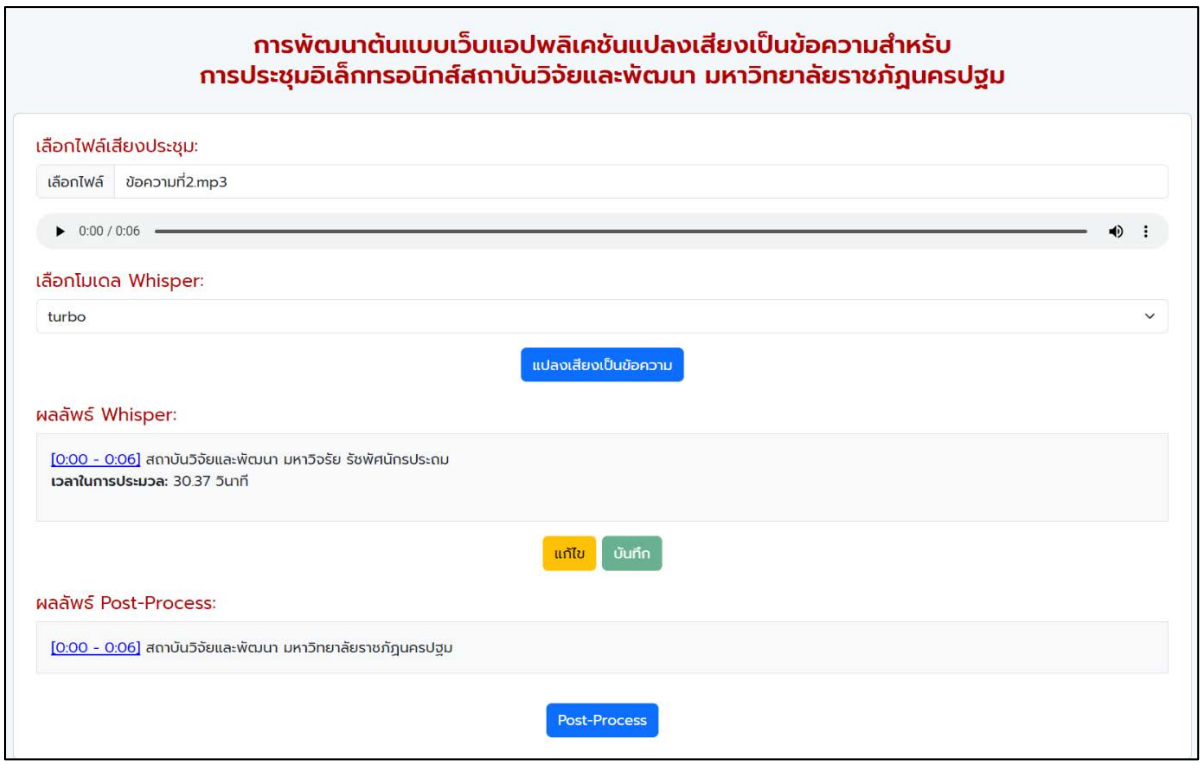
5. ผลการวิจัย

การทดลองเปรียบเทียบประสิทธิภาพของโมเดล Whisper ทั้ง 6 ขนาด ได้แก่ 1) Tiny 2) Base 3) Small 4) Medium 5) Large และ 6) Turbo มีผลการวิจัย ดังนี้

5.1 ผลการพัฒนาต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความสำหรับการประชุมอิเล็กทรอนิกส์ สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม ดังภาพที่ 2

5.2 ผลการทดลองแปลงเสียงเป็นข้อความจำนวน 3 ข้อความ โดยแต่ละข้อความถูกทดสอบข้อความละ 3 รอบ และพบว่าจำนวนตัวอักษรที่แปลงถูกต้องมีค่าเท่ากันทั้ง 3 รอบ แต่ระยะเวลาในการประมวลผลแตกต่างกัน ดังตารางที่ 3 - 5

5.3 ผลการเปรียบเทียบค่าเฉลี่ยความแม่นยำของโมเดล Whisper และการเปรียบเทียบค่าเฉลี่ยระยะเวลาในการประมวลผลของโมเดล Whisper ดังภาพที่ 3 - 4



