วารสารวิทยาศาสตร์บูรพา ปีที่ 27 (ฉบับที่ 1) มกราคม – เมษายน พ.ศ. 2565
BURAPHA SCIENCE JOURNAL Volume 27 (No.1) January – April 2022 บทความวิจัย

ความเป็นธรรมชาติของเสียงสังเคราะห์ : กรณีศึกษาเปรียบเทียบเชิงจิตวิสัยด้วยการประยุกต์ใช้วิธีการประเมินแบบ ACR ระหว่าง Siri และ Google Translate ด้วยเนื้อหาข่าว

Naturalness of Synthesized Speech: A Subjective-Comparative Study Utilizing ACR Listening-Opinion Tests between Siri and Google Translate Using News Content

พิสิฐ พรพงศ์เตชวาณิช[1] และ เทอดพงษ์ แดงสี[2*]

Phisit Pornpongtechavanich[1] and Therdpong Daengsi[2*]

[1] *สาขาวิชาเทคโนโลยีสารสนเทศ คณะอุตสาหกรรมและเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลรัตนโกสินทร์*

[2] *สาขาวิชาวิศวกรรมการจัดการอุตสาหกรรมเพื่อความยั่งยืน คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลพระนคร*

[1] *Information Technology Department, Faculty of Industry and Technology,*

Rajamangala University of Technology Rattanakosin

[2] *Sustainable Industrial Management Engineering Department, Faculty of Engineering,*

Rajamangala University of Technology Phra Nakhon

Received : 9 February 2021

Revised : 23 July 2021

Accepted : 13 August 2021

บทคัดย่อ

การสังเคราะห์เสียงจากข้อความถือเป็นหนึ่งในเทคโนโลยีที่สำคัญบนพื้นฐานของการประมวลผลภาษามนุษย์ในปัจจุบัน อย่างไรก็ตามหนึ่งในประเด็นที่น่าสนใจเกี่ยวกับเสียงสังเคราะห์คือความเป็นธรรมชาติของเสียง บทความนี้จึงนำเสนอการประยุกต์ใช้วิธีการประเมินคุณภาพเสียงเพื่อใช้ในการประเมินความเป็นธรรมชาติของเสียงสังเคราะห์ภาษาไทยที่สร้างจากระบบสังเคราะห์เสียงจากข้อความที่ได้รับความนิยม 2 ระบบ ได้แก่ Siri และ Google Translate สำหรับวิธีดำเนินการวิจัย เสียงสังเคราะห์ภาษาไทยที่มีเนื้อหาเกี่ยวกับข่าวในพระราชสำนัก 2 ข่าว และข่าวทั่วไปเกี่ยวกับไวรัสโคโรน่าสายพันธุ์ใหม่ 2019 (โควิด-19) ที่สร้างโดย Siri และ Google Translate ได้ถูกประเมินโดยอาสาสมัครชาย 16 คน และอาสาสมัครหญิง 16 คน พบว่า ในภาพรวมค่าเฉลี่ยคะแนนความคิดเห็นของเสียงสังเคราะห์ภาษาไทยที่สร้างจาก Google Translate ซึ่งมีค่าเท่ากับ 3.53 ± 0.67 มีค่าสูงกว่า 3.16 ± 0.77 ที่สร้างขึ้นโดย Siri ยิ่งไปกว่านั้น เมื่อวิเคราะห์ทางสถิติด้วย t-test พบว่า ได้ค่า p-value เท่ากับ 0.037 จึงสรุปได้ว่า ส่วนที่ทำหน้าที่สังเคราะห์เสียงใน Google Translate สามารถสร้างเสียงที่มีความเป็นธรรมชาติมากกว่าส่วนที่ทำหน้าที่สังเคราะห์เสียงใน Siri อย่างมีนัยสำคัญ (p-value น้อยกว่า 0.05) ดังนั้นวิธีดำเนินการวิจัยในบทความนี้สามารถนำไปประยุกต์ใช้ในการประเมินระดับความเป็นธรรมชาติของแอปพลิเคชัน/บริการ/ระบบอื่น ๆ เพื่อพัฒนาคุณภาพเสียงสังเคราะห์

คำสำคัญ : ทีทีเอส ; เอ็มโอเอส ; ความเป็นธรรมชาติ ; เสียงสังเคราะห์

Abstract

Text-To-Speech synthesis (TTS) is one of the most important technologies based on human language processing at present. However, one of interesting open issues about synthesized speech is naturalness of speech. This article presents the application of a speech quality assessment method to assess the naturalness of Thai synthesized speech from two popular TTS systems, Siri and Google Translate. For methodology, Thai synthesized speech, associated with two royal news and two general news (COVID-19), provided by Siri and Google Translate have been assessed by sixteen Thai male volunteers and sixteen Thai female volunteers. It has been found the overall result that the value of the naturalness - Mean Opinion Score (MOS) of Thai synthesized speech provided by Google Translate is 3.53 ± 0.67, which is higher than the value of 3.16 ± 0.77 provided by Siri. Furthermore, after the statistical analysis using t-test, it is been found that the p-value is 0.037. In conclusion, the speech synthesis engine in Google Translate provides better naturalness than the one in Siri significantly. Therefore, the methodology in this article can be applied to assess naturalness level of other applications/services/systems in order to improve synthesized speech quality.

Keywords : TTS ; MOS ; naturalness ; synthesized speech

***Corresponding author. E-mail :** therdpong.d@rmutp.ac.th

บทนำ

ในปัจจุบันนี้เทคโนโลยีการสังเคราะห์เสียงจากข้อความ (Text-To-Speech Synthesis: TTS) กลายเป็นเทคโนโลยีที่เข้ามามีบทบาทในหลาย ๆ ด้าน ได้แก่ (Wutiwiwatchai *et al*., 2017; ReadSpeaker, 2020)

(1) การเข้าถึงระบบด้วยเสียง เช่น อุปกรณ์สื่อสารทางเลือก (Augmentative and Alternative Communication: AAC) และหนังสือเสียงดิจิทัลเดซี่ (DAISY digital talking books) เป็นต้น

(2) ด้านการเงินการธนาคาร เช่น ระบบเทเลโฟนแบงก์กิ้ง (telephone banking) และการติดตามหนี้ เป็นต้น

(3) ด้านการกระจายเสียงและสื่อ เช่น สมาร์ตทีวี และการพยากรณ์อากาศ เป็นต้น

(4) ด้านบันเทิง เช่น การประยุกต์ใช้ในวิดีโอเกมต่าง ๆ และตุ๊กตาพูดได้ เป็นต้น

(5) ด้านสุขภาพ เช่น ระบบแจ้งเตือนการนัดหมายอัตโนมัติ ระบบเรียกคิวในโรงพยาบาล เป็นต้น

(6) ด้านการศึกษา เช่น การเรียนภาษา การทดสอบมาตรฐาน และการฝึกอบรมแบบจำลอง เป็นต้น

(7) ระบบแจ้งเตือนและประกาศในที่สาธารณะ เช่น ระบบแจ้งเตือนกรณีเกิดเหตุฉุกเฉิน เป็นต้น

(8) ด้านสิ่งพิมพ์และสื่อ เช่น หนังสือเสียง และระบบแจ้งข่าวสาร เป็นต้น

(9) ด้านการสื่อสารโทรคมนาคม เช่น ระบบคอลเซ็นเตอร์ และระบบตอบรับอัตโนมัติ เป็นต้น

(10) ด้านอินเทอร์เน็ตของสรรพสิ่ง (Internet of Thing: IoT) เช่น การประยุกต์ใช้ในอุปกรณ์สำหรับบ้านอัจฉริยะ เมืองอัจฉริยะ และระบบรักษาความปลอดภัยภายในอาคาร เป็นต้น

