โมเดลวิเคราะห์เสียงยิ้มภาษาไทยด้วยเทคนิคการเรียนรู้เชิงลึก

โดย

นายนริศร์ พรหมบุตร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ แบบ 1.1 ระดับปริญญาดุษฎีบัณฑิต

ภาควิชาคอมพิวเตอร์

มหาวิทยาลัยศิลปากร

ปีการศึกษา 2566

โมเดลวิเคราะห์เสียงยิ้มภาษาไทยด้วยเทคนิคการเรียนรู้เชิงลึก

โดย
นายนริศร์ พรหมบุตร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ แบบ 1.1 ระดับปริญญาดุษฎีบัณฑิต
ภาควิชาคอมพิวเตอร์
มหาวิทยาลัยศิลปากร
ปีการศึกษา 2566

THAI SMILE VOICE CLASSIFICATION MODEL USING CONVOLUTION NEURAL NETWORK

By

MR. Naris PROMBUT

A Thesis Submitted in Partial Fulfillment of the Requirements

for Doctor of Philosophy INFORMATION TECHNOLOGY

Department of COMPUTER SCIENCE

Academic Year 2023

หัวข้อ	โมเดลวิเคราะห์เสียงยิ้มภาษาไทยด้วยเทคนิคการเรียนรู้เชิงลึก
โดย	นายนริศร์ พรหมบุตร
สาขาวิชา	เทคโนโลยีสารสนเทศ แบบ 1.1 ระดับปริญญาดุษฎีบัณฑิต
อาจารย์ที่ปรึกษาหลัก	ผู้ช่วยศาสตราจารย์ ดร. ณัฐโชติ พรหมฤทธิ์
อาจารย์ที่ปรึกษาร่วม	อาจารย์ ดร. สัจจาภรณ์ ไวจรรยา

คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร ได้รับพิจารณาอนุมัติให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปรัชญาดุษฎีบัณฑิต

.....คณบดีคณะวิทยาศาสตร์
(ผู้ช่วยศาสตราจารย์ ดร. นรงค์ ฉิมพาลี)

พิจารณาเห็นชอบโดย

.....ประธานกรรมการ
(รองศาสตราจารย์ ดร. ปานใจ ธารทัศนวงศ์)

.....อาจารย์ที่ปรึกษาหลัก
(ผู้ช่วยศาสตราจารย์ ดร. ณัฐโชติ พรหมฤทธิ์)

.....อาจารย์ที่ปรึกษาร่วม
(อาจารย์ ดร. สัจจาภรณ์ ไวจรรยา)

.....ผู้ทรงคุณวุฒิภายใน
(ผู้ช่วยศาสตราจารย์ ดร. อรวรรณ เชาวลิต)

.....ผู้ทรงคุณวุฒิภายนอก
(รองศาสตราจารย์ ดร. เยาวดี เต็มธนาภัทร์)

60309803 : เทคโนโลยีสารสนเทศ แบบ 1.1 ระดับปริญญาดุษฎีบัณฑิต
คำสำคัญ : Data Augmentation, 2DCNN, Deep Learning, เสียงยิ้ม

นาย นริศร์ พรหมบุตร: โมเดลวิเคราะห์เสียงยิ้มภาษาไทยด้วยเทคนิคการเรียนรู้เชิงลึก
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก : ผู้ช่วยศาสตราจารย์ ดร. ณัฐโชติ พรหมฤทธิ์

เสียงยิ้ม (Smile Voice) คือเสียงที่ผู้ฟังได้ฟังแล้วเกิดความรู้สึกว่าผู้พูดกำลังยิ้มอยู่ ซึ่งการฝึกเสียงยิ้ม เป็นสิ่งที่พนักงานคอลเซ็นเตอร์ต้องฝึกฝน เนื่องจากพนักงานคอลเซ็นเตอร์ คือด่านแรกของการติดต่อ สอบถาม แจ้งปัญหาจากลูกค้า และผู้รับบริการ ซึ่งในการสนทนาพนักงานคอลเซ็นเตอร์ต้องควบคุมอารมณ์ และเลือกใช้บทสนทนาที่ส่งผลกับภาพลักษณ์ของสินค้า และบริการนั้น ๆ ในทางบวก การพูดด้วยเสียงยิ้มพนักงานคอลเซ็นเตอร์ต้องมีการฝึกอบรมการใช้เสียง และการแสดงออกทางสีหน้า การฝึกออกเสียงทำโดยฝึกที่หน้ากระจก และมีพนักงานพี่เลี้ยงคอยให้คำแนะนำเกี่ยวกับน้ำเสียง และใบหน้าที่มีรอยยิ้ม และในขณะที่ปฏิบัติงานพนักงานคอลเซ็นเตอร์จะต้องวางกระจกตรงหน้าของตน เพื่อสังเกตตนเองขณะที่สนทนากับลูกค้าด้วย วิทยานิพนธ์นี้สร้างโมเดลวิเคราะห์เสียงพูดภาษาไทยด้วยรอยยิ้ม โดยใช้เทคนิคการเรียนรู้เชิงลึก เพื่อมุ่งให้เกิดการพัฒนาต่อยอดเป็นเครื่องมือช่วยฝึกการพูดด้วยเสียงยิ้ม โดยทดลองสร้างโมเดลจำแนกเสียงยิ้มในภาษาไทย และนำโมเดลมาปรับใช้กับชุดข้อมูลเสียงของคอลเซ็นเตอร์ โดยนำชุดข้อมูลมาทำ Data Augmentation พบว่า โมเดลที่มีประสิทธิภาพดีที่สุดคือ 2D CNN MFCC ร่วมกับ Augmentation ได้ค่าความถูกต้อง 75.61% และมีการสร้างต้นแบบโปรแกรมเว็บประยุกต์ (Web Application Prototype) เพื่อเป็นเครื่องมือช่วยในการฝึกฝน โดยให้ผู้ใช้บันทึกคลิปวิดีโอ และนำไปวิเคราะห์หาเสียงยิ้มได้

60309803 : Major INFORMATION TECHNOLOGY
Keyword : Smile Voice, 2DCNN, Data Augmentation, Deep Learning

MR. Naris PROMBUT : Thai Smile Voice Classification Model using Convolution Neural Network Thesis advisor : Assistant Professor Nuttachot Promrit, Ph.D.

"Smile Voice" refers to a distinctive vocal quality that conveys the impression of a speaker smiling. This practice of infusing a smiling quality into one's voice is essential for call center professionals. As the initial point of contact, inquiry, and issue resolution for customers and service recipients, call center employees must adeptly manage their emotions and employ conversations that positively reflect the product and service image. Speaking with a smile entails specialized training, including voice modulation and facial expression control.

Pronunciation exercises often involve practicing in front of a mirror with guidance from a coach who provides feedback on tone and facial expressions. Additionally, during their work, call center staff should keep a mirror handy to self-monitor their interactions with customers. This research project introduces a model for analyzing Thai speech with a smiling quality, focusing on facial emotions, employing deep learning techniques.

The primary goal is to develop a tool that aids in practicing speaking with a smile. And this model is applied to call center voice data with Data Augmentation further improves the model's performance, with the 2D CNN MFCC model coupled with Augmentation achieving a 75.61% accuracy.

Additionally, a prototype web application is developed to be a tool to help in training by allowing users to record video clips and can be used to analyze the sound of smiles.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สามารถสำเร็จไปได้ด้วยดีนั้น ต้องขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. ณัฐโชติ พรหมฤทธิ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก และ อาจารย์ ดร. สัจจาภรณ์ ไวจรรยา อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม เป็นอย่างยิ่งสำหรับความเมตตาในสั่งสอนตั้งแต่เริ่มต้นเข้าศึกษาในระดับปริญญาเอก อีกทั้งคอยแนะนำให้คำปรึกษาถึงแนวทางการทำวิจัย และข้อแนะนำทั้งในด้านแนวคิดในกระบวนการทำวิจัยตลอดจนโอกาสและประสบการณ์ ที่ท่านอาจารย์มอบให้ซึ่งเป็นประโยชน์อย่างมากที่สุดสำหรับการทำวิทยานิพนธ์ และยังเป็นแนวทางในการดำรงชีวิตตลอดการศึกษาในระดับปริญญาเอก ที่มหาวิทยาลัยศิลปากร อีกทั้งขอขอบพระคุณประธานกรรมการและกรรมการสอบวิทยานิพนธ์ทุกท่าน ที่ได้กรุณาให้ข้อแนะนำเพิ่มเติมเพื่อให้วิทยานิพนธ์มีความสมบูรณ์มากยิ่งขึ้น

สุดท้ายนี้ ขอกราบขอบพระคุณ ครอบครัวพรหมบุตร และญาติพี่น้องทุกท่าน สำหรับความช่วยเหลือ ส่งเสริมสนับสนุนในทุกๆ ด้าน รวมถึงกำลังใจ และคำแนะนำที่ดีมาโดยตลอด ส่งผลทุกอย่างสำเร็จลุล่วงไปได้ด้วยดี

นริศร์ พรหมบุตร

สารบัญ

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

คอลเซ็นเตอร์ หรือ คอนแทคเซ็นเตอร์ ถือเป็นส่วนสำคัญที่ช่วยให้ลูกค้ามีความพึงพอใจในสินค้าหรือบริการขององค์กร เพราะถือเป็นหน่วยงานแรกเมื่อลูกค้ามีการโทรศัพท์ติดต่อเข้ามา ยิ่งในยุคที่ผู้บริโภคมีทางเลือกมากขึ้น การทำให้ผู้บริโภคมีความพึงพอใจในสินค้าและบริการย่อมส่งผลดีต่อความแข็งแกร่งของแบรน รวมถึงการกลับมาใช้งานอีกครั้ง หรือชวนคนอื่น ๆ มาใช้บริการ คอลเซ็นเตอร์จึงถือเป็นหน้าด่านในกรณีที่ลูกค้าต้องการสอบถามหรือติชมสินค้า ซึ่งน้ำเสียง และอารมณ์ของเจ้าหน้าที่ ถือเป็นส่วนสำคัญอย่างยิ่งในการทำงาน เพราะน้ำเสียงเป็นส่วนหนึ่งที่ช่วยสร้างความประทับใจ และความไว้วางใจให้กับลูกค้า ซึ่งโดยปกติน้ำเสียงและอารมณ์จะไปในแนวทางเดียวกัน นั่นคือถ้าอารมณ์ดี น้ำเสียงก็จะออกมาดี ซึ่งส่งผลทำให้ผู้ฟังรู้สึกได้ถึงอารมณ์ที่พร้อมจะบริการของพนักงานคนนั้น ซึ่งถ้าเจ้าหน้าที่คอลเซ็นเตอร์อารมณ์ดี และมีน้ำเสียงที่ยิ้มแย้มแจ่มใส จะสร้างความไว้วางใจ และพึงพอใจให้กับลูกค้าเพิ่มขึ้น เพราะเหตุนี้ที่โต๊ะของคอลเซ็นเตอร์ จะมีกระจกวางอยู่ที่ด้านหน้าตัวเอง เพื่อให้ได้มองเห็นใบหน้าตัวเองในขณะพูด เพื่อใช้ตรวจสอบสีหน้าตัวเองอยู่ตลอดเวลาว่าสีหน้าเป็นอย่างไรบ้าง เพราะในการทำงานสีหน้าของคอลเซ็นเตอร์ต้องร่าเริงแจ่มใส ซึ่งจะถ่ายทอดไปถึงน้ำเสียงที่ใช้ในการสื่อสารกับลูกค้า ซึ่งเราเรียกเสียงที่มาจากใบหน้าที่ยิ้มแย้มว่า เสียงที่มีรอยยิ้ม หรือ “Smile Voice”

การแสดงออกทางอารมณ์สามารถเชื่อมโยงไปถึงความน่าดึงดูดใจ และลักษณะในแง่บวกอื่นๆ ด้วย ในงานวิจัยของ (Torre, 2013) ได้พูดถึงอารมณ์และการแสดงออกของสีหน้านั้น มีความสัมพันธ์กับความน่าเชื่อถือ และความน่าร่วมมือด้วย ซึ่งการยิ้มเป็นการแสดงอารมณ์ที่เป็นในด้านบวก ซึ่งการยิ้มสามารถรับรู้ได้จากเสียง แม้แต่ไม่ได้เห็นผู้พูด หรือเราเรียกว่าเสียงยิ้ม “ Smiling Voice” ในงานวิจัยของ (Shor, 1978) เป็นการแสดงให้เห็นถึงความสัมพันธ์ของการยิ้มที่มีผลกับเสียงที่พูด นั่นคือการยิ้มจะมีผลต่อเสียงที่พูดออกมาโดยจะมีเสียงที่สั้นลง เมื่อเทียบกับเสียงพูดที่เหมือนกันแต่ไม่ได้ยิ้ม และในงานวิจัยของ (Fagel, 2010) ได้เปรียบเทียบความถี่ของเสียงที่เกิดการยิ้ม ซึ่งการยิ้มทำให้พื้นฐานของความถี่ (Fundamental Frequencies) และความถี่สั่นพ้อง (Formant) เพิ่มขึ้น ในหลายงานวิจัย (Drahota et al., 2008) (Emond & Laforest, 2013) ได้แสดงให้เห็นถึงเสียงยิ้มสามารถตรวจจับได้จากสัญญาณเสียง โดยเป็นการตรวจหาเสียงยิ้ม โดยใช้วิธีแบบ Manual นั่นคือใช้

วิธีการเปิดเสียงพูดที่บันทึกไว้ ซึ่งมีทั้งเสียงพูดที่เป็นเสียงยิ้ม และเสียงพูดที่ไม่เป็นเสียงยิ้ม จากนั้นเปิดให้อาสาสมัครฟัง และให้อาสาสมัครเป็นคนบอกว่าเสียงที่ได้ยินเป็นเสียงยิ้มหรือไม่

งานวิจัยนี้เสนอการประยุกต์ใช้เทคนิคการเรียนรู้เชิงลึกสร้างโมเดลวิเคราะห์เสียงพูดภาษาไทยที่พูดด้วยด้วยน้ำเสียงที่ยิ้ม นั่นคือน้ำเสียงที่ผู้ฟัง ฟังแล้วรู้สึกได้ว่าเรายิ้มให้เขาอยู่ ซึ่งสามารถนำไปประยุกต์ใช้ในการฝึกอบรมการใช้เสียงของคอลเซ็นเตอร์ ซึ่งสามารถช่วยวิเคราะห์เสียงยิ้มได้เองโดยไม่ต้องอาศัยผู้มีทักษะในการใช้เสียงมาช่วยฟัง

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาและพัฒนาเทคนิคการเรียนรู้เชิงลึก สำหรับการวิเคราะห์เสียงยิ้มในภาษาไทย
2. เพื่อเป็นแนวทางพัฒนาระบบวิเคราะห์เสียงยิ้มในภาษาไทย

1.3 ขอบเขตงานวิจัย

วิทยานิพนธ์นี้จะศึกษาและพัฒนาต้นแบบการวิเคราะห์เสียงยิ้ม โดยใช้ข้อมูลจากวิดีโอ ที่มีทั้งภาพและเสียง และหาความสัมพันธ์ระหว่างอารมณ์บนใบหน้ากับเสียงที่เป็นเสียงยิ้ม โดยขอบเขตของการวิจัยแบ่งเป็น ขอบเขตของกระบวนการในงานวิจัย ขอบเขตของเครื่องมือ ขอบเขตการทดลอง และขอบเขตการวัดและประเมินผล โดยอธิบายได้ดังนี้

1.3.1 ขอบเขตของกระบวนการในงานวิจัย

1) สร้างโมเดลการเรียนรู้เชิงลึก เพื่อวิเคราะห์อารมณ์จากเสียงพูด ได้แก่ อารมณ์ปกติ โกรธ ดีใจ เสียใจ ฉุนเฉียว โดยใช้คลังข้อมูล ภาษาอังกฤษ และภาษาไทย

2) สร้างโมเดลการเรียนรู้เชิงลึก เพื่อวิเคราะห์เสียงยิ้ม

3) สร้างโมเดลการเรียนรู้เชิงลึก เพื่อวิเคราะห์ใบหน้าที่สอดคล้องกับเสียงพูด โดยใช้คลังข้อมูล รูปภาพหรือภาพเคลื่อนไหว

4) สร้างเครื่อง มือสำหรับตรวจสอบหาเสียงยิ้ม โดยใช้ภาพและเสียงจากวิดีโอ

1.3.2 ขอบเขตของเครื่องมือ

1) วิทยานิพนธ์นี้นำเสนอโมเดลการวิเคราะห์เสียงยิ้ม โดยเริ่มต้นจากการทดสอบสร้างโมเดลเสียงยิ้มจากอาสาสมัคร และสร้างโมเดลเสียงยิ้มจากคนที่ทำงานด้านคอลเซ็นเตอร์

2) โมเดลที่สร้างขึ้นมา สามารถวิเคราะห์เสียงที่ได้ยิน ว่าเป็นเสียงของอารมณ์อะไร และสามารถบอกได้ว่าเป็นเสียงยิ้ม หรือเสียงไม่ยิ้มหรือไม่

1.3.3 ขอบเขตการทดลอง

1) สร้างโมเดลวิเคราะห์อารมณ์ของเสียง โดยสามารถวิเคราะห์ได้ทั้งหมด 5 อารมณ์ คือ ปกติ โกรธ ดีใจ เสียใจ ฉุนเฉียว โดยเปรียบเทียบเทคนิคการแยกเสียง โดย Mel-Frequency Cepstrum Coefficients (MFCC) และ Mel-Spectrogram

2) สร้างโมเดลวิเคราะห์เสียงยิ้ม โดยสามารถวิเคราะห์ เป็นเสียงยิ้ม (Smile Voice) หรือ เสียงไม่ยิ้ม (Non-Smile Voice) โดยเปรียบเทียบเทคนิคการแยกเสียง ทั้งแบบ MFCC และ Mel-Spectrogram

3) สร้างเครื่องมือที่วิเคราะห์ความสัมพันธ์กันของอารมณ์บนใบหน้า เปรียบเทียบกับเสียงยิ้ม ว่ามีความสัมพันธ์กันอย่างไร

1.3.4 ขอบเขตการวัดและประเมินผล

1) ประเมินผลลัพธ์ที่ได้จากโมเดลที่สร้างขึ้น โดยนำมาเปรียบเทียบความถูกต้องของเสียงที่เลเบลมาจากคอลเซ็นเตอร์ทั้งที่เป็นเสียงยิ้ม และเสียงไม่ยิ้ม

1.4 ขั้นตอนการดำเนินการวิจัย

1. ศึกษางานวิจัยที่เกี่ยวข้องเกี่ยวกับการวิเคราะห์ อารมณ์ของเสียง และเสียงที่มีรอยยิ้ม
2. เก็บข้อมูลเสียง จากกลุ่มคอลเซ็นเตอร์ โดยให้คอลเซ็นเตอร์พูดประโยค ที่เป็นเสียงยิ้ม และพูดโดยเป็นเสียงไม่ยิ้ม
3. เปรียบคุณลักษณะของการนำคลังข้อมูลมาใช้ในวิธีที่นำมาใช้โดยตรง และโดยการประยุกต์ใช้เทคนิคการเพิ่มจำนวนของคลังข้อมูล (Data Augmentation)
4. สร้างเว็บไซต์เพื่อให้ผู้เชี่ยวชาญเข้ามาตรวจสอบเสียงยิ้ม และเสียงไม่ยิ้มที่เก็บมาจากคอลเซ็นเตอร์
5. วัดผลโมเดลเสียงยิ้ม และปรับพารามิเตอร์ต่าง ๆ เพื่อพัฒนาปรับปรุงให้ผลดีขึ้น
6. ออกแบบและพัฒนาเครื่องมือที่แสดงความสัมพันธ์ของเสียงยิ้ม และอารมณ์บนใบหน้า

1.5 ประโยชน์ที่คาดว่าจะได้รับ

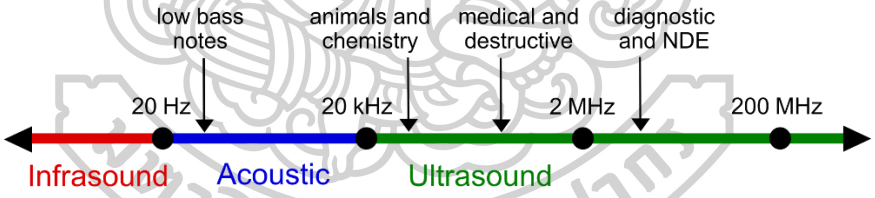
1. โมเดลที่สร้างขึ้นสามารถวิเคราะห์เสียงที่มีรอยยิ้ม และเสียงที่ไม่มีรอยยิ้มจากเสียงภาษาไทย
2. สามารถใช้วิเคราะห์เสียงที่ไม่ใช่เสียงที่เกิดจากรอยยิ้ม เพื่อสามารถช่วยเป็นแนวทางปรับปรุงการใช้เสียง
3. สามารถใช้ฝึกการออกเสียงของคอลเซ็นเตอร์ที่ต้องใช้เสียงที่มีรอยยิ้มสื่อสารกับลูกค้า

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการวิเคราะห์เสียงยิ้มโดยใช้เทคนิคการเรียนรู้เชิงลึกมีรายละเอียดดังต่อไปนี้

2.1 เสียง (Sound)

ในทางฟิสิกส์ เสียง คือการสั่นสะเทือนที่แพร่กระจายเป็นคลื่นเสียง โดยต้องอาศัยตัวกลางในการส่งผ่าน เช่น ก๊าซ ของเหลว หรือของแข็ง ส่วนในทางด้านสรีรวิทยา และจิตวิทยาของมนุษย์ เสียงคือการรับสัญญาณของคลื่นดังกล่าว โดยสมองทำหน้าที่รับรู้คลื่นดังกล่าว (S. Wikipedia, 2023) และโดยส่วนใหญ่ของมนุษย์จะรับรู้ หรือได้ยินเฉพาะคลื่นอะคูสติกที่มีความถี่อยู่ประมาณ 20 เฮิรตซ์ ถึง 20 กิโลเฮิรตซ์ ซึ่งเป็นช่วงความถี่เสียงมนุษย์สามารถรับรู้ได้โดยการได้ยิน ในส่วนของคลื่นเสียงที่สูงกว่า 20 kHz เรียกว่าอัลตราซาวนด์ ซึ่งมนุษย์ไม่ได้ยินได้ และคลื่นเสียงที่ต่ำกว่า 20 Hz เราจะเรียกความถี่ในย่านนี้ว่าอินฟราซาวด์ โดยในสัตว์ชนิดต่างๆ จะมีช่วงความถี่ในการได้ยินที่แตกต่างกัน จากภาพที่ 2.1 แสดงให้เห็นถึงช่วงของความถี่ที่แตกต่างกันในแต่ละช่วง



ภาพที่ 2.1 ความถี่ใน อินฟราซาวด์ อะคูสติก และอัลตราซาวนด์
(S. Wikipedia, 2023)

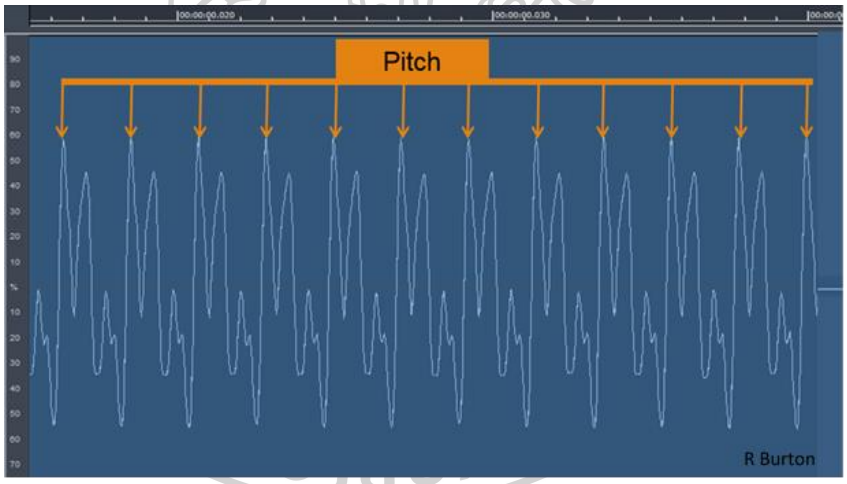
หรือในอีกทางหนึ่ง เสียง คือการคลื่อนที่ของอากาศ หรือตัวกลางที่ยืดหยุ่นได้ และเสียงยังเป็นตัวกระตุ้นกลไกการได้ยิน และการรับรู้ ทำให้ส่งผลไปถึงความรู้สึกด้วย

คลื่นเสียง (Sound Wave) เสียงสามารถเคลื่อนที่ผ่านตัวกลางได้หลายรูปแบบ ไม่ว่าจะเป็นวัตถุของแข็ง ของเหลว หรือก๊าซ เมื่อวัตถุเกิดการเคลื่อนที่หรือถูกกระทำด้วยแรงจากภายนอก จนก่อให้เกิดการสั่นสะเทือนของโมเลกุลภายในวัตถุนั้น ซึ่งส่งผลไปยังอนุภาคของอากาศหรือตัวกลางที่อยู่บริเวณโดยรอบ ก่อให้เกิดการรบกวนหรือการถ่ายโอนพลังงาน ผ่านการสั่นและการกระทบกัน ทำ

ให้อนุภาคของอากาศมีการเคลื่อนที่มากระทบกันจนเกิด การบีบอัด (Compression) และเมื่ออนุภาคของอากาศพยายามเคลื่อนที่กลับตำแหน่งเดิมจะเกิด การยืดขยาย (Rarefaction) ดังนั้น คลื่นเสียงจึงเรียกว่าคลื่นความดัน (Pressure wave) เพราะอาศัยการผลักดันกันของโมเลกุลในตัวกลางในการเคลื่อนที่

คลื่นเสียงนั้น มีคุณสมบัติเช่นเดียวกับคลื่นอื่นๆ เช่น แอมพลิจูด (Amplitude) ความเร็ว (Velocity) หรือ ความถี่ (Frequency) ซึ่งลักษณะเฉพาะของเสียง ในแต่ละเสียงจะต่างกัน ไม่ว่าจะเป็นเสียงสูง-เสียงต่ำ, เสียงดัง-เสียงเบา รวมถึงคุณภาพของเสียง และแหล่งกำเนิดของเสียงด้วย

ความถี่ หรือ Pitch เราจะรับรู้เป็น เสียงสูง เสียงต่ำ โดยสิ่งที่ทำให้เสียงแต่ละเสียงสูงต่ำ ไม่เท่ากัน ขึ้นอยู่ที่ความเร็วในการสั่นสะเทือนของวัตถุ วัตถุที่สั่นเร็วเสียงจะมีความถี่สูง กว่าวัตถุที่สั่นช้า สำหรับเสียงธรรมดา ระดับเสียงจะสัมพันธ์กับความถี่ของการสั่นที่ช้าที่สุดในเสียง (ฮาร์มอนิกพื้นฐาน) ในกรณีของเสียงมีความซับซ้อนมากขึ้น การรับรู้ของระดับเสียงอาจแตกต่างกันไป ดังภาพที่ 2.2

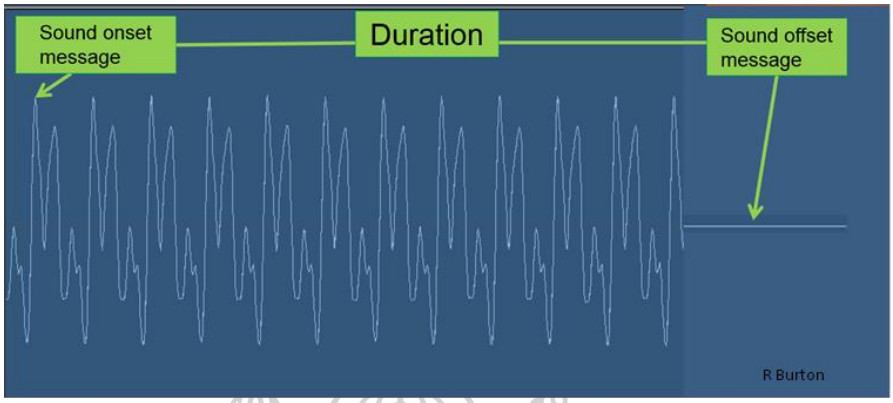


ภาพที่ 2.2 ความถี่ของคลื่นเสียง (Pitch)

(S. Wikipedia, 2023)

ความยาวของเสียง (Duration) เป็นตัวบอกว่าเป็นเสียงที่ “สั้น” หรือ “ยาว” ซึ่งมีความสัมพันธ์เริ่มตั้งแต่ สัญญาณเริ่มต้น และออฟเซ็ตที่เกิดจากการตอบสนองของเส้นประสาทต่อเสียง ซึ่งระยะเวลาของเสียงจะเริ่มตั้งแต่เวลาที่รับรู้ถึงเสียงในครั้งแรกจนกระทั่งเสียงนั้นมีการเปลี่ยนแปลงหรือหยุดไป แต่ในบางครั้งสิ่งนี้ อาจจะไม่ได้เกี่ยวข้องโดยตรงกับระยะเวลาของเสียงจริงๆ

เช่น ในสภาพแวดล้อมที่มีเสียงดัง หรือการหายไปของเสียงเนื่องจากถูกรบกวน ซึ่งในเสียงจริงๆ นั้นเป็นเสียงที่ต่อเนื่องกัน ดังภาพที่ 2.3



ภาพที่ 2.3 ความยาวของคลื่นเสียง (Duration)

(S. Wikipedia, 2023)

2.2 การยิ้ม (Smiling)

การยิ้มเกิดจากการกระตุ้นการทำงานของ zygomaticus major ซึ่งเป็นกล้ามเนื้อใบหน้าที่เชื่อมต่อออกมาจากกระดูกแก้มส่วนล่างของใบหน้าและเชื่อมต่อกับมุมปาก การกระตุ้นกล้ามเนื้อชุดนี้จะดึงมุมริมฝีปากเข้าหากระดูกแก้มในทิศทางเฉียงขึ้น (Ekman & Friesen, 1978) และส่งผลให้เกิดหน้าตาที่เรามักเรียกกันว่า "รอยยิ้ม"

ในปี 1982 Ekman และ Frisen ได้จัดกลุ่มของรอยยิ้มประเภทต่างๆ (Ekman & Friesen, 1982) โดยอ้างอิงบนพื้นฐานของ Facial Action Coding Systems ซึ่งเป็นระบบของการวัด ที่เขาทำขึ้นมาในงานวิจัยก่อนหน้านี้ (Ekman & Friesen, 1978) ซึ่งจุดประสงค์หลัก คือต้องการ แยกแยะและแสดงให้เห็นความแตกต่างระหว่างรอยยิ้มที่เกิดจาก "ความรู้สึก (felt)" และ "รอยยิ้มที่แกล้งทำ (false)" รอยยิ้มที่เกิดจากความรู้สึก หรือเรียกอีกอย่างหนึ่งว่า Duchenne Smile ซึ่งเชื่อกันว่าเป็นการแสดงออกถึงอารมณ์เชิงบวกโดยธรรมชาติ และมีลักษณะเฉพาะคือการทำงานของกล้ามเนื้อใบหน้า 2 มัดประกอบกัน คือ Zygomaticus major และ Orbicularis oculi (pars lateralis และ pars medialis)

ซึ่งในทางตรงข้ามกับรอยยิ้มที่เกิดจากความรู้สึก คือรอยยิ้มที่แกล้งทำ จะเห็นว่าไม่มีความสัมพันธ์กับส่วนของดวงตา และถือเป็นความพยายามโดยเจตนาที่จะสื่อสารความรู้สึกเชิงบวกทั้งๆ ที่จริงๆ แล้วไม่ได้รู้สึกเลย (Ekman & Friesen, 1982)

การยิ้มเป็นสิ่งที่คลุมเครืออย่างมาก และโดยทั่วไปจะสังเกตได้ในบริบทต่างๆ มากมาย ไม่ว่าจะเป็น สัญญาณของความเคารพ ความสบายใจ และความเป็นมิตร (van Hooff, 1972) และในงานวิจัยหลายๆ งาน มีแนวโน้มว่าความแตกต่างระหว่างรอยยิ้มประเภทต่างๆ อาจช่วยให้เราเข้าใจการทำงานของรอยยิ้มที่เกี่ยวกับในความสัมพันธ์ทางสังคมได้ เช่น เกิดการผ่อนปรนจากผู้ตัดสิน (Forgas, 1987) (LaFrance & Hecht, 1995) หรือ ผลประโยชน์ที่มีค่าประเภทต่างๆ (Brown & Moore, 2002) ซึ่งแนวความคิดนี้ได้รับการสนับสนุนในหลายงานวิจัย เช่น รอยยิ้มของ Duchenne Smile จะถูกแสดงในอัตราที่สูงกว่าโดยผู้คนที่มีการแบ่งปันกัน ไม่ว่าจะเป็น สิ่งของ หรือทรัพยากรต่างๆ และจะมี Duchenne Smile น้อยลง ในกลุ่มของผู้คนเมื่ออยู่ในสภาวะที่ถูกควบคุม (Mehu et al., 2007) นอกจากนี้ รอยยิ้มของ Duchenne Smile ยังมีความสัมพันธ์เชิงบวกกับการแสดงความมีน้ำใจต่อคู่รัก หรือในมุมของผู้รับ รอยยิ้มของ Duchenne Smile ยังได้ช่วยทำให้ ผู้รับมีความรู้สึกถึงความมีน้ำใจและความเป็นกันเองที่เกิดขึ้นบนใบหน้าของคนแปลกหน้าอย่างมาก (Mehu et al., 2007)