ภาพที่ 2 ต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความสำหรับการประชุมอิเล็กทรอนิกส์ สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม

ตารางที่ 3 ผลการทดลองแปลงเสียงเป็นข้อความ “สวัสดีครับ”

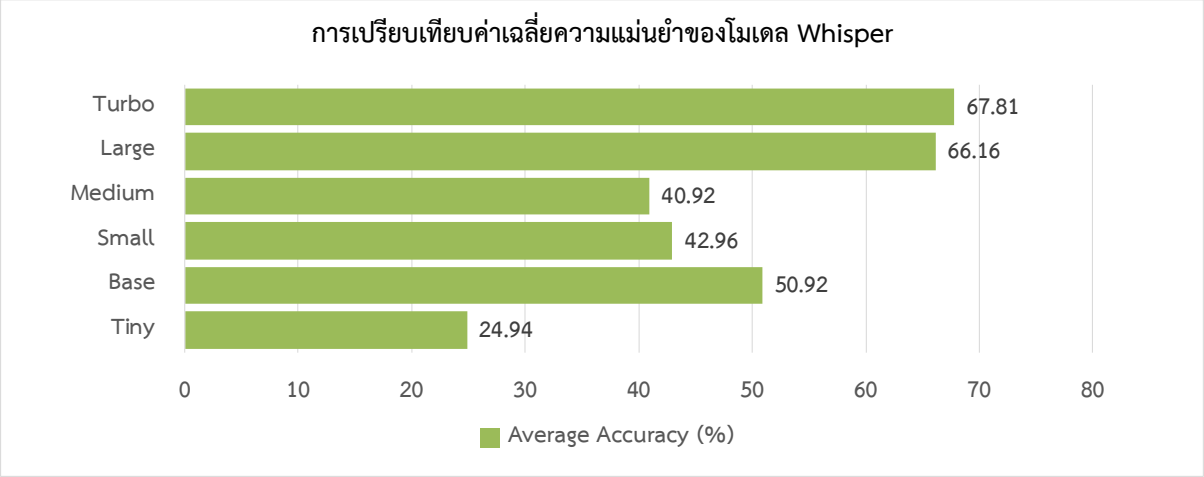
Size Whisper	จำนวนตัวอักษรทั้งหมด	จำนวนตัวอักษรที่แปลงถูกต้อง	เปอร์เซ็นต์ (%)	ระยะเวลาการประมวลผล / (วินาที)			
				ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	ค่าเฉลี่ย
tiny	10	5	50%	4.19	4.12	4.11	4.14
base	10	10	100%	6.47	5.13	5.18	5.59
small	10	10	100%	9.66	9.26	9.16	9.36
medium	10	10	100%	26.16	26.19	25.34	25.90
large	10	10	100%	109.81	114.40	114.72	112.98
turbo	10	10	100%	28.54	28.92	28.69	28.72

ตารางที่ 4 ผลการทดลองแปลงเสียงเป็นข้อความ “สถาบันวิจัยและพัฒนามหาวิทยาลัยราชภัฏนครปฐม”