ในการพัฒนาเทคโนโลยีการสังเคราะห์เสียงจากข้อความยุคแรก ๆ เป็นการออกแบบและพัฒนาขึ้นโดยอิงอยู่กับภาษาอังกฤษเป็นส่วนใหญ่ อย่างไรก็ดี ระบบสังเคราะห์เสียงพูดจากข้อความภาษาไทย เริ่มมีการศึกษาและพัฒนาอย่างจริงจังเมื่อช่วงทศวรรษ 2000 และถึงแม้ระบบสังเคราะห์เสียงพูดจากข้อความภาษาไทยจะได้รับการพัฒนาและปรับปรุงประสิทธิภาพมาอย่างต่อเนื่อง แต่ก็ยังมีหลายประเด็นที่ยังต้องการการปรับปรุงให้มีประสิทธิภาพดียิ่งขึ้น ซึ่งหนึ่งในประเด็นที่สำคัญและมีนักวิจัยหลายคนให้ความสนใจก็คือ เรื่องความเป็นธรรมชาติของเสียงสังเคราะห์ เพราะเป็นสิ่งที่ผู้ฟังเสียงสังเคราะห์สามารถรับรู้ได้โดยง่าย (Wutiwiwatchai *et al*., 2017; Sornlertlamvanich *et al*., 2000; Cardoso *et al*., 2015; Capes., 2017; Csapo, 2012; Dinh *et al*., 2020)

ดังนั้น คณะผู้วิจัยซึ่งสนใจประเด็นนี้เช่นกันจึงได้ทำการศึกษาโดยมีกรอบแนวคิดว่าควรทำการศึกษาเปรียบเทียบความเป็นธรรมชาติของระบบสังเคราะห์เสียงพูดจากข้อความภาษาไทยที่ทำงานในผลิตภัณฑ์ของบริษัทผู้พัฒนาเทคโนโลยีสองรายใหญ่ ได้แก่ Siri ของบริษัท Apple และ Google Translate ของบริษัท Google ซึ่งเป็นแนวคิดที่พัฒนาต่อจากงานวิจัยเดิมของ Daengsi & Pornpongtechavanich (2020) โดยที่งานวิจัยนี้มีการใช้เนื้อหาเกี่ยวกับข่าวในพระราชสำนักและข่าวทั่วไปในการศึกษา แล้วทำการประเมินเชิงจิตวิสัยกับกลุ่มอาสาสมัครรวม 30 คน และใช้การวิเคราะห์ทางสถิติ เพื่อให้ทราบถึงประสิทธิภาพด้านความเป็นธรรมชาติของเสียงสังเคราะห์ของระบบสังเคราะห์เสียงพูดจากข้อความภาษาไทยของทั้งสองระบบว่ามีคุณภาพและมีความแตกต่างกันอย่างไรหรือไม่ ซึ่งจากการศึกษานี้พบว่า เสียงสังเคราะห์ภาษาไทยที่สร้างจาก Google Translate มีความเป็นธรรมชาติของเสียงแตกต่างจากเสียงสังเคราะห์ภาษาไทยที่สร้างจาก Siri อย่างมีนัยสำคัญ

สำหรับโครงสร้างของบทความวิจัยนี้ หลังจากที่ได้มีการกล่าวบทนำแล้ว คณะผู้วิจัยได้บรรยายถึงการทบทวนวรรณกรรม ทฤษฎีที่เกี่ยวข้องไม่ว่าจะเป็น หลักภาษาไทยเบื้องต้น ระบบสังเคราะห์เสียงพูดจากข้อความภาษาไทย การประเมินคุณภาพเสียงเชิงจิตวิสัย Siri และ Google Translate จากนั้นคณะผู้วิจัยได้อธิบายวิธีดำเนินการวิจัย นำเสนอผลการวิจัย ทำการวิจารณ์ผลการวิจัยดังกล่าว แล้วทำการสรุปผลการวิจัย ซึ่งคณะผู้วิจัยได้มีการให้ข้อเสนอแนะสำหรับผู้สนใจเอาไว้ด้วย

การทบทวนวรรณกรรม

ในการดำเนินการวิจัยนี้ คณะผู้วิจัยได้ทำการศึกษางานวิจัยที่มีลักษณะคล้ายกันและอยู่ในแขนงเดียวกันที่มีการศึกษาทั้งในและต่างประเทศ โดยคณะผู้วิจัยมุ่งเน้นศึกษางานวิจัยหรือการศึกษาเรื่องความเป็นธรรมชาติของเสียงสังเคราะห์ พบว่า มีงานวิจัยหลายชิ้นที่ศึกษาความเป็นธรรมชาติของการพูดสังเคราะห์โดยใช้การทดสอบหรือประเมินเชิงจิตวิสัย (ดังแสดงในตารางที่ 1) (Dinh *et al*., 2020; Siri Team, 2017; Dall *et al*., 2014; Shirali-Shahreza & Penn, 2018; Martin *et al*., 2020)

อย่างไรก็ตามจากตารางดังกล่าวจะเห็นได้ว่า การศึกษาส่วนใหญ่เป็นการศึกษากับระบบสังเคราะห์เสียงพูดสำหรับภาษาอื่นที่ไม่ใช่ภาษาไทย ยกเว้นงานวิจัยของ Janyoi & Seresangtakul (2017) และ Chao-angthong *et al*. (2017) ที่มีการศึกษากับเสียงสังเคราะห์ภาษาไทย-อีสาน และภาษาไทย-เหนือ (แต่ก็ยังไม่มีการศึกษากับภาษาไทยมาตรฐานหรือภาษากลาง) และมีงานวิจัยของ Daengsi & Pornpongtechavanich (2021) ที่ทำการศึกษาเบื้องต้นเกี่ยวกับเสียงสังเคราะห์ภาษาไทยมาตรฐาน แต่เนื้อหาข่าวในการศึกษานั้นก็ไม่ได้ครอบคลุมถึงข่าวในพระราชสำนักซึ่งมีคำราชาศัพท์ที่อ่านยากหรือมักมีการอ่านผิด

ทฤษฎี

ก่อนที่จะมีการอธิบายในหัวข้อวิธีดำเนินการวิจัย มีหลายหัวข้อเกี่ยวกับทฤษฎีที่เกี่ยวข้องกับงานวิจัยนี้ที่จำเป็นต้องอธิบายไว้ในหัวข้อย่อยต่าง ๆ ดังนี้

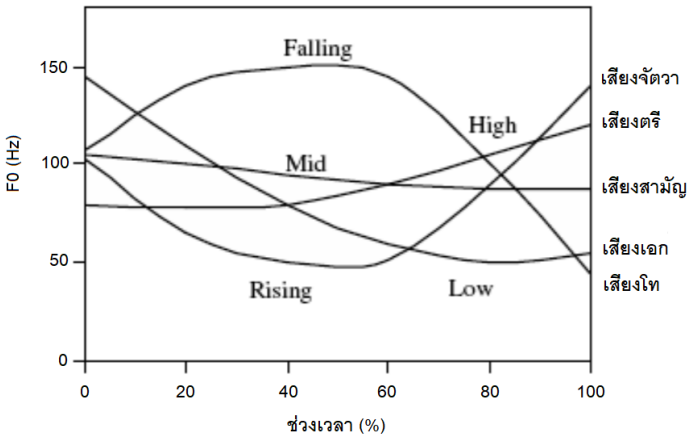
1. หลักภาษาไทยเบื้องต้น

ภาษาไทยเป็นภาษาราชการและเป็นภาษาประจำชาติของคนไทย ถือเป็นภาษาที่มีเอกลักษณ์และมีความแตกต่างจากภาษาตะวันตกเป็นอย่างมาก เช่น ภาษาไทยเป็นภาษาที่มีวรรณยุกต์ (ดังภาพที่ 1) กล่าวคือ เมื่อมีการผันเสียงหรือเปลี่ยนโทนเสียง ก็จะได้เป็นคำใหม่ที่มีความหมายใหม่ ในขณะที่ภาษาตะวันตกส่วนใหญ่ไม่มีคุณสมบัตินี้ ดังนั้น การเรียนรู้เรื่องวรรณยุกต์ภาษาไทยจึงถือเป็นเรื่องที่ค่อนข้างยากสำหรับชาวตะวันตกที่ต้องการเรียนภาษาไทย (Daengsi & Preechayasomboon, 2012; Daengsi *et al*., 2013)