2.3 คอลเซ็นเตอร์

คอลเซ็นเตอร์ (Call Center) หรือศูนย์บริการทางโทรศัพท์คือทีมผู้เชี่ยวชาญด้านการบริการลูกค้าที่ช่วยรับสายโทรศัพท์จากลูกค้าที่มีคำถามเกี่ยวกับบริการหรือผลิตภัณฑ์ของบริษัท (Zendesk, 2023) ทีมงานของคอลเซ็นเตอร์ในหลายแห่ง มุ่งเน้นที่ความพึงพอใจของลูกค้า และให้การสนับสนุนที่ครอบคลุมในหลากหลายบริการ บางรายของ คอลเซ็นเตอร์ อาจมีการตั้งเป้าหมาย เพื่อเพิ่มการสร้างความสนใจในตัวสินค้า หาลูกค้าใหม่ หรือเป็น One Stop Service คือเริ่มตั้งแต่การสั่งซื้อไปจนถึงการรับชำระเงิน

เพราะฉะนั้นคอลเซ็นเตอร์จึงมีบทบาทสำคัญในการมุ่งสร้างประสบการณ์ที่ดีให้กับลูกค้า และเพื่อรักษาความสัมพันธ์นี้ พวกเขาจึงต้องรักษาระดับการบริการไว้ในระดับสูงตลอดเวลา ซึ่งหมายความว่าตัวแทนของคอลเซ็นเตอร์จะต้องมีความรู้ อดทน และให้ความช่วยเหลือในการโต้ตอบกับลูกค้า และสิ่งที่สำคัญของคอลเซ็นเตอร์คือข้อมูลความถูกต้องที่ตอบให้ลูกค้า น้ำเสียงที่พูดกับลูกค้า ซึ่งต้องเป็นน้ำเสียงที่ยิ้มอยู่เสมอ โดยปกติที่โต๊ะทำงานของคอลเซ็นเตอร์จะมีกระจกให้ดูว่าตอนกำลังพูด ว่ากำลังยิ้มอยู่หรือไม่ ซึ่งน้ำเสียงที่ยิ้มนี้ช่วยเพิ่มความพึงพอใจให้กับลูกค้าในทางอ้อมด้วย ซึ่งในงานวิจัยนี้เป็นการใช้เสียงและหน้าของผู้พูดมาตรวจสอบหา น้ำเสียงที่ยิ้ม (The Voice With a Smile) และ คอลเซ็นเตอร์ ไม่เหมือนกับ ศูนย์ติดต่อ (Contact Center) ศูนย์ติดต่อมีความแตกต่างตรงที่ มีการจัดการการสื่อสารกับลูกค้าผ่านช่องทางต่างๆ รวมถึงอีเมล แชท แอพส่งข้อความ หรือทางโซเชียลมีเดีย

2.4 การรู้จำอารมณ์บนใบหน้า (Facial Expression Recognition)

การจดจำการแสดงออกทางสีหน้า เป็นการจำแนกการแสดงออกบนใบหน้า ซึ่งสามารถแบ่งออกเป็นประเภทต่างๆ เช่น ความโกรธ ความกลัว ความประหลาดใจ ความเศร้า ความสุข และอื่น ๆ ผ่านซอฟต์แวร์ ซึ่งการจดจำการแสดงออกทางสีหน้าเป็นเทคโนโลยีที่ใช้เครื่องหมายไบโอเมทริกซ์ (Biometrics) คือลักษณะของมนุษย์ที่สร้างเอกลักษณ์ของแต่ละบุคคล เช่น ลักษณะบนใบหน้า ดวงตา เป็นต้น เพื่อตรวจจับอารมณ์บนใบหน้ามนุษย์ แม่นยำยิ่งขึ้น เทคโนโลยีนี้เป็นเครื่องมือวิเคราะห์ความรู้สึก และสามารถตรวจจับการแสดงออกทั้ง 6 ขั้นพื้นฐานได้โดยอัตโนมัติ ได้แก่ ความสุข ความเศร้า ความโกรธ ความประหลาดใจ ความกลัว และความขยะแขยง หรือสามารถพัฒนาเป็นอารมณ์ต่างๆ ตามรูปแบบหรือโดเมน ของการศึกษาวิจัยที่ต่างกันออกไปตามบริบท

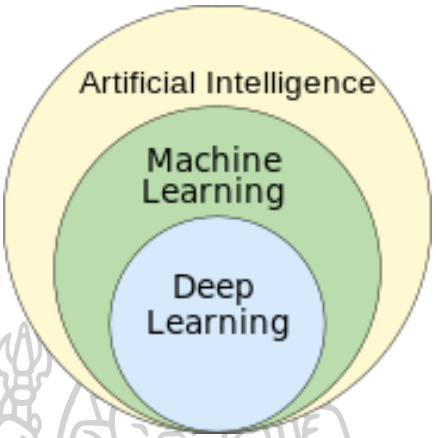
การทำงานของการรู้จำอารมณ์บนใบหน้านั้น จะทำงานบนระบบจดจำการแสดงออกทางสีหน้าผ่านเทคโนโลยีคอมพิวเตอร์ (Huang et al., 2023) ดังนั้นจึงใช้อัลกอริทึมสำหรับตรวจจับใบหน้า การเขียนคำสั่งการแสดงออกทางสีหน้า และรับรู้สภาวะทางอารมณ์แบบเรียลไทม์ โดยการวิเคราะห์ใบหน้าจากรูปภาพหรือวิดีโอ ผ่านกล้องที่ติดตั้งบนเครื่องคอมพิวเตอร์ โทรศัพท์มือถือ (Suk & Prabhakaran, 2014) และระบบ digital การวิเคราะห์ใบหน้าผ่านกล้องที่ใช้คอมพิวเตอร์โดยทั่วไปมี 3 ขั้นตอน ดังนี้

1) การตรวจจับใบหน้า คือการระบุตำแหน่งใบหน้าจากภาพหรือวิดีโอ เพื่อค้นหาใบหน้าและนำไปสู่ขั้นตอนต่อไป
2) การตรวจจับจุดสังเกตบนใบหน้า คือ ดึงข้อมูลเกี่ยวกับลักษณะใบหน้าที่ตรวจพบ ตัวอย่างเช่น การตรวจจับรูปร่างของส่วนประกอบบนใบหน้า หรือการอธิบายพื้นผิวของผิวหนังในบริเวณใบหน้า
3) การจำแนกสีหน้าและอารมณ์ คือ การวิเคราะห์การเคลื่อนไหวของลักษณะใบหน้า หรือการเปลี่ยนแปลงลักษณะที่ปรากฏของใบหน้า จากนั้นจำแนกข้อมูลนี้เป็นหมวดหมู่ที่สื่อถึงการแสดงออก เช่น การกระตุ้นกล้ามเนื้อใบหน้า รอยยิ้มหรือขมวดคิ้ว หมวดหมู่อารมณ์ความสุขหรือความโกรธ เป็นต้น

2.5 การเรียนรู้เชิงลึก (Deep Learning)

การเรียนรู้เชิงลึกเป็นส่วนหนึ่งของปัญญาประดิษฐ์ (AI : Artificial Intelligence) โดยเป็นการประมวลผลข้อมูลคล้ายกับสมองของมนุษย์ และสร้างรูปแบบเพื่อใช้ในการตัดสินใจ การเรียนรู้เชิงลึกยังเป็นส่วนหนึ่งของการเรียนรู้ของเครื่อง (ML : Machine Learning) ซึ่งเป็นส่วนหนึ่งของ AI เหมือนกัน โดยเป็นวิธีการที่เครื่องคอมพิวเตอร์สร้างโครงข่ายประสาทเทียมขึ้นมา โดยอาจจะมีหลาย

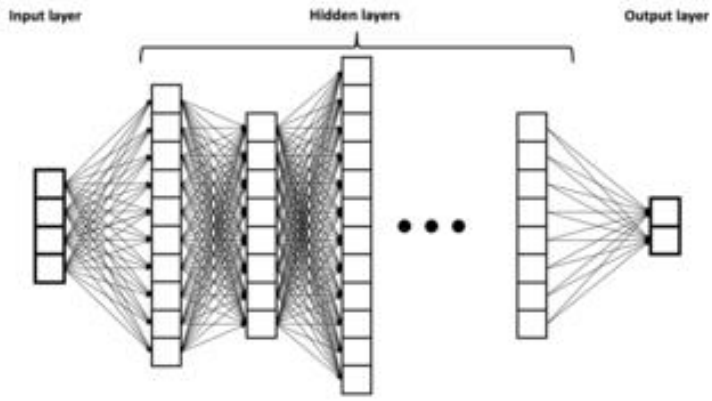
ชั้นต่อๆ กัน เปรียบเสมือนการทำงานของสมองมนุษย์ที่มีเครือข่ายสมองที่ซับซ้อน ซึ่งเรียกว่า Deep Learning หรือ Deep Neural Network (Bonner, 2019) โดยเป็น Algorithm ที่ได้รับการสร้างขึ้นมาจากการฝึกฝนโมเดล ชุดของข้อมูล ดังภาพที่ 2.4



ภาพที่ 2.4 การเรียนรู้เชิงลึก (Deep Learning)

(Wikipedia, 2023)

จากภาพที่ 2.4 จะเห็นได้ว่า Deep Learning เป็นส่วนหนึ่งของ Machine Learning (Team, 2023) ซึ่งได้มีการทำงานคล้ายกับโครงสร้างของสมองมนุษย์ โดยอัลกอริทึมการเรียนรู้เชิงลึกจะพยายามสร้างข้อสรุปจากข้อมูลที่มี ซึ่งเป็นการทำงานคล้ายกับการทำงานของสมองมนุษย์ โดยจะทำโดยการวิเคราะห์ข้อมูลอย่างต่อเนื่องด้วยโครงสร้างเชิงตรรกะที่ได้กำหนดไว้ เพื่อให้ได้ผลลัพธ์จากข้อมูลที่ได้เรียนรู้ไป โดยใช้โครงสร้างของอัลกอริทึมจากไม่กี่ชั้น ไปจนถึงโครงสร้างที่มีความซับซ้อนหลายๆ ชั้นมาต่อกัน ซึ่งที่เรียกว่าโครงข่ายประสาทเทียม ดังภาพที่ 2.5



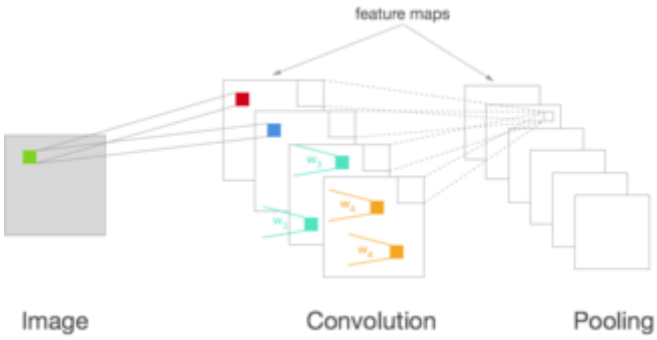
ภาพที่ 2.5 Deep Learning

(BrunelloN, 2021)

2.6 โครงข่ายประสาทแบบคอนโวลูชั่น (Convolution Neural Network: CNN)

โครงข่ายประสาท Convolutional หรือที่เรียกว่า ConvNets ได้รับการแนะนำครั้งแรกในปี 1980 (Yalçın, 2021) นักวิจัยด้านวิทยาการคอมพิวเตอร์ได้สร้างต่อยอดมาจากจากผลงานของนักวิทยาศาสตร์ชาวญี่ปุ่นซึ่งก่อนหน้านี้ได้ประดิษฐ์ neocognitron (Fukushima, 2014) ซึ่งเป็นเครือข่ายประสาทรับรู้ภาพขั้นพื้นฐาน CNN สามารถจดจำตัวเลขที่เขียนด้วยลายมือได้ CNN นำไปประยุกต์ใช้งานด้านการบริการธนาคารและไปรษณีย์ในส่วนของการอ่านรหัสไปรษณีย์บนซองจดหมายและตัวเลขบนเช็ค ถึงแม้จะมีความฉลาด แต่ ConvNets แต่ยังไม่สามารถปรับขนาดได้ CNN ต้องการข้อมูลจำนวนมากและทรัพยากรในการประมวลผลเพื่อให้ทำงานได้อย่างมีประสิทธิภาพสำหรับรูปภาพขนาดใหญ่ ในขณะนั้น เทคนิคนี้ใช้ได้กับภาพที่มีความละเอียดต่ำเท่านั้น จนกระทั่งในปี 2012 AlexNet (Alom et al., 2018) แสดงให้เห็นว่า AI ที่ใช้โครงข่ายประสาทเทียมแบบหลายชั้น ความพร้อมใช้งานของชุดข้อมูลขนาดใหญ่ เช่น ชุดข้อมูล ImageNet (Deng et al., 2009) ที่มีรูปภาพหลายล้านภาพติดป้ายกำกับ และทรัพยากรในการประมวลผลจำนวนมากช่วยให้นักวิจัยสร้าง CNN ที่ซับซ้อนซึ่งสามารถทำงานด้านการมองเห็นด้วยคอมพิวเตอร์ที่ไม่เคยทำได้มาก่อน

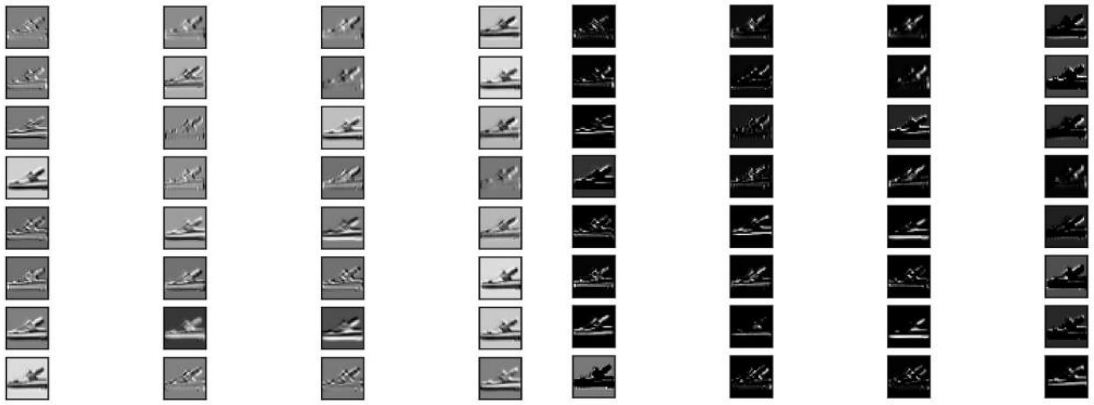
โครงข่ายประสาท Convolutional ประกอบด้วยเซลล์ประสาทเทียมหลายชั้น (Dickson, 2020) เซลล์ประสาทเทียมซึ่งเป็นการเลียนแบบเซลล์ทางชีววิทยาพื้นฐานเป็นฟังก์ชันทางคณิตศาสตร์ที่คำนวณผลรวมถ่วงน้ำหนักของอินพุตหลายตัวและส่งออกค่าทางสถิติ ดังภาพที่ 2.6



ภาพที่ 2.6 โครงข่ายประสาท Convolutional

(Rath, 2019)

พฤติกรรมของเซลล์ประสาทแต่ละเซลล์ถูกกำหนดน้ำหนัก ผ่านการป้อนด้วยค่าพิกเซล เซลล์ประสาทเทียมของ CNN จะเลือกลักษณะภาพต่างๆ และป้อนรูปภาพลงใน Convolutional Network แต่ละเลเยอร์จะสร้างแผนที่เปิดใช้งานหลายแผนที่ แผนที่การเปิดใช้งานจะเน้นคุณสมบัติที่เกี่ยวข้องของรูปภาพ เซลล์ประสาทแต่ละเซลล์ใช้แพทช์ของพิกเซลเป็น Input นำมาคำนวนร่วมกับค่าสีด้วยน้ำหนักแล้วนำผลมารวมเข้าด้วยกัน และเรียกใช้ผ่านฟังก์ชันการเปิดใช้งาน โดยมักจะตรวจจับคุณสมบัติพื้นฐาน เช่น ขอบแนวนอน แนวตั้ง และแนวทแยง ดังนั้นเอาต์พุตของเลเยอร์แรกจะถูกป้อนเป็นอินพุตของเลเยอร์ถัดไปซึ่งจะดึงคุณสมบัติที่ซับซ้อนมากขึ้น เมื่อโครงข่ายประสาทเทียมเลเยอร์ต่างๆ จะเริ่มตรวจจับคุณสมบัติแล้วนั้น จะดำเนินการคูณค่าพิกเซลด้วยน้ำหนักและรวมเข้าด้วยกันเรียกว่า "การแปลง" ด้วยเหตุนี้จึงเรียกว่าโครงข่ายประสาทเทียมแบบ Convolutional ซึ่งจะประกอบด้วยเลเยอร์การแปลงหลายชั้น และมีส่วนประกอบอื่น ๆ ด้วย เลเยอร์สุดท้ายของ CNN คือการจำแนกเลเยอร์ ดังภาพที่ 2.7 เป็นการแสดงผลลัพทธ์ที่ได้จากเลเยอร์ต่างๆ



ภาพที่ 2.7 การจำแนกเลเยอร์

(Promrit, 2020)

ดังนั้นการออกแบบโครงข่ายประสาทเทียมเป็นการเลียนแบบโครงสร้างของสมองมนุษย์ เช่นเดียวกับการใช้สมองในการจำแนกประเภทของข้อมูลต่างๆ นอกจากนี้ยังสามารถแสดงผลในแต่ละชั้นของเครือข่ายประสาทเทียมเพื่อเป็นตัวกรองประเภทหนึ่งที่ทำงานตั้งแต่ขั้นต้นไปจนถึงขั้นที่อยู่ในระดับที่ต่ำที่สุด เพื่อเพิ่มโอกาสในการตรวจจับและให้ผลลัพธ์ที่ถูกต้อง นอกจากนั้นสมองของมนุษย์ทำงานในทำนองเดียวกัน เมื่อใดก็ตามที่ได้รับข้อมูลใหม่สมองจะพยายามเปรียบเทียบกับวัตถุที่รู้จัก แนวคิดเดียวกันนี้ยังใช้กับโครงข่ายประสาทเทียมแบบลึก (Artificial neural networks: ANN)

โครงข่ายประสาทเทียมสามารถทำงานได้หลายอย่าง เช่นการทำคลัสเตอร์การจัดประเภทหรือการถดถอย ด้วยโครงข่ายประสาทเทียม สามารถจัดกลุ่มหรือจัดเรียงข้อมูลที่ไม่มีป้ายกำกับ (Label) ตามความคล้ายคลึงกันระหว่างกลุ่มตัวอย่างในข้อมูลนี้ หรือในกรณีของการจำแนกประเภทยังสามารถฝึกฝนโมเดล ชุดข้อมูลที่มีป้ายกำกับ (Label) เพื่อจัดประเภทตัวอย่างในชุดข้อมูลนี้เป็นหมวดหมู่ต่างๆ โดยทั่วไป โครงข่ายประสาทเทียมสามารถทำงานได้เหมือนกับอัลกอริทึมคลาสสิกของการเรียนรู้ของเครื่อง (Machine learning)

2.7 การตรวจจับวัตถุ TensorFlow (TensorFlow Object Detection)

Tensorflow เป็น Open Source Machine Learning Framework ที่ถูกพัฒนามาโดยทีมพัฒนาของ Google (Tensorflow.org, 2015) สำหรับการเขียนโปรแกรมเกี่ยวกับการจัดการข้อมูลในงานต่างๆ และรองรับการคำนวนของ Node ในกราฟทางคณิตศาสตร์ของข้อมูลหลายมิติที่มีความ

เชื่อมโยงกัน รวมถึง Tensorflow ยังเป็นเทคนิคทางคอมพิวเตอร์ สำหรับช่วยในการตรวจจับ ระบุตำแหน่ง และติดตามวัตถุจากภาพนิ่ง หรือวิดีโอ ทำให้เราได้เข้าใจรายละเอียดของข้อมูล เช่นรูปภาพได้อย่างลึกซึ้งมากยิ่งขึ้น โดยวิธีการทำงานของการตรวจจับวัตถุโดยใช้ Tensorflow เป็นเทคนิคการทำให้คอมพิวเตอร์มองเห็น ซึ่งการมองเห็นนี้ทำได้โดยใช้การทำงานด้วยระบบซอฟต์แวร์คอมพิวเตอร์ทำให้สามารถตรวจจับค้นหาและติดตามวัตถุจากภาพนิ่งหรือวิดีโอที่กำหนด หรือการตรวจจับวัตถุที่ระบุอยู่ใน Class ของวัตถุ เช่น บุคคล โต๊ะ หรือเก้าอี้ เป็นต้น พร้อมกับสามารถกำหนดพิกัดเฉพาะตำแหน่งบนภาพ หรือระบุตำแหน่งโดยการวาดกรอบล้อมรอบวัตถุ โดยกล่องขอบเขตอาจระบุตำแหน่งของวัตถุได้อย่างแม่นยำหรือไม่ ขึ้นอยู่กับความสามารถในการค้นหาวัตถุภายในภาพ ซึ่งขึ้นอยู่กับประสิทธิภาพของอัลกอริทึมที่ใช้ในการตรวจจับ โดยการตรวจจับบนใบหน้าเป็นหนึ่งในตัวอย่างการตรวจจับวัตถุ และถือเป็นส่วนเริ่มต้นของการตรวจจับอารมณ์บนใบหน้าอีกด้วย

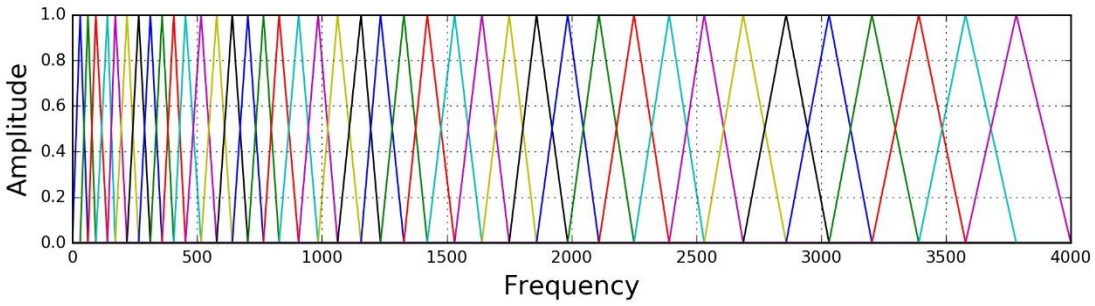
ปัจจุบัน TensorFlow เป็นซอฟต์แวร์ไลบรารีที่ได้รับความนิยมมากที่สุด มีแอปพลิเคชันการเรียนรู้เชิงลึก (Deep Learning) นำมาประยุกต์ใช้งานเป็นจำนวนมาก เนื่องจาก TensorFlow เป็นไลบรารีโอเพนซอร์สสำหรับ deep learning และ machine learning จึงถูกใช้ได้กับแอปพลิเคชันในรูปแบบของแบบข้อความ การจดจำภาพ การค้นหาด้วยเสียง และอื่นๆ ตัวอย่างเช่น การใช้งาน DeepFace (Serengil & Ozpinar, 2020) ระบบจดจำรูปภาพของ Facebook ใช้ TensorFlow สำหรับการจดจำรูปภาพ หรือ Siri ของ Apple ใช้สำหรับการจดจำเสียง หรือแอป Google ในหลายๆ แอปประยุกต์ใช้ความสามารถ TensorFlow เป็นต้น

2.8 Mel-Frequency Cepstrum Coefficients (MFCC)

สัมประสิทธิ์เซปสตรัลบนสเกลเมล (Abdul & Al-Talabani, 2022) ค่าสัมประสิทธิ์เซปสตรัมเป็นค่าลักษณะสำคัญที่นิยมมาก ทั้งในระบบรู้จำผู้พูดและเสียงพูด โดยพื้นฐานแล้วเซปสตรัม (Cepstrum) สามารถคำนวณได้จาก การแปลงโคซายน์แบบไม่ต่อเนื่อง (Discrete Cosine Transform) ของค่าลอการิธึมจากสเปกตรัมในช่วงสั้นๆ ซึ่งสัมประสิทธิ์เซปสตรัลบนความถี่เมลเป็นเทคนิคที่ปรับปรุงมาจากเซปตรัม คือการผ่านสเปกตรัมของสัญญาณเสียงเข้าไปในกลุ่มของตัวกรอง (Mel-Frequency Filter Bank) ซึ่งกระจายอยู่บนสเกลความถี่ที่ไม่สม่ำเสมอ เช่น การกระจายตามสเกลเมล ซึ่งออกแบบมาให้เหมาะสมกับการรับฟังของหู ซึ่งค่าพลังงานของสเปกตรัมของเสียงที่ได้จากตัวกรองแต่ละตัวจะถูกนำมาใช้คำนวณค่าสัมประสิทธิ์เซปตรัม แทนค่าสเปกตรัมปกติ ค่าสัมประสิทธิ์เซปตรัมที่ได้จากการกระทำเช่นนี้จึงเรียกว่า สัมประสิทธิ์เซปสตรัลบนสเกลเมล

โดยสเปกตรัมของสัญญาณเสียงสามารถหาได้โดยการแปลงฟูเรียร์แบบไม่ต่อเนื่อง หรือการแปลงฟูเรียร์แบบรวดเร็ว โดยขั้นตอนของการแปลงสัญญาณดังกล่าวอยู่บนแนวความคิดที่ว่าสเปกตรัมของสัญญาณเสียงกำเนิดจาก 2 ส่วนคือ เอนโวลอปของสเปกตรัม และโครงสร้างรายละเอียดของสเปกตรัม ทั้ง 2 ส่วนสามารถแยกกันได้โดยการใส่ อัลกอริทึ่ม สัมประสิทธ์เซบตรัมเป็นการแทนที่ของสัญญาณในส่วนเอนเวลอปสเปกตรัมเท่านั้น

Mel-frequency filter bank เป็น ขั้นตอนการหาค่าสัมประสิทธิ์ เซปสตรัมบนสเกลเมล เริ่มต้นจากการนำสัญญาณเสียงมาผ่านการประมวลผลสัญญาณเสียง จากนั้นส่งสัญญาณไปยังตัวกรองฟิลเตอร์แบงค์ (Filter Bank) เพื่อเน้นความสำคัญของความถี่ที่อยู่ในช่วงกลางของชุดตัวกรองแต่ละตัวกรอง ชุดตัวกรองฟิลเตอร์แบงค์ มีลักษณะดังภาพที่ 2.8



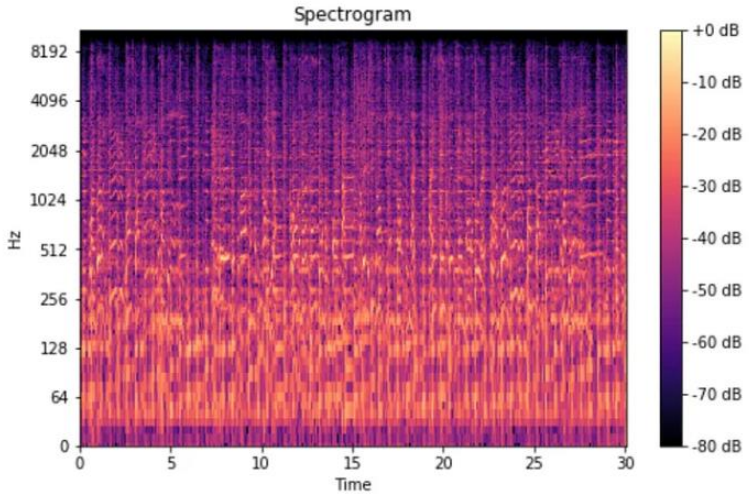
ภาพที่ 2.8 Filter Bank บน Mel-Scale

(Fayek, 2016)

2.9 Mel-Spectrogram

Spectrogram คือการแสดงรูปแบบของสเปกตรัมของสัญญาณความถี่ออกมาเป็นภาพ โดยแปรผันตามเวลา (Roberts, 2020) และเมื่อมีการนำไปใช้กับสัญญาณเสียง บางครั้งจะถูกเรียกว่า sonographs voiceprints หรือ voicegrams และเมื่อข้อมูลถูกนำเสนอในรูปแบบ 3 มิติ บางครั้งอาจจะเรียกว่า การแสดงผลแบบ Waterfall โดย Spectrogram ถูกนำมาใช้อย่างกว้างขวางใน เรื่องเกี่ยวกับดนตรี ภาษาศาสตร์ โซนาร์ เรดาร์ การประมวลผลคำพูด ตรวจจับแผ่นดินไหว และอื่นๆ Spectrogram ของเสียงสามารถใช้เพื่อระบุคำพูดตามสัทศาสตร์ และวิเคราะห์เสียงต่างๆ ของสัตว์ได้ด้วย Spectrogram สามารถสร้างขึ้นได้โดย Optical Spectrometer ร่วมกับกลุ่มตัวกรอง band-pass โดยการแปลงฟูเรียร์หรือโดยการแปลง wavelet

โดยทั่วไป Spectrogram จะแสดงรูปแบบเหมือนแผนที่ความร้อน (Heat map) กล่าวคือเป็นภาพที่มีความเข้มและจางของสี ซึ่งแสดงโดยการเปลี่ยนสีหรือความสว่างดังภาพที่ 2.9 เป็นการแสดงผลที่เหมือนกับแผนที่ความร้อนของ Spectrogram

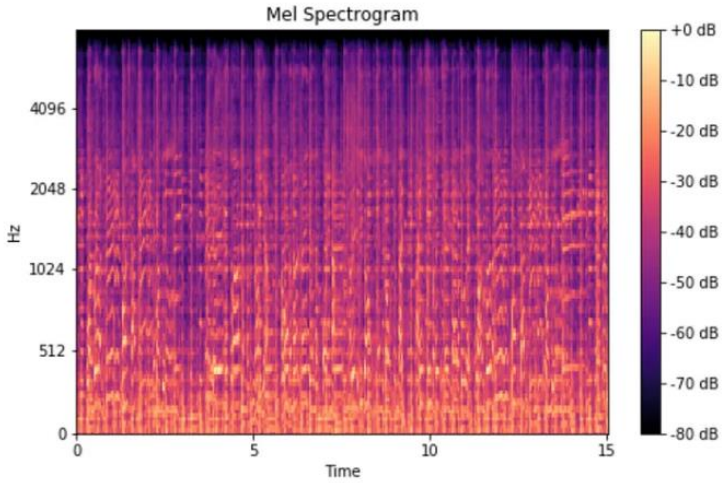


ภาพที่ 2.9 การแสดงผลสัญญาณเสียงแบบ Spectrogram

(Roberts, 2020)

Mel Scale คือระดับการรับรู้ของระดับเสียงที่เหมาะสมสำหรับการรับฟังเสียงของมนุษย์ จุดอ้างอิงระหว่างการวัดสเกลนี้และความถี่ปกติ ถูกกำหนดโดยการกำหนดระดับเสียงการรับรู้ที่ 1,000 เมล ต่อกับโทนเสียง 1,000 เฮิร์ตซ์ ซึ่งสูงกว่าเกณฑ์ของผู้ฟัง 40 เดซิเบล

จากการทำงานและแสดงผลของ Spectrogram และ Mel Scale สามารถนำมาอธิบายในส่วนของ Mel-Spectrogram นั่นคือการนำ Spectrogram มาแสดงบน Mel-Scale ในที่นี้คือแกน Y ดังภาพที่ 2.10 จะสังเกตได้ว่าความถี่ที่แสดงบน Mel-Scale จะเป็นส่วนหนึ่งของความถี่ทั้งหมด



ภาพที่ 2.10 การแสดงผลสัญญาณเสียงแบบ Mel-Spectrogram

(Roberts, 2020)

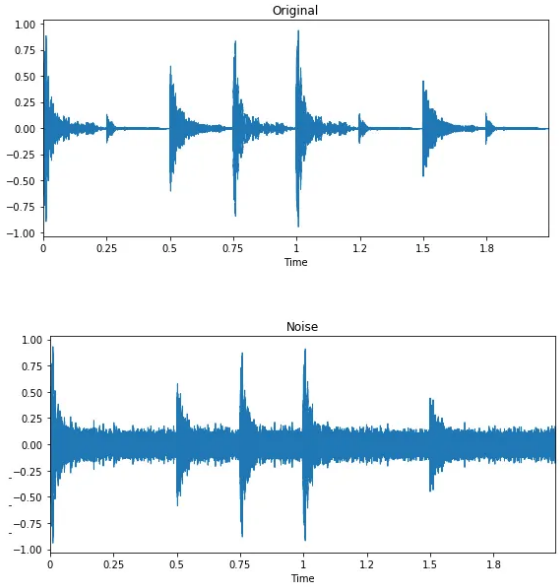
2.10 Data Augmentation

การเพิ่มข้อมูล หรือ Data Augmentation เป็นเทคนิคที่ใช้เพิ่มจำนวนชุดข้อมูลของที่ใช้ในการสร้างแบบจำลอง โดยเพิ่มข้อมูลใหม่จากการปรับแก้ชุดข้อมูลเดิม ในทางปฏิบัติแล้วการใช้ Data Augmentation จะใช้เพื่อป้องกันการเกิด Overfit หรือชุดข้อมูลเริ่มต้นมีขนาดเล็กเกินกว่าจะฝึกได้ หรือแม้กระทั่งหากต้องการเพิ่มประสิทธิภาพของโมเดล โดยการเพิ่มจำนวนของข้อมูลที่มีอยู่แล้วให้มีขนาดใหญ่ขึ้น ซึ่งโดยทั่วไปการมีชุดข้อมูลขนาดใหญ่มีความสำคัญต่อประสิทธิภาพของทั้งโมเดล Machine Learning (ML) และ Deep Learning (DL)