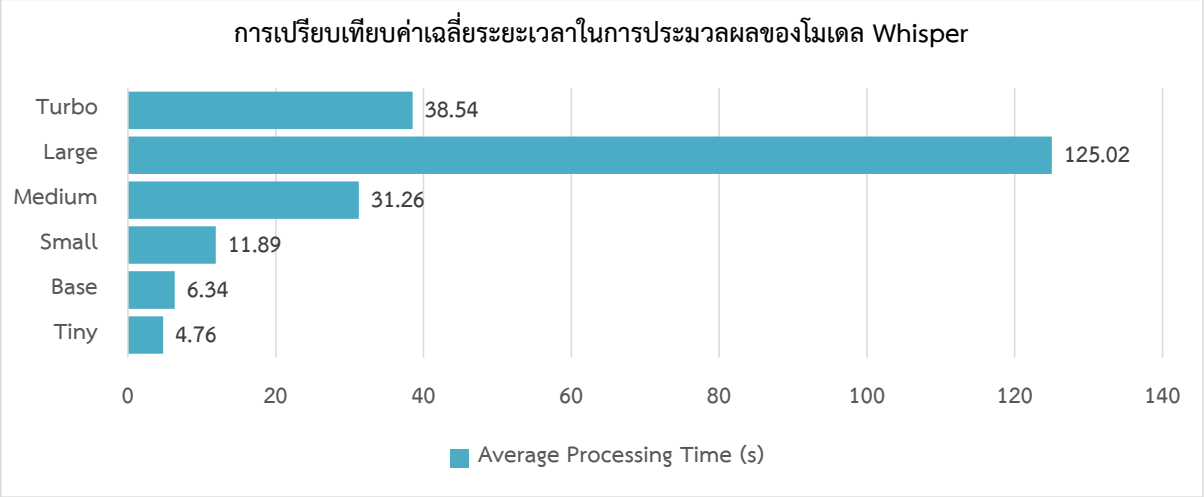
Size Whisper	จำนวนตัวอักษรทั้งหมด	จำนวนตัวอักษรที่แปลงถูกต้อง	เปอร์เซ็นต์ (%)	ระยะเวลาการประมวลผล / (วินาที)			
				ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	ค่าเฉลี่ย
tiny	42	8	19.05%	5.15	4.53	4.57	4.75
base	42	18	42.86%	5.83	5.86	5.78	5.82
small	42	9	21.43%	11.54	11.08	11.33	11.32
medium	42	4	9.52%	30.67	30.07	29.78	30.17
large	42	24	57.14%	111.78	119.51	118.42	116.57
turbo	42	24	57.14%	31.07	30.59	29.57	30.41

ตารางที่ 5 ผลการทดลองแปลงเสียงเป็นข้อความ “การพัฒนาต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความสำหรับการประชุมอิเล็กทรอนิกส์สถาบันวิจัยและพัฒนามหาวิทยาลัยราชภัฏนครปฐม”

Size Whisper	จำนวนตัวอักษรทั้งหมด	จำนวนตัวอักษรที่แปลงถูกต้อง	เปอร์เซ็นต์ (%)	ระยะเวลาการประมวลผล / (วินาที)			
				ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	ค่าเฉลี่ย
tiny	121	7	5.79%	5.31	5.33	5.49	5.38
base	121	12	9.92%	7.34	7.96	7.51	7.60
small	121	9	7.44%	14.71	14.44	15.84	15.00
medium	121	16	13.22%	37.76	36.79	38.56	37.70
large	121	50	41.32%	148.36	143.9	144.3	145.52
turbo	121	56	46.28%	68.15	58.93	42.41	56.50



ภาพที่ 3 การเปรียบเทียบค่าเฉลี่ยความแม่นยำของโมเดล Whisper



ภาพที่ 4 การเปรียบเทียบค่าเฉลี่ยระยะเวลาในการประมวลผลของโมเดล Whisper

6. สรุปผลการวิจัย ข้อเสนอแนะในการวิจัยครั้งต่อไป

6.1. สรุปผลการวิจัย

การทดลองนี้มีวัตถุประสงค์เพื่อเปรียบเทียบความแม่นยำของโมเดล Whisper 6 ขนาด ได้แก่ 1) Tiny 2) Base 3) Small 4) Medium 5) Large และ 6) Turbo ในการแปลงเสียงเป็นข้อความจำนวน 3 ข้อความ โดยใช้เสียงจากผู้พูดคนเดียวกัน ด้วยอุปกรณ์บันทึกเสียงและสถานที่ห้องประชุมเดียวกัน และทดลองประมวลผลบนคอมพิวเตอร์ส่วนบุคคลที่มีทรัพยากรจำกัด