ในหลักการเขียนภาษาไทย จะเขียนจากซ้ายไปขวาเช่นเดียวกันกับการเขียนภาษาอังกฤษ แต่ไม่มีการเว้นช่องว่างระหว่างคำเหมือนภาษาอังกฤษและภาษาตะวันตกอื่น ๆ ส่วนใหญ่จะมีการเว้นวรรคเมื่อมีการเว้นประโยคเป็นหลัก ภาษาไทยเป็นภาษาที่มีโครงสร้างประโยคไม่ซับซ้อน ประโยคทั่ว ๆ ไปจะประกอบด้วย “ประธาน + กิริยา + กรรม” โดยไม่มีการเปลี่ยนรูปของคำหรือรูปแบบการเขียนประโยคตามการเปลี่ยนแปลงของเวลา ปริมาณ หรือเพศ และไม่มีคำเอกพจน์ คำพหูพจน์ เป็นต้น

ตารางที่ *1* งานวิจัยที่เกี่ยวข้องกับการประเมินความเป็นธรรมชาติของสังเคราะห์จากข้อความเชิงจิตวิสัย (Daengsi & Pornpongtechavanich, 2021)

คณะผู้วิจัย	ระบบ/แอปพลิเคชัน/บริการ	ภาษา	เครื่องมือประเมิน	จำนวนผู้ทดสอบ (คน)
Dinh *et al*. (2020)	Google, Microsoft, Ivona, Loquendo, Espeak, Pico, AT&T, และ Nuance	สเปน	จิตวิสัย (การประเมินค่าเฉลี่ยคะแนนความคิดเห็นเปรียบเทียบ (CMOS tests))	125
Martin *et al*. (2020)	แบบจำลอง Novel asymmetric bilinear ที่ใช้ NMF	อังกฤษ	จิตวิสัย (การเปรียบเทียบเสียงสังเคราะห์แบบคู่)	125
Shirali-Shahreza & Penn (2018)	4 ระบบจากงาน The 2013 Blizzard Challenge	อังกฤษ-อังกฤษ และ อินเดียน-อังกฤษ	จิตวิสัย (การประเมินค่าเฉลี่ยคะแนนความคิดเห็นด้วยการเปรียบเทียบแบบคู่หรือการทดสอบแบบเอบี (AB tests))	139
Janyoi & Seresangtakul (2017	Isarn Dialect HMM-based TTS	ไทย-อีสาน	จิตวิสัย (การประเมินค่าเฉลี่ยคะแนนความคิดเห็นด้วยการฟัง)	20
Chao-angthong *et al*. (2017)	Northern Thai Dialect TTS	ไทย-เหนือ	จิตวิสัย (การประเมินค่าเฉลี่ยคะแนนความคิดเห็นด้วยการฟัง)	20
Siri Team (2017)	Siri	อเมริกัน-อังกฤษ	จิตวิสัย (การประเมินค่าเฉลี่ยคะแนนความคิดเห็นด้วยการเปรียบเทียบแบบคู่หรือการทดสอบแบบเอบี (AB tests))	30
		อังกฤษ-อังกฤษ		30
		ออสเตรเลี่ยน-อังกฤษ		30
		สเปน		30
		อิตาเลี่ยน		30
		รัสเซี่ยน		30
		จีน		10+30
Dall *et al*. (2014)	General and Conversational TTS	อังกฤษ	จิตวิสัย (การประเมินค่าเฉลี่ยคะแนนความคิดเห็นด้วยการฟังเสียง)	32



ภาพที่ 1 ตัวอย่างแสดงเส้นโค้งของความถี่พื้นฐานเมื่อมีการออกเสียงวรรณยุกต์ภาษาไทย (Daengsi *et al*., 2012)

อย่างไรก็ดี เนื่องจากรูปแบบการเขียนประโยคภาษาไทยเป็นการเขียนคำติดกันถ้าอยู่ในประโยคเดียวกัน ระบบสังเคราะห์เสียงพูดจากข้อความภาษาไทย จึงจำเป็นต้องมีความสามารถในการตัดคำที่มีความแม่นยำสูง มิฉะนั้นจะส่งผลให้การสังเคราะห์เสียงภาษาไทยเกิดความผิดพลาดได้ (Wutiwiwatchai *et al*., 2017; Sornlertlamvanich *et al*., 2000)

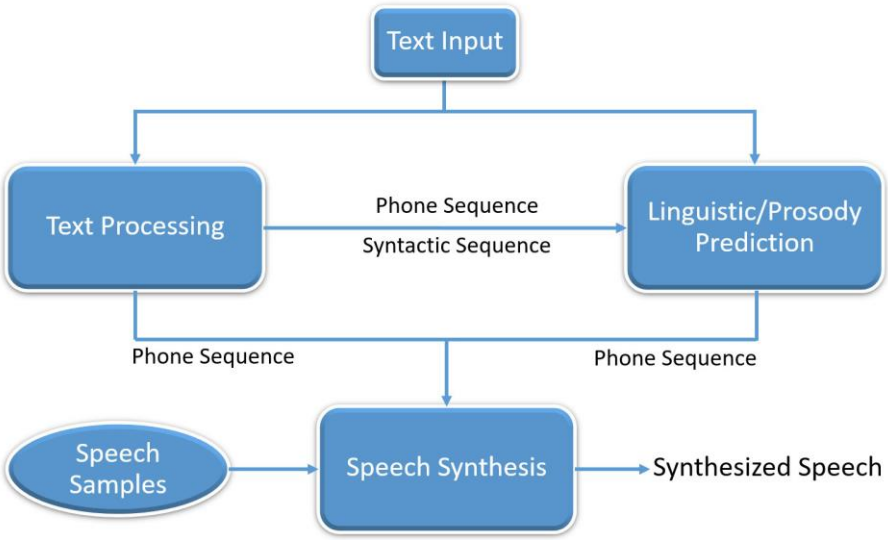
2. *<u>ระบบสังเคราะห์เสียงพูดจากข้อความภาษาไทย</u>*

โดยทั่วไประบบสังเคราะห์เสียงพูดจากข้อความภาษาไทย จะมีหลักการทำงานดังแสดงในภาพที่ 2 โดยเริ่มจากการข้อความเข้ามาแล้วผ่านกระบวนการต่าง ๆ จนถึงขั้นตอนสุดท้ายที่เป็นกระบวนการสังเคราะห์เสียง อย่างไรก็ตาม มีหลายประเด็นที่พบในกระบวนการสังเคราะห์เสียงจากข้อความภาษาไทย เช่น ไม่มีขอบเขตของคำที่ชัดเจน และในบางครั้งมีการใช้คำประสม ไม่มีขอบเขตของประโยคที่ชัดเจน และในบางครั้งมีการเขียนรูปประโยคที่ซับซ้อน มีการออกเสียงของคำบางคำที่ไม่เป็นไปตามหลักเกณฑ์ปกติ (มีข้อยกเว้น) และการออกเสียงเป็นประโยคหรือวลีมีความซับซ้อนในทางเทคนิค เนื่องจากมีประเด็นเรื่องความถี่พื้นฐานและข้อจำกัดเกี่ยวกับเสียงพยางค์ (Wutiwiwatchai *et al*., 2017)

โดยเฉพาะอย่างยิ่ง เมื่อพิจารณาที่โมดูลสังเคราะห์เสียง (Speech synthesis module) พบว่า มีประเด็นเรื่องความราบรื่นหรือความลื่นไหลในการเชื่อมคำ (Smoothness of connected points) ซึ่งส่งผลกระทบต่อความเป็นธรรมชาติของเสียงสังเคราะห์โดยตรง (Wutiwiwatchai *et al*., 2017) อย่างไรก็ดี แม้ว่าจะมีงานวิจัยที่พยายามแก้ไขปัญหานี้ เช่น Kertkeidkachorn *et al*., (2014) แต่ก็ยังมีอีกหลายประเด็นที่ยังเปิดกว้างสำหรับการศึกษาวิจัยเกี่ยวกับความเป็นธรรมชาติของเสียงสังเคราะห์ภาษาไทย