โดยทั่วไป Data Augmentation มักใช้ในการสร้างแบบจำลอง Deep Learning โดยข้อมูลที่สามารถนำมาทำ Data Augmentation ได้ เช่น เสียง, ข้อความ, รูปภาพ และ ข้อมูลประเภทอื่นๆ ในส่วนของการทำ Data Augmentation ที่เกี่ยวกับเสียง (Ma, 2019) เราสามารถใช้วิธีการเพิ่ม หรือปรับเปลี่ยน คุณลักษณะ ของเสียงเช่น Noise injection, Shifting time, เปลี่ยนค่า Pitch และ Speed

Noise Injection รูปภาพที่ 2.11 แสดงสัญญาณเสียงต้นฉบับ และสัญญาณเสียงที่ผ่านทางเพิ่ม Noise เข้าไป ซึ่งเป็นเทคนิคที่ง่ายที่สุดในการทำ Data Augmentation ของเสียง วิธีนี้เป็นการเพิ่มสัญญาณเสียงใหม่เข้าไปรวมกับสัญญาณเสียงเดิม จะสังเกตได้ว่าความถี่และความดัง (Amplitude) ของสัญญาณเสียงเดิมไม่ได้การเปลี่ยนแปลง แต่ถ้ามีสัญญาณเสียงใหม่ที่เพิ่มเข้าไปมีค่า

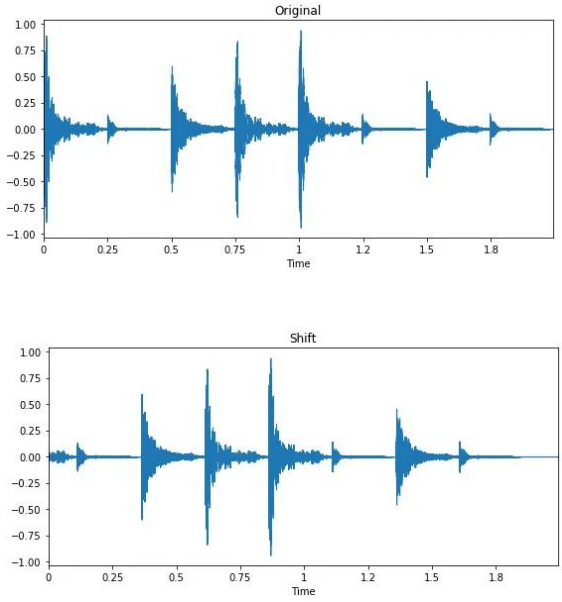
มากเกินไป จนทับความถี่เสียงและควาดังของสัญญาณเสียงเดิม จะทำให้สัญญาณเสียงไม่ได้คงไว้ซึ่งสัญญาณเสียงเดิม



ภาพที่ 2.11 สัญญาณเสียงต้นฉบับ และเสียงที่เพิ่ม Noise

(Ma, 2019)

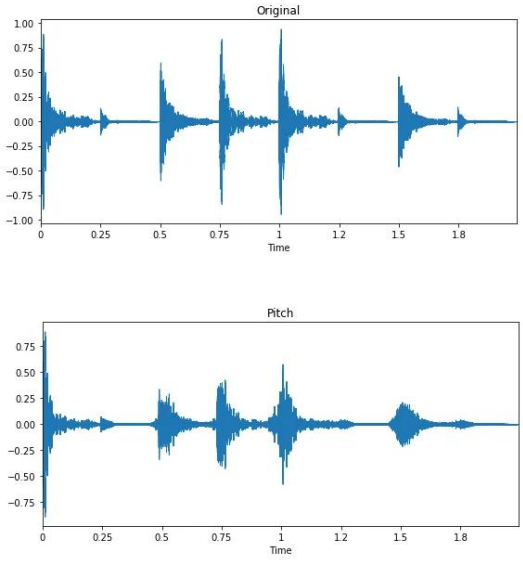
Shifting Time รูปภาพที่ 2.12 แสดงสัญญาณเสียงต้นฉบับ และสัญญาณเสียงที่ทำ Shifting Time โดยเป็นการเลื่อนสัญญาณเสียงไปทางซ้าย หรือทางขวา ในจำนวนเวลาที่กำหนด โดยเริ่มจากแกน x ที่เป็น 0 และเปลี่ยนตัวเลขโดยเพิ่มขึ้น หรือลดลง จะเห็นได้ว่าสัญญาณเสียงจะเหมือนกับต้นฉบับ ทั้งความถี่ และความดัง ซึ่งไม่ได้มีการเปลี่ยนแปลง จะเปลี่ยนเพียงเวลาเริ่มต้นของสัญญาณเสียง และเวลาสุดท้ายของสัญญาณเสียงเท่านั้น



ภาพที่ 2.12 สัญญาณเสียงต้นฉบับ และเสียงที่ผ่านการทำ Shifting Time

(Ma, 2019)

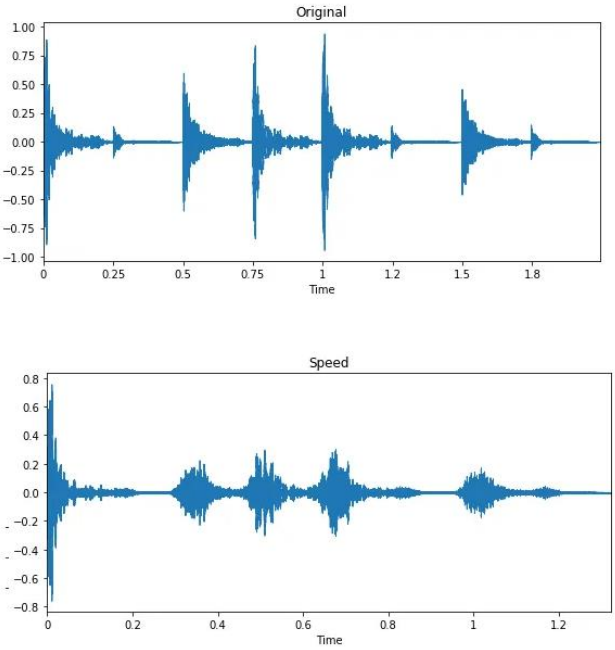
Changing Pitch รูปภาพที่ 2.13 แสดงสัญญาณเสียงต้นฉบับ และสัญญาณเสียงที่ผ่านการเปลี่ยนแปลงค่า Pitch ซึ่งคือการเปลี่ยนแปลงขนาดของสัญญาณเสียงให้เพิ่มขึ้น หรือลดลง



ภาพที่ 2.13 สัญญาณเสียงต้นฉบับ และเสียงที่ผ่านการทำ Shifting Time

(Ma, 2019)

Changing Speed รูปภาพที่ 2.14 แสดงสัญญาณเสียงต้นฉบับ และสัญญาณเสียงที่ผ่านการเปลี่ยนแปลงความเร็วของสัญญาณเสียง โดยเป็นการยืดสัญญาณเสียง เข้าและออกตามค่าที่กำหนด



ภาพที่ 2.14 สัญญาณเสียงต้นฉบับ และเสียงที่ผ่านการทำ Shifting Time

(Ma, 2019)

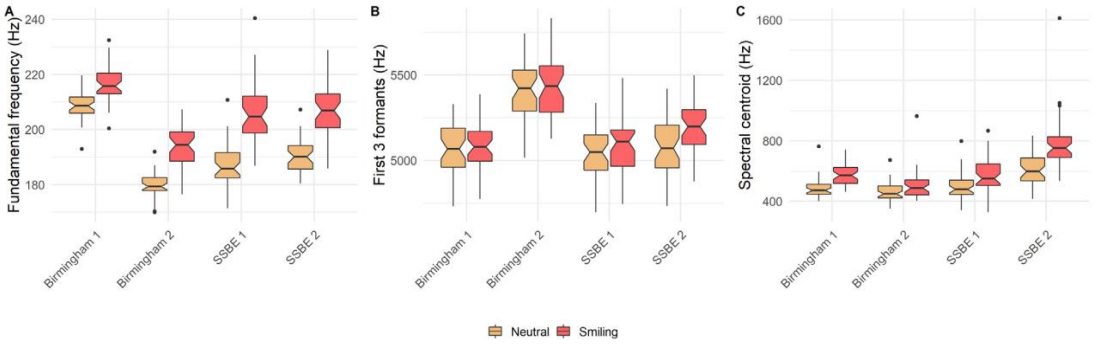
2.11 งานวิจัยที่เกี่ยวข้อง

รอยยิ้มเป็นสิ่งที่คนทั่วไปเข้าใจในความหมาย ไม่ว่าจะเป็นชนชาติ หรือเชื้อชาติใดๆ ก็เข้าใจความหมายไปในทางเดียวกัน โดยในงานวิจัยหลายๆ งานที่พูดถึงเสียงที่มีรอยยิ้ม (Smile Voice) เช่น Speech-Smile (Kohler, 2008) หรือ Smiled Speech (Emond & Laforest, 2013) ในงานวิจัยนี้จะใช้ชื่อว่า Smile Voice ซึ่งในทางตรงกันข้ามกับ Smile Voice ก็คือ Non-Smile Voice ซึ่งเป็นการหมายถึงการพูดที่ไม่ได้มีการเคลื่อนไหวของกล้ามเนื้อบนใบหน้าที่สัมพันธ์กับการยิ้ม

การยิ้มเป็นการแสดงออกที่เป็นสากล โดยผู้ที่เห็นทุกคนจะเข้าใจได้ว่าคือการยิ้ม ซึ่งพบได้ทั่วไปในมนุษย์และสัตว์อื่นๆ (Mehu & Dunbar, 2008) (van Hooff, 1972)

ในงานวิจัยหลายงาน แสดงให้เห็นว่า สามารถตรวจพบรอยยิ้มได้ ในสัญญาณเสียงพูดเพียงอย่างเดียว (Tartter & Braun, 1994) (Drahota et al., 2008) (Haddad et al., 2015) ในงานวิจัย (Torre, 2013) ได้ให้คน 4 มาพูดประโยคที่เหมือนกัน หลายๆประโยค โดยรอบแรกให้พูดในเสียงปกติ

และรอบที่สองให้พูดด้วยอารมณ์ยิ้มแย้มโดยมีการเปิดวิดีโอที่สร้างความขบขันให้ดูซึ่งจุดประสงค์คือให้ยิ้ม และมีอารมณ์ที่ไปในทางบวกตอนพูด ซึ่งจากภาพจะแสดงให้เห็นว่าเมื่อเกิดเสียงยิ้มจะมีความถี่ของเสียงที่เพิ่มขึ้น (Torre et al., 2020) ซึ่งการเพิ่มขึ้นจะแตกต่างกันทั้งในสำเนียงและผู้พูด ในภาพที่ 2.15 แสดงให้เห็นถึงความถี่ที่แตกต่างกันของผู้พูดที่พูดด้วยเสียงปกติ และพูดเมื่อยิ้มไปด้วยจะสังเกตุได้ว่าความถี่ของเสียงที่พูดด้วยรอยยิ้มจะมีความถี่ที่สูงกว่านิดหน่อย ซึ่งในการพูดแต่ละสำเนียงได้ผลเหมือนกัน



ภาพที่ 2.15 แสดงความถี่ของเสียงพูดปกติ และเสียงที่มีรอยยิ้ม ในสำเนียงที่แตกต่างกัน

(Torre et al., 2020)

ความท้าทายในการทำ Speech Emotion Recognition (SER) ในงานวิจัยต่างๆ เช่นการเลือกวิธีการดึงข้อมูล Feature ออกมา หรือโทนเสียงที่เปลี่ยนไป สไตล์ในการพูด และน้ำหนักของเสียง และยังรวมไปถึงการแสดงออกของอารมณ์ในคนแต่ละคน และสภาพแวดล้อมในขณะที่พูด ซึ่งระบบ SER ที่มีความคงที่ ส่วนที่มีความสำคัญคือ การแยกคุณลักษณะ (Extract Feature) ของเสียงพูด ซึ่งการการแยกคุณลักษณะ ของ Speech ก็จะมีหลายแบบด้วยกันไม่ว่าจะเป็น Prosodic feature (Singh et al., 2012) Spectral features (Hu et al., 2007) Mel-frequency cepstral coefficients (MFCC) (Huang et al., 2018) และ Spectrogram (Satt et al., 2017) (Li et al., 2019) ในงานวิจัยเกี่ยวกับ SER ตั้งแต่ปี 2009 (INTERSPEECH 2009 Emotion Challenge) เริ่มมีการใช้การเรียนรู้ในเชิงลึกเข้ามาในงานวิจัย มีการใช้วิธี Extract feature ที่หลากหลายโดยเฉพาะ MFCC

ในงานวิจัยของ (Li et al., 2019) ในปี ซึ่ง 2019 ได้ทำงานวิจัยโดยมองหาจุดเด่นของน้ำเสียงที่สื่อถึงอารมณ์ในขณะนั้น โดยใช้การเรียนรู้เชิงลึกในหลายแบบ ไม่ว่าจะเป็น CNN, BLSTM, Self Attention โดยงายวิจัยนี้ใช้ Dataset ของ IEMOCAP (Busso et al., 2008) โดยมีทั้งเสียงของ

ผู้ชายและ ผู้หญิง โดยพูดด้วยอารมณ์ที่แตกต่างกัน มี โกรธ รังเกียจ ตื่นเต้น หวาดกลัว ผิดหวัง มีความสุข ปกติ เสียใจ ประหลาดใจ โดยงานวิจัยนี้ จำกลุ่มของอารมณ์โดย ตื่นเต้น และมีความสุข เข้าด้วยกัน และใช้แค่ 4 กลุ่มคือ โกรธ มีความสุข ปกติ และเสียใจ

ดังนั้นจากการทบทวนวรรณกรรมและงานที่เกี่ยวข้องพบว่างานวิจัยที่ผ่านมานั้นนำเอาเทคโนโลยีมาประยุกต์ใช้ในด้านเสียงกันอย่างแพร่หลาย แต่ยังไม่มีงานวิจัยไหนวิเคราะห์เสียงยิ้มในภาษาไทย โดยใช้เทคนิคการเรียนรู้เชิงลึก รวมไปถึงนำไปประยุกต์หาความสัมพันธ์ กันระหว่างอารมณ์บนใบหน้า และอารมณ์ของเสียงที่พูดออกมา และยิ่งในเฉพาะเสียงภาษาไทย เพราะการออกเสียงของอารมณ์ในภาษาต่างๆ จะแตกต่างกันตามเชื้อชาติและภาษาด้วย ซึ่งงานวิจัยนี้ ได้นำเสนอโมเดลการเรียนรู้เชิงลึกในการวิเคราะห์หาเสียงยิ้มที่เป็นภาษาไทย

บทที่ 3
ขั้นตอนและวิธีดำเนินการ

จากการศึกษางานที่เกี่ยวข้องกับขอบเขตการวิจัยในการพัฒนาเครื่องมือที่ใช้ในการวิเคราะห์เสียงยิ้ม พบว่ายังไม่พบงานวิจัยที่วิเคราะห์เสียงยิ้มจากเสียงพูดภาษาไทย ดังนั้นเพื่อให้ได้โมเดลต้นแบบในการวิเคราะห์เสียงพูดภาษาไทยด้วยร้อยยิ้ม และแสดงความสัมพันธ์กับอารมณ์บนใบหน้า ผู้วิจัยจึงกำหนดขั้นตอนการดำเนินงานในวิทยานิพนธ์นี้ออกเป็น 9 ขั้นตอน คือ

1. เตรียมคลังข้อมูลเพื่อสร้างโมเดลวิเคราะห์อารมณ์จากเสียง
2. ออกแบบ และสร้างโมเดลวิเคราะห์อารมณ์จากเสียง
3. สร้างคลังข้อมูลเสียงจากอาสาสมัคร เพื่อใช้ในการวิเคราะห์เสียงยิ้ม
4. สร้างคลังข้อมูลเสียงจากคอลเซ็นเตอร์ เพื่อใช้ในการวิเคราะห์เสียงยิ้ม
5. วิเคราะห์เสียงยิ้มโดยผู้เชี่ยวชาญ
6. ออกแบบ และสร้างโมเดลวิเคราะห์เสียงยิ้ม
7. ทดสอบประสิทธิภาพและปรับจูนพารามิเตอร์
8. ออกแบบขั้นตอนการทำงาน (Process flow) สำหรับการวิเคราะห์เสียงยิ้ม
9. ออกแบบการประยุกต์โมเดลวิเคราะห์เสียงยิ้ม เพื่อแสดงความสัมพันธ์ของเสียงและใบหน้า

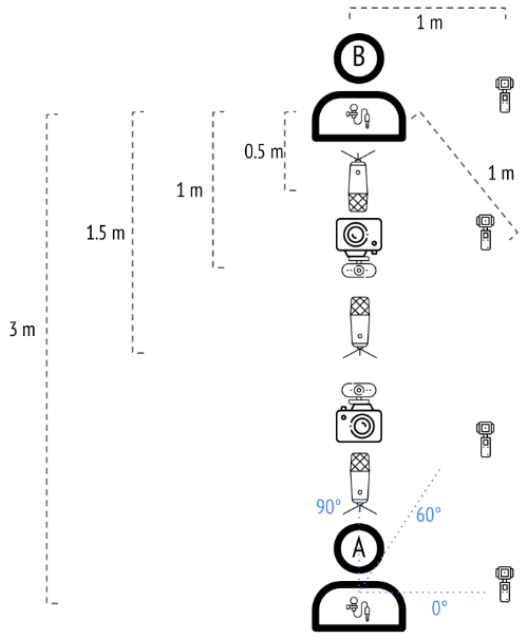
โดยรายละเอียดอธิบายในแต่ละหัวข้อดังต่อไปนี้

3.1 เตรียมคลังข้อมูลเพื่อสร้างโมเดลวิเคราะห์อารมณ์จากเสียง

ผู้วิจัยใช้คลังข้อมูลเสียงภาษาไทยจากสถาบันวิทยสิริเมธี (Vidyasirimedhi Institute of Science and Technology: VISTEC) ซึ่งได้มีการเผยแพร่คลังข้อมูลเสียงภาษาไทย พร้อมกับผลความแม่นยำของโมเดลที่พัฒนาขึ้นเพื่อวิเคราะห์อารมณ์จากเสียง เมื่อปี พ.ศ. 2564 (VISTEC, 2021) โดยคลังข้อมูลนี้เรียกว่า THAI SER (Thai Speech Emotion Recognition) โดยมีวัตถุประสงค์เพื่อใช้ในการพัฒนาโมเดลการรู้จำอารมณ์จากเสียงพูด ซึ่งเป็นคลังข้อมูลเสียงด้านอารมณ์ชุดแรกที่เป็นภาษาไทย โดยมีอารมณ์ทั้งหมด 5 อารมณ์ ได้แก่ โกรธ เศร้า สุข หงุดหงิด และ ปกติ โดยผู้วิจัยได้นำคลังข้อมูลชุดนี้ เป็นตัวตั้งต้นในการ ออกแบบ และสร้างโมเดลวิเคราะห์เสียงในภาษาไทย และนำโมเดลที่ได้ ไปประยุกต์ในขั้นตอนต่อไปเพื่อต่อยอดไปยังโมเดลวิเคราะห์เสียงยิ้มในภาษาไทย

วิทยานิพนธ์นี้จึงทดลองแนวคิดโดยการสร้างโมเดลวิเคราะห์อารมณ์ด้วยวิธีการที่ต่างจากหน่วยงานเจ้าของข้อมูล โดยเริ่มจากการศึกษา และทำความเข้าใจชุดข้อมูล

คลังข้อมูล Thai Emotion มีจำนวนเสียงทั้งหมด 27,854 เสียง โดยแบ่งเป็น 15,874 เสียงที่เป็นการพูดโดยไม่มีบท (No Script) และ 11,980 เป็นการพูดตามบท (Script) โดยการบันทึกเสียงใช้ในห้องสตูดิโอ (Studio) ซึ่งเป็นห้องที่ควบคุมองค์ประกอบ และ โปรแกรม Video Conference (Zoom) ซึ่งในห้องสตูดิโอการจัดวางตำแหน่งของผู้พูด และอุปกรณ์เป็นดังภาพที่ 3.1 โดยเป็นการบันทึกเสียงทั้งหมด 100 ครั้ง แบ่งเป็น Zoom 20 ครั้ง และห้อง Studio 80 ครั้ง และมีการใช้ไมค์ในการบันทึกเสียงหลายแบบ เช่น คอนเด็นเซอร์ไมค์ในห้องสตูดิโอ และไมค์จาก Computer (ในการณีที่ผ่าน Video Conference) โดยทั้งหมดนี้เป็นการออกเสียงใน 5 อารมณ์คือ โกรธ ฉุนเฉียว ปกติ เสียใจ และ ดีใจ โดยมีอาสาสมัครในการบันทึกเสียงทั้งหมด 200 คน



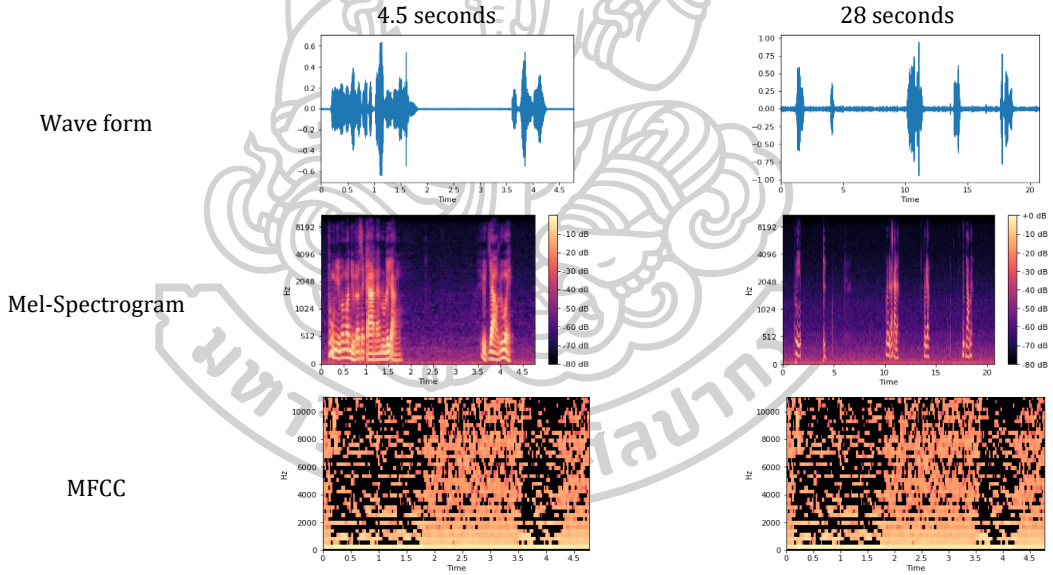
ภาพที่ 3.1 ตำแหน่งของผู้พูด และระยะการจัดวางของอุปกรณ์ในห้องสตูดิโอ

(VISTEC, 2021)

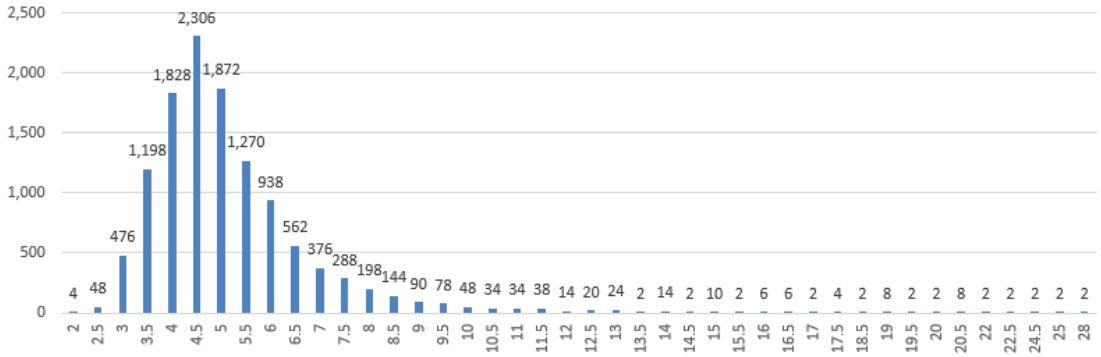
ในการสร้างโมเดลวิเคราะห์อารมณ์ และเพื่อควบคุมรูปแบบของประโยค และอารมณ์ ที่จะนำเข้าไปสร้างโมเดลวิเคราะห์อารมณ์ ผู้วิจัยเลือกใช้เฉพาะคลังข้อมูลที่เป็นการพูดแบบมีบทพูดเท่านั้น โดยจำนวนเสียงที่พูดตามบทมีทั้งหมด 11,980 เสียง และบทที่ใช้ในการพูดจะมี ทั้งหมด 3 ประโยค ตามบทด้านล่าง

บทพูด
ประโยค 1: พรุ่งนี้มันวันหยุดราชการนะรู้รึยัง หยุดยาวด้วย
ประโยค 2: อ่านหนังสือพิมพ์วันนี้รึยัง รู้ไหมเรื่องนั้นกลายเป็นข่าวใหญ่ไปแล้ว
ประโยค 3: ก่อนหน้านี้ก็ยังเห็นทำตัวปกติดี ใครจะไปรู้หล่ะ ว่าเค้าคิดแบบนั้น

โดยการพูดจะเป็นการพูดประโยคละ 4 ครั้งใน 1 อารมณ์ ซึ่ง 2 ครั้งแรกจะเป็นการพูดตามอารมณ์ที่กำหนด และอีก 2 ครั้งจะเป็นการพูดตามอารมณ์ที่กำหนดแต่ให้เน้นอารมณ์มากขึ้น โดยทางผู้วิจัยเลือกเอาเฉพาะ เสียงที่เป็น Script มาใช้เท่านั้น ซึ่งจำนวนเสียงที่นำมาใช้มีทั้งหมด 11,980 เสียง ทั้งนี้เสียงพูดในประโยคเดียวกันโดยผู้พูดต่างกัน ผู้พูดจะมีเทคนิคและวิธีการพูดที่แตกต่างกัน ซึ่งทำให้ได้รูปร่างของเสียง และเวลาที่ต่างกัน เช่น ภาพที่ 3.2 ที่แสดง Wave form ในแบบต่างๆ ที่เกิดจากประโยคในการพูดเดียวกัน แต่ใช้คนละคนในการพูด และความยาวในการพูดแตกต่างกัน



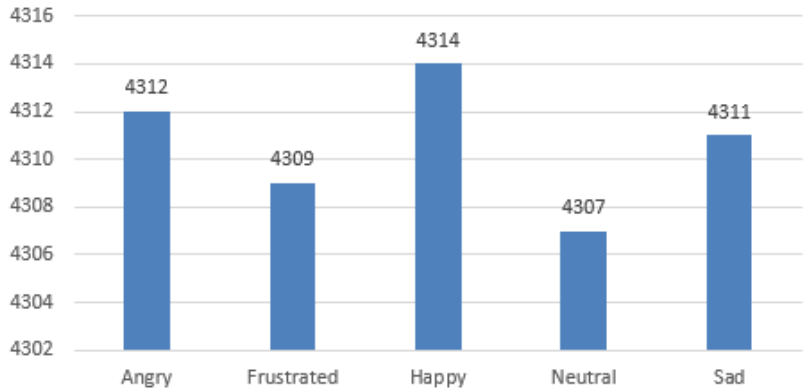
ภาพที่ 3.2 รูปแบบของ Wave form ในประโยคเดียวกัน คนพูดต่างกัน และเวลาไม่เท่ากัน



ภาพที่ 3.3 จำนวนของเสียงพูด ในระยะเวลาต่าง ๆ

จากภาพที่ 3.3 แสดงให้เห็นภาพรวมของเวลาที่ใช้ในการพูดของแต่ละประโยคของทุกคน ซึ่งจะเห็นได้ว่า เวลาโดยส่วนใหญ่อยู่ที่ เวลา 3.5-6 วินาที โดยไฟล์เสียงชุดนี้จะถูกนำเข้าไปฝึกโดยใช้โมเดล CNN อย่างไรก็ตามอินพุตของ CNN ต้องมีขนาดเท่ากันทั้งหมด เพราะฉะนั้นไฟล์เสียงที่นำเข้าไปต้องมีระยะเวลาเท่ากันด้วย ทางผู้วิจัยจึงเลือกความยาวเสียงที่ 4 วินาที เป็นค่าที่ใช้เป็นอินพุตของโมเดล CNN แต่จากรูปจะเห็นได้ว่าเสียงส่วนใหญ่อยู่ที่เวลา 4.5 วินาที และลำดับถัดมาคือ 5 และ 4 วินาที ซึ่งขั้นตอนต่อไปเป็นการทำเสียงให้มีเวลาเท่ากัน จะเห็นว่าถ้าใช้ที่เวลา 4.5 วินาที จะต้องเพิ่มเสียงว่างเปล่าเข้าไปที่เสียงที่ต่ำกว่า 4.5 วินาที ซึ่งมีจำนวนค่อนข้างมาก และทางผู้วิจัยต้องการให้มีการเพิ่มเสียงว่างเปล่าให้น้อยที่สุด ในทางกลับกันก็จะมีการตัดเสียงออกให้น้อยที่สุดเช่นกัน ซึ่งจะเห็นได้ว่าเสียงที่มีเวลามากกว่า 10 วินาที มีจำนวนน้อยมากจึงไม่ได้มีผลกระทบโดยตรงต่อการฝึกโมเดล ทางผู้วิจัยจึงเลือกเวลาที่ 4 วินาทีเป็นเวลาที่จะนำเข้าเพื่อฝึกโมเดล

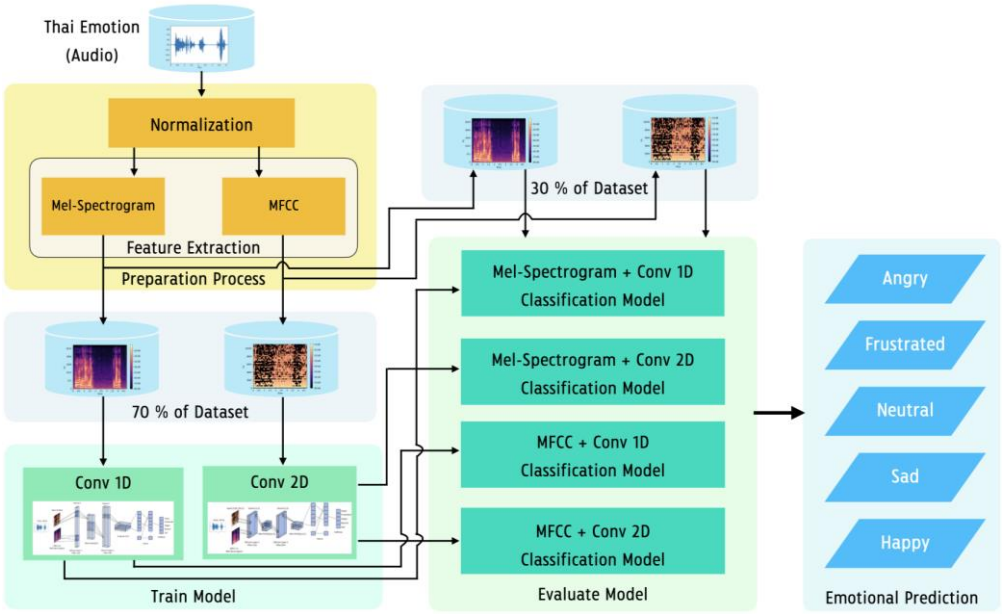
โดยในขั้นตอนการเตรียมข้อมูลไฟล์ที่มีเวลาน้อยกว่า 4 วินาที จะมีการเพิ่มเสียงว่างเปล่าเข้าไปให้เป็น 4 วินาที และในส่วนของไฟล์เสียงที่มีเวลา มากกว่า 4 วินาที ทางผู้วิจัยจะลดความยาวของไฟล์เสียงให้มีขนาด 4 วินาที เพื่อให้ข้อมูลมีขนาดเท่ากันทั้งหมดก่อนนำเข้ามาฝึกฝนในโมเดล CNN โดยประโยคแต่ละประโยคมีการพูดในอารมณ์ ที่แตกต่างกัน คือ ปกติ โกรธ ดีใจ เสียใจ ฉุนเฉียว และในส่วนของจำนวนของเสียงในแต่ละอารมณ์จะเป็นไปดังภาพที่ 3.4



ภาพที่ 3.4 จำนวนของเสียงพูด ในแต่ละอารมณ์ (ปกติ โกรธ ดีใจ เสียใจ ฉุนเฉียว)

3.2 ออกแบบ และสร้างโมเดลวิเคราะห์อารมณ์จากเสียง

การออกแบบ และสร้างโมเดลสำหรับวิเคราะห์อารมณ์จากเสียง ด้วยคลังข้อมูลของ Vistec ซึ่งเป็นคลังข้อมูลที่จำแนกอารมณ์จากเสียงพูดภาษาไทย ซึ่งเป็นเสียงพูดในภาษาไทยที่สื่ออารมณ์ทั้งหมด 5 อารมณ์ ได้แก่ โกรธ เศร้า สุข หงุดหงิด และปกติ โดยแบบจำลองที่สร้างขึ้น จะวิเคราะห์อารมณ์ได้ 5 อารมณ์เช่นกัน คือ โกรธ เศร้า สุข หงุดหงิด และปกติ และเป็นการสร้างโมเดลในอีกรูปแบบ โดยขั้นตอนของการสร้างโมเดลวิเคราะห์อารมณ์จากเสียง แสดงดังภาพที่ 3.5