จากผลการทดลองพบว่าโมเดลขนาด Turbo ให้ผลลัพธ์ที่ดีที่สุดในภาพรวม โดยมีค่าความแม่นยำเฉลี่ยสูงสุดอยู่ที่ 67.81% และใช้ระยะเวลาในการประมวลผลเฉลี่ย 38.54 วินาที แสดงให้เห็นถึงประสิทธิภาพความแม่นยำและระยะเวลาในการประมวลผล แม้จะใช้งานบนเครื่องคอมพิวเตอร์ส่วนบุคคลที่มีทรัพยากรจำกัด รองลงมาคือโมเดลขนาด Large มีค่าความแม่นยำเฉลี่ยอยู่ที่ 66.16% ซึ่งใกล้เคียงกับโมเดลขนาด Turbo แต่ใช้ระยะเวลาในการประมวลผลเฉลี่ยสูงถึง 125.02 วินาที แสดงถึงข้อจำกัดในการใช้งานจริง เนื่องจากต้องอาศัยทรัพยากรในการประมวลผลที่สูงกว่ามาก รองลงมาคือโมเดลขนาด Base มีค่าความแม่นยำเฉลี่ยอยู่ที่ 50.92% และใช้ระยะเวลาในการประมวลผลเฉลี่ย 6.34 วินาที แม้จะใช้ระยะเวลาประมวลผลที่น้อยกว่า แต่ยังมีค่าความแม่นยำต่ำกว่าโมเดลขนาด Turbo โดยเฉพาะเมื่อนำไปใช้กับข้อความที่มีความยาว หรือซับซ้อนมากขึ้น รองลงมาคือโมเดลขนาด Small มีค่าความแม่นยำเฉลี่ยอยู่ที่ 42.96% และใช้ระยะเวลาในการประมวลผลเฉลี่ย 11.89 วินาที ซึ่งไม่เหมาะกับงานที่ข้อความที่มีความยาว หรือซับซ้อนมาก ๆ รองลงมาคือโมเดลขนาด Medium มีค่าความแม่นยำเฉลี่ย 40.92% และใช้ระยะเวลาในการประมวลผลเฉลี่ย 31.26 วินาที ซึ่งเป็นโมเดลขนาดกลาง แต่ความแม่นยำน้อยกว่าโมเดลขนาด Small และยังใช้ระยะเวลาประมวลผลมากขึ้นแสดงให้เห็นว่ายังไม่เหมาะสมกับการใช้งานที่ต้องการความคุ้มค่าทั้งด้านเวลาและผลลัพธ์ และสุดท้ายคือโมเดลขนาด Tiny ซึ่งแม้จะใช้เวลาในการประมวลผลเฉลี่ยเพียง 4.76 วินาที แต่มีค่าความแม่นยำเฉลี่ยต่ำสุดที่ 24.94% แสดงให้เห็นว่ายังไม่เหมาะสมกับข้อความทั้งขนาดสั้นและยาว

จากผลการวิเคราะห์ข้างต้นสามารถสรุปได้ว่าโมเดลขนาด Turbo เป็นตัวเลือกที่เหมาะสมที่สุดสำหรับการนำไปใช้ในเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความภาษาไทย เนื่องจากสามารถให้ค่าความแม่นยำที่สูงควบคู่ไปกับระยะเวลาในการประมวลผลที่อยู่ในเกณฑ์ที่ยอมรับได้ ขณะที่โมเดลขนาดเล็กสามารถประมวลผลได้เร็วกว่า แต่ยังไม่สามารถตอบโจทย์ความแม่นยำของข้อความที่มีความยาวมากขึ้นในระดับที่ใช้งานจริงอย่างมีประสิทธิภาพ

จากนั้นผู้วิจัยได้นำโมเดล Whisper ขนาด Turbo ไปใช้งานในการพัฒนาต้นแบบเว็บแอปพลิเคชันแปลงเสียงเป็นข้อความภาษาไทย โดยใช้เทคนิคการปรับปรุงผลลัพธ์หลังการประมวลผล (Post-Processing) เพื่อเพิ่มความแม่นยำของผลลัพธ์ โดยจัดทำฐานข้อมูลคำศัพท์ที่เก็บคู่ของ “ข้อความที่โมเดลแปลงผิด” กับ “ข้อความที่ถูกต้อง” ไว้เพื่อทำการแทนที่ข้อความผิดให้ถูกต้อง และนำไปใช้งานได้จริงต่อไป