3. *การประเมินคุณภาพเสียงเชิงจิตวิสัย*

การประเมินคุณภาพเสียงนั้น เดิมทีเป็นวิธีการที่ใช้ในการวัดคุณภาพเสียงในโครงข่ายโทรคมนาคม เช่น ระบบโทรศัพท์ โดยมีพื้นฐานมาจากการใช้กลุ่มตัวอย่างประมาณ 30 คน ในการประเมินด้วยการฟังเสียงในแต่ละเงื่อนไขที่ทำการทดสอบ แล้วให้แต่ละคนทำการประเมินด้วยการให้คะแนน โดยใช้เกณฑ์คะแนน 1-5 (เมื่อ 5 คือ ดีเยี่ยม และ 1 คือ แย่) จากนั้นจึงนำคะแนนที่ได้จากทุกคนมาหาค่าเฉลี่ยคะแนนความคิดเห็น (Mean Opinion Score: MOS) ซึ่งต่อไปนี้จะเรียกว่าค่า MOS โดยที่ค่า MOS จะมีความสัมพันธ์โดยตรงกับคุณภาพเสียงที่ผู้ฟังได้รับ การประเมินด้วยการหาค่า MOS ได้รับการยอมรับอย่างกว้างขวางและมีการพัฒนาเป็นวิธีการวัดคุณภาพเสียงเชิงวัตถุวิสัย (Objective Measurements) เช่น E-model และ POLQA อย่างไรก็ดีการประเมินคุณภาพเสียงเชิงจิตวิสัยก็ยังเป็นวิธีการที่ได้รับความนิยม เนื่องจากเป็นวิธีการที่ใช้ค่าใช้จ่ายน้อย ปัจจุบันนี้มีการประยุกต์ใช้ค่า MOS ในการประเมินคุณภาพด้านต่าง ๆ อย่างแพร่หลาย ไม่ว่าจะเป็นการประเมินคุณภาพของวิดีโอ และการประเมินคุณภาพมัลติมีเดียอื่น ๆ (Daengsi & Wuttidittachotti, 2019; ITU-T, 1996; ITU-T, 2007; Pornpongtechavanich & Daengsi, 2019; ITU-T, 2016)

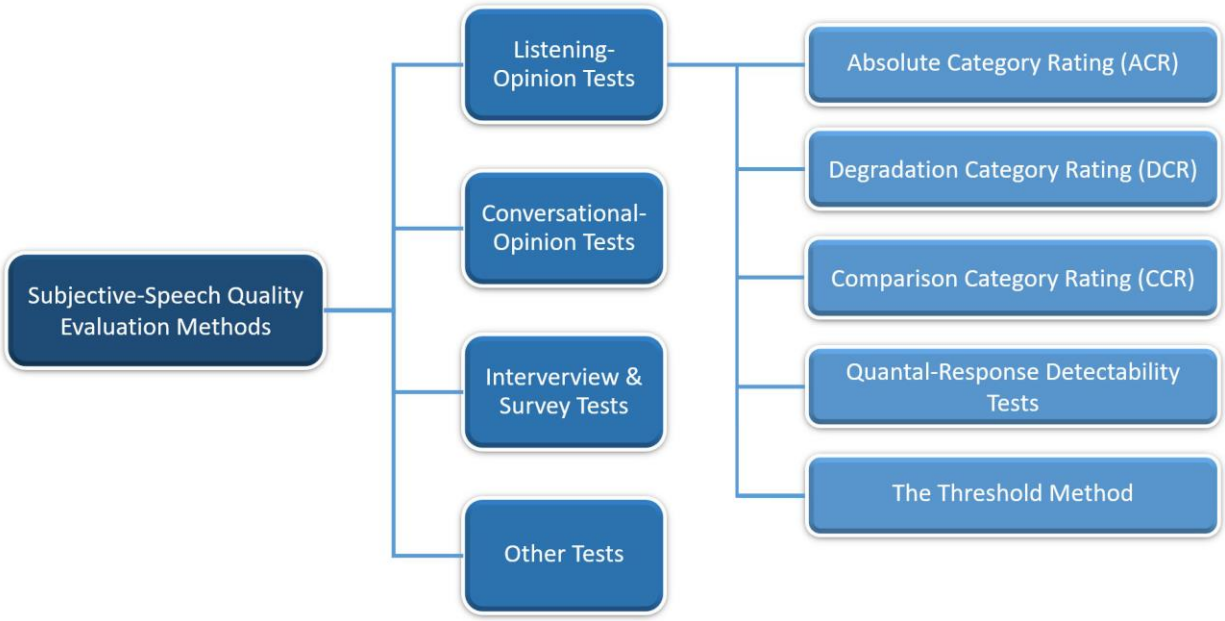


ภาพที่ 2 หลักการทำงานเบื้องต้นของระบบสังเคราะห์เสียงพูดจากข้อความ (Wutiwiwatchai *et al.*, 2017)

การประเมินคุณภาพเสียงเชิงจิตวิสัยสามารถทำได้ทั้งด้วยวิธีการฟัง วิธีการสนทนา และวิธีการสัมภาษณ์และสำรวจ (ดังภาพที่ 3) อย่างไรก็ตาม Daengsi *et al.* (2014) ได้ทำการศึกษาและรายงานว่า วิธีการประเมินด้วยวิธีการฟังแบบ ACR (ย่อมาจาก Absolute Category Rating) เป็นวิธีการประเมินที่มีความน่าเชื่อถือสูงที่สุดเมื่อเทียบกับวิธีการสนทนา และวิธีการสัมภาษณ์ เนื่องจากเป็นวิธีการที่เอื้อให้ผู้ประเมินมีสมาธิดีที่สุดเมื่อเทียบกับวิธีการอื่น

4. สิริ (Siri)

Siri เป็นแอปพลิเคชันที่พัฒนาขึ้นโดยบริษัท Apple เพื่อทำหน้าที่เป็นเสมือนผู้ช่วยส่วนตัวให้กับผู้ใช้งานโทรศัพท์เคลื่อนที่หรืออุปกรณ์พกพาบางชนิด เช่น เครื่องแท็บเล็ต ยี่ห้อ Apple Siri ได้รับการพัฒนาขึ้นภายใต้โครงการ Siri TTS ซึ่งเป็นโครงการพัฒนาระบบสังเคราะห์เสียงพูดจากข้อความที่สามารถรองรับได้หลากหลายภาษา โดยเริ่มเปิดให้ใช้งานได้ในโทรศัพท์ iPhone ที่ใช้ระบบปฏิบัติการ iOS 10 ในปี 2016 แล้วหลังจากนั้นก็ได้รับความนิยมและมีการใช้งานอย่างแพร่หลายไปทั่วโลกในระบบปฏิบัติการ iOS รุ่นต่อ ๆ มาจนถึงปัจจุบัน เทคโนโลยีที่ทำงานอยู่เบื้องหลังของ Siri เป็นการเทคโนโลยีระบบสังเคราะห์เสียงพูดจากข้อความที่ได้รับการพัฒนาขึ้นบนพื้นฐานของการเรียนรู้เชิงลึก (Deep learning) ที่มีการฝึกระบบกับข้อมูลการฝึกพูดที่มีคุณภาพเสียงสูง และเป็นระบบที่มีความสามารถในการทำนายการกระจายของคุณลักษณะเป้าหมายของคำพูด (ระยะเวลา ระดับเสียง และสเปกตรัม จึงทำให้เสียงพูดของ Siri มีความเป็นธรรมชาติ นอกจากนี้คลังข้อมูลที่ใช้สำหรับ Siri ยังประกอบด้วยมีคำศัพท์ภาษาต่าง ๆ รวมกันมากกว่า 200 ล้านคำ รวมทั้งภาษาไทยด้วย (Capes., 2017; Siri Team, 2017)