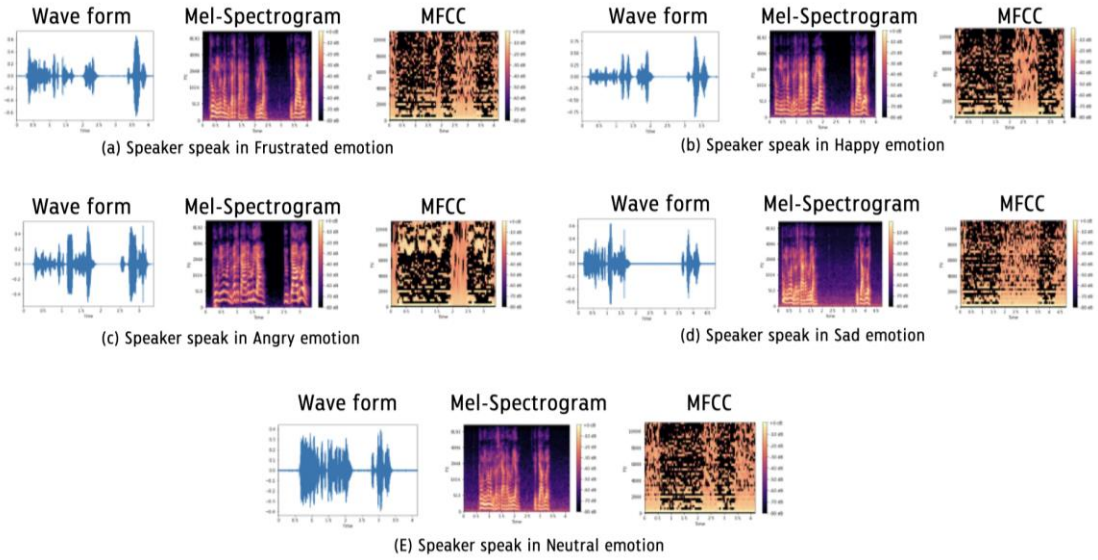


ภาพที่ 3.5 ขั้นตอนของการสร้างโมเดลวิเคราะห์อารมณ์จากเสียง

(Prombut et al., 2021)

จากภาพที่ 3.5 การนำเข้าข้อมูลเสียงจะผ่านการทำให้เสียงมีขนาดความยาวของเสียงเท่ากันทั้งหมดก่อน (Normalization) ซึ่งขั้นตอนของการทำ Normalization ของเสียงโดยการใช้ python library ชื่อว่า librosa ซึ่งเป็น library ที่ใช้เพื่อจัดการเรื่องเสียง ซึ่งเสียงที่ผ่านกระบวนการ Normalization แล้วจะมีความยาว 4 วินาที ซึ่งเป็นความยาวโดยเฉลี่ยที่พบจากการสำรวจข้อมูล และเสียงที่ได้จะถูกนำไปสกัด Feature ทั้งแบบ Mel-Spectrogram และแบบ Mel frequency cepstrum coefficient (MFCC) จากนั้นข้อมูลจากการสกัด Feature จะถูกแบ่งออกเป็นข้อมูลชุดฝึกฝน 30% และชุดทดสอบ 70%

ผู้วิจัยพัฒนาโมเดล เพื่อวิเคราะห์อารมณ์จากเสียงแบบ Convolutional Neural Network 1D (Conv 1D) และ Convolutional Neural Network 2D (Conv 2D) โดยทั้ง 2 โมเดลใช้ข้อมูลนำเข้าเป็นเสียงที่สกัด Feature แล้วทั้ง 2 แบบ เพื่อเปรียบเทียบหาวิธีการที่เหมาะสม ทั้งในการสกัด Feature และสร้างโมเดล โดยในภาพที่ 3.6 จะแสดงให้เห็นถึงรูปแบบของ Wave form ที่แตกต่างกันที่นำไปสกัด Feature แต่ละแบบ

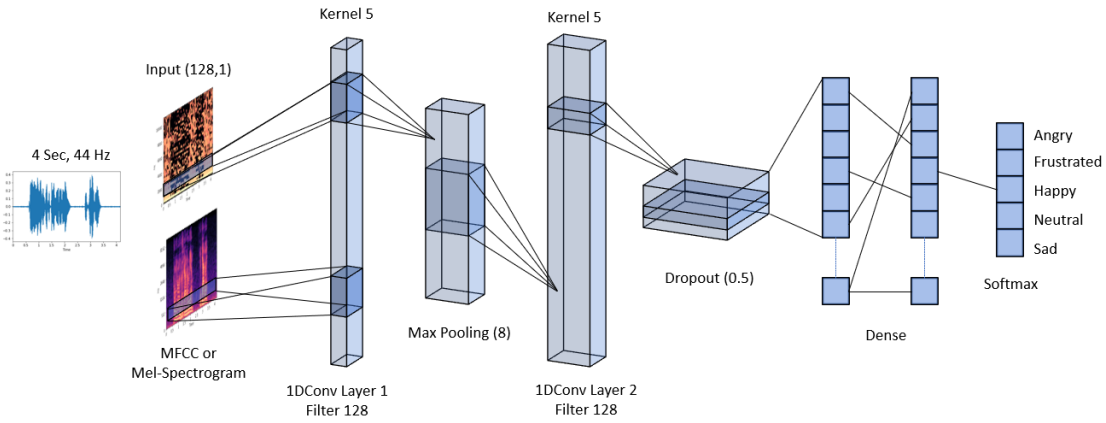


ภาพที่ 3.6 Wave form ที่แตกต่างกันที่นำไปสกัด Feature Mel-Spectrogram และ MFCC

(Prombut et al., 2021)

3.2.1 การสร้างโมเดลแบบ Convolutional Neural Network 1D (Conv 1D)

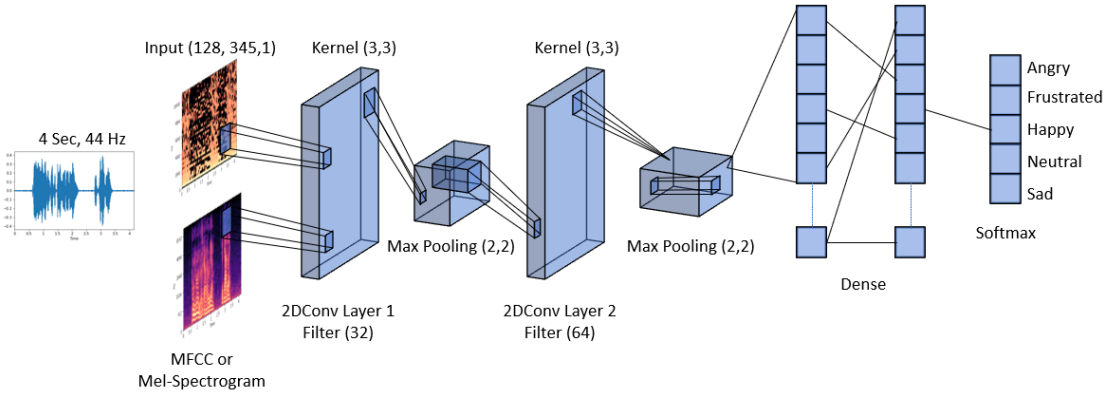
ในส่วนของขั้นตอนการออกแบบจำลอง 1D CNN จะมีการ Extract Feature แบบ Mel-Spectrogram และแบบ MFCC จากนั้นจะนำข้อมูลผ่านเข้าไปสู่ชั้น 1D CNN Layer แรก โดยมีขนาดของฟิลเตอร์เท่ากับ 128 และ ขนาดของ Kernel เท่ากับ 5 เมื่อผ่าน Layer แรก เป็นการ Dropout ด้วยค่าเท่ากับ 0.5 เพื่อป้องกันการเกิด Overfit และต่อด้วย Max pooling ซึ่งมีขนาด 8 จากนั้นข้อมูลจะถูกส่งผ่านไป 1D CNN Layer ที่ 2 โดยมีขนาดของฟิลเตอร์เท่ากับ Layer แรกคือ 128 และ Kernel เท่ากับ 5 ตามด้วย Dropout 0.5 และจะเข้าไปที่ Fully Connected Layer ซึ่งเป็น Layer สุดท้าย และผ่าน Softmax Function เพื่อให้ได้ผลลัพธ์เป็น 5 Class ดังแสดงใน ภาพที่ 3.7



ภาพที่ 3.7 1D CNN Model

3.2.2 การสร้างโมเดลแบบ Convolutional Neural Network 2D (Conv 2D)

ในส่วนขั้นตอนของการออกแบบจำลอง 2D CNN จะมีการ Extract Feature แบบ Mel-Spectrogram และแบบ MFCC จากนั้นจะนำข้อมูลผ่านเข้าไปสู่ชั้น 2D CNN Layer แรก โดยมีขนาดของฟิลเตอร์เท่ากับ 128 และ ขนาดของ Kernel เท่ากับ 3x3 ตามด้วย Dropout ด้วยค่าเท่ากับ 0.5 เพื่อป้องกันการเกิด Overfit และต่อด้วย Max pooling ซึ่งมีขนาด 2x2 จากนั้นข้อมูลจะถูกส่งผ่านไป 2D CNN Layer ที่ 2 โดยมีขนาดของฟิลเตอร์เท่ากับ Layer แรกคือ 128 และ Kernel 3x3 ตามด้วย Dropout 0.5 และจะเข้าไปที่ Fully Connected Layer ซึ่งเป็น Layer สุดท้าย และผ่าน Softmax Function เพื่อให้ได้ผลลัพธ์ออกมาเป็น 5 Class ดังแสดงใน ภาพที่ 3.8



ภาพที่ 3.8 2D CNN Model

3.3 สร้างคลังข้อมูลเสียงจากอาสาสมัคร เพื่อใช้ในการวิเคราะห์เสียงยิ้ม

เมื่อสร้างโมเดลต้นแบบและทดลองโดยการวิเคราะห์อารมณ์จากเสียงที่มาจากชุดข้อมูลของ Thai Emotion แล้ว ผู้วิจัยได้รวบรวมเสียงจากอาสาสมัครที่เป็นนักศึกษาระดับ ปริญญาตรี อายุประมาณ 18-20 ปี จำนวน 7 คน โดยเป็นผู้หญิง 4 คน และ ผู้ชาย 3 คน ซึ่งการอัดเสียงทำในห้องประชุมที่ ไม่มีเสียงรบกวน และให้อาสาสมัครถือไมค์เอาไว้ใกล้ตัวในระหว่างที่อัดเสียง ดังภาพที่ 3.9 โดยจะมีบทให้พูด และให้อาสาสมัครพูดต่อเนื่องไปเรื่อยๆ และเว้นระยะประมาณ 2-3 วินาทีในการเริ่มพูดประโยคใหม่ โดยผู้วิจัยได้บันทึก แยกในแต่ละอาสาสมัครว่าได้พูดประโยคใดไปแล้วบ้าง และมีประโยคที่ต้องแก้ไขหรือไม่

ภาพที่ 3.9 การจัดวางรูปแบบของการอัดเสียงจากอาสาสมัคร

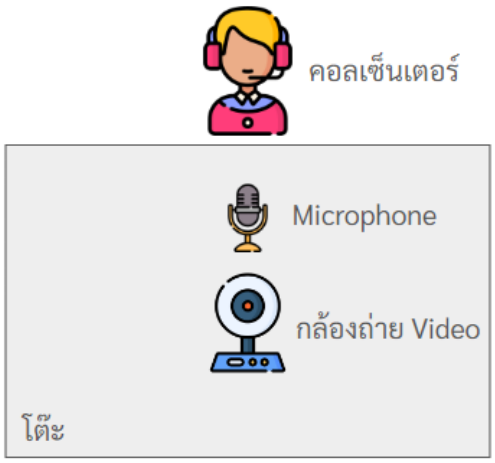
โดยบทที่ใช้พูดจะมีทั้งหมด 6 ประโยค แต่ละประโยคจำนวน 4 ครั้ง โดย 2 ครั้งแรกเป็นเสียง Smile Voice และอีก 2 ครั้งเป็น Non-Smile Voice โดยการเลือกประโยคที่ใช้ ทางผู้วิจัยเลือกใช้เหตุการณ์ของการซื้อของออนไลน์ ที่ลูกค้าโทรเข้ามาสอบถามเกี่ยวกับการซื้อของ และประโยคเกี่ยวกับปัญหาที่คาดว่าจะเกิดขึ้น

บทพูด
ประโยค 1: ของที่ลูกค้าสั่งไว้ยังมาไม่ถึงประเทศไทย
ประโยค 2: ค่าโทรศัพท์ของคุณเลยกำหนดชำระเงินแล้ว
ประโยค 3: ขอบคุณที่ชำระค่าบริการบัตรเครดิตในเดือนนี้
ประโยค 4: อีกสักครู่จะมีรหัสเข้าใช้งาน เป็น SMS ส่งไปหา
ประโยค 5: สินค้าที่สั่งไว้ ยังทำไม่เสร็จ คงอีกสัก 2-3 วัน
ประโยค 6: เรายังให้คำตอบไม่ได้ในตอนนี้ เนื่องจากกำลังศึกษาปัญหาอยู่

จากการอัดเสียงจะได้ชุดข้อมูลทั้งหมด 168 เสียง แบ่งเป็นเสียงผู้หญิง 96 เสียง และเสียงผู้ชาย 72 เสียง โดยมีความยาวของเสียงรวมทั้งหมดประมาณ 8.5 นาที โดยเป็นเสียง Smile Voice จำนวน 84 เสียง และ Non-Smile Voice จำนวน 84 เสียง

3.4 สร้างคลังข้อมูลเสียงจากคอลเซ็นเตอร์ เพื่อใช้ในการวิเคราะห์เสียงยิ้ม

เนื่องจากในปัจจุบันยังไม่มีคลังข้อมูลเสียงยิ้มที่เป็นภาษาไทย ทางผู้วิจัยได้ติดต่อทีมงานคอลเซ็นเตอร์มืออาชีพ จากธนาคารแห่งหนึ่ง ซึ่งต้องมีการใช้งานเสียงยิ้ม หรือเสียงพูดที่ทำให้ลูกค้าพึงพอใจอยู่ตลอดเวลาในการทำงานจริง โดยทางผู้วิจัยได้ตั้งเครื่องมือถ่ายวิดีโอ ไว้บนโต๊ะ และไมค์โครโฟนอยู่ด้านหน้าผู้อัดเสียง (คอลเซ็นเตอร์) และในส่วนคอลเซ็นเตอร์ที่บันทึกเสียงจะนั่งอยู่ในโต๊ะเดียวกัน และหันหน้าเข้ากล้อง โดยการจัดวางเป็นดัง ภาพที่ 3.10



ภาพที่ 3.10 การจัดวางตำแหน่ง ของอุปกรณ์ในขั้นตอนการอัดเสียง

โดยขั้นตอนการถ่ายวิดีโอจะเป็นการถ่ายเฉพาะส่วนใบหน้า จนถึงส่วนคอของอาสาสมัคร และอาสาสมัครจะหันหน้าเข้าหากล้องถ่ายวิดีโอเป็นหน้าตรงเสมอ ในส่วนบนโต๊ะจะมี Condenser ไมค์วางไว้อยู่ด้านหน้า เพื่อนำมาสร้างคลังข้อมูลเป็น Smile Voice และ Non-Smile Voice โดยมีรายละเอียดดังนี้

อาสาสมัครคอลเซ็นเตอร์มีทั้งหมด 15 คน แบ่งเป็นผู้ชาย 3 คน และผู้หญิง 12 คน โดยทุกคนจะพูดประโยคทั้งหมด 7 ประโยค ทั้งเสียงยิ้ม (Smile Voice) และเสียงไม่ยิ้ม (Non-Smile Voice) เพราะฉะนั้นใน 1 ประโยคจะพูดทั้งหมด 2 ครั้ง โดยมีประโยคที่เตรียมไว้ทั้งหมด 7 ประโยค ที่พูดเรียงตามลำดับ

บทพูด

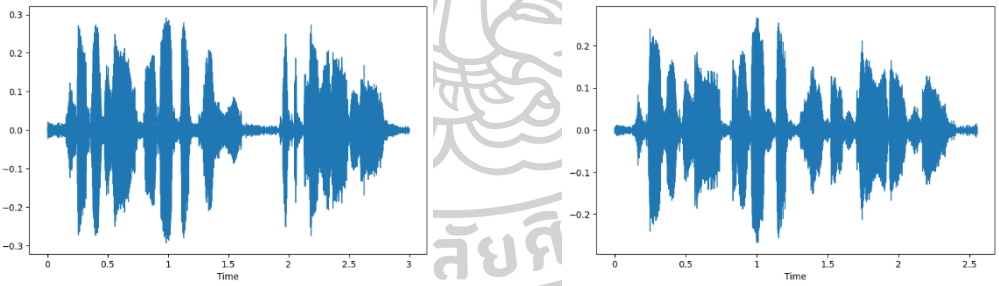
---

1. สวัสดี (ครับ/คะ) ต้องการสอบถามว่าอย่างไร (ครับ/คะ)
2. ขออภัยที่ถือสายรอ (ครับ/คะ) จากการตรวจสอบข้อมูล มียอดเข้ามาแล้ว (ครับ/คะ) จำนวน 2,000 บาท
3. ขออภัย (ครับ/คะ) เอกสารที่ถืออยู่ ระบุข้อความว่าอย่างไร (ครับ/คะ)
4. ขอบคุณ (ครับ/คะ) เอกสารที่ได้รับคือ ใบเสร็จที่ลูกค้าชำระเงินเมื่อวันที่ 20 กพ. 2566 (ครับ/คะ)
5. สามารถสร้าง QR Code ผ่านช่องทาง Application ผ่านหน้าเว็ปไซต์ หรือ line ad (ครับ/คะ)
6. ยินดี (ครับ/คะ) มีข้อมูลด้านอื่นๆ สอบถามเพิ่มเติมไหม (ครับ/คะ)

7. ก่อนวางสาย รบกวนเวลาสักครู่ ช่วยกดผลประเมินความพึงพอใจ ในการให้บริการ ขอบพระคุณที่ใช้บริการ สวัสดี (ครับ/คะ)

โดยประโยคที่ใช้พูดจะเป็นแนวทางในการพูดของอาสาสมัคร และเนื่องจากแต่ละคนมีประสบการณ์ในงานคอลเซ็นเตอร์ที่แตกต่างกัน เพื่อความเป็นธรรมชาติ การพูดของแต่ละคนอาจจะปรับเปลี่ยนไปตามความถนัด และความคุ้นชินของผู้พูด ซึ่งประโยคที่เรียงลำดับดังกล่าว ประกอบกันเป็นเรื่องราว เสมือนว่ามีคนโทรมาสอบถามจริง ๆ โดยเริ่มตั้งแต่กล่าวคำทักทาย สอบถามเกี่ยวกับปัญหาที่เกิดขึ้น รวมไปถึงทำความเข้าใจในคำถามของผู้โทรเข้ามาสอบถาม และตอบคำถาม ก่อนจบประโยคการสนทนา ด้วยประโยคปิดท้ายให้ประเมินความพึงพอใจก่อนวางสาย ซึ่งตัวอย่างของรูปแบบ Wave form ที่มาจากผู้พูดคนเดียวกัน ในแต่ละประโยคตั้งแต่ประโยคที่ 1 – 7 ข้างต้น ทั้งในแบบ Smile Voice และ Non Smile Voice จะมีรูปแบบดังแสดงใน ภาพที่ 3.11 – ภาพที่ 3.17

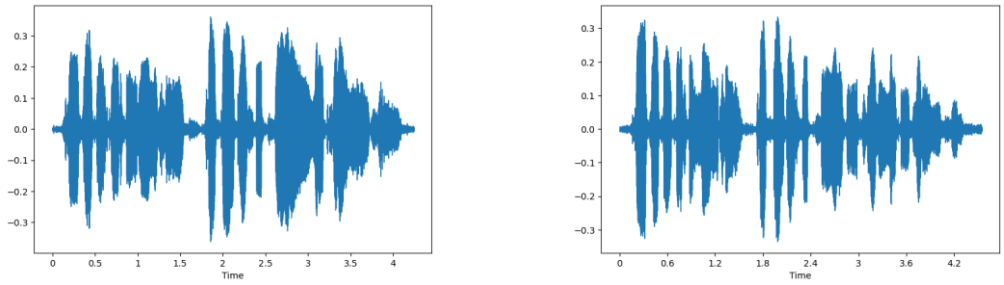
รูปแบบ Wave form จากประโยคที่ 1 “สวัสดีคะ ต้องการสอบถามว่าอย่างไรคะ” แสดงดังภาพที่ 3.11 โดย (ก) คือ Smile Voice และ (ข) คือ Non Smile Voice



(ก) ประโยคที่ 1 แบบ Smile Voice (ข) ประโยคที่ 1 แบบ Non Smile Voice

ภาพที่ 3.11 Wave form จากประโยคที่ 1

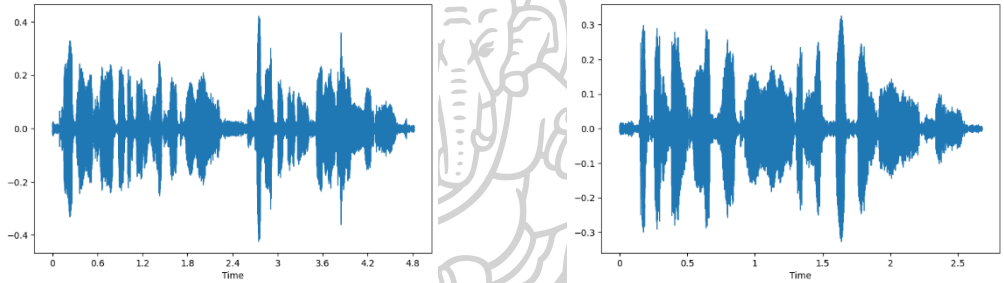
รูปแบบ Wave form จากประโยคที่ 2 “ขออภัยที่ถือสายรอ (ครับ/คะ) จากการตรวจสอบข้อมูล มียอดเข้ามาแล้ว (ครับ/คะ) จำนวน 2,000 บาท” แสดงดังภาพที่ 3.12



(ก) ประโยคที่ 2 แบบ Smile Voice (ข) ประโยคที่ 2 แบบ Non Smile Voice

ภาพที่ 3.12 Wave form จากประโยคที่ 2

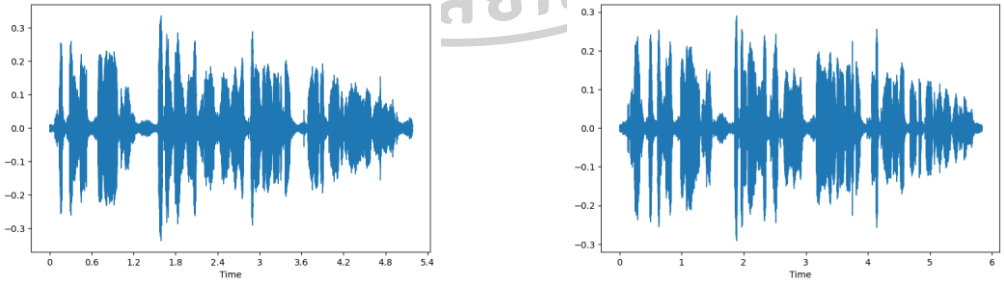
รูปแบบ Wave form จากประโยคที่ 3 “ขออภัย (ครับ/คะ) เอกสารที่ถืออยู่ ระบุข้อความว่าอย่างไร (ครับ/คะ)” แสดงภาพที่ 3.13



(ก) ประโยคที่ 3 แบบ Smile Voice (ข) ประโยคที่ 3 แบบ Non Smile Voice

ภาพที่ 3.13 Wave form จากประโยคที่ 3

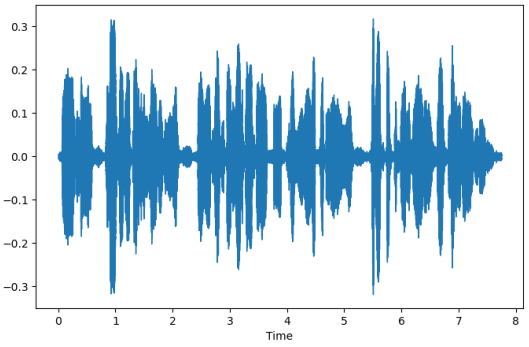
รูปแบบ Wave form จากประโยคที่ 4 “ขอบคุณ (ครับ/คะ) เอกสารที่ได้รับคือ ใบเสร็จที่ลูกค้าชำระเงินเมื่อวันที่ 20 กพ. 2566 (ครับ/คะ)” แสดงดังภาพที่ 3.14



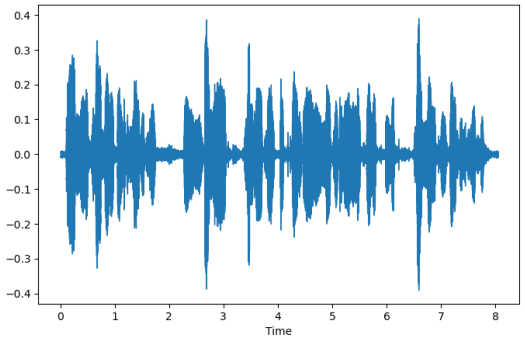
(ก) ประโยคที่ 4 แบบ Smile Voice (ข) ประโยคที่ 4 แบบ Non Smile Voice

ภาพที่ 3.14 Wave form จากประโยคที่ 4

รูปแบบ Wave form จากประโยคที่ 5 “สามารถสร้าง QR Code ผ่านช่องทาง Application ผ่านหน้าเว็ปไซต์ หรือ line ad (ครับ/คะ)” แสดงดังภาพที่ 3.15



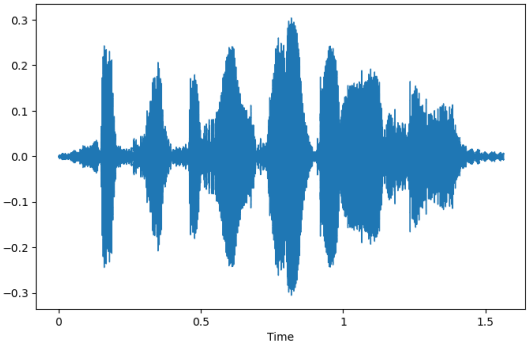
3.12 (ก) ประโยคที่ 5 แบบ Smile Voice



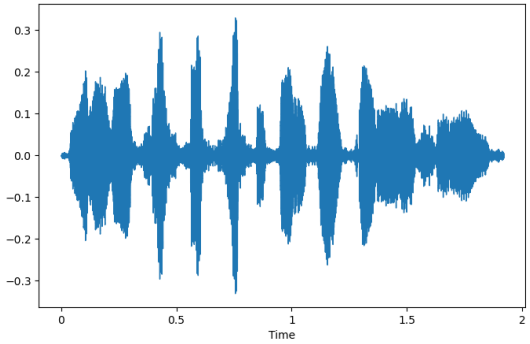
3.12 (ข) ประโยคที่ 5 แบบ Non Smile Voice

ภาพที่ 3.15 Wave form จากประโยคที่ 5

รูปแบบ Wave form จากประโยคที่ 6 “ยินดี (ครับ/คะ) มีข้อมูลด้านอื่นๆ สอบถามเพิ่มเติมไหม (ครับ/คะ)” แสดงดังภาพที่ 3.16



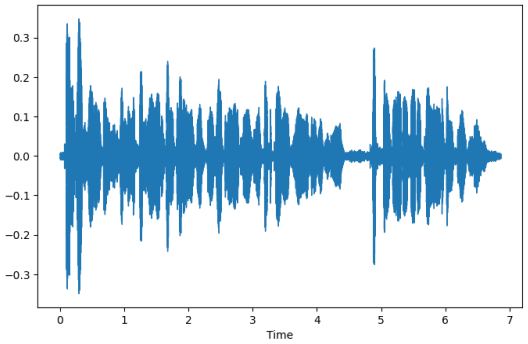
(ก) ประโยคที่ 6 แบบ Smile Voice



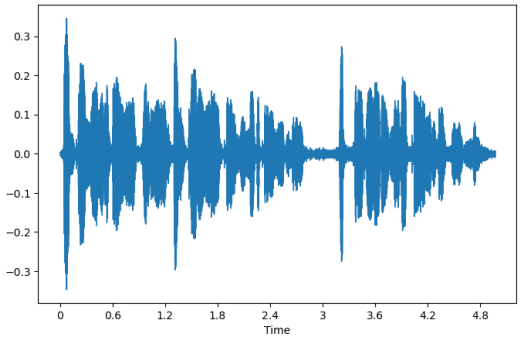
(ข) ประโยคที่ 6 แบบ Non-Smile Voice

ภาพที่ 3.16 Wave form จากประโยคที่ 6

รูปแบบ Wave form จากประโยคที่ 7 “ก่อนวางสาย รบกวนเวลาสักครู่ ช่วยกดผลประเมินความพึงพอใจ ในการให้บริการ ขอบพระคุณที่ใช้บริการ สวัสดี (ครับ/คะ)” แสดงดังภาพที่ 3.17



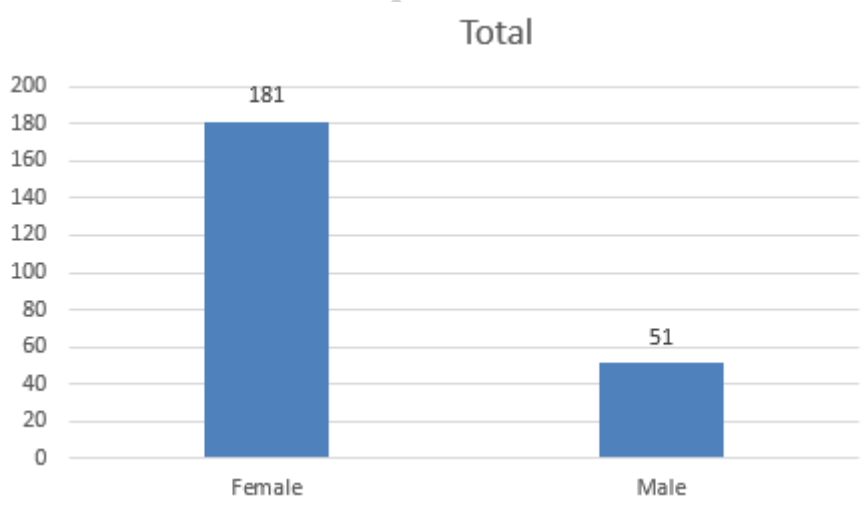
(ก) ประโยคที่ 7 แบบ Smile Voice



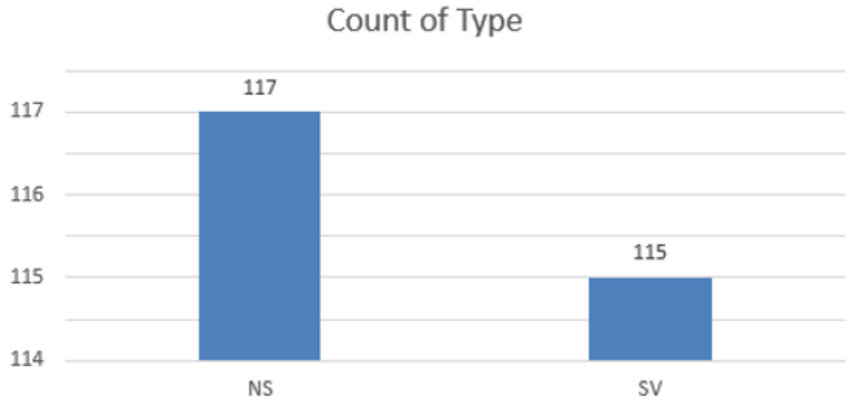
(ข) ประโยคที่ 7 แบบ Non Smile Voice

ภาพที่ 3.17 Wave form จากประโยคที่ 7

จากประโยคข้างต้น ที่พูดทั้งในแบบ Smile Voice และ Non-Smile Voice จะได้ไฟล์เสียงมาทั้งหมด 232 ไฟล์ โดยในแต่ละคนจะไม่ได้พูดทั้งหมด 14 ประโยคทุกคน คือ Smile Voice 7 และ Non-Smile Voice 7 เนื่องจากคอลเซ็นเตอร์แต่ละคนมีเทคนิควิธีการพูดที่แตกต่างกัน บางคนจึงมีการพูดที่เป็นการรวมประโยคเข้าด้วยกัน หรือบางคนอาจจะมีการพูดเพิ่มเติมจากประโยคที่เตรียมไว้แต่ยังคงรูปแบบการถามตอบเหมือนเดิม ทำให้จำนวนของไฟล์เสียงในการพูดไม่ได้เป็น 14 ประโยคทุกคน โดยไฟล์เสียงที่ได้จะแบ่งเป็น เสียงผู้ชาย 51 และเสียงผู้หญิง 181 เสียงดังภาพที่ 3.18 โดยแบ่งเป็นเสียง Smile Voice 115 เสียง และ Non Smile Voice 117 เสียงดังภาพที่ 3.19