6.2 ข้อเสนอแนะในการวิจัยครั้งต่อไป

6.2.1 ควรเพิ่มเติมทดลองข้อมูลเสียงอื่น ๆ เช่น มีผู้พูดหลายคน มีเสียงรบกวน หรือข้อความที่มีความยาวมากขึ้น เป็นต้น

6.2.2 ควรนำระบบที่พัฒนาขึ้นไปให้ผู้เชี่ยวชาญทางด้านคอมพิวเตอร์หรือผู้เชี่ยวชาญด้านภาษาตรวจสอบอย่างน้อย 3 คน

6.2.3 ควรเพิ่มการวิเคราะห์ทางสถิติ เช่น ANOVA หรือ t-test เพื่อแสดงความแตกต่างที่มีนัยสำคัญระหว่างโมเดล ซึ่งจะทำให้การเลือกโมเดลมีความน่าเชื่อถือยิ่งขึ้น

7. กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณมหาวิทยาลัยราชภัฏนครปฐม ที่ให้การสนับสนุนทุนอุดหนุนการวิจัยโครงการวิจัยสถาบันบูรณาการงานวิจัยกับงานประจำ R to R ประจำปีงบประมาณ 2568 [RR_68_02] และขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.เดชธรรมศิริ ที่เสียสละเวลาอันมีค่าเป็นที่ปรึกษาโครงการวิจัยที่ดีตลอดมา

8. เอกสารอ้างอิง (References)

[1] Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning, 202, 28492 - 28518. Proceedings of Machine Learning Research. https://proceedings.mlr.press/v202/radford23a.html

[2] Ayon, Z. A. H., Husain, G., Bisoi, R., Rahman, W., & Osborn, T. (2025). An efficient approach to represent enterprise web application structure using Large Language Model in the service of Intelligent Quality Engineering. arXiv. https://arxiv.org/abs/2501.06837

[3] Labied, M., Belangour, A., & Banane, M. (2025). Speech translation from Darija to Modern Standard Arabic: Fine-tuning Whisper on the Darija-C corpus across six model versions. IEEE Access, 13, 48656–48671. https://doi.org/10.1109/ACCESS.2025.3551229

[4] K. Tripathi, R. Gothi & P. Wasnik. (2025). Enhancing Whisper’s Accuracy and Speed for Indian Languages through Prompt-Tuning and Tokenization. ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5. https://doi.org/10.1109/ICASSP49660.2025.10887595

[5] R. E. Zezario, S. M. Siniscalchi, H. -M. Wang & Y. Tsao. (2025). A study on zero-shot non-intrusive speech assessment using large language models. ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5. https://doi.org/10.1109/ICASSP49660.2025.10889809

[6] Andreyev, A. (2025). Quantization for OpenAI's Whisper Models: A Comparative Analysis. arXiv preprint arXiv:2503.09905.

[7] P. Guo, X. Chang, H. Lv, S. Watanabe & L. Xie. (2024). SQ-Whisper: Speaker-Querying Based Whisper Model for Target-Speaker ASR in IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 175-185, 2025. https://doi.org/10.1109/TASLP.2024.3513835.

[8] Shah, H., Patel, M., & Gupta, R. K. (2024). Enhancing multimedia accessibility: Automated video captioning and translation system using OpenAI Whisper. In Proceedings of the 2024 Asian Conference on Intelligent Technologies (ACOIT) (pp. 1–6). IEEE. https://doi.org/10.1109/ACOIT62457.2024.10939144

[9] Aung, Z. H., Thavornmongkol, T., Boribalburephan, A., Tangsriworakan, V., Pipatsrisawat, K., & Achakulvisut, T. (2024). Thonburian Whisper: Robust Fine-tuned and Distilled Whisper for Thai. In Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024) (pp. 149–156). Trento: Association for Computational Linguistics.

[10] Chaovarit Janpirom, Wachirawictch Nilsook, Tippawan Rattanathamrongpun, Jirawat Chotikhun and Prattana Thankaew (2022). Developing a System to Convert Video and Audio Files to Text from Envelope Pages with Image Processing. Journal of Mass Communication Technology, RMUTP, 7(1), 33–46. (In Thai)