ภาพที่ 3 วิธีการประเมินคุณภาพเสียงเชิงจิตพิสัย (Daengsi *et al*., 2014)

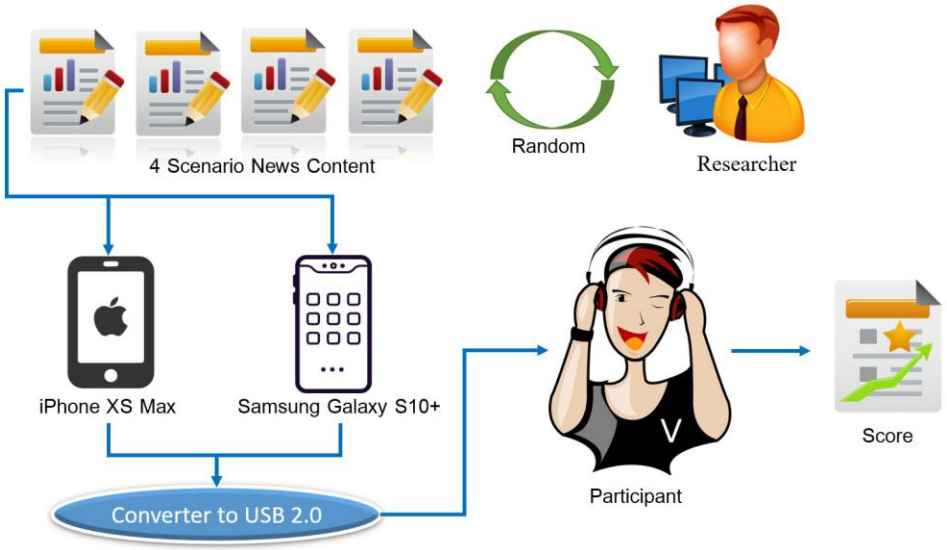
5. กูเกิลทรานสเลต (Google Translate)

Google Translate ถือเป็นหนึ่งในบริการยอดนิยมของ Google ที่สามารถใช้ได้ทั้งในรูปแบบของบริการผ่านเว็บไซต์และในรูปแบบของแอปพลิเคชัน Google Translate ได้รับการออกแบบและพัฒนาโดยบริษัท Google เพื่อแปลคำ ข้อความและประโยคที่อยู่ในภาษาหนึ่งให้เป็นอีกภาษาหนึ่งได้ทันที ในฐานข้อมูลการแปลภาษาของ Google Translate บรรจุคำและ

ข้อความจำนวนมหาศาล ดังนั้นจึงสามารถแปลเป็นภาษาต่าง ๆ ได้มากกว่า 100 ภาษาสำหรับโหมดออนไลน์และประมาณ 60 ภาษาในโหมดออฟไลน์ผ่านแอปพลิเคชัน นอกจากนี้ ยังมีความสามารถในการรองรับฟังก์ชันแปลงข้อความเป็นเสียงพูด โดยใช้เครื่องมือสังเคราะห์เสียงพูดจากข้อความของ Google โดยผู้ใช้งานสามารถแปลงข้อความเป็นเสียงพูดหรือแปลภาษาได้สูงสุดครั้งละ 5,000 อักขระ (Google Play, 2020; Martín, 2017)

วิธีดำเนินการวิจัย

ระบบที่ใช้ในการวิจัยประกอบด้วยโทรศัพท์เคลื่อนที่ 2 เครื่อง (ดังแสดงในภาพที่ 4) โดยหนึ่งเครื่องใช้ระบบปฏิบัติการ iOS สำหรับทดสอบความเป็นธรรมชาติของเสียงสังเคราะห์ที่สร้างจาก Siri และอีกหนึ่งเครื่องใช้ระบบปฏิบัติการ Android สำหรับทดสอบความเป็นธรรมชาติของเสียงสังเคราะห์ที่สร้างจาก Google Translate โดยใช้ WiFi ในการเชื่อมต่อโทรศัพท์เคลื่อนที่ทั้งสองเครื่องเข้ากับเครือข่ายอินเทอร์เน็ตที่มีความเร็วสูงสุด 1 Gbps ของคณะอุตสาหกรรมและเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลรัตนโกสินทร์ วิทยาเขตวังไกลกังวล โดยมีอาสาสมัครซึ่งเป็นนักศึกษาของคณะดังกล่าว เข้าร่วมการประเมินความเป็นธรรมชาติของเสียงสังเคราะห์ที่สร้างจาก Siri และ Google Translate จำนวนทั้งหมด 32 คน (Wuttidittachotti *et al*, 2017; Daengsi and Wutiwiwatchai, 2013) (แบ่งเป็นชาย 16 คน และหญิง 16 คน) แล้วทำการประเมินด้วยการประยุกต์ใช้วิธีการประเมินด้วยการฟังแบบ ACR ซึ่งเป็นวิธีการประเมินคุณภาพเสียงที่มีความน่าเชื่อถือสูงที่สุดเมื่อเทียบกับวิธีการอื่น (Daengsi *et al*., 2014) โดยอาสาสมัครกลุ่มดังกล่าวนี้มีอายุระหว่าง 18-21 ปี (เฉลี่ย 19.00±1.14 ปี) และมีการแบ่งกลุ่มอาสาสมัครออกเป็น 4 กลุ่มย่อยแบบสุ่ม (คละชาย-หญิง) แล้วให้ทำการทดลองฟังเสียงสังเคราะห์ภาษาไทยที่สังเคราะห์จาก Siri และ Google Translate แบบสุ่มโดยที่อาสาสมัครผู้เข้าร่วมทดสอบจะไม่ทราบว่ากำลังฟังเสียงสังเคราะห์ภาษาไทยที่จาก Siri และ Google Translate ทั้งนี้อาสาสมัครที่เข้าร่วมทดสอบเป็นกลุ่มตัวอย่างที่มีคุณลักษณะคล้ายคลึงกันทั้งช่วงอายุและพื้นฐานการศึกษา เนื่องจากเป็นนักศึกษาสาขาวิชาเดียวกัน (เพื่อหลีกเลี่ยงประเด็นเรื่องความแปรปรวนของข้อมูลหรือผลการทดสอบ) โดยเนื้อหาที่ใช้ในการทดสอบหรือทดลองเป็นเนื้อหาข่าวในพระราชสำนัก 2 ข่าวและข่าวทั่วไป 2 ข่าว ดังแสดงในตารางที่ 2 ซึ่งใช้เวลาในการฟังเนื้อหาข่าวรวมกันประมาณ 2:20 นาที (หากเนื้อหายาวอาสาสมัครอาจเกิดความรู้สึกเบื่อหน่าย และอาจส่งผลต่อคะแนนการประเมินของอาสาสมัคร) โดยผู้ควบคุมการทดลองให้อาสาสมัครทดลองฟังทีละคน และต้องทำการประเมินความเป็นธรรมชาติของเสียงสังเคราะห์ภาษาไทยทันทีที่ได้ฟังเสียงสังเคราะห์จบ โดยใช้เกณฑ์คะแนน 1-5 เมื่อ 5 คือมีความเป็นธรรมชาติในระดับดีเยี่ยม และ 1 คือมีความเป็นธรรมชาติในระดับแย่ หรือไม่มีความเป็นธรรมชาติ ซึ่งวิธีการประเมินที่ใช้ในวิธีดำเนินการวิจัยนี้ถือเป็นวิธีการประเมินด้วยวิธีการฟังแบบ ACR อย่างไม่เป็นทางการหรืออย่างง่าย