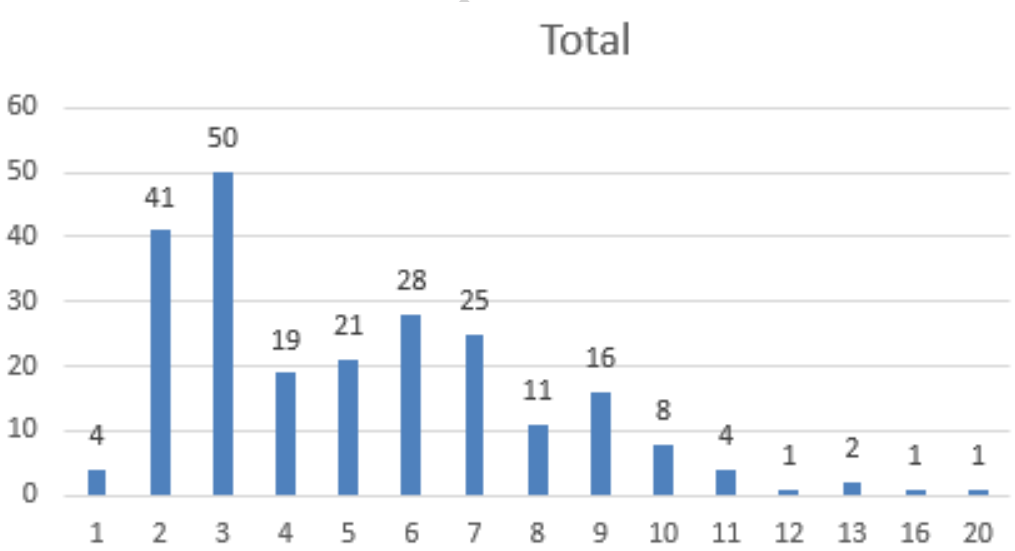


ภาพที่ 3.18 จำนวนเสียงพูดของคอลเซ็นเตอร์ ผู้ชายและผู้หญิง



ภาพที่ 3.19 จำนวนเสียงพูด Smile Voice และ Non-Smile Voice

โดยจากไฟล์เสียงที่ได้มา จะเห็นได้ว่าแต่ละคนจะมีรูปแบบ และเทคนิคในการพูด ที่ไม่เหมือนกัน ขึ้นอยู่กับประสบการณ์ และความถนัดของประโยค รวมถึงเทคนิคในการพูดของแต่ละคน รวมทั้งเวลาที่ใช้ในการพูดในแต่ละประโยคของแต่ละคนไม่เท่ากัน โดยจะมีเวลาที่ใช้ในการพูดของแต่ละประโยค อยู่ในช่วงตั้งแต่ 1-20 วินาที ซึ่งเวลาเฉลี่ยในการพูดส่วนใหญ่จะอยู่ที่ ประมาณ 2-3 วินาที โดยเวลาเฉลี่ย 3 วินาที มีจำนวน 50 เสียง และ 2 วินาที มีจำนวน 41 เสียง และเวลาเฉลี่ยเสียงที่สั้นที่สุดคือ 1 วินาที มีจำนวน 4 เสียง และเวลาเฉลี่ยของเสียงที่นานที่สุดคือ 20 วินาที มีจำนวน 1 เสียง โดยภาพที่ 3.20 แสดงจำนวนประโยคที่ใช้ในการพูดในแต่ละช่วงเวลา



ภาพที่ 3.20 จำนวนประโยค เมื่อเทียบกับระยะเวลาในการพูด

เนื่องจาก Dataset ในส่วนของ Smile Voice และ Non-Smile Voice มีจำนวน 232 ไฟล์ และมีความยาวแตกต่างกัน ซึ่งขั้นตอนของการนำคลังข้อมูลมาฝึกฝนโมเดล เพื่อให้มีค่าความถูกต้องที่สูง จำนวนของคลังข้อมูลจะต้องมีจำนวนมากพอ ซึ่งในกรณีนี้คลังข้อมูลเสียงมีจำนวน 232 ไฟล์ ซึ่งค่อนข้างน้อย อาจจะทำให้ประสิทธิภาพในการฝึกโมเดลได้ไม่ดีพอ ทางผู้วิจัยได้เพิ่มจำนวนไฟล์เสียง โดยแบ่ง File เป็น 3 กลุ่ม ดังนี้

1. **ไฟล์ต้นฉบับ**

เป็นไฟล์ต้นฉบับที่เก็บมาจากกลุ่มคอลเซ็นเตอร์ จำนวน 232 File ที่ไม่ได้เปลี่ยนแปลง ทำเพียงตัดเสียงที่ว่างเปล่าออก และทำให้เป็นไฟล์เสียงตรงตามแต่ละประโยคเท่านั้น ซึ่งไฟล์ชุดนี้จะใช้เป็นแม่แบบในการต่อยอดไปในไฟล์อีก 2 กลุ่มที่เหลือ

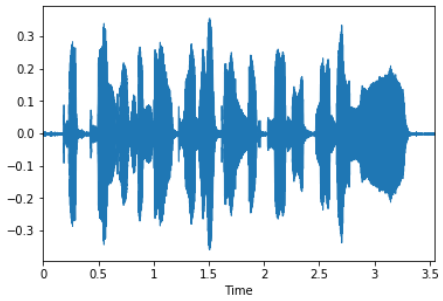
2. ไฟล์ที่มีความยาว 4 วินาที

เป็นการนำไฟล์ต้นฉบับมาแยกออกเป็น File ย่อย ที่มีขนาด 4 วินาที ซึ่งถ้าเป็นไฟล์ที่มีขนาดน้อยกว่า 4 วินาที จะเป็นการเพิ่มความยาวให้ครบ 4 วินาที โดยเพิ่มเสียงเงียบเข้าไป ซึ่งถ้าดูจากไฟล์เสียงทั้งหมด ไฟล์ที่มีความยาว มากกว่า 4 วินาที มีประมาณครึ่งหนึ่งของไฟล์เสียงทั้งหมด กลุ่มนี้จะเป็นการ นำ File ที่มีขนาดยาวกว่า 4 วินาทีมาแบ่งออกเป็น File ย่อย ซึ่งทำให้มีจำนวน Dataset ทั้งหมด 362 ไฟล์

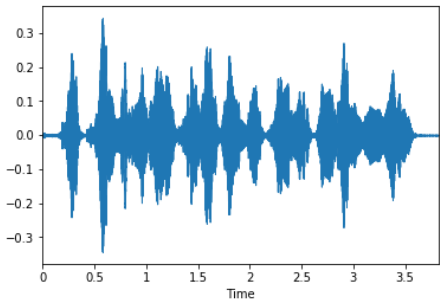
3. Data Augmentation

ผู้วิจัยนำ Dataset จากกลุ่มที่ 2 จำนวน 362 ไฟล์ มาทำ Data Augmentation เพื่อเพิ่มจำนวนของไฟล์ ซึ่งมีการทำ Data Augmentation ทั้งหมด 4 แบบ คือ 1) Time stretching 2) Pitch shifting 3) Adding noise 4) Downsample โดยในแต่ละแบบจะมีการเปลี่ยนค่า Parameter ใน Range ที่แตกต่างกัน นั่นคือ แบบที่ 1. Time Stretching จะมีการปรับ Parameter มีค่าอยู่ระหว่าง 0.8 - 1.2 แบบที่ 2. Pitch Shifting ปรับ Parameter มีค่าอยู่ระหว่าง -3 ถึง 3 แบบที่ 3. Adding Noise ปรับ Parameter มีค่าอยู่ระหว่าง 0 - 0.01 และ แบบที่ 4. Downsample ปรับ Parameter มีค่าอยู่ระหว่าง 0.95 - 1.2 ซึ่งในแต่ละแบบจะมีการทำหนด Range เพื่อต้องการคงรูปแบบของ Wave form ให้เหมือนเดิมมากที่สุด โดยถ้ามีการปรับนอกเหนือจาก Range ที่กำหนดจะทำให้ Wave form เปลี่ยนแปลงจากรูปแบบเดิมมากเกินไป จนไม่สามารถฟังเข้าใจว่าไฟล์เสียงนั้นพูดในประโยคอะไร

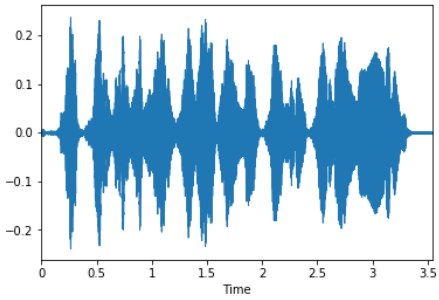
โดยการปรับเปลี่ยนค่า Parameter ทั้งหมด 5 แบบข้างต้น ทำให้หลังจากการทำ Data Augmentation จะได้ Dataset เพิ่มขึ้นจาก 362 ไฟล์ เป็น 7,240 ไฟล์ โดยตัวอย่างของ Wave form ของประโยคหนึ่งที่ทั้งแบบไม่ทำ Data Augmentation และทำ Data Augmentation ทั้ง 4 แบบ แล้วจะมีรูปแบบที่แตกต่างกันไป ดังภาพที่ 3.21 โดย เป็นภาพที่ (ก) ถึง (จ) โดยจะสังเกตได้ว่ารูปแบบของ Wave form มีการเปลี่ยนแปลงไป แต่ยังคงรูปแบบไกล้คียงกับ Wave form ก่อนที่จะมีการทำ Data Augmentation



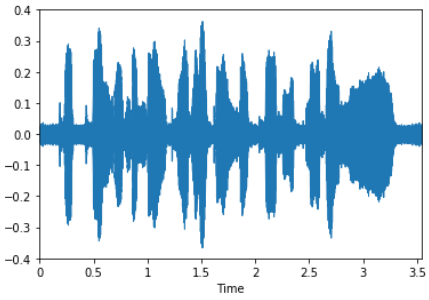
(ก) Normal Wave



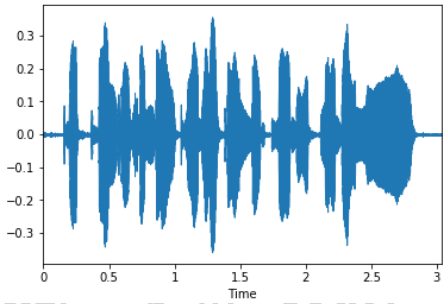
(ข) Time stretching



(ค) Pitch shifting



(ง) Adding noise



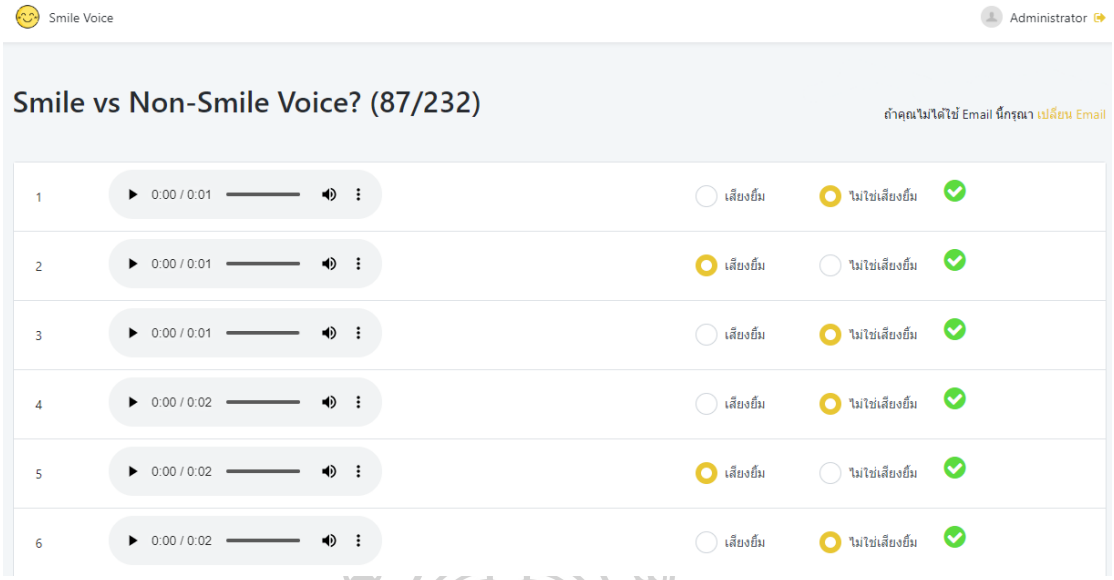
(จ) Downsample

ภาพที่ 3.21 รูปแบบ Wave Form จากการทำ Data Augmentation แบบต่างๆ

3.5 วิเคราะห์เสียงยิ้มโดยผู้เชี่ยวชาญ

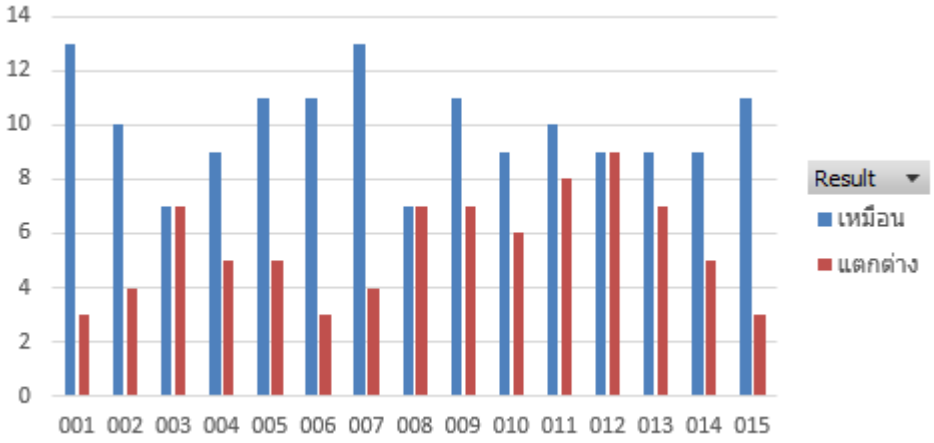
จากคลังข้อมูลที่ได้เก็บรวมรวมจากอาสาสมัครคอลเซ็นเตอร์ จำนวน 15 คน เสียงที่รวบรวมค่อนข้างมีความหลากหลายในเรื่องของระยะเวลา และลักษณะของเสียง อันเนื่องจากประสบการณ์การทำงานในด้านคอลเซ็นเตอร์ของแต่ละคน โดยอาสาสมัครคอลเซ็นเตอร์ที่มาสร้างคลังข้อมูลมีประสบการณ์การทำงานเริ่มตั้งแต่ 3 เดือน ไปจนถึงอายุงานมากกว่า 7 ปี ซึ่งบางคนมีประสบการณ์การเป็นคอลเซ็นเตอร์จากที่อื่นด้วย

เพื่อให้เห็นความแตกต่าง ในมุมของประสบการณ์การทำงานคอลเซ็นเตอร์ ในคนที่มีประสบการณ์เยอะและประสบการณ์น้อย มีความแตกต่างของการออกเสียง Smile Voice แตกต่างกันอย่างไร ผู้วิจัยได้ทำ Blind test กับ ผู้เชี่ยวชาญทางด้านคอลเซ็นเตอร์ ทั้งหมด 5 คนมา ซึ่งเป็นผู้ที่มีประสบการณ์การทำงานคอลเซ็นเตอร์มามากกว่า 5 ปี ซึ่งผู้วิจัยไม่ได้บอกให้ทราบว่า เสียงไหนถูกออกเสียงด้วยอารมณ์อะไร โดยผู้วิจัยได้สร้างหน้า Web Application ดังภาพที่ 3.22 เพื่อให้ผู้เชี่ยวชาญ สามารถเข้ามากดฟังและวิเคราะห์ ผ่านทาง Online โดยรวมทั้งผู้เชี่ยวชาญแต่ละคน จะไม่สามารถเห็นข้อมูลการ Label ของผู้เชี่ยวชาญคนอื่น ๆ เช่นกัน



ภาพที่ 3.22 Application Blind test Smile Voice และ Non-Smile Voice

ผลลัพธ์ที่ได้แสดงให้เห็นว่า เสียงที่คอลเซ็นเตอร์ออกเสียงตามเลเบล กับเสียงที่วิเคราะห์โดยผู้เชี่ยวชาญ ค่อนข้างมีความหลากหลาย ทั้งแยกแยะได้เหมือนเลเบลต้นฉบับ และแยกแยะได้ต่างกัน ดังผลลัพธ์ที่แสดงใน ภาพที่ 3.23



ภาพที่ 3.23 เปรียบเทียบเลเบลต้นฉบับของคอลเซ็นเตอร์แต่ละคน

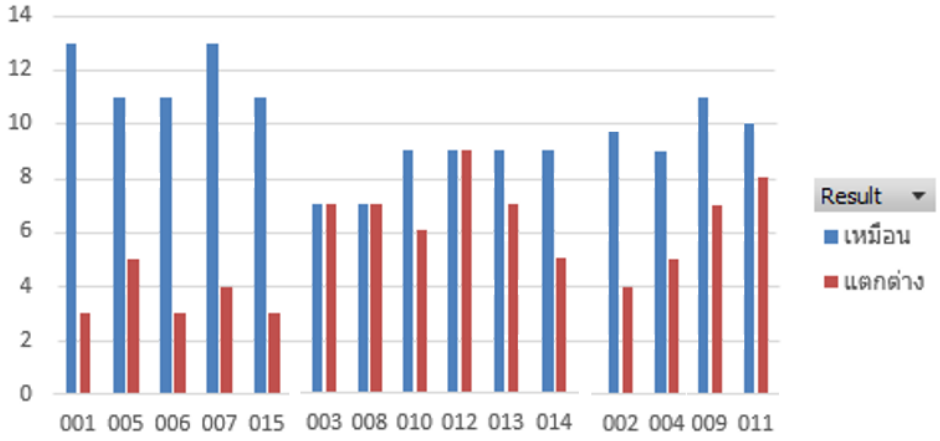
จากผลลัพธ์ในภาพที่ 3.23 เราสามารถแบ่งได้เป็น 3 ส่วนคือ 1. เหมือน มากกว่าแตกต่างเกิน 50% 2. เหมือน เท่ากับแตกต่าง หรือแตกต่างกันไม่เกิน 20% 3. เหมือนมากกว่าแตกต่างประมาณ 20-30%

จะเห็นได้ว่า กลุ่มที่ 1 คือ เหมือน มากกว่าแตกต่าง เกิน 50% เช่น คอลเซ็นเตอร์รหัส 001 005 006 007 015 ซึ่งในกลุ่มนี้จะเป็นคอลเซ็นเตอร์ที่มีประสบการณ์มากกว่า 2 ปีขึ้นไป

ในกลุ่มที่ 2 คือ เหมือน เท่ากับแตกต่าง หรือไกล้เคียงกัน เช่น คอลเซ็นเตอร์รหัส 003 008 012 013 014 ในกลุ่มนี้ จะเป็นในกลุ่มของคอลเซ็นเตอร์ที่มีประสบการณ์มากกว่า 2 ปีขึ้นไป เช่นกัน

และในกลุ่มสุดท้ายคือ กลุ่มที่ 3 คือ เหมือนมากกว่าแตกต่าง 20-30% เช่น คอลเซ็นเตอร์รหัส 002 004 009 011 ในกลุ่มนี้จะเป็นคอลเซ็นเตอร์ที่มีประสบการณ์น้อยกว่า 1 ปี

จากภาพที่ 3.24 แสดงการจัดวางกลุ่มของคอลเซ็นต์เตอร์ ตามกลุ่มของประสบการณ์ในการทำงานตามลำดับ

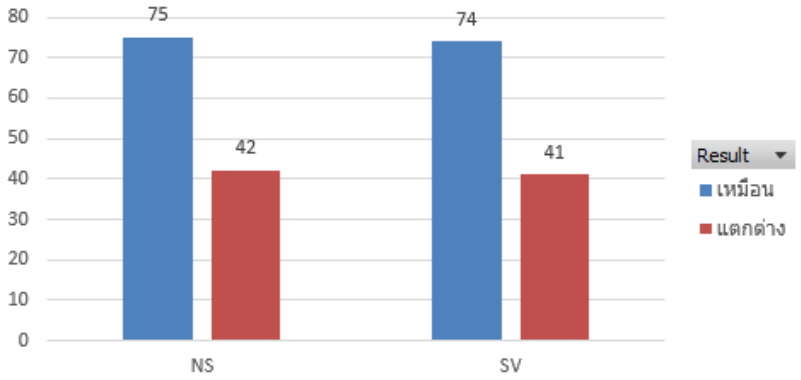


ภาพที่ 3.24 แสดงการจัดกลุ่มของคอลเซ็นเตอร์ตามกลุ่มของประสบการณ์

โดยในการพูดคุยกับอาสาสมัคร ถึงวิธีการในการออกเสียง Smile Voice และ Non-Smile Voice ในกลุ่มของผู้ที่มีประสบการณ์มากกว่า 2 ปี พวกเขาจะออกเสียงโดยธรรมชาติ ด้วยความคุ้นชินในการใช้เสียง ทำให้เสียงที่ได้ค่อนข้างไกล้เคียงกันโดยที่ไม่ต้องใช้ความพยายามในการดัดเสียงเข้ามาช่วยมาก หรือเป็นการออกเสียงโดยธรรมชาตินั่นเอง ในทางตรงกันข้ามคอลเซ็นเตอร์ที่มีประสบการณ์น้อยกว่า 1 ปี จะควบคุมการใช้เสียงได้ไม่ดีพอ ทำให้อาจจะมีการดัดเสียงในบางจังหวะมากเกินไปเพื่อให้ได้เสียงที่ต้องการ ซึ่งไม่ได้เกิดจากการพูดออกเสียงโดยธรรมชาติเหมือนกับคนที่มีประสบการณ์ ทำให้ผลลัพธ์ที่วิเคราะห์โดยผู้เชี่ยวชาญจะเป็นเสียงที่ค่อนข้างกำกวม (แตกต่าง 20-30%)

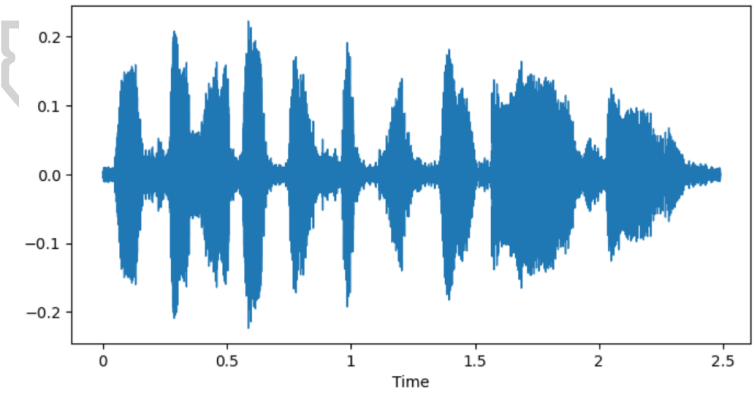
ในภาพรวมของเสียง Smile Voice และ Non-Smile Voice ที่วิเคราะห์โดยผู้เชี่ยวชาญ จากภาพที่ 3.25 จะเห็นได้ว่าผู้เชี่ยวชาญวิเคราะห์เสียงได้ตรงกับเลเบลต้นฉบับได้มากกว่า ทั้ง Smile

Voice และ Non-Smile Voice เนื่องจากอาสาสมัครคอลเซ็นเตอร์จำนวนกว่า 70% มีประสบการณ์มามากกว่า 2 ปี และในเสียงที่วิเคราะห์แล้วแตกต่างจากเลเบลต้นฉบับ ส่วนใหญ่เกิดจากอาสาสมัครกลุ่มที่มีประสบการณ์น้อยกว่า 2 ปี



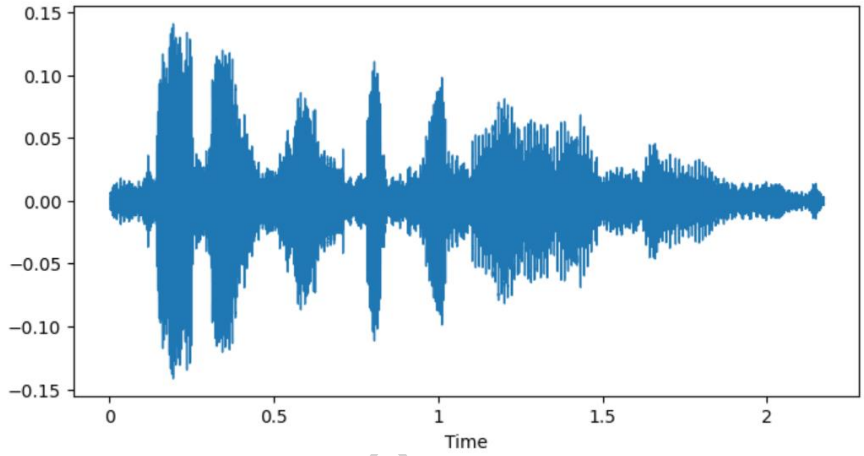
ภาพที่ 3.25 เลเบลต้นฉบับของคอลเซ็นเตอร์ เปรียบเทียบกับเสียงที่วิเคราะห์โดยผู้เชี่ยวชาญ

ตัวอย่างรูปแบบเสียงที่ผู้เชี่ยวชาญทั้ง 5 คนวิเคราะห์ไม่ตรงกับเลเบลต้นฉบับ โดยต้นฉบับเป็น Non-Smile Voice แต่ผู้เชี่ยวชาญทั้งหมดวิเคราะห์ว่าเป็น Smile Voice ในภาพที่ 3.26 เป็นตัวอย่างสัญญาณเสียง Non-Smile Voice ของประโยค “ยินดีค่ะ มีข้อมูลด้านอื่นๆ สอบถามเพิ่มเติมไหม คะ”



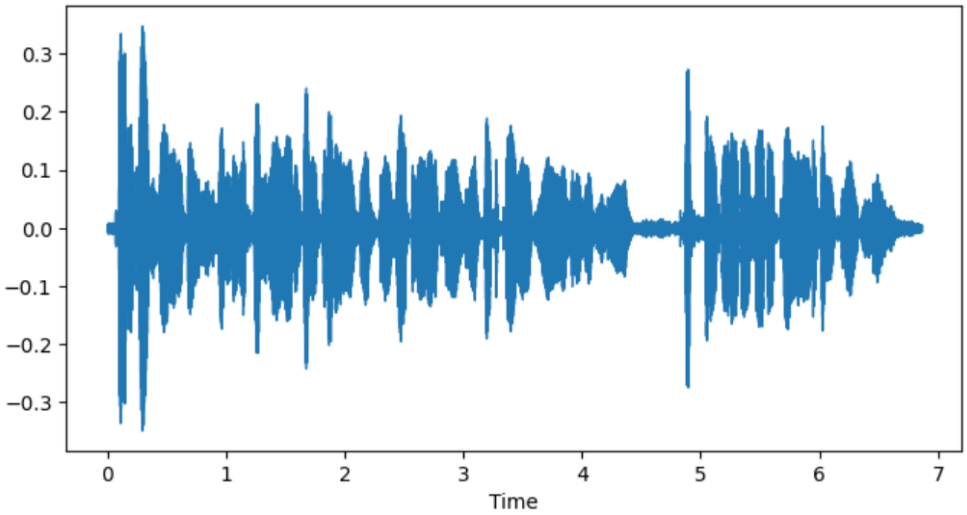
ภาพที่ 3.26 รูปแบบเสียงที่ผู้เชี่ยวชาญวิเคราะห์เป็นเสียงตรงกันข้าม

ตัวอย่างรูปแบบเสียงที่ผู้เชี่ยวชาญทั้ง 5 คน วิเคราะห์แตกต่างกัน นั่นคือส่วนหนึ่งจะเหมือนกับเลเบลต้นฉบับ และส่วนหนึ่งตรงข้ามกับเลเบลต้นฉบับ โดยรูปแบบสัญญาณเสียง เป็นเสียงเป็นของประโยค “สวัสดีค่ะ ต้องการสอบถามว่าอย่างไรคะ” ดังภาพที่ 3.27



ภาพที่ 3.27 รูปแบบเสียงที่ผู้เชี่ยวชาญวิเคราะห์เป็นเสียงที่แตกต่างกัน

ตัวอย่างรูปแบบเสียงที่ผู้เชี่ยวชาญทั้ง 5 คนวิเคราะห์เหมือนกับเลเบลต้นฉบับทั้งหมด ซึ่งในรูปแบบของสัญาณเสียงนี้เป็นเสียง Smile Voice ของประโยค “ก่อนวางสาย รบกวนเวลาสักครู่ ช่วยกดผลประเมินความพึงพอใจ ในการให้บริการ ขอบพระคุณที่ใช้บริการ สวัสดีค่ะ” ดังภาพที่ 3.28

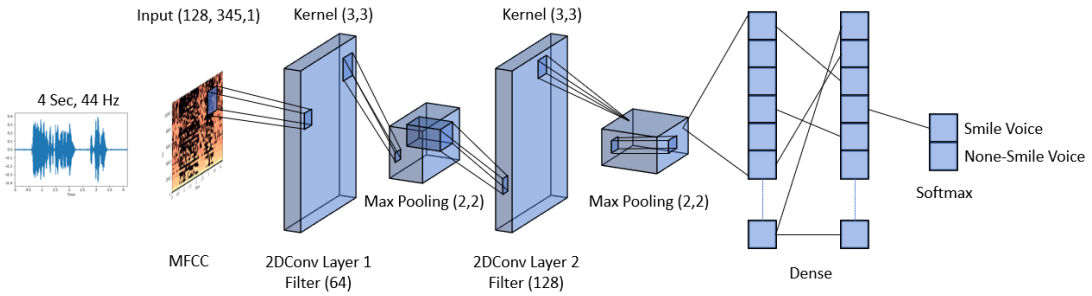


ภาพที่ 3.28 รูปแบบเสียงที่ผู้เชี่ยวชาญวิเคราะห์เป็นเสียงตรงกับเลเบลต้นฉบับ

จากรูปแบบเสียงทั้ง 3 แบบด้านบน จะเห็นได้ว่าเราไม่สามารถแยกความแตกต่างของสัญญาณเสียงด้วยตาเปล่าว่ารูปแบบเสียงนั้นเป็น Smile Voice หรือ Non-Smile Voice จากรูปภาพได้เลย ซึ่งผู้เชี่ยวชาญวิเคราะห์ออกมาได้ด้วยความรู้สึกจากการฟัง เพราะฉะนั้นถ้าไม่ใช้มนุษย์ในการวิเคราะห์ จึงจำเป็นที่จะต้องใช้ระบบคอมพิวเตอร์มาวิเคราะห์ความแตกต่างของเสียงโดยใช้การเรียนรู้เชิงลึกเข้ามาช่วยในการวิเคราะห์

3.6 ออกแบบ และสร้างโมเดลวิเคราะห์เสียงยิ้ม

โมเดล สำหรับวิเคราะห์เสียงยิ้ม โดยใช้ Dataset ของ อาสาสมัคร ทั้งที่เป็นคอลเซ็นเตอร์ มืออาชีพ และไม่ได้เป็นคอลเซ็นเตอร์ โดยในแบบจำลองนี้ จะวิเคราะห์ข้อมูลออกมาเป็น เสียงยิ้มและเสียงไม่ยิ้ม ซึ่งโครงสร้างของ โมเดล จะคล้ายแบบแรก แต่จะมี 2 Output คือ Smile และ Non-Smile ในส่วนของ การทำ Feature Extract จะทำทั้ง 2 แบบคือ Mel-Spectrogram และ MFCC โดย นำไปใช้กับ โมเดล ทั้งที่ เป็น 1D CNN และ 2D CNN เพื่อเปรียบเทียบประสิทธิภาพ ดังภาพที่ 3.29 เป็นโครงสร้างของ โมเดล Smile Voice และเนื่องจากข้อมูลมีขนาดไม่มากก่อนที่จะนำ Data เข้ามาในโมเดล ผู้วิจัยจึงได้ทำ Data Augmentation กับข้อมูล ก่อนส่งเข้าไปฝึกฝนโมเดล ซึ่งทำทั้ง 2 Dataset คือ Dataset ที่เป็น อาสาสมัคร และ Dataset ที่เป็นคอลเซ็นเตอร์จริง ๆ



ภาพที่ 3.29 Smile Voice โมเดล

3.7 ทดสอบประสิทธิภาพและปรับจูนพารามิเตอร์

การปรับจูนพารามิเตอร์กับแบบจำลอง โดยผู้วิจัยให้ความสำคัญของค่า accuracy, f1-score, precision และ recall โดยการหาค่าประสิทธิภาพที่แม่นยำที่สุด โดยในการทำ Feature Extraction ในส่วนของเสียงจะมี 2 รูปแบบ คือ Mel-Spectrogram และ MFCC ดังภาพที่ 3.30 การทำ Feature Extraction แบบ MFCC และภาพที่ 3.31 การทำ Feature Extraction แบบ Mel-Spectrogram

```
def feature_extract(audio_y, sr):
    y = librosa.util.fix_length(audio_y, size=sr*4)
#1D
    feature = np.mean(librosa.feature.mfcc(y=y, sr=sr, n_mfcc=128).T,axis=0)

    return feature
```

ภาพที่ 3.30 การทำ Feature Extraction แบบ MFCC

```python
def feature_extract(audio_y, sr):
    y = librosa.util.fix_length(audio_y, size=sr*4)

#1D
    feature = np.mean(librosa.feature.melspectrogram(y=y, sr=sr, n_fft=1024, hop_length=256, n_mels=128).T, axis = 0)
    return feature
```

ภาพที่ 3.31 การ ทำ Feature Extraction แบบ Mel-Spectrogram

และเพื่อลดขั้นตอนในการทำ Feature Extraction ทุกครั้งที่มีการเปลี่ยนแปลงโมเดล เมื่อทำ Feature Extraction แต่ละแบบเสร็จเรียบร้อยแล้ว จะ Save Feature เป็น File เก็บไว้เพื่อให้โมเดลต่าง ๆ มาดึงไปใช้ทุก ๆ ครั้งที่มีการปรับจูน และฝึกฝนโมเดลใหม่ ซึ่งการเตรียม File Feature และการ Save ข้อมูลแสดงในภาพที่ 3.32 เป็นการ save file โดยใช้ component pickle

```python
#for item in files[0:10]:
#Label 1:Smile Voice, 2:None Smile Voice
lst_feature = []
label = 1
counter = 1

for item in voice_path:
    if item[35:37] == "SV":
        label = 1
    elif item[35:37] == "NS":
        label = 2

    y, sr = librosa.load(item, res_type='kaiser_fast')
    feature = feature_extract(y, sr)
    lst_feature.append((feature, label))
    print("{0} : {1} : {2}".format(counter,label,item))
    counter = counter + 1

#Save Feature to File
import pickle
pickle.dump(lst_feature, open(path_pickle + model_file, "wb"))
```

ภาพที่ 3.32 ขั้นตอนการ Save file หลังจากทำ Feature Extraction

ในส่วนของการปรับจูนหาค่าประสิทธิภาพที่เหมาะสมโดยการทดสอบเปลี่ยนแปลงค่า ขนาดของ Filter ขนาดของ Kernal ใน CNN โมเดล และจำนวนชั้นความลึกของการฝึกฝนโมเดล รวมถึงการปรับ learning rate ดังตัวอย่าง Code ในภาพที่ 3.33 โดยในส่วน Filter มีการทดลองปรับขนาดตั้งแต่ 512 256 128 และ 64 ซึ่งในแต่ละชั้นก็จะมีการทดลองปรับขนาดที่แยกจากกัน ในส่วนของ Kernal มีการทดลองปรับขนาดตั้งแต่ 6 ลงมาเรื่อยๆ จนถึง 3 ในส่วนของความลึกมีการเพิ่มและลดความลึกจาก 4 3 และ 2 ชั้น และ learning rate ทดลองปรับเปลี่ยนจาก 0.0001 ไปเป็น 0.00001

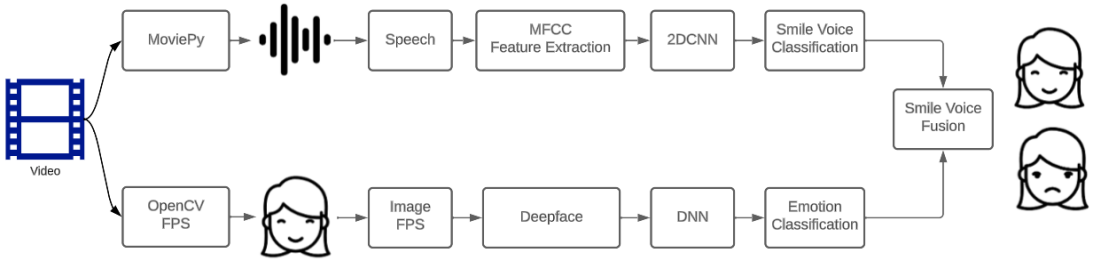
```
model.add(Conv2D(64, kernel_size=(3, 3), strides=(1, 1),activation='relu', input_shape=(120, 345, 1)))
model.add(Dropout(0.5))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Conv2D(64, kernel_size=(3, 3), activation='relu', kernel_regularizer=regularizers.l1_l2(l1=0.001, l2=0)))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.5))
model.add(Flatten())
model.add(Dense(2))
model.add(Activation('softmax'))

model.summary()

lr_schedule = tf.keras.optimizers.schedules.ExponentialDecay(
    initial_learning_rate=0.00001,
    decay_steps=30,
    decay_rate=0.9)

opt = tf.keras.optimizers.SGD(learning_rate=lr_schedule)
```

ภาพที่ 3.33 การปรับ Kernal size และ Learning rate

3.8 ออกแบบ Flow ของระบบต้นแบบสำหรับการวิเคราะห์เสียงและใบหน้า

ในการออกแบบ Flow การทำงาน เพื่อที่จะให้ง่ายต่อการนำโมเดลไปใช้งาน จะเป็นการเริ่มต้นจากไฟล์วิดีโอที่ต้องการตรวจเสียงยิ้ม นำมาผ่านกระบวนการวิเคราะห์เสียงโดยใช้วิธีการเรียนรู้เชิงลึก และแสดงผลลัพธ์เป็นรูปภาพ และผลของการวิเคราะห์เสียงตาม Timeline ของวิดีโอที่ส่งเข้าไปมาในระบบ

ในขั้นตอนการทำงาน สามารถอธิบายได้จากภาพที่ 3.34 โดยเริ่มจากการนำไฟล์วิดีโอการฝึกอบรมการพูดด้วยเสียงที่มีรอยยิ้ม ของผู้ที่เราต้องการตรวจสอบมาแยกภาพและเสียงออกจากกัน เพื่อนำทั้งข้อมูลทั้ง 2 แบบไปแยกแยะอารมณ์จากเสียงพูด และใบหน้า โดยไฟล์ภาพจะนำไปผ่านกระบวนการวิเคราะห์หารอยยิ้ม ส่วนไฟล์เสียงจะนำไปผ่านกระบวนการวิเคราะห์หาเสียงยิ้ม จากนั้นนำผลลัพธ์ที่ได้จากทั้งภาพและเสียงมาประกอบกันเพื่อหาความสัมพันธ์กันระหว่างเสียงยิ้มกับใบหน้าในกรณีถ้าวิเคราะห์เสียงได้เป็นเสียงยิ้ม ใบหน้าจะยิ้มหรือไม่



ภาพที่ 3.34 Process Flow สำหรับการวิเคราะห์เสียงและภาพ

3.9 การออกแบบและพัฒนาระบบ

ในขั้นตอนนี้จะแบ่งออกเป็น 2 ส่วนคือ ส่วนที่เป็นวิดีโอและส่วนที่เป็นเสียงพูด โดยการแยกวิดีโอออกมาเป็นไฟล์ภาพให้สอดคล้องกับช่วงเวลาต่างๆ ในการแยกภาพออกมาจากวิดีโอ ผู้วิจัยใช้ OpenCV เพื่อแยกภาพ Keyframe ออกมาจากไฟล์วิดีโอ โดยจะแยกภาพออกมาทุกๆ keyframe ที่สอดคล้องกับวินาที ซึ่งจำนวน keyframe ในแต่ละวินาที ขึ้นอยู่กับคุณภาพของวิดีโอที่นำมาใช้ เช่น วิดีโอที่ 10 fps (frame per second) หมายถึง ใน 1 วินาที จะมีจำนวน 10 keyframe โดยในงานวิจัยนี้เลือกใช้ความละเอียดวิดีโอที่ 30 fps และในทุกๆ 4 วินาที จะดึงรูปภาพจำนวนมาจำนวน 4 ภาพ นั่นคือ วินาทีละภาพ และนำรูปภาพที่ได้ไปเก็บไว้ในพื้นที่ของระบบที่เตรียมไว้ เพื่อนำไปใช้วิเคราะห์อารมณ์บนใบหน้าในลำดับถัดไป ดังภาพที่ 3.35 เป็นตัวอย่างของรูปภาพที่ดึงมาจาก keyframe ที่สอดคล้องกับเวลา 1 วินาที

ซึ่งในส่วนของการที่ดึงรูปภาพในแต่ละเวลา 1 วินาที เนื่องจากในการพูดของคอลเซ็นเตอร์จะต้องเป็นการพูดด้วยความเร็วปกติจนค่อนไปทางช้า เพราะต้องสื่อสาร ให้คนฟังให้เข้าใจและได้ยินชัดเจนที่สุด และในระหว่างที่พูดกล้ามเนื้อบนใบหน้าซึ่งมีความสัมพันธ์กันก็จะเปลี่ยนแปลงตามไปด้วย เพราะฉะนั้นเมื่อพูดในความเร็วปกติ การเปลี่ยนแปลงของใบหน้าจึงไม่ได้เปลี่ยนแปลงเร็วมาก ทางผู้วิจัยจึงเลือกดึงรูปภาพในแต่ละเวลา 1 วินาทีมาใช้งาน เพราะสามารถเห็นการเปลี่ยนแปลงได้ชัดเจน ตามภาพที่ 3.35

ภาพที่ 3.35 ตัวอย่างภาพจากวิดีโอ ที่ดึงรูปภาพออกมาทุก 1 วินาที

การวิเคราะห์อารมณ์บนใบหน้า (Model Facial Emotions) ในงานวิจัยนี้จะประยุกต์ใช้ Deepface (Serengil & Ozpinar, 2020) ในการตรวจสอบอารมณ์ของใบหน้า

DeepFace เป็นไลบรารีการจดจำใบหน้าและการวิเคราะห์คุณลักษณะใบหน้าที่มีขนาดเล็กสำหรับ Python ไลบรารี DeepFace ใช้โมเดลสำหรับการจดจำใบหน้าหลายแบบ คือ VGG-Face, Google FaceNet, OpenFace, Facebook DeepFace, DeepID, ArcFace, Dlib และ SFac ซึ่งสามารถเลือกใช้ได้ผ่านทาง Parameter โดยไลบรารี DeepFace มีความสามารถต่าง ๆ ดังนี้

1. การตรวจสอบใบหน้า : โดยการเปรียบเทียบใบหน้ากับใบหน้าอื่นเพื่อตรวจสอบว่าตรงกันหรือไม่
2. การจดจำใบหน้า : หมายถึงการค้นหาใบหน้าในฐานข้อมูลรูปภาพ ที่ทาง Deepface มีเก็บไว้
3. การวิเคราะห์ลักษณะใบหน้า : หมายถึงการอธิบายคุณสมบัติการมองเห็นของภาพใบหน้า เช่น อายุ เพศ อารมณ์
4. การวิเคราะห์ใบหน้าแบบเรียลไทม์ : สามารถจดจำใบหน้า และการวิเคราะห์ลักษณะใบหน้าด้วย วิดีโอแบบเรียลไทม์

ในการแยกไฟล์เสียงออกมาจาก File Video จะใช้ python ไลบรารี moviepy ซึ่งจะทำหน้าที่ในการแยกเสียงออกมาจากไฟล์ภาพเคลื่อนไหว และแปลงเป็นไฟล์นามสกุล .wav จากนั้นจะแบ่งไฟล์เสียงออกเป็นไฟล์ย่อย ๆ ตามช่วงเวลา โดยใช้ python ไลบรารี pydub Audio Segment โดยแยกเสียงเป็นออกไฟล์ย่อย ๆ เพื่อเตรียมข้อมูลก่อนนำไปใช้กับเทคนิคการเรียนรู้เชิงลึก

3.10 ประยุกต์โมเดลวิเคราะห์เสียงยิ้ม เพื่อแสดงความสัมพันธ์ของภาพและเสียง

ในการประยุต์ใช้โมเดลวิเคราะห์เสียงยิ้ม ในส่วนของเสียงทางผู้วิจัยได้นำผลลัพธ์ที่ได้จากการฝึกฝนโมเดล ที่ได้ประสิทธิภาพดีที่สุด มาทำ API เชื่อมต่อให้สามารถเรียกใช้งานได้ผ่านโปรแกรม โดยเป็นการส่งข้อมูลของวิดีโอที่ที่เปลี่ยนรูปแบบให้เป็นรูปแบบ Base64 ส่งเข้ามาผ่านจาก Internet ซึ่งภายใน API จะทำหน้าที่ แยกภาพและเสียง เพื่อส่งไปประมวลผลตาม Process ในภาพที่ 3.34 ซึ่งแสดงให้เห็นการ Process ในขั้นตอนต่างๆ จนกระทั่งส่งข้อมูลกลับมาเป็นผลลัพธ์ของโมเดล โดยเป็นการส่งรูปภาพที่แยกออกมา พร้อมกับอารมณ์ของรูปภาพที่เป็นรูปแบบ Base64 และอารมณ์ของเสียงตาม Timeline ที่บันทึกไฟล์วิดีโอส่งเข้าไป จากภาพที่ 3.36 เป็นการรับข้อมูลจาก API และประมวลผมแยกรูปภาพ และเสียงก่อนนำไปเข้า โมเดล ที่เตรียมไว้ ซึ่ง ทางผู้วิจัยใช้ Python FastAPI ในการสร้าง API แล้วสร้าง Method Smile เพื่อรับ Video ที่เป็น Base64 เข้ามาประมวลผล

```python
@app.post("/smile_video/", response_model=list[EmoOut])
async def smile(base64: Base64):

    now = datetime.now(tz=ZoneInfo("Asia/Bangkok"))
    # dd/mm/YY H:M:S
    dt_string = now.strftime("%m%d_%H%M")
    parent_dir = "SmileVoice_Temp"
    path = os.path.join(parent_dir, dt_string)
    os.mkdir(path)

    #save video & extract
    save_video_to_file_and_extract(base64.video_base64, path, base64.file_name)

    #face emotion
    df = pd.DataFrame(extract_face_emotion(path))
    lst = []
    for index, row in df.iterrows():
        print (row["File_name"])
        lst.append(EmoOut(file_name=row["File_name"], emotion=row["Emo"], base64=row["Base64"]))

    return lst
```

ภาพที่ 3.36 Code การรับข้อมูลจาก API และประมวลผล

ในการเรียกใช้งาน API ผู้วิจัยได้สร้าง Web Application ขึ้นมาเพื่อให้สามารถ Upload Video ที่เป็นการฝึกพูดเสียงยิ้ม โดยจะต้องถ่ายหน้าตรง ให้เห็นใบหน้าชัดเจน แล้วพูดกับกล้อง จากนั้นกด Upload เข้ามาใน Application ก่อนที่จะส่งผลลัพธ์กลับไปแสดงผลที่ Web Application เดียวกัน

บทที่ 4
ผลการดำเนินงานวิจัย

จากขั้นตอนการดำเนินงานที่ได้กล่าวถึงในบทที่ 3 วิทยานิพนธ์นี้ได้ออกแบบ และสร้างโมเดลแบบ Convolution Neural Network หลายรูปแบบ โดยปรับจูนโมเดลสำหรับชุดข้อมูล 3 ชุด ดังนั้นเพื่อรายงานผลการดำเนินงานวิจัย จึงแบ่งเนื้อหาในส่วนของผลการดำเนินงานวิจัย ดังนี้

4.1) ผลการทดสอบและเปรียบเทียบประสิทธิภาพของโมเดลวิเคราะห์อารมณ์จากเสียงพูด (Thai SER Model)

4.2) ผลการนำเสียงของคอลเซ็นเตอร์ ทดสอบวิเคราะห์อารมณ์ผ่าน Thai SER Model

4.3) ผลการทดสอบและเปรียบเทียบประสิทธิภาพของโมเดล Smile Voice โดยชุดข้อมูลเสียงอาสาสมัคร

4.4) ผลการทดสอบและเปรียบเทียบประสิทธิภาพของ โมเดล Smile Voice โดยชุดข้อมูลเสียงคอลเซ็นเตอร์

4.5) การประยุกต์ใช้งาน โมเดล Smile Voice

แต่ละการดำเนินงานข้างต้น มีรายละเอียดดังต่อไปนี้

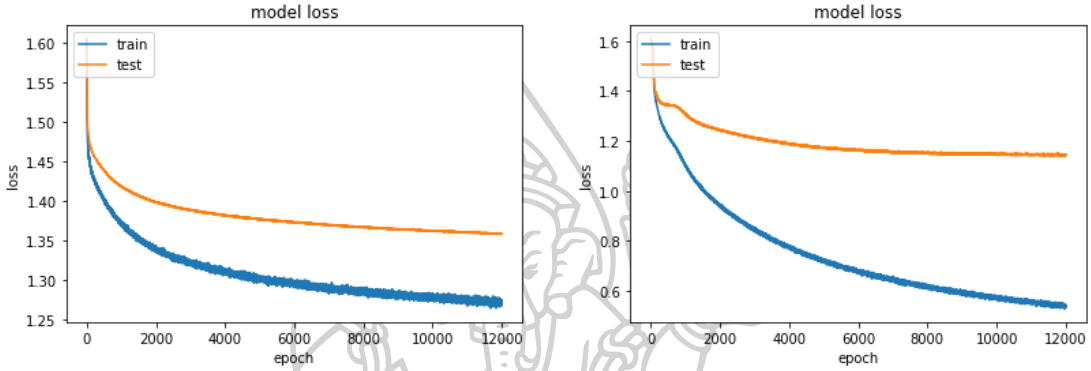
4.1 ผลการทดสอบและเปรียบเทียบประสิทธิภาพของโมเดลวิเคราะห์อารมณ์จากเสียงพูด (Thai SER Model)

การสร้างแบบจำลอง Thai SER จากคลังข้อมูลของสถาบันวิทยสิริเมธี (Vidyasirimedhi Institute of Science and Technology: VISTEC) โดยใช้ 1DCNN และ 2DCNN โดยทั้ง 2 โมเดล มีการทำ Feature Extraction 2 แบบคือ Mel-Spectrogram และ Mel-Frequency Cepstrum Coefficients (MFCC) นั่นคือ 1DCNN + Mel-Spectrogram, 1DCNN + MFCC และ 2DCNN + Mel-Spectrogram, 2DCNN + MFCC รวมทั้งหมด 4 แบบจำลอง ซึ่งผลลัพธ์ได้นำมาเปรียบเทียบผ่านทางค่า Accuracy, Loss และ Confusion Matrix เพื่อดูประสิทธิภาพของแบบจำลองแต่ละแบบ

4.1.1 การประเมินผลแบบจำลอง Thai SER โดยใช้ Loss Value

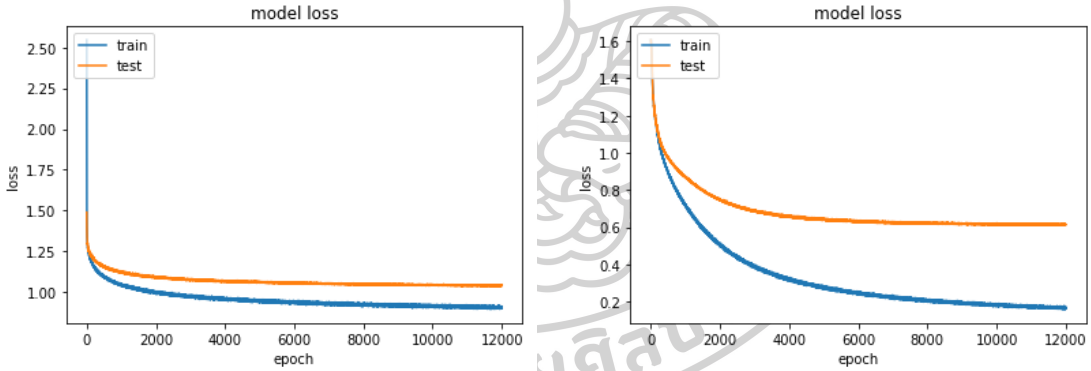
ค่า Loss ของ โมเดล Mel-Spectrogram + 1D Conv แสดงให้เห็นว่ามีนาวโน้มลดลง แต่ค่า Loss ยังค่อนข้างสูงเมื่อทียบกับโมเดลอื่น ใน โมเดล Mel-Spectrogram + 2D Conv ค่า Val loss

จะมีเพิ่มขึ้นในช่วง EPOCHS 400 และหลังจากจาก EPOCH ที่ 8,000 ค่า Test loss ลดลงน้อยมาก ในส่วนของ โมเดล MFCC + 2D Conv ค่า loss เริ่มลดลงที่ EPOCH 1,000 แต่ค่า test loss ยังคงอยู่ที่ ประมาณ 1 อย่างไรก็ตามค่า Loss ของ โมเดล MFCC + 2D Conv ยังน้อยกว่า 2 โมเดลข้างต้น และใน โมเดล MFCC + 2D Conv ค่า loss ลงลงอย่างเห็นได้ชัดตั้งแต่ 0 - 6000 จากนั้นค่า loss ค่อนข้างค่าที่ ทั้งค่า loss train และ val โดยโมเดล นี้ให้ค่า loss ที่ดีที่สุด จากภาพที่ 4.1 แสดงให้เห็นถึงค่าประสิทธิภาพค่า Loss ของ โมเดล ทั้งหมด



(ก) Mel-Spectrogram + 1DConv

(ข) Mel-Spectrogram + 2DConv



(ค) MFCC + 1DConv

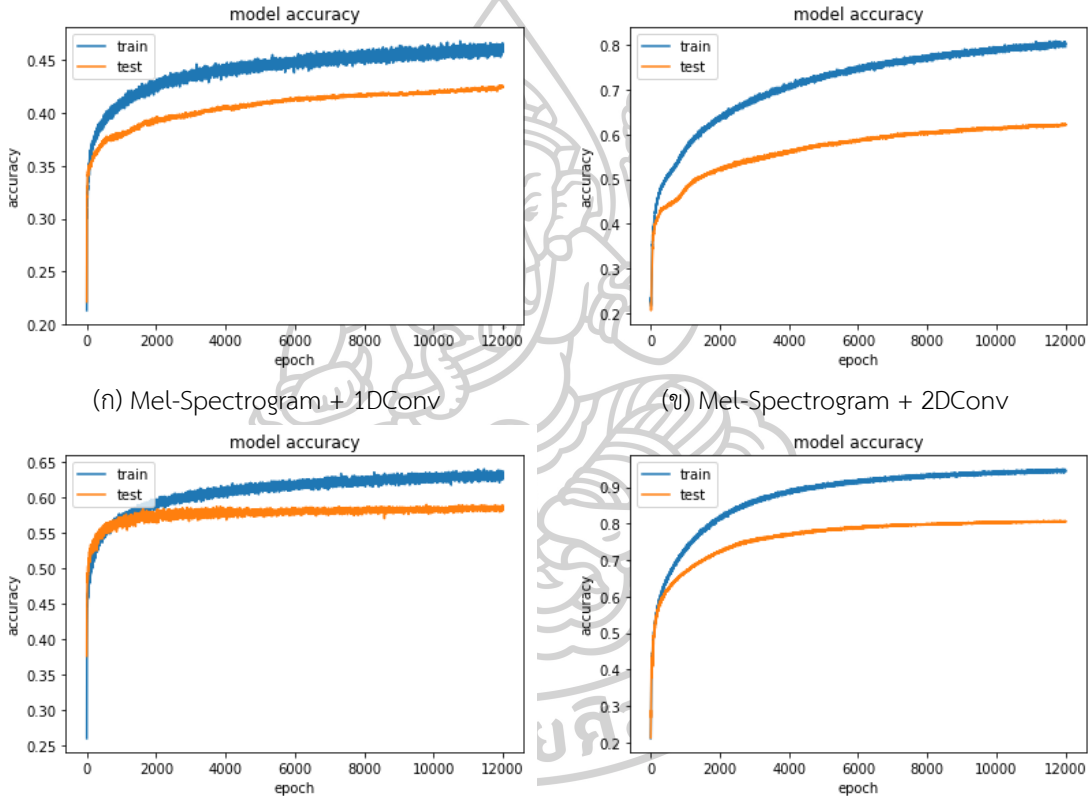
(ง) MFCC + 2DConv

ภาพที่ 4.1 ค่า Loss ของ Thai SER ในแต่ละ โมเดล

4.1.2 การประเมินผลแบบจำลอง Thai SER โดยใช้ Accuracy Value

จากภาพที่ 4.2 เป็นการเปรียบเทียบค่า Accuracy โมเดล ทั้ง 4 แบบ ค่า Accuracy ของ โมเดล Mel-Spectrogram + 1D CNN มีแนวโน้มเพิ่มขึ้นแต่ก็ยังถือว่าน้อย เมื่อเทียบกับ โมเดล อื่นๆ โมเดล Mel-Spectrogram + 2D CNN ค่า Accuracy มีแนวโน้มเพิ่มขึ้นอย่างเห็นได้ชัดโดยเริ่มตั้งแต่ epoch 100 จนถึง 1,800 และยังเพิ่มขึ้นต่อไป แต่ความแตกต่างระหว่างค่าของ train และ val ยัง

ค่อนข้างสูงเมื่อเทียบกับโมเดลอื่น รวมถึงยังมากกว่าโมเดล Mel-Spectrogram + 1D CNN ด้วย สำหรับ MFCC + 1D Conv ค่า Accuracy test เริ่มคงที่ประมาณ epoch 3,000 แล้วเพิ่มขึ้นไม่มาก จนถึง epoch 12,000 โดยค่า train และ val ดีกว่า Mel-Spectrogram + 1D CNN แต่ยังต่ำกว่า Mel-Spectrogram + 2D CNN โดยค่าของ Accuracy ที่ดีที่สุดคือ โมเดล MFCC + 2D CNN โดยมีค่าเพิ่มขึ้นเริ่มตั้งแต่ epoch ที่ 100 – 8,000 จากนั้นการเพิ่มขึ้นค่อนข้างคงที่ ทั้งค่า train และ val โดยที่ accuracy ของ โมเดล MFCC + 2D CNN มีค่าดีที่สุดใน โมเดล ที่ผ่านมา จากภาพที่ 4.2 แสดงให้เห็นถึงค่าประสิทธิภาพค่า Loss ของ โมเดล ทั้งหมด



(ก) Mel-Spectrogram + 1DConv (ข) Mel-Spectrogram + 2DConv

(ค) MFCC + 1DConv (ง) MFCC + 2DConv

ภาพที่ 4.2 ค่า Accuracy ของ Thai SER ในแต่ละ โมเดล

การเปรียบเทียบค่า Accuracy และ Loss ในแต่ละแบบ ดังแสดงในตารางที่ 1 จะเห็นได้ว่า โมเดล Mel-Spectrogram + 1D Conv มีประสิทธิภาพต่ำที่สุดนั่นคือ ค่า Accuracy 0.4251 ซึ่งน้อยกว่าทุกโมเดล และค่า Loss สูงสุด 1.3579 แต่เมื่อมีการเปลี่ยนเป็น 2D Conv พบว่าประสิทธิภาพของโมเดลดีขึ้น และในทางเดียวกับ เมื่อเปลี่ยนการทำสกัด Feature ของเสียงเป็น

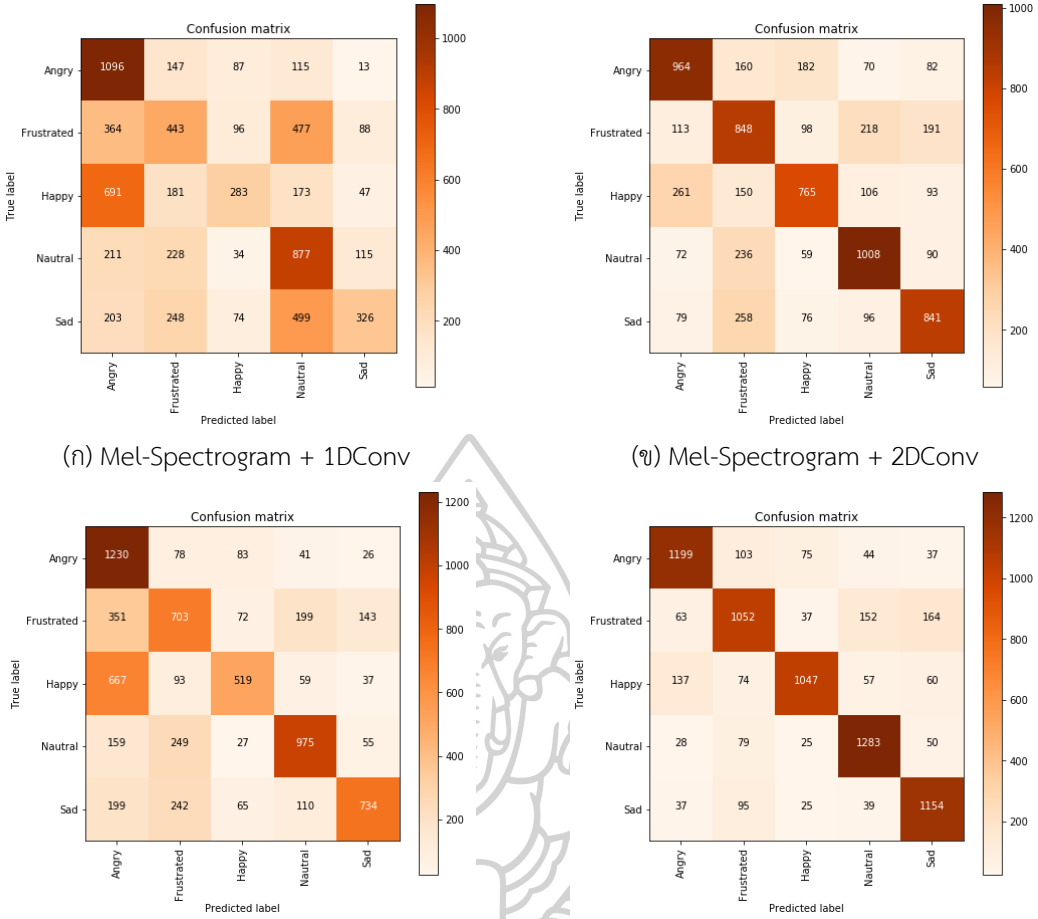
MFCC โดยใช้ 1D Conv เหมือน ก็ทำให้ประสิทธิภาพเพิ่มขึ้น จะสังเกตได้ว่า ในมุมของ Model ดีสุด คือ 2D Conv และ การสกัด Feature ดีที่สุดคือ MFCC เพราะฉะนั้นเมื่อนำ 2 วิธีมารวมกัน จะเป็น MFCC + 2D Conv ทำให้เป็นโมเดลที่ทำประสิทธิภาพดีที่สุดคือได้ค่า Accuracy สูงสุดนั่นคือ 0.8059 และ ค่า Loss ต่ำที่สุดคือ 0.6140 ดังแสดงใน ตารางที่ 1

ตารางที่ 1 เปรียบเทียบค่า Accuracy และ Loss ของ โมเดล ThaiSER แต่ละแบบ

Model	Accuracy	Loss
Mel-Spectrogram + 1D Conv	0.4251	1.3579
Mel-Spectrogram + 2D Conv	0.6220	1.1464
MFCC + 1D Conv	0.5847	1.0418
MFCC + 2D Conv	0.8059	0.6140

4.1.3 การประเมินผลแบบจำลอง Thai SER โดยใช้ Confusion Matrix Value

ภาพที่ 4.3 แสดงการเปรียบเทียบ Confusion Matrix ของ โมเดล ทั้ง 4 แบบ โดยไล่เรียงจาก โมเดล ที่มีค่าการ Classify น้อยที่สุดไล่ไปจนถึงดีที่สุดคือ Mel-Spectrogram + 1D CNN, MFCC + 1D CNN, Mel-Spectrogram + 2D CNN และ โมเดล ที่มีค่า Confusion Matrix ดีที่สุดคือ MFCC + 2D CNN โดยดูจากค่าความถูกต้องของการวิเคราะห์ในแต่ละอารมณ์ โดยเสียงอารมณ์ โกรธ จะมีค่าการวิเคราะห์ได้ถูกต้องค่อนข้างมาก และเสียงอารมณ์มีความสุข จะได้ความถูกต้องน้อยที่สุด



(ก) Mel-Spectrogram + 1DConv

(ข) Mel-Spectrogram + 2DConv

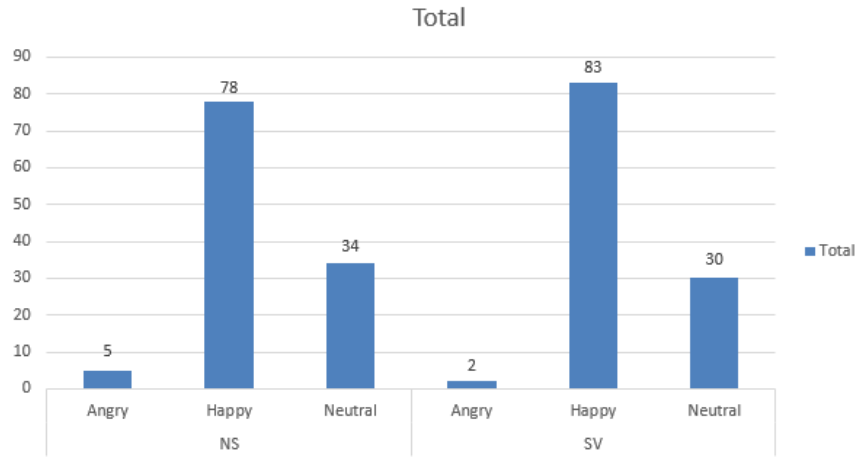
(ค) MFCC + 1DConv

(ง) MFCC + 2DConv

ภาพที่ 4.3 ค่า Confusion Matrix ของ Thai SER ในแต่ละ โมเดล

4.2 ผลการนำเสียงของคอลเซ็นเตอร์ ทดสอบวิเคราะห์อารมณ์ผ่าน Thai SER Model

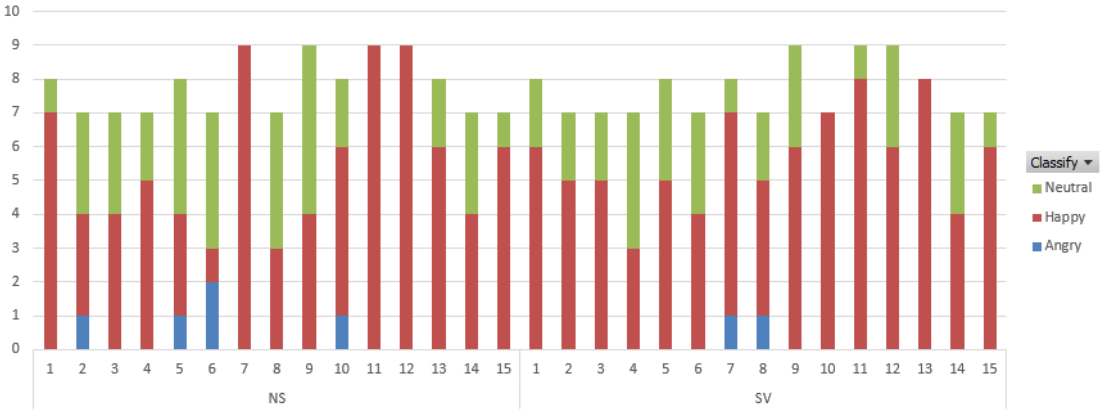
จากที่ผู้วิจัยได้เก็บชุดข้อมูลจากทีมงานคอลเซ็นเตอร์ ซึ่งมีเลเบล เป็น Smile Voice และ Non-Smile Voice โดยขั้นตอนการอัดเสียงผู้วิจัยจะให้คอลเซ็นเตอร์ พูดประโยคที่เตรียมไว้เป็นเสียงยิ้มและไม่เป็นเสียงยิ้ม จากนั้นทางผู้วิจัยได้นำชุดข้อมูลที่เก็บ มาวิเคราะห์กับโมเดล Thai SER ซึ่งเป็นโมเดล วิเคราะห์อารมณ์ในภาษาไทยที่ผู้วิจัยได้พัฒนาขึ้นมา เพื่อวิเคราะห์เสียงในชุดข้อมูลว่าสามารถตรวจสอบได้เป็นอารมณ์อะไร โดยได้มีระบุความยาวของเสียงเป็น 4 วินาที ในส่วนของไฟล์เสียงที่มีขนาดน้อยกว่า 4 วินาที จะเป็นการ เพิ่ม silence ให้ได้ 4 วินาที และไฟล์ที่มีความยาวมากกว่า 4 วินาที จะเลือกใช้เฉพาะ 4 วินาทีแรกเท่านั้น ซึ่งจำนวนของไฟล์เสียงที่นำไปวิเคราะห์มีจำนวน 232 ไฟล์ ซึ่งผลการวิเคราะห์เป็นได้แสดงในรูปภาพที่ 4.4