ภาพที่ 4 ภาพรวมกระบวนการประเมินความเป็นธรรมชาติของเสียงสังเคราะห์ในการศึกษานี้

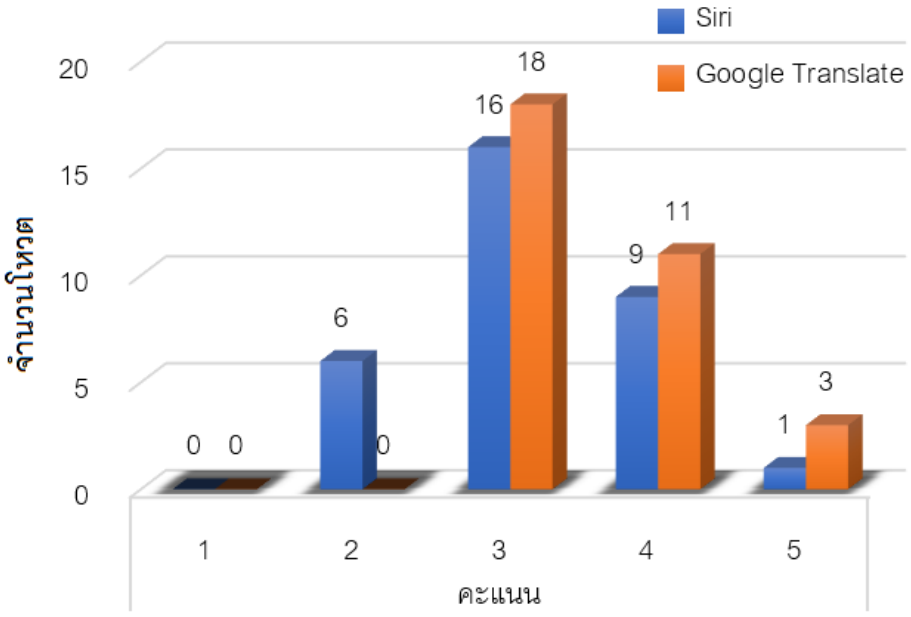
ตารางที่ 2 หัวข้อข่าวและเนื้อหาข่าวที่ใช้ในการประเมินความเป็นธรรมชาติของเสียงสังเคราะห์ภาษาไทยที่สร้างจาก Siri และ Google Translate ในการศึกษานี้

หัวข้อข่าว	เนื้อหาข่าว
ข่าวในพระราชสำนัก 1	เมื่อวันที่ 3 มีนาคม พระบาทสมเด็จพระเจ้าอยู่หัว ทรงพระกรุณาโปรดเกล้าโปรดกระหม่อมให้ส่งข้อความพระราชสาส์นอำนวยพรไปยังประธานาธิบดีแห่งสาธารณรัฐบัลแกเรีย ในโอกาสวันชาติของสาธารณรัฐบัลแกเรีย ซึ่งตรงกับวันที่ 3 มีนาคม 2563
ข่าวในพระราชสำนัก 2	สมเด็จพระเจ้าน้องนางเธอ เจ้าฟ้าจุฬาภรณวลัยลักษณ์ อัครราชกุมารี กรมพระศรีสวางควัฒน วรขัตติยราชนารี เสด็จไปทรงสักการะสังเวชนียสถาน และเป็นประธานฝ่ายฆราวาสในพิธีลาสิกขาบทพระนวกะ “โครงการบรรพชาอุปสมบทหมู่และปฏิบัติธรรม ถวายเป็นพระกุศล ศาสตราจารย์ ดร.สมเด็จพระเจ้าน้องนางเธอ เจ้าฟ้าจุฬาภรณวลัยลักษณ์ อัครราชกุมารี กรมพระศรีสวางควัฒน วรขัตติยราชนารี เนื่องในโอกาสครบรอบ 10 ปี โรงพยาบาลจุฬาภรณ์” ณ พระอุโบสถ วัดไทยกุสินาราเฉลิมราชย์ เมืองกุสินารา รัฐอุตรประเทศ สาธารณรัฐอินเดีย
ข่าวทั่วไป 1	สถานการณ์ไวรัสโคโรน่าสายพันธุ์ใหม่ 2019 ที่แพร่ระบาดหนักในประเทศจีน และอีกหลายระเทศทั่วโลก รวมถึงประเทศไทยด้วย "กรุงเทพธุรกิจออนไลน์" ได้รวบรวมข้อมูลต่าง ๆ เพื่ออัพเดทสถานการณ์ล่าสุดและเรื่องราวที่เกี่ยวกับ "ไวรัสโคโรน่า" สายพันธุ์ใหม่นี้ (COVID-19) ณ วันที่ 4 มีนาคม 2563
ข่าวทั่วไป 2	นายแพทย์ธนรักษ์ ผลิพัฒน์ รองอธิบดีกรมควบคุมโรค โพสต์เฟซบุ๊กส่วนตัวระบุว่า โรคติดเชื้อไวรัสโคโรนา 2019 จะแพร่ระบาดกว้างขวางแค่ไหนในประเทศไทย ขึ้นอยู่กับอะไรบ้าง ขึ้นอยู่กับ 3 ปัจจัย ได้แก่ 1.ปัจจัยที่เกี่ยวกับตัวเชื้อ เชื้อตัวนี้เรารู้กันเป็นอย่างดีว่าสามารถแพร่จากคนไปสู่คนได้อย่างมีประสิทธิภาพ ผู้ป่วย 1 คน สามารถแพร่เชื้อให้กับผู้ที่ยังไม่ติดเชื้อได้มากกว่า 2 คน

ผลการวิจัย

ผลที่ได้จากการประเมินความเป็นธรรมชาติของเสียงสังเคราะห์ที่สร้างจาก Siri และ Google Translate ในรูปแบบกราฟแท่งแสดงความถี่ของคะแนนที่ได้จากการประเมิน แสดงดังภาพที่ 5 ซึ่งจะเห็นได้ว่าอาสาสมัครที่ประเมินหรือโหวตให้คะแนนความเป็นธรรมชาติของเสียงสังเคราะห์ภาษาไทยที่ได้จาก Google Translate ในช่วง 3 ถึง 5 คะแนน มีคะแนนสูงกว่าคะแนนความเป็นธรรมชาติของเสียงสังเคราะห์ภาษาไทยที่ได้จาก Siri ทุกกรณี และเมื่อนำไปคำนวณหาค่าความเป็นธรรมชาติของเสียงสังเคราะห์ (ต่อไปนี้เรียกว่า Natural-MOS หรือสามารถเขียนสั้น ๆ เป็น nMOS) และแสดงในตารางที่ 3 (บางส่วนเป็นข้อมูลที่ใช้วิเคราะห์ใน (Daengsi & Pornpongtechavanich, 2021)) ซึ่งจะเห็นได้ว่าค่าเฉลี่ยรวมของความเป็นธรรมชาติของเสียงสังเคราะห์จาก Google Translate (nMOS = 3.53 ± 0.67) มีค่าเฉลี่ยรวมมากกว่าค่าความเป็นธรรมชาติของเสียงสังเคราะห์จาก Siri (nMOS = 3.16 ± 0.77)

อย่างไรก็ดี เพื่อวิเคราะห์ว่าค่า nMOS สำหรับความเป็นธรรมชาติของเสียงสังเคราะห์ที่สร้างจาก Siri และ Google Translate มีความแตกต่างกันอย่างมีนัยสำคัญหรือไม่ จึงได้มีการวิเคราะห์ทางสถิติด้วย t-test แล้วแสดงผลการวิเคราะห์ในตารางที่ 4 ซึ่งจากผลการวิเคราะห์ด้วย t-test พบว่า ความเป็นธรรมชาติของเสียงสังเคราะห์ที่ได้จาก Siri และ Google Translate โดยรวมแล้ว มีความแตกต่างกันอย่างมีนัยสำคัญที่ช่วงความเชื่อมั่น 95% เนื่องจากมีค่า p-value น้อยกว่า 0.05 โดยมีค่าเท่ากับ 0.037 (แม้ว่าเมื่อทำการวิเคราะห์แยกผลการประเมินตามเพศแล้วพบว่า ผลการวิเคราะห์คะแนนประเมินที่ได้จากเพศหญิงจะไม่แตกต่างกันอย่างมีนัยสำคัญก็ตาม (p-value = 0.289)



ภาพที่ 5 กราฟแท่งแสดงความถี่ของคะแนนที่ได้จากการประเมิน

ตารางที่ 3 ค่าคะแนนจากผลการประเมินความเป็นธรรมชาติของเสียงสังเคราะห์ที่ได้จากการศึกษานี้