ภาพที่ 4.4 การวิเคราะห์ Dataset ด้วย Thai SER

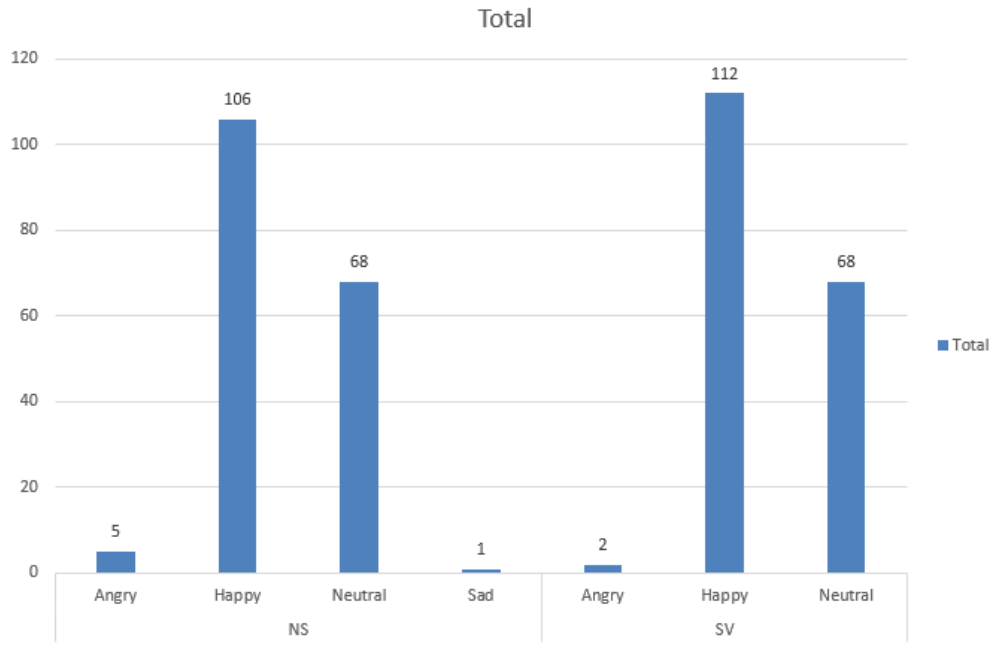
โดยจากภาพที่ 4.4 จะสังเกตได้ว่าส่วนใหญ่จะเป็นอารมณ์ Happy และ Neutral และมี Angry จำนวน 2 ไฟล์ ซึ่งเกิดขึ้นใน Dataset กับเสียงที่เป็น Smile Voice และในส่วนของ Smile Voice แบบจำลอง Thai SER จะวิเคราะห์ได้ Emotion Happy มากกว่าทางด้าน Non-Smile Voice

และถ้าวิเคราะห์ แยกตามคอลเซ็นเตอร์ แต่ละคน จะเห็นได้ว่า ทุกคนก็จะมีทั้ง Emotion ที่เป็น Neutral และ Happy ผสมกันไปทั้งใน Smile Voice และ Non-Smile Voice ไม่ได้ กระจุกตัวอยู่ที่คนใดคนนึงเป็นพิเศษ ตามภาพที่ 4.5



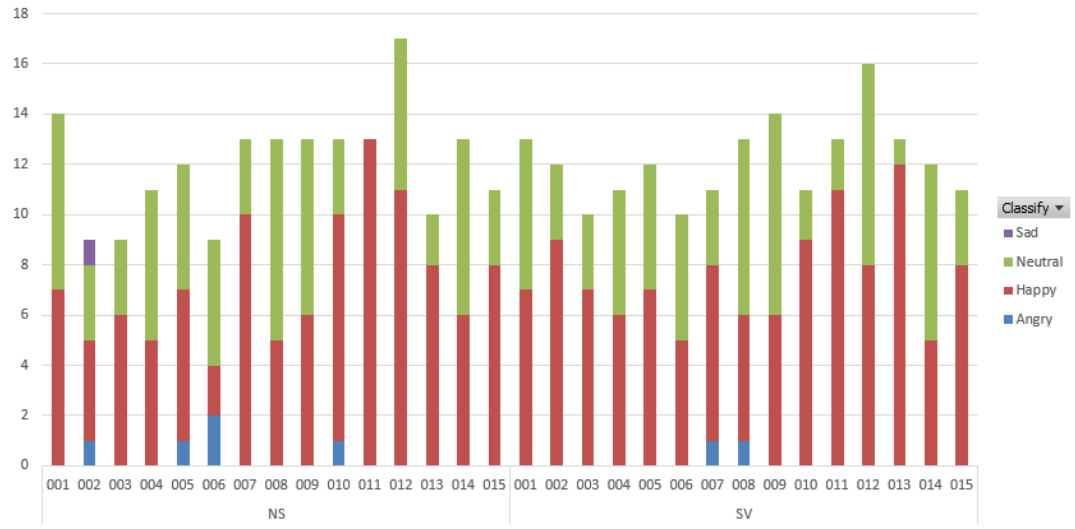
ภาพที่ 4.5 การวิเคราะห์ Dataset ด้วย Thai SER แยกรายคน

จากภาพที่ 4.6 เป็นการนำเสียงทั้งประโยคเข้ามาวิเคราะห์ด้วย โดยนำประโยคที่มีความยาวมากกว่า 4 วินาทีมาแบ่งเป็นประโยคย่อย ความยาวประโยคละ 4 วินาที เช่น ประโยคที่มีความยาว 12 วินาที จะแบ่งได้เป็น 3 ไฟล์ ซึ่งหลังจากการแบ่งเป็นไฟล์ย่อย จะทำให้มีไฟล์เพิ่มขึ้นเป็น 362 ไฟล์ และนำมาวิเคราะห์อารมณ์กับโมเดล Thai SER จะเห็นได้ว่า เสียงที่ Classify ว่าเป็น Happy ในส่วนของ Smile Voice จะมีจำนวนมากกว่าเสียงที่ Classify ว่าเป็น Happy ใน Non-Smile Voice และเสียงที่เป็น Angry ในส่วนของ Non-Smile Voice จะมากกว่า Smile Voice และในส่วนของ Neutral จะมีพอๆ กันทั้ง Smile Voice และ Non-Smile Voice ซึ่งจะสอดคล้องกับ ภาพที่ 4.4 นั่นคือ อารมณ์ของเสียงในแต่ละประโยคไม่ว่าจะสั้นหรือยาว จะยังคงอารมณ์ในประโยคไว้เหมือนเดิม ทั้งนี้ เนื่องจากประโยคที่พูดไม่ได้มีความยาวมาก



ภาพที่ 4.6 ผลลัพธ์การ Classify ด้วย Thai SER ของเสียงทั้งประโยค

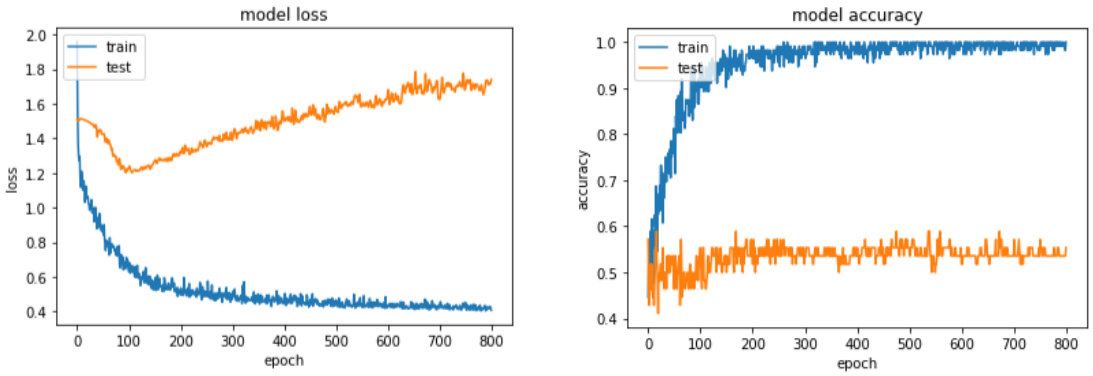
ในส่วนขอผลลัพธ์การวิเคราะห์ Emotion ของคอลเซ็นเตอร์ แต่ละคนผลลัพธ์ ของ Dataset ทั้งประโยค ค่าที่ได้จะคล้ายกับการใช้ประโยคเพียง 4 วินาทีแรก นั่นคือทุกคนที่พูดจะมีทั้ง Happy และ Neutral ไม่ว่าจะเป็นเสียงที่ Smile Voice และเสียง Non-Smile Voice ดังภาพที่ 4.7 ซึ่งจะเห็นได้ว่าข้อมูลสอดคล้องกับ ภาพที่ 4.5 เช่นกัน นั้นคือไม่ว่าจะนำประโยคมาใช้ทั้งประโยค ทุกคนก็จะมีทั้ง Emotion ที่เป็น Neutral และ Happy ผสมกันไปทั้งใน Smile Voice และ Non-Smile Voice ไม่ได้ กระจุกตัวอยู่ที่คนใดคนนึงเป็นพิเศษ



ภาพที่ 4.7 ผลลัพธ์การ Classify ด้วย Thai SER ของเสียงทั้งประโยค รายคน

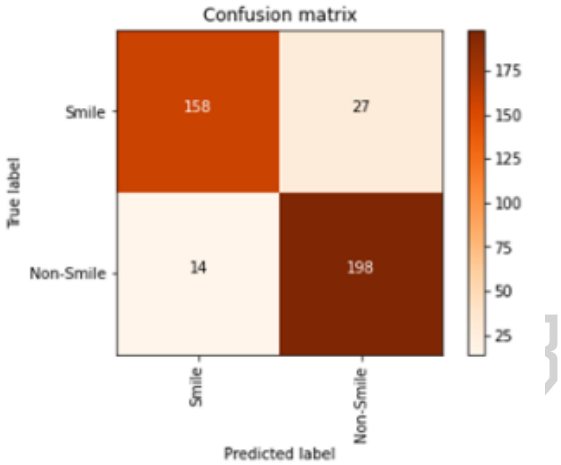
4.3 ผลการทดสอบและเปรียบเทียบประสิทธิภาพของโมเดล Smile Voice โดยชุดข้อมูลเสียงอาสาสมัคร

ในการทดสอบประสิทธิภาพของโมเดล Smile Voice จะเป็นก่อน และหลังการทำ Data Augmentation โดย Dataset ชุดที่ใช้ เป็นอาสาสมัครที่มาออกเสียง Smile Voice และ Non-Smile Voice โดยจะนำเข้าโมเดล แบบเดียวกัน แต่จะแตกต่างกันในจำนวนของ Dataset ที่นำเข้า นั่นคือ รอบแรกจะเป็น Dataset ที่ไม่มีการทำ Data Augmentation และรอบที่สองจะใช้ Dataset ที่ผ่านการทำ Data Augmentation โดย Dataset ทั้ง 2 ชุด ในขั้นตอนการสกัด Feature จะใช้ MFCC และใช้ 2DCNN เพื่อเทรนโมเดล และจากผลลัพธ์ที่ได้จากการทดลองก่อนหน้า วิธี 2DCNN + MFCC จะได้ค่าที่มีประสิทธิภาพดีที่สุด ซึ่งในการทดลองนี้มุ่งเน้นที่การเพิ่มประสิทธิภาพ โดยการเพิ่มจำนวนของ Dataset ด้วยการทำ Data Augmentation ซึ่งจะเห็นความแตกต่างได้ ดังภาพที่ 4.8 เป็นประสิทธิภาพของ โมเดล ที่เกิดจากทางใช้ Dataset อาสาสมัครที่ยังไม่ได้ ทำ Data Augmentation จะเห็นได้ว่าในโมเดล Loss เกิด overfit ขึ้นที่ epoch ที่ 100 เท่านั้น และมีค่าเพิ่มขึ้นอย่างต่อเนื่อง โดยค่า Loss เพิ่มขึ้นจนถึง epoch สุดท้าย และยังมีแนวโน้มที่จะเพิ่มขึ้นต่อไปได้เรื่อยๆ และในโมเดล Accuracy ค่าที่ได้มีค่าขึ้นลง อยู่ระหว่าง 0.5 – 0.6 และไม่มีแนวโน้มเพิ่มขึ้นจนถึง epoch สุดท้าย



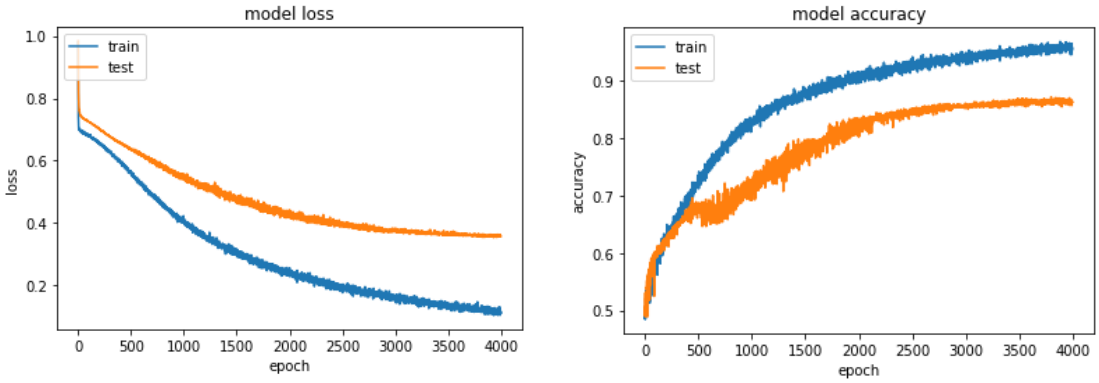
ภาพที่ 4.8 ค่า Loss และ Accuracy Smile Voice โมเดล Non-Call Center

Confusion Matrix ของผลลัพธ์ที่ได้จาก Dataset ที่ไม่ได้ทำ Data Augmentation ดังแสดงดังภาพที่ ภาพที่ 4.9



ภาพที่ 4.9 ค่า Confusion Matrix ของ Smile Voice Model Non-Call Center

ในส่วนของ Dataset อาสาสมัคร ชุดเดียวกัน แต่นำไปทำ Data Augmentation จะเห็นได้ว่าประสิทธิภาพเพิ่มขึ้นอย่างเห็นได้ชัด นั่นคือได้ค่า Accuracy เป็น 0.86 และค่า loss เป็น 0.22 ซึ่งแตกต่างจากก่อนทำ Data Augmentation ที่ได้ค่า Accuracy 0.5 และค่า loss เป็น 0.7 โดยภาพที่ 4.10 เป็นค่าประสิทธิภาพ เมื่อมีการทำ Data Augmentation กับ Dataset ชุดเดียวกัน



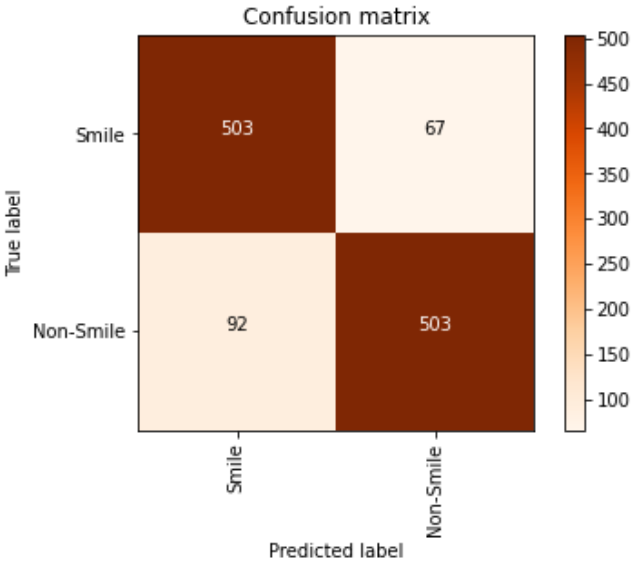
ภาพที่ 4.10 ค่า Loss และ Accuracy Smile Voice Model Non-Call Center ที่มีการทำ Data Augmentation

จากตารางที่ 2 จะเห็นได้ว่าการนำ Dataset ที่มีขนาดไม่ใหญ่มาก มาเพิ่มจำนวนด้วยวิธีการทำ Data Augmentation สามารถช่วยเพิ่มประสิทธิภาพของโมเดลได้ ดังจะเห็นได้ว่าค่า Accuracy ก่อนและหลังทำ Data Augmentation เพิ่มขึ้นอย่างเห็นได้ชัดจาก 0.50 ก่อนทำ Data Augmentation เพิ่มขึ้นเป็น 0.86 หลังจากทำ Data Augmentation และเช่นเดียวกับค่า Loss ก่อนและหลังทำก็ลดลงอย่างเห็นได้ชัดเช่นกันคือ ก่อนทำ Data Augmentation 0.70 ลดลงเหลือ 0.22 หลังจากทำ Data Augmentation

ตารางที่ 2 เปรียบเทียบค่า Accuracy และ Loss โมเดล Smile Voice โดยชุดข้อมูลเสียงอาสาสมัคร

Model	Accuracy	Loss
2DCNN + MFCC	0.50	0.70
2DCNN + MFCC + Data Augmentation	0.86	0.22

จากรูปภาพที่ 4.11 แสดงผลของ Confusion Matrix ของ Model ที่นำเข้าชุดข้อมูลของอาสาสมัคร ที่ผ่านการทำ Data Augmentation



ภาพที่ 4.11 ค่า Confusion Matrix ของ Smile Voice Model Non-Call Center ที่มีการทำ Data Augmentation

4.4 ผลการทดสอบและเปรียบเทียบประสิทธิภาพของ โมเดล Smile Voice โดยชุดข้อมูลเสียงคอลเซ็นเตอร์

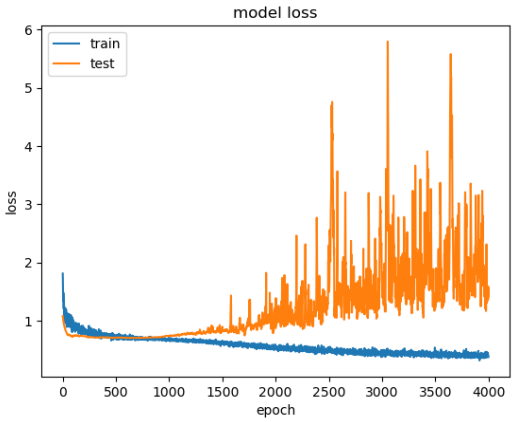
ในการทดสอบประสิทธิภาพจะแบ่งตามประเภทของ CNN ทั้งโมเดลแบบ 1D และ 2D โดยในแต่ละ CNN โมเดล จะมีการสกัด Feature เป็น 2 แบบคือ แบบ Mel-Spectrogram และ MFCC โดย Dataset ที่ใช้จะใช้เหมือนกันในทุกๆ แบบของการฝึกฝนโมเดล โดยการทดสอบจะมีทั้งหมด 4 แบบคือ

1. CNN 1D Mel-Spectrogram Feature Extraction
2. CNN 2D Mel-Spectrogram Feature Extraction
3. CNN 1D MFCC Feature Extraction
4. CNN 2D MFCC Feature Extraction

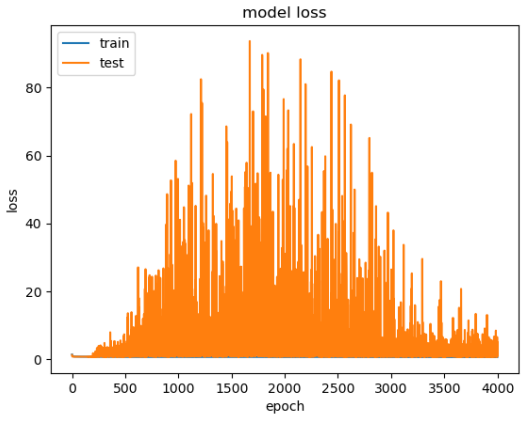
4.4.1 การประเมินผลแบบจำลองโดยใช้ Loss Value

จากผลลัพธ์ของค่า Loss ที่ได้ของแบบจำลองแต่ละแบบแสดงในภาพที่ 4.12 จะเห็นได้ว่าแบบจำลอง CNN 2D MFCC ที่ใช้ Dataset ที่ถูกเพิ่มจำนวนผ่านการทำ Data Augmentation จะได้ค่า loss ต่ำที่สุด รองลงมาจะเป็น CNN 2D MFCC ที่ใช้ Dataset ที่ผ่านการทำ Data Augmentation เช่นกัน แต่ในส่วนของ แบบจำลองแบบ 1D ทั้ง 3 แบบ คือ CNN 1D MFCC และ

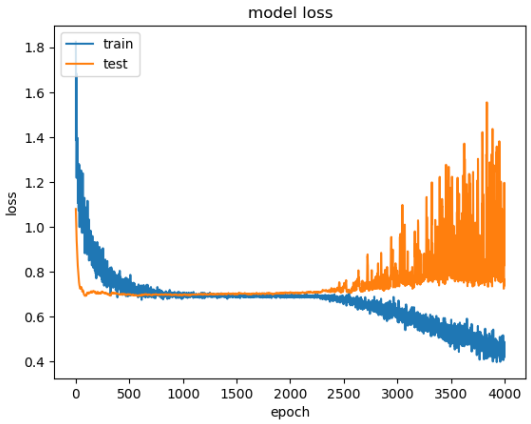
CNN 1D Mel-Spectrogram และ CNN 1D Mel-Spectrogram Dataset Augmentation จะให้ค่า Loss ที่ค่อนข้างสูงและเกิด Overfit ในทุกแบบจำลอง ยกเว้นแบบจำลองที่ใช้ MFCC Feature Extraction ทำให้เห็นได้ว่าเมื่อเปรียบเทียบกับข้อมูล Dataset ชุดเดียวกัน และการสกัด Feature แบบเดี่ยว แบบจำลอง CNN 1D จะมีประสิทธิภาพต่ำกว่าแบบจำลอง CNN 2D อย่างเห็นได้ชัด



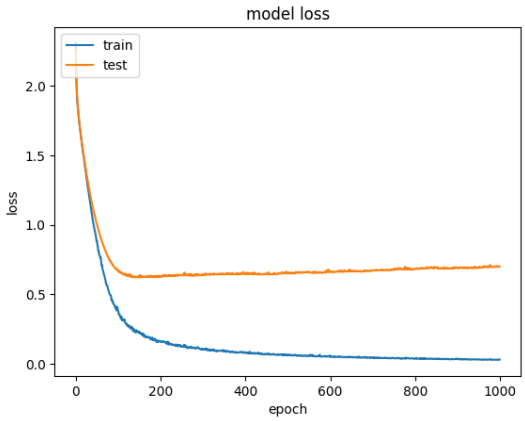
CNN 1D Mel-Spec



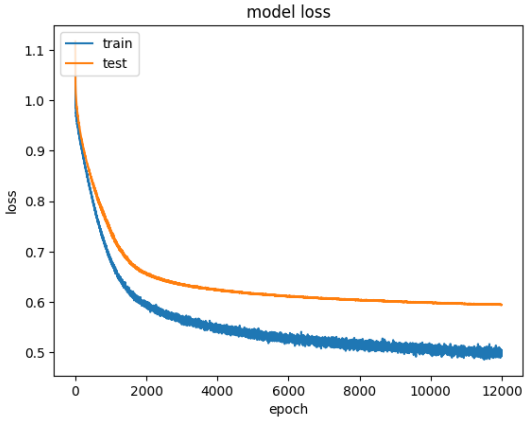
CNN 1D Mel-Spec Data Augmentation



CNN 1D MFCC



CNN 2D MFCC

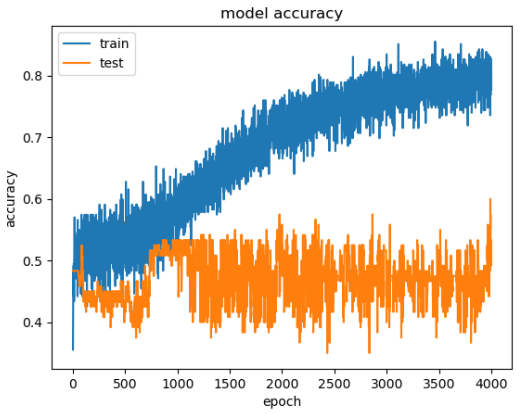


CNN 2D MFCC Data Augmentation

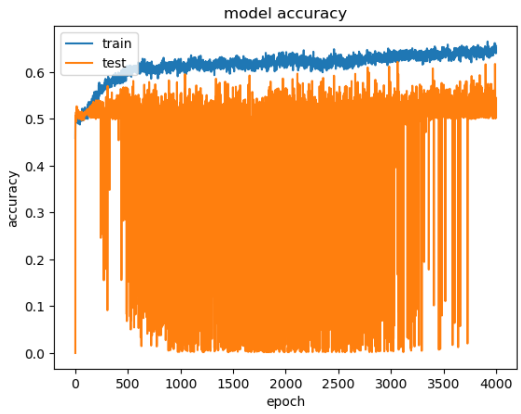
ภาพที่ 4.12 ค่า Loss ของแต่ละ Smile Voice Model โดยชุดข้อมูลเสียงคอลเซ็นเตอร์

4.4.2 การประเมินผลแบบจำลองโดยใช้ Accuracy Value

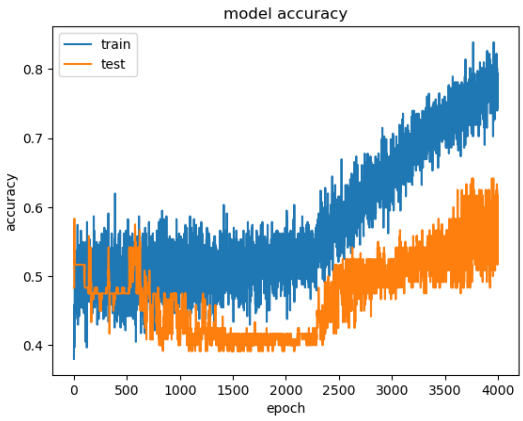
แบบจำลอง CNN 2D MFCC ที่ใช้ Dataset Augmentation จะได้ค่า Accuracy ที่ดีที่สุด รองลงมาคือ CNN 1D MFCC แต่มีค่า val และ train เปลี่ยนแปลงค่อนข้างมาก และในส่วนของ CNN 2D MFCC มีความแต่ต่างระหว่าง val และ train ค่อนข้างสูง และในส่วนที่ประสิทธิภาพต่ำที่สุดจะอยู่ใน Model แบบ CNN 1D ที่ใช้ Mel-Spectrogram ทั้งในแบบ ทำ Data Augmentation และไม่ทำ Data Augmentation ดังภาพที่ 4.13



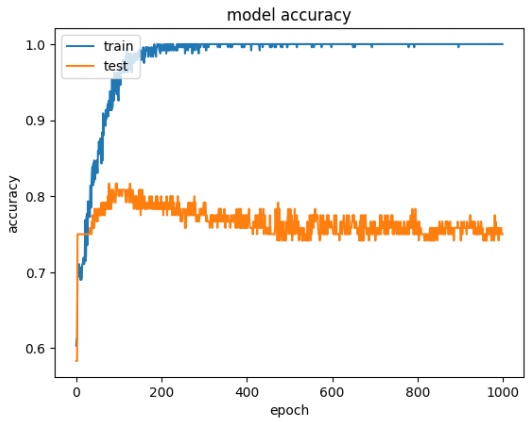
CNN 1D Mel-Spec



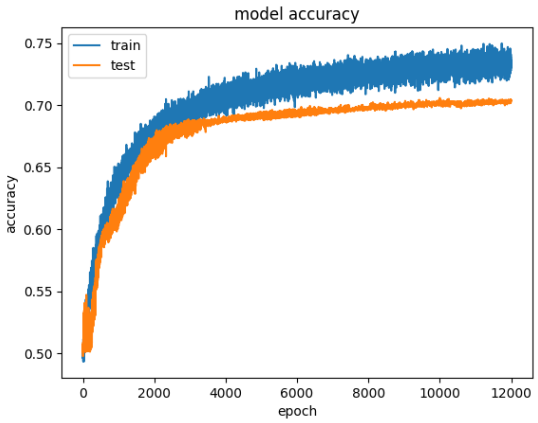
CNN 1D Mel-Spec Data Augmentation



CNN 1D MFCC



CNN 2D MFCC

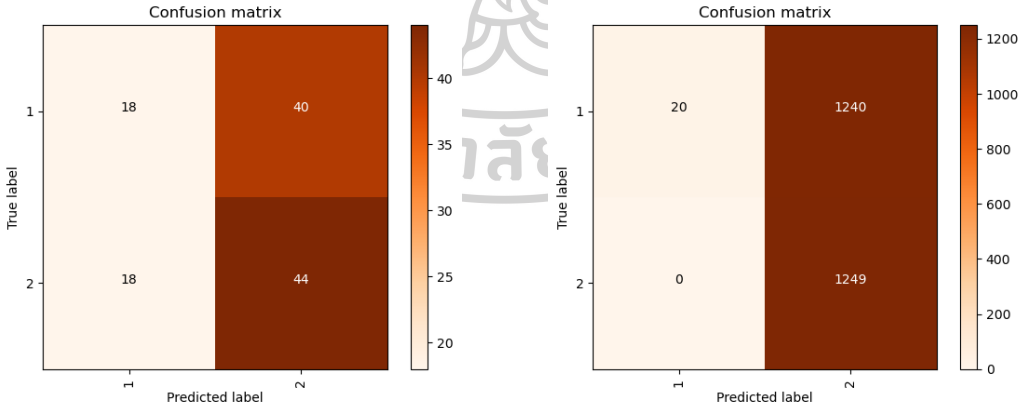


CNN 2D MFCC Data Augmentation

ภาพที่ 4.13 ค่า Accuracy ของแต่ละ Smile Voice Model โดยชุดข้อมูลเสียงคอลเซ็นเตอร์

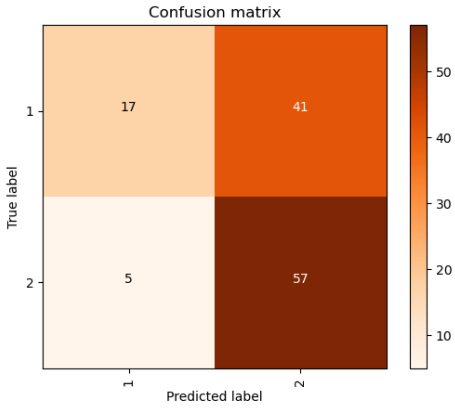
4.4.3 การประเมินผลแบบจำลองโดยใช้ Confusion matrix

ผลของ Confusion Matrix ในภาพที่ 4.14 โดยส่วนใหญ่ ของ Model จะ Classify ข้อมูลระหว่าง Non-Smile Voice ผิดไปเป็น Smile Voice ใน Model CNN 1D Mel-Spec และ CNN1D Mel-Spec แบบ Data Augmentation ในส่วนของ CNN 1D MFCC เริ่ม Classify Smile Voice ได้เพิ่มขึ้น และ ดีขึ้นใน CNN 2D MFCC และได้ค่า การ Classify ถูกต้องมากที่สุดคือ CNN 2D MFCC ร่วมกับคลังข้อมูลที่ทำ Data Augmentation