อาสาสมัคร	จำนวน	nMOS ± SD	
		Siri	Google Translate
ชาย	16	3.00 ± 0.73	3.44 ± 0.51
หญิง	16	3.31 ± 0.79	3.63 ± 0.81
รวม	32	3.16 ± 0.77	3.53 ± 0.67

หมายเหตุ : SD คือค่าเบี่ยงเบนมาตรฐาน

ตารางที่ 4 ผลการวิเคราะห์ทางสถิติ

สมมุติฐาน		p-value	หมายเหตุ
H1 (ชาย)	$H1_0$: $nMOS_{Siri} = nMOS_{GoogleTranslate}$	0.048*	มีนัยสำคัญ
	$H1_1$: $nMOS_{Siri} \neq nMOS_{GoogleTranslate}$		
H2 (หญิง)	$H2_0$: $nMOS_{Siri} = nMOS_{GoogleTranslate}$	0.289	ไม่มีนัยสำคัญ
	$H2_1$: $nMOS_{Siri} \neq nMOS_{GoogleTranslate}$		
H3 (รวมชายและหญิง)	$H3_0$: $nMOS_{Siri} = nMOS_{GoogleTranslate}$	0.037*	มีนัยสำคัญ
	$H3_1$: $nMOS_{Siri} \neq nMOS_{GoogleTranslate}$		

วิจารณ์ผลการวิจัย

จากการศึกษานี้ที่ใช้เนื้อหาข่าวในพระราชสำนักซึ่งมีคำราชาศัพท์และข่าวทั่วไปในการทดสอบ และได้ผลการวิจัยภาพที่ 5 จะเห็นได้ว่า คะแนนโหวตของอาสาสมัคร (จำนวน 36 คน) ในการทดสอบกับ Google Translate กระจายอยู่ในช่วง 3-5 คะแนน โดยระดับคะแนน 3 ถูกโหวตมากที่สุดคือ 18 คน ระดับคะแนน 4 ถูกโหวต 11 คน และระดับคะแนน 5 ถูกโหวต 3 คน ในขณะที่คะแนนโหวตการทดสอบกับ Google Translate กระจายอยู่ในช่วง 2-5 คะแนน โดยระดับคะแนน 3 มีจำนวนโหวตมากที่สุดคือ 16 คน ระดับคะแนน 4 ถูกโหวต 9 คน และระดับคะแนน 5 ถูกโหวตเพียง 1 คน นอกจากนี้ยังมีระดับคะแนน 2 ถูกโหวตด้วย 6 คน ซึ่งคิดเป็นสัดส่วนถึง 1 ใน 6 ของอาสาสมัคร

เมื่อพิจารณาผลการวิจัยในตารางที่ 3 จะเห็นได้ว่า ค่า nMOS หรือค่าความเป็นธรรมชาติของเสียงสังเคราะห์ภาษาไทยที่สร้างจาก Google Translate สูงกว่า Siri ในทุกเงื่อนไข ไม่ว่าจะเป็นเงื่อนไขของอาสาสมัครชายที่พบว่า

$nMOS_{GoogleTranslate}$ มีค่าเท่ากับ 3.44 ± 0.51 ส่วน $nMOS_{Siri}$ มีค่าเท่ากับ 3.00 ± 0.73 เงื่อนไขของอาสาสมัครหญิง พบว่า $nMOS_{GoogleTranslate}$ มีค่าเท่ากับ 3.63 ± 0.81 ส่วน $nMOS_{Siri}$ มีค่าเท่ากับ 3.31 ± 0.79 และในเงื่อนไขที่รวมอาสาสมัครทั้งชายและหญิงที่พบว่า $nMOS_{GoogleTranslate}$ และ $nMOS_{Siri}$ มีค่าเท่ากับ 3.53 ± 0.67 และ 3.16 ± 0.77 ตามลำดับ ยิ่งไปกว่านั้นเมื่อทำการวิเคราะห์ทางสถิติด้วย t-test ในเงื่อนไขที่รวมอาสาสมัครทั้งชายและหญิง (ดังตารางที่ 4) พบว่ามีค่า p-value เท่ากับ 0.037 ซึ่งน้อยกว่า 0.05 จึงยืนยันได้ว่า ค่าความเป็นธรรมชาติของเสียงสังเคราะห์ภาษาไทยที่สร้างจาก Google Translate สูงกว่า Siri อย่างมีนัยสำคัญที่ช่วงความเชื่อมั่น 95%

อย่างไรก็ตาม จากค่าความเป็นธรรมชาติของเสียงสังเคราะห์ภาษาไทยที่ได้จากการศึกษานี้ สามารถกล่าวได้ว่าเสียงสังเคราะห์ภาษาไทยที่สร้างจาก Google Translate และ Siri มีคุณภาพความเป็นธรรมชาติของเสียงอยู่ในระดับปานกลางเท่านั้น โดยเฉพาะอย่างยิ่งเสียงสังเคราะห์ของ Google Translate ที่ได้จากการศึกษานี้ มีค่าความเป็นธรรมชาติของเสียงสังเคราะห์ใกล้เคียงกับค่าที่เคยมีการศึกษาในงานวิจัยของ Martin et al. (2020) ที่มีค่าเท่ากับ 3.79 จากการศึกษากับเสียงสังเคราะห์ภาษาสเปน

สรุปผลการวิจัย

จากการศึกษานี้ ซึ่งใช้วิธีการประเมินเชิงจิตวิสัยในการศึกษาเปรียบเทียบความเป็นธรรมชาติของเสียงสังเคราะห์ภาษาไทยที่สร้างจาก Siri และ Google Translate พบว่า เสียงสังเคราะห์ภาษาไทยที่สร้างจาก Google Translate มีแนวโน้มของการมีความเป็นธรรมชาติมากกว่าเสียงสังเคราะห์ภาษาไทยที่สร้างจาก Siri อย่างมีนัยสำคัญ แม้ว่าคุณภาพความเป็นธรรมชาติของเสียงที่สร้างจากทั้ง Siri และ Google Translate อยู่ในระดับปานกลางก็ตาม ทั้งนี้ผู้พัฒนาระบบสังเคราะห์เสียงจากข้อความ สามารถนำวิธีการประเมินที่ใช้ในการศึกษานี้ไปประยุกต์ใช้ในการประเมินระบบสังเคราะห์เสียงจากข้อความระบบอื่น ๆ และสามารถนำไปช่วยในการปรับปรุงและพัฒนาระบบสังเคราะห์เสียงจากข้อความเพื่อให้ได้เสียงสังเคราะห์ที่มีความเป็นธรรมชาติมากขึ้นได้

อย่างไรก็ตาม การศึกษานี้จำกัดเฉพาะเนื้อหาข่าวเท่านั้น ยังไม่ครอบคลุมการศึกษาเปรียบเทียบด้วยเนื้อหาอื่น ๆ เช่น นวนิยาย เรื่องสั้น และบทความวิทยาศาสตร์ เป็นต้น นอกจากนี้การศึกษานี้ยังจำกัดเฉพาะการประเมินเชิงจิตวิสัยเท่านั้น ผู้วิจัยจึงมีแนวคิดที่จะดำเนินการวิจัยเพิ่มเติมในระยะถัดไปโดยนำการประเมินเชิงวัตถุวิสัยมาประยุกต์กับงานวิจัยเกี่ยวกับความเป็นธรรมชาติของเสียงสังเคราะห์ภาษาไทยด้วย

กิตติกรรมประกาศ

ขอขอบคุณคณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลพระนคร และคณะอุตสาหกรรมและเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลรัตนโกสินทร์ ที่ให้การสนับสนุนการวิจัยในครั้งนี้

เอกสารอ้างอิง

Capes, T., Coles, P., Conkie, A., Golipour, L., Hadjitarkhani, A., Hu, Q., Huddleston, N., Hunt, M., Li, J., Neeracher, M., Prahallad, K., Raitio, T., Rasipuram, R., Townsend, G., Williamson, B., Winarsky, D., Wu, Z., & Zhang, H. (2017). Siri On-Device Deep Learning-Guided Unit Selection Text-to-Speech System. In *Proc. of INTERSPEECH*. (pp. 4011-4015). Stockholm, Sweden. Retrieved February 9, 2021, from: https://pdfs.semanticscholar.org/702e/aa99bcb366d08d7f450ed7e354f9f6920b23.pdf

Cardoso, W., Smith, G., & Fuentes C. G. (2015). Evaluating text-to-speech synthesizers. In *Proc. of EUROCALL*. (pp. 108-113). Padova, Italy. Retrieved February 9, 2021, from: https://files.eric.ed.gov/fulltext/ED564181.pdf

Csapo, T. G. (2020). *Increasing the Naturalness of Synthesized Speech*. Retrieved February 9, 2021, from: http://smartlab.tmit.bme.hu/csapo/downloads/Csapo-phonetician2012-paper.pdf

Daengsi, T. & Pornpongtechavanich, P. (2021). Quality of Experience: Comparison of Synthesized Speech Naturalness Between Apple's Siri and Google Translate Referring to Thai Language. In *Proc of ICCCI 2021)*. Coimbatore, INDIA.

Daengsi, T., Preechayasomboon, A., Sukparungsee, S., & Wutiwiwatchai, C. (2012). Thai Text Resource: A Recommended Thai Text Set for Voice Quality Measurements and Its Comparative Study. *KKU Science Journal*, *40*(4), 1114-1127. Retrieved February 9, 2021, from: http://scijournal.kku.ac.th/files/Vol_40_No_4_P_1114-1127.pdf

Daengsi, T., Wutiwiwatchai, C., Preechayasomboon, A., & Sukparungsee, S. (2014). IP Telephony: Comparison of Subjective Assessment Methods for Voice Quality Evaluation. *Walailak Journal of Science and Technology*, *11*(2), 87-92. Retrieved February 9, 2021, from: https://wjst.wu.ac.th/index.php/wjst/article/view/577/353

Daengsi, T., & Wuttidittachotti, P. (2019). QoE Modeling for Voice over IP: Simplified E-model Enhancement Utilizing the Subjective MOS Prediction Model – A Case of G.729 and Thai Users. *Journal of Network and Systems Management*, 27(4), 837–859. Retrieved February 9, 2021, from: https://link.springer.com/article/10.1007/s10922-018-09487-4

Daengsi, T. Yochanang, K., & Wuttidittachotti, P. (2013). A Study of Perceptual VoIP Quality Evaluation with Thai Users and Codec Selection Using Voice Quality - Bandwidth Tradeoff Analysis. In *Proc. of 4th ICTC*. (pp. 691-696). Jeju Island, Korea. Retrieved February 9, 2021, from: http://www2.it.kmutnb.ac.th/teacher/FileDL/Kiattisak1712255614465.pdf

Dall, R., Yamagishi, J., & King, S. (2014). Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation. In *Proc. of Conference contribution*. (pp. 1-5). Dublin, Ireland. Retrieved February 9, 2021, from: https://core.ac.uk/download/pdf/24060899.pdf

Dinh, T., Kain, A., Samlan, R., Cao B., & Wang, J. (2020). Increasing the Intelligibility and Naturalness of Alaryngeal Speech Using Voice Conversion and Synthetic Fundamental Frequency. In *Proc. of INTERSPEECH*. (pp. 4781-4785) Shanghai, China. Retrieved February 9, 2021, from: https://isca-speech.org/archive/Interspeech_2020/pdfs/1196.pdf

Google Play. (2020). *Google Translate*. Retrieved February 9, 2021, from: https://play.google.com/store/apps/details?id=com.google.android.apps.translate&hl=en

ITU-T Recommendation P.800. (1996). *Methods for subjective determination of transmission quality*. Retrieved February 9, 2021, from: http://www.itu.int/rec/T-REC-P.800-199608-I

ITU-T Recommendation P.800.2 (2016). *Mean opinion score interpretation and reporting*. Retrieved February 9, 2021, from: https://www.itu.int/rec/T-REC-P.800.2/en

ITU-T Recommendation P.805. (2007). *Subjective evaluation of conversational quality*. Retrieved February 9, 2021, from: https://www.itu.int/rec/T-REC-P.805/en

Janyoi, P., & Seresangtakul, P. (2017). An Isarn dialect HMM-based text-to-speech system. Proc. INCIT 2017, doi: 10.1109/INCIT.2017.8257873

Kertkeidkachorn, N., Chanjaradwichai, S., Punyabukkana, P., & Suchato, A. (2014). CHULA TTS: A modularized text-to-speech framework. In *Proc. of PACLIC*. (pp. 414–421). Phuket, Thailand. Retrieved February 9, 2021, from: https://www.aclweb.org/anthology/Y14-1048.pdf

Martin, A. F., Malfaz, M., Castro-González, A., Castillo, C. J., & Salichs, A. M. (2020). Four-Features Evaluation of Text to Speech Systemsfor Three Social Robots," *Electronics*, *9*(2), 1-23. Retrieved February 9, 2021, from: https://www.mdpi.com/2079-9292/9/2/267/pdf

Martín, B. S. (2017). Translation Quality Assessment of Google Translate and Microsoft Bing Translator. Thesis, Universidad de Valladolid, spain. Retrieved February 9, 2021, from: http://uvadoc.uva.es/bitstream/handle/10324/22596/TFG_F_2017_7.pdf?sequence=1&isAllowed=y

Pornpongtechavanich, P. & Daengsi, T. (2019). Video Telephony - Quality of Experience: A Simple QoE Model to Assess Video Calls Using Subjective Approach. *Multimedia Tools and Applications*, *78*(22), 31987-32006. Retrieved February 9, 2021, from: https://link.springer.com/article/10.1007/s11042-019-07928-z

ReadSpeaker. (2020). *TTS Software Use Cases*. Retrieved February 9, 2021, from: https://www.readspeaker.com/tts-software-use-cases/

Shirali-Shahreza, S., & Penn, G. (2018). MOS Naturalness and the quest for human-like speech. In *Proc. of IEEE SLT Workshop*. (pp. 346-352) Athens, Greece. Retrieved February 9, 2021, from: https://doi.org/10.1109/SLT.2018.8639599

Siri Team. (2017). *Deep Learning for Siri's Voice: On-device Deep Mixture Density Networks for Hybrid Unit Selection Synthesis*. Retrieved May 10, 2020, from: https://machinelearning.apple.com/research/siri-voices

Sriwongchai, S., Setthee, P., & Prasongsook, S. (2017). Study on Behavior of Participation in Solid Waste Management of Burapha University Sakaeo Campus's Students and Personnel. *Burapha Scuence Journal, 22*(2), 288-299. Retrieved February 9, 2021, from: http://science.buu.ac.th/ojs246/index.php/sci/article/download/1505/1448

Sornlertlamvanich, V., Potipiti, T., Wutiwiwatchai C., & Mittrapiyanuruk P. (2020). The State of the Art in Thai Language Processing. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. (pp. 1-2). Bangkok, Thailand. Retrieved February 9, 2021, from: https://dl.acm.org/doi/pdf/10.3115/1075218.1075296

Wutiwiwatchai, C., Hansakunbuntheung, C., Rugchatjaroen, A., Saychum, S., Kasuriya S. & Chootrakool P. (2017). Thai Text-to-Speech Synthesis: A Review. *Journal of Intelligent Informatics and Smart Technology*, *2*, 1-8. Retrieved February 9, 2021, from: https://jiist.aiat.or.th/assets/uploads/1507618319428sW8WxJ002.1_Chai_ThaiTTS.pdf

Wuttidittachotti, P., Khaoduang, P., and Daengsi, T. (2018). MOS Estimation Model Development Using ACR Listening-Opinion Tests with Thai Users Referring to Loss Effects: A Case of G.726 and G.729. *Multimedia Systems*, *Vol. 24*(3), pp. 285–295.