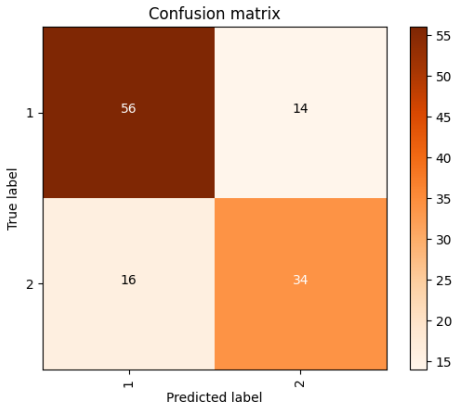


CNN 1D Mel-Spec

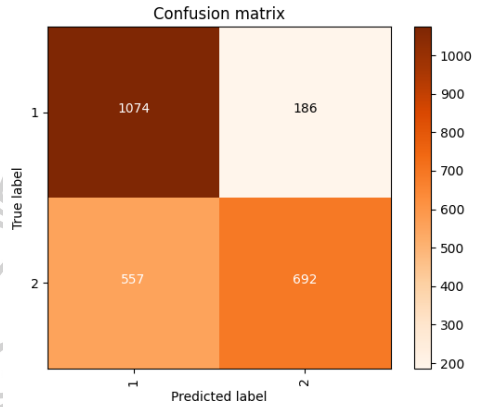
CNN 1D Mel-Spec Data Augmentation



CNN 1D MFCC



CNN 2D MFCC



CNN 2D MFCC Data Augmentation

ภาพที่ 4.14 ค่า Confusion Matrix ของแต่ละแบบจำลอง

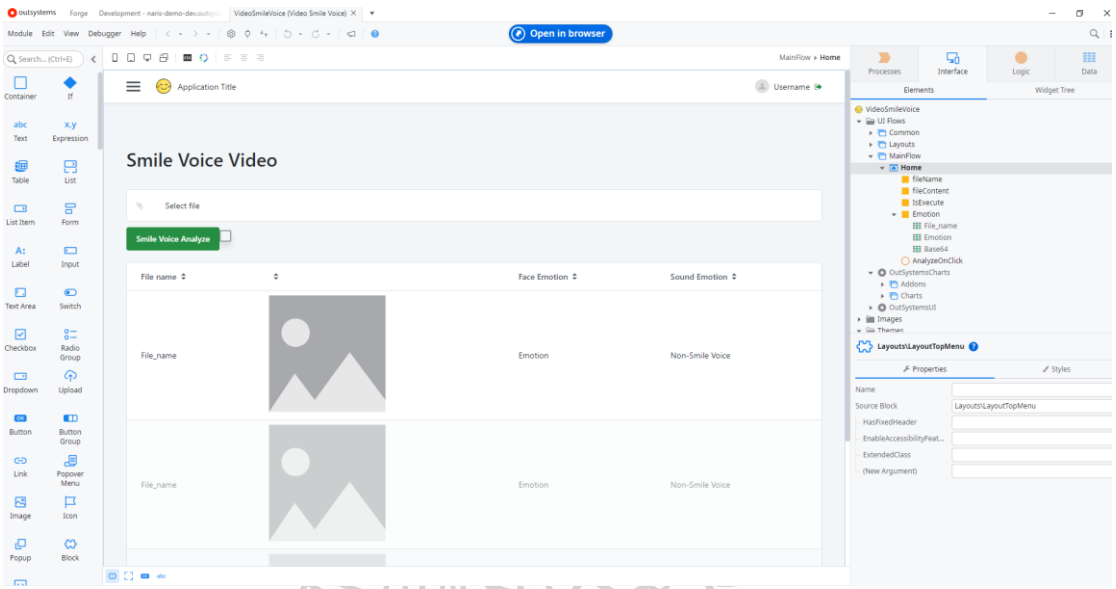
การเปรียบเทียบค่า Accuracy และ Loss ในแต่ละ โมเดล แสดงดัง ตารางที่ 3 จะเห็นได้ว่า โมเดล CNN 2D MFCC ให้ค่า Accuracy สูงสุด ในขณะที่ CNN 2D MFCC Data Augmentation จะมีค่า Loss ต่ำที่สุด

ตารางที่ 3 เปรียบเทียบค่า Accuracy และ Loss ของ โมเดล Smile Voice แต่ละแบบ

Model	Accuracy	Loss
CNN 1D Mel-Spectrogram	0.6083	0.7066
CNN 1D Mel-Spec Data Augmentation	0.6218	0.6065
CNN 1D MFCC	0.6417	0.6933
CNN 2D MFCC	**0.8167**	0.6211
CNN 2D MFCC Data Augmentation	0.7561	**0.5269**

4.5 การประยุกต์ใช้งาน โมเดล Smile Voice

ในการดูความสัมพันธ์กันของ Smile Voice ทั้งในภาพและเสียง ทางผู้วิจัยได้สร้าง Web Application ขึ้นมาเพื่อให้ผู้ที่ต้องการทดสอบการวิเคราะห์เสียง Smile Voice สามารถ Upload Video เข้าไปในระบบ โดย Video ที่ Upload จะต้องเป็นการอัดภาพหน้าตรงให้เห็นแค่ส่วนหัว พร้อมพูดด้วยเสียง Smile Voice โดยมีความยาวของวิดีโอไม่เกิน 20 วินาที เนื่องจากข้อจำกัดทางทรัพยากรของเครื่องที่ใช้ในการประมวลผล โดยในหน้าของเครื่องมือที่ใช้พัฒนา Web Application Front End จะเป็นดังภาพที่ 4.15



ภาพที่ 4.15 หน้าจอพัฒนา Web Application

ในส่วน Backend ใช้ Python ทำงานผ่าน Jupyter Notebook โดยแบ่งโครงสร้างของโปรแกรมออกเป็น 2 ส่วนคือ ส่วนที่เป็น Business Logic ใช้ในการแยก File รูปภาพ และ แยกเสียงออกมาจากภาพเคลื่อนไหว รวมถึงส่งเข้าไปวิเคราะห์ใน โมเดล ทั้ง ภาพและเสียง ดังภาพที่ 4.16 เป็นการแสดงผลลัพธ์ที่ได้ จากการนำ Video เข้าไปแยกเป็น File รูปภาพ และส่งเข้า Deepface เพื่อวิเคราะห์

```
File Location : SmileVoice_Temp/0925_1206/WIN_20230108_20_25_55_Pro.mp4
frames :1321.0, fps : 30.0
duration in seconds: 44
video time: 0:00:44
frame rate for cut :120
Creating...SmileVoice_Temp/0925_1206/frame_00000.jpg
Creating...SmileVoice_Temp/0925_1206/frame_00120.jpg
Creating...SmileVoice_Temp/0925_1206/frame_00240.jpg
Creating...SmileVoice_Temp/0925_1206/frame_00360.jpg
Creating...SmileVoice_Temp/0925_1206/frame_00480.jpg
Creating...SmileVoice_Temp/0925_1206/frame_00600.jpg
Creating...SmileVoice_Temp/0925_1206/frame_00720.jpg
Creating...SmileVoice_Temp/0925_1206/frame_00840.jpg
Creating...SmileVoice_Temp/0925_1206/frame_00960.jpg
Creating...SmileVoice_Temp/0925_1206/frame_01080.jpg
Creating...SmileVoice_Temp/0925_1206/frame_01200.jpg
Create image keyframe finished
11
Action: emotion: 100%|██████████| 1/1 [00:00<00:00,  3.14it/s]
neutral
Action: emotion: 100%|██████████| 1/1 [00:00<00:00, 17.41it/s]
neutral
Action: emotion: 100%|██████████| 1/1 [00:00<00:00, 16.18it/s]
happy
Action: emotion: 100%|██████████| 1/1 [00:00<00:00, 18.25it/s]
happy
Action: emotion: 100%|██████████| 1/1 [00:00<00:00, 16.04it/s]
neutral
Action: emotion: 100%|██████████| 1/1 [00:00<00:00, 15.33it/s]
happy
Action: emotion: 100%|██████████| 1/1 [00:00<00:00, 15.96it/s]
happy
Action: emotion: 100%|██████████| 1/1 [00:00<00:00, 15.68it/s]
neutral
Action: emotion: 100%|██████████| 1/1 [00:00<00:00, 11.84it/s]
neutral
Action: emotion: 100%|██████████| 1/1 [00:00<00:00, 15.52it/s]
neutral
```

ภาพที่ 4.16 หน้าจอ Config เพื่อเชื่อมต่อ Internet

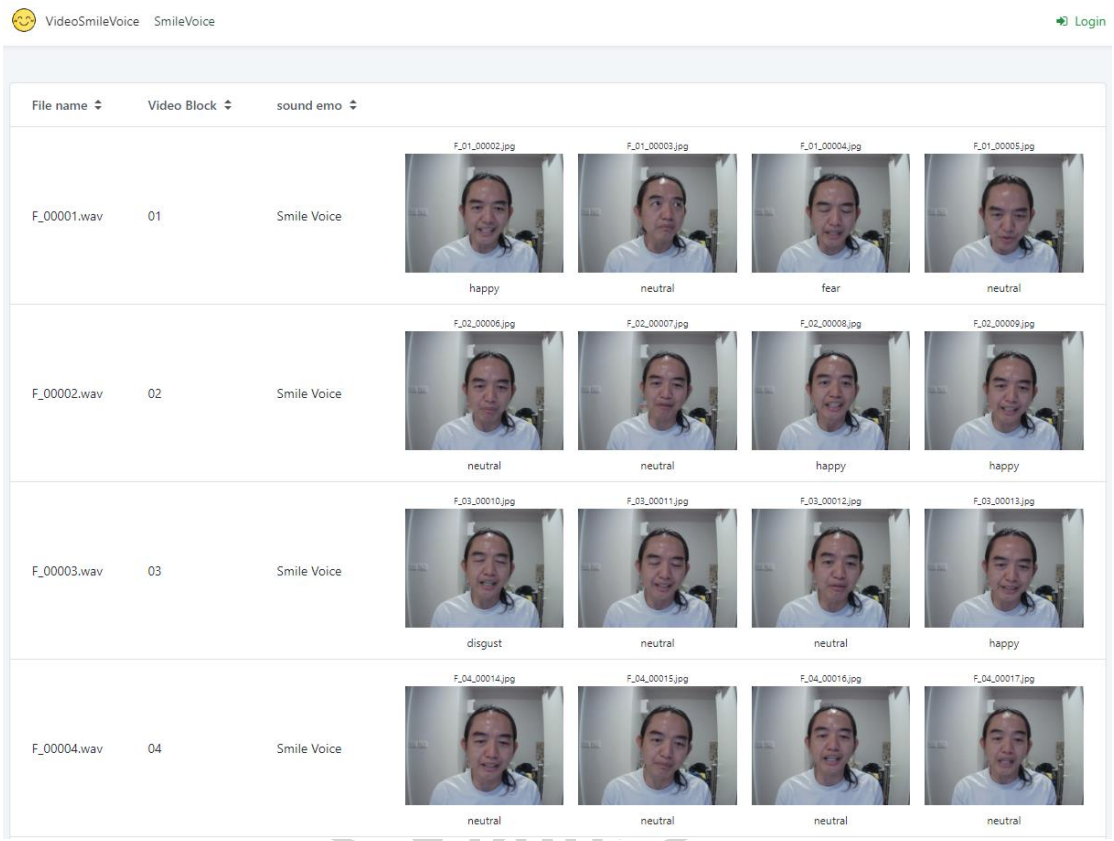
เพื่อสร้าง API Call ให้ Web Application รับส่งข้อมูล ระหว่าง FroneEnd และ Backend ได้ในส่วนของ API จะใช้ FastAPI นำมาใช้งานร่วมกับ Ngrok เพื่อให้สามารถเชื่อมต่อใช้งานผ่านทาง Internet ได้ โดยการ Config Ngrok เป็นดังภาพที่ 4.17

```
import nest_asyncio
from pyngrok import ngrok
import uvicorn

!ngrok authtoken 2EACvVkjVBqDZW6FpHLxRcj5lz7_84XjQeQPy7e4a
ngrok_tunnel = ngrok.connect(8000)
print('Public URL:', ngrok_tunnel.public_url)
nest_asyncio.apply()
uvicorn.run(app, port=8000)
```

ภาพที่ 4.17 หน้าจอ Config เพื่อเชื่อมต่อ Internet

การทำงานของ Application จะเป็นการ Import Video เข้ามาและแยก Video ออกมาเป็นรูปภาพ ซึ่งการที่จะแยกภาพออกมาตามเวลาที่ระบุ จะใช้ OpenCV เพื่อช่วยในการแยกภาพ โดยจะเป็นการระบุ keyframe ที่ต้องการ และ ภาพใน keyframe จะต้องมีความสัมพันธ์กับเวลาของเสียงที่ Classify ด้วยซึ่งคือเวลา 4 วินาที ซึ่งวิดีโอที่ถ่ายมาอาจจะมี framerate ไม่เท่ากัน (fps) เราจำเป็นต้องหา เวลาของ video และค่า fps ออกมาก่อน จึงจะ แยก keyframe ออกมาให้มีความสัมพันธ์กับเวลาของภาพที่ต้องการ ผลลัพธ์ที่ได้ใน 1 video จะมีรูปภาพที่แตกออกมาหลาย

รูปภาพ เพื่อนำไปเข้า โมเดล Face Emotion Deepface ซึ่งเป็น Opensource เพื่อหาอารมณ์ของใบหน้าในรูป และ โมเดล Smile Voice เพื่อวิเคราะห์เสียงที่เป็น Smile Voice และแสดงผลความสัมพันธ์กันของหน้า และเสียง โดย Application จะเก็บข้อมูลรูปภาพ และเสียงที่แยกออกมาแล้ว เพื่อนำมาใช้ในการวิเคราะห์ในภายหลังได้ ในกรณีที่มีการพัฒนา โมเดล ใหม่ ขึ้นมาและมีประสิทธิภาพมากกว่าเดิม ภาพที่ 4.18 แสดงตัวอย่าง Web Application ที่ใช้ในการตรวจสอบเสียงยิ้ม



ภาพที่ 4.18 การแสดงความสัมพันธ์กันของภาพและเสียง สำหรับ Smile Voice

ซึ่งผลของการทดสอบโปรแกรมที่ประยุกต์ใช้งาน Smile Voice พบว่า มีรูปแบบที่เกิดขึ้น 2 แบบคือ หน้าและเสียงมีความสัมพันธ์กัน นั่นคือตรวจพบอารมณ์บนใบหน้าเป็น happy และเสียงเป็น Smile Voice และในอีกรูปแบบคือ หน้าและเสียงที่มีความสัมพันธ์กัน นั่นคือตรวจพบอารมณ์บนใบหน้าที่ม่าได้เป็น happy และเสียงเป็น Smile Voice ซึ่งในกรณีแรกที่หน้าและเสียงมีความสัมพันธ์กัน ส่วนใหญ่จะพบในคอลเซ็นเตอร์ที่มีอายุงานน้อยกว่า 2 ปี และอีกกรณีคือหน้าและเสียงที่ไม่มีความสัมพันธ์กันส่วนใหญ่จะพบได้ในคอลเซ็นเตอร์ที่มีอายุงานมากกว่า 2 ปีขึ้นไป จากผลการทดลอง

นี้ผู้วิจัยได้สัมภาษณ์คอลเซ็นเตอร์ทั้ง 2 กลุ่ม โดยในกลุ่มแรกคือกลุ่มที่มีอายุงานน้อยกว่า 2 ปี เกี่ยวกับการออกเสียง Smile Voice ต้องมีการใช้ความพยายามในการพูดให้เป็นเสียงยิ้ม รวมไปถึงต้องทำหน้ายิ้มเพื่อให้เสียงออกมาเป็น Smile Voice ซึ่งจะแตกต่างจากในกลุ่มที่มีอายุงานมากกว่า 2 ปี การออกเสียง Smile Voice สามารถพูดออกมาได้เป็นเสียงธรรมชาติโดยไม่ต้องใช้ความพยายามในการออกเสียง รวมไปถึงไม่ว่าจะทำหน้าตาเป็นอารมณ์ไหนก็สามารถออกเสียง Smile Voice ได้

บทที่ 5
สรุปผลการดำเนินงานวิจัยและข้อเสนอแนะ

งานวิจัยนี้เป็นการสร้างโมเดลตรวจสอบเสียงยิ้ม โดยทางผู้วิจัยได้เริ่มต้นทดลองสร้างโมเดลวิเคราะห์อารมณ์จากเสียง (Thai SER – Thai Speech Emotion Recognition) โดยใช้ชุดข้อมูลเสียงภาษาไทย ที่สร้างโดยสถาบันวิทยสิริเมธี (Vidyasirimedhi Institute of Science and Technology: VISTEC) และถือเป็นชุดข้อมูลเสียงที่มีการระบุอารมณ์ และมีขนาดใหญ่ที่สุดในภาษาไทย ซึ่งมีจำนวนเสียงทั้งหมด 27,854 เสียง โดยแบ่งเป็น 15,874 เสียงที่เป็นการพูดโดยไม่มีบท (No Script) และ 11,980 เป็นการพูดตามบท (Script) โดยมีอาสาสมัครในการบันทึกเสียงทั้งหมด 200 คน โดยทั้งหมดนี้เป็นการออกเสียงใน 5 อารมณ์คือ โกรธ ฉุนเฉียว ปกติ เสียใจ และ ดีใจ โดยในการสร้างโมเดลวิเคราะห์อารมณ์นี้ ผู้วิจัยเลือกใช้เฉพาะคลังข้อมูลที่เป็นการพูดแบบมีบทพูดเท่านั้น ซึ่งมีจำนวน 11,980 เนื่องจากมีรูปแบบของการพูดที่คล้ายกัน เนื่องจากถูกกำหนดโดยบทพูด ซึ่งมีทั้งหมด 3 ประโยค และผู้พูดทุกคนต้องพูดประโยคนี้เหมือนกันทั้งหมด ซ้ำๆ กันและพูดในอารมณีที่แตกต่างกันในแต่ละประโยค และในแต่ละคนจะมีรูปแบบ หรือเทคนิคในการพูดที่แตกต่างกัน

จากข้อมูลคลังข้อมูลเริ่มต้นชุดนี้ ผู้วิจัยได้สร้างแบบจำลอง 1DCNN และ 2DCNN และทำการสกัด Feature เป็น 2 แบบคือ Mel-Spectrogram และ Mel frequency cepstrum coefficient (MFCC) ซึ่งเป็นการสกัด Feature ที่นิยมใช้เกี่ยวกับเสียง จากนั้นผู้วิจัยได้สร้างแบบจำลองขึ้นทั้งหมด 4 แบบ เพื่อเปรียบเทียบประสิทธิภาพกัน นั้นคือ 1) 1DCNN + Mel-Spectrogram 2) 1DCNN + MFCC 3) 2DCNN + Mel-Spectrogram และ 4) 2DCNN + MFCC ซึ่งจากผลการทดสอบ แบบจำลองที่สร้างขึ้นโดยใช้ 2DCNN + MFCC ได้ประสิทธิภาพดีกว่าแบบจำลอง ทั้ง 3 แบบอย่างเห็นได้ชัด โดยได้ค่า Accuracy ที่ 0.8059 และ ค่า Loss ที่ 0.6140 โดยแบบจำลองอื่นๆ จะมีค่า Accuracy อยู่ประมาณ 0.4 - 0.5 และค่า Loss อยู่ที่ 1.0 – 1.3 โดยถ้าเรียงตามแบบจำลองที่มีประสิทธิภาพดีที่สุดไปจนถึงต่ำสุดจะเป็น 1) 2DCNN + MFCC 2) 2DCNN + Mel-Spectrogram 3) 1DCNN + MFCC และแบบจำลองที่ประสิทธิภาพต่ำที่สุดคือ 4) 1DCNN + Mel-Spectrogram ซึ่งค่าของ Accuracy และ Loss ในแต่ละแบบจำลอง ดังแสดงใน ตารางที่ 1

การทดลองถัดมาเป็นการนำเสียงของคอลเซ็นเตอร์มาวิเคราะห์โดยโมเดล Thai SER เพื่อตรวจสอบว่าได้เสียงเป็นอารมณ์ใดบ้าง โดยต้องมีการเตรียมไฟล์เสียงก่อนที่จะนำไปใช้ในโมเดล ซึ่งเป็นการกดหนดเสียงให้เป็น 4 วินาที โดยถ้าไฟล์เสียงใดม่ความยาวมากกว่า 4 วินาที จะตัดให้เหลือ 4 วินาที และไฟล์เสียงที่มีความยาวน้อยกว่า 4 วินาที จะเป็นการเพิ่มเสียง silence ให้ได้ 4 วินาที ซึ่งไฟล์เสียงที่นำมาวิเคราะห์มีจำนวน 232 ไฟล์ โดยผลลัพธ์ที่ได้จากโมเดล Thai SER พบว่า เสียงของ

คอลเซ็นเตอร์ที่เลเบลเป็น Smile Voice จะวิเคราะห์เป็นอารมณ์ มีความสุข มากกว่าเสียงที่เลเบล Non-Smile Voice อย่างไรก็ตาม ผลลัพธ์ที่ได้ไม่ได้แตกต่างกันอย่างเห็นได้ชัดเนื่องจาก โมเดลที่ใช้วัดเป็นการวิเคราะห์อารมณ์ ทำให้เห็นได้ว่าการพูดเสียง Smile Voice และ Non-Smile Voice ในคอลเซ็นเตอร์ไม่ได้แตกต่างกันอย่างเห็นได้ชัด จากการที่ได้สัมภาษณ์คอลเซ็นเตอร์การออกเสียง 2 รูปแบบนี้ให้แตกต่างกัน หรือไม่แตกต่างกันอยู่ที่ประสบการณ์ในการใช้เสียงของคอลเซ็นเตอร์ ถ้าเป็นคนที่มีประสบการณ์มากกว่า 2 ปีขึ้นไป การออกเสียงเป็นธรรมชาติ และใช้การปรับโทนเสียงเพียงนิดเดียวเพื่อให้เกิด Smile Voice และในทางกลับกันคอลเซ็นเตอร์ที่ไม่มีประสบการณ์มากอาจจะต้องใช้พยายามมากขึ้นในการปรับโทนเสียงให้เป็น Smile Voice ซึ่งจากข้อมูลระดับรายบุคคล โมเดล Thai SER จะวิเคราะห์เป็นอารมณ์ Happy สำหรับเสียงที่เลเบลเป็น Smile Voice ได้มากในบุคคลที่มีประสบการณ์คอลเซ็นเตอร์มากกว่า 2 ปีขึ้นไป

เมื่อได้แนวทางของการสร้างแบบจำลอง และนำไปทดสอบในคลังข้อมูลทั้ง 2 แบบ ข้างต้นแล้ว ผู้วิจัยได้ทดลองสร้างคลังข้อมูลเสียง Smile Voice และ Non-Smile Voice ขึ้นมา โดย เก็บรวมรวมเสียงจากอาสาสมัครที่เป็นนักศึกษาระดับ ปริญญาตรี อายุประมาณ 18-20 ปี จำนวน 7 คน โดยเป็นผู้หญิง 4 คน และ ผู้ชาย 3 คน โดยใช้สถานที่อัดเสียงเป็นห้องประชุม และเลือกเป็นห้องประชุมที่ไม่มีเสียงรบกวน วิธีการอัดเสียงเป็นการใช้ไมโครโฟนมีสายต่อเข้ากับมือถือ เพื่ออัดเสียง และมีการเตรียมบทให้พูดทั้งหมด 6 ประโยค โดยจะต้องพูดตามบท แต่ละประโยคจำนวน 4 ครั้ง โดย 2 ครั้งแรกเป็นเสียง Smile Voice และอีก 2 ครั้งเป็น Non-Smile Voice

โดยข้อมูลไฟล์เสียงที่ได้มีจำนวน 168 ไฟล์ เป็นไฟล์เสียง Smile Voice 84 และ Non-Smile Voice 84 จากนั้นนำไฟล์เสียงที่ได้ไป ทำให้เหลือความยาว 4 วินาที ตามรูปแบบข้างบน และนำไปสกัด Feature ในรูปแบบ MFCC และนำไปเข้าแบบจำลองแบบ 1DCNN และ 2DCNN ที่ได้เคยทดลองผ่านมา จากนั้นปรับจูนพารามิเตอร์ เช่น ความลึกของแบบจำลอง ขนาดของ Kernal ปริมาณของการ Dropout เป็นต้น โดยใช้แบบจำลองชุดแรกเป็นต้นแบบ ที่ต้องปรับจูนพารามิเตอร์เพิ่มเติมเพราะแบบจำลองชุดแรกใช้สำหรับวิเคราะห์อารมณ์ของผู้พูด ในส่วนแบบทดลองชุดนี้เป็นการวิเคราะห์เสียงที่เป็น Smile Voice และ Non-Smile Voice

ซึ่งในการทดลองชุดแรกกับไฟล์เสียงจำนวน 168 ไฟล์ ได้เกิดการ overfit ขึ้นอย่างมาก โดยเริ่มตั้งแต่ epoch ที่ 100 และไม่มีแนวโน้มที่จะลดลง สาเหตุเนื่องจากข้อมูลของชุดข้อมูลมีจำนวนน้อยเกินไป เพื่อทำให้จำนวนไฟล์เสียงเดิมมีขนาดมากขึ้น ผู้วิจัยได้ใช้ Data Augmentation เทคนิคเข้ามาช่วย โดยทำทั้งหมด 4 แบบคือ 1) Time stretching 2) Pitch shifting 3) Adding noise 4) Downsample หลังจากทำ Data Augmentation เพื่อเพิ่มจำนวนของไฟล์เสียง พร้อมกับปรับจูนพารามิเตอร์เพิ่มเติมเพื่อให้มีประสิทธิภาพดีที่สุด ในผลลัพธ์แสดงพบว่าเกิด Overfit ลดน้อยลงอย่าง

เห็นได้ชัด รวมถึงค่า Accuracy ที่เพิ่มขึ้นและ Loss ที่ลดลง ซึ่งผู้วิจัยจะนำเทคนิค Data Augmentation ไปใช้ในการทดลองลำดับถัดไป

หลังจากได้ทดลองการสร้างชุดข้อมูลจากเสียงอาสาสมัครที่เป็นนักศึกษาแล้ว ทางผู้วิจัยได้ติดต่อทางธนาคารที่มีหน่วยงานคอลเซ็นเตอร์อยู่ เพื่อขอเข้าไปเก็บเสียง รวมถึงภาพเคลื่อนไหวจากคนที่เป็นคอลเซ็นเตอร์จริงๆ ซึ่งได้รับความร่วมมือเป็นอย่างดีจากธนาคาร โดยให้เข้าไปเก็บเสียงของคอลเซ็นเตอร์ จริงๆ โดยมี อาสาสมัครที่เป็นคอลเซ็นเตอร์ เข้ามาอัดเสียงและภาพเคลื่อนไหวจำนวน 15 คน ในครั้งนี้ผู้วิจัยได้ใช้คอนเด็นเซอร์ไมค์ เพื่อให้ได้เสียงที่ดีที่สุด โดยวางไว้ที่หน้าคอลเซ็นเตอร์ที่พูด และมีบทพูดเพื่อเป็นแนวทางในการพูดเท่านั้น เพื่อให้ทางผู้พูดปรับเปลี่ยนรูปแบบไปตามที่ตนเองถนัด จากการอัดเสียง ทางผู้วิจัยได้ไฟล์เสียงมาทั้งหมด 232 ไฟล์ ซึ่งมีทั้ง แบบ Smile Voice และ Non-Smile Voice

เนื่องจากทางผู้วิจัยต้องการทำ Blind test กับเสียงของคอลเซ็นเตอร์ เพื่อจะดูรูปแบบของการออกเสียงที่เป็น Smile Voice และ Non-Smile Voice สามารถเกิดจากปัจจัยอื่นอีกหรือเปล่า โดยเฉพาะประสบการณ์การทำงานเกี่ยวกับคอลเซ็นเตอร์ และเสียงที่ทางอาสาสมัครคอลเซ็นเตอร์ออกเสียงมาส่วนใหญ่ตรงกับ Label ที่วางเอาไว้หรือไม่ ทางผู้วิจัยได้สร้าง Web Application ขึ้นมาเพื่อให้ทางผู้เชี่ยวชาญด้านคอลเซ็นเตอร์จำนวน 5 คน เข้ามาตรวจสอบ โดยทางผู้เชี่ยวชาญจะไม่รู้ว่าเสียงที่ได้ฟังถูกพูดด้วยเสียงอะไร และจะไม่สามารถเห็นข้อมูลการระบุเสียงจากผู้เชี่ยวชาญคนอื่นด้วย โดยเมื่อฟังเสียงแล้วทางผู้เชี่ยวชาญจะระบุเสียงที่ได้ยินกว่าเป็นเสียง Smile Voice หรือ Non-Smile Voice ซึ่งจากผลที่ได้จากการทำ Blind test มากกว่า 50% ผู้เชี่ยวชาญสามารถระบุเสียงตรงกับที่อาสาสมัครคอลเซ็นเตอร์ ออกเสียงได้ โดยที่เสียงส่วนใหญ่ที่ระบุใกล้เคียงกับเลเบลต้นฉบับที่สุด จะเป็นเสียงที่มาจากคอลเซ็นเตอร์ที่มีประสบการณ์มากกว่า 2 ปีขึ้นไป ในส่วนเสียงที่ระบุไม่ตรงกับเลเบลต้นฉบับ ส่วนใหญ่จะเป็นเสียงที่พูดโดยคอลเซ็นเตอร์ที่มีประสบการณ์น้อยกว่า 2 ปี โดยมีอัตราที่ผู้เชี่ยวชาญระบุเสียงเหมือนต้นฉบับ โดยเฉลี่ยเท่ากับหรือน้อยกว่า 50%

จากชุดข้อมูลที่ได้จากคอลเซ็นเตอร์ ทางผู้วิจัยได้สร้างแบบจำลองเสียงยิ้มโดยอาศัยแนวทางจากการทดลองครั้งที่ 1 และครั้งที่ 2 เป็นตัวตั้งต้น และใช้เป็นแนวทางในการสร้างแบบจำลอง ทั้ง 2 แบบคือ 1DCNN และ 2DCNN โดยสกัด Feature เป็น Mel-Spectrogram และ MFCC โดยจะแบ่งการทดลองออกเป็น 4 แบบ คือ 1) 1DCNN + Mel-Spectrogram 2) 2DCNN + Mel-Spectrogram 3) 1DCNN + MFCC และ 4) 2DCNN + MFCC ซึ่งเมื่อเทียบประสิทธิภาพจากโมเดลทั้ง 4 แบบแล้ว โมเดล ที่ได้ค่าประสิทธิภาพดีที่สุดคือ 2DCNN + MFCC โดยค่า Accuracy 0.8167 และค่า Loss เป็น 0.6211 อย่างไรก็ตามผลการทดลองเกิด overfit ขึ้นในทุกแบบจำลอง โดยเฉพาะโมเดลที่สกัด Feature แบบ Mel-Spectrogram ทางผู้วิจัยจึงได้เพิ่มจำนวนข้อมูลด้วย Data

Augmentation ในทุกๆ แบบของการทดลอง และใช้โมเดลทั้ง 4 แบบในการทดลองอีกครั้งโดยไม่ปรับเปลี่ยนค่าใดๆ ยกเว้นจำนวนข้อมูลที่เพิ่มขึ้น ซึ่งโมเดลที่มีประสิทธิภาพดีที่สุดหลังจาก Data Augmentation คือ CNN2D+MFCC ซึ่งได้ค่า Accuracy เป็น 0.7561 และค่า Loss 0.5269

ขั้นตอนสุดท้ายเป็นการนำโมเดล Smile Voice มาประยุกต์ใช้งาน โดยเป็นการสร้าง Application เพื่อให้ผู้ใช้สามารถ Upload วิดีโอที่ต้องการวิเคราะห์ Smile Voice โดยในการทดลองนี้ นอกจากการวิเคราะห์ Smile Voice และ Non-Smile Voice แล้ว ผู้วิจัยได้นำอารมณ์บนใบหน้าของผู้พูดมาวิเคราะห์ร่วมด้วย โดยในส่วนของการตรวจสอบอารมณ์บนใบหน้า ผู้วิจัยเลือกใช้ Component Deepface ซึ่งต้องใช้ภาพใบหน้าของผู้พูดมาส่งเข้าไปใน Deepface ในขั้นตอนนี้จะเป็นการนำวิดีโอที่ต้องการวิเคราะห์ Smile Voice มาแยกเสียงออกจากวิดีโอ และส่งข้อมูลเสียงเข้าไปในโมเดล Smile Voice ข้างต้น ในส่วนของวิดีโอผู้วิจัยใช้ component moviepy เพื่อสกัดวิดีโอออกมาเป็นภาพนิ่ง ซึ่งผู้วิจัยเลือกการสกัดรูปภาพที่ทุกเวลา 1 วินาที เนื่องจากปกติในการพูดของคอลเซ็นเตอร์จะพูดด้วยความเร็วปกติจนถึงช้า เพื่อให้ผู้ฟังเข้าใจ ทำให้กกล้ามเนื้อบนใบหน้าไม่ได้มีการเปลี่ยนแปลงมากในช่วงเวลาหนึ่ง ผู้วิจัยจึงเลือกใช้ที่เวลา 1 วินาที และเมื่อได้รูปภาพมาแล้ว จะส่งเข้าไปใน Deepface เพื่อวิเคราะห์อารมณ์บนใบหน้าออกมา และนำมาวิเคราะห์ร่วมกับผลลัพธ์ที่ได้จาก Smile Voice เพื่อดูความสัมพันธ์กันของใบหน้า และ Smile Voice

จากผลการทดลองใช้งานผ่าน Application ผมลัพทธ์ที่ได้จะเห็นได้ว่า การที่วิเคราะห์เสียงเป็น Smile Voice เมื่อเปรียบเทียบความสัมพันธ์ของอารมณ์บนใบหน้าจะพบได้ 2 รูปแบบ คือ Smile Voice + อารมณ์ Happy หรือ Neutral และ Smile Voice + ไม่พบ อารมณ์ Happy และ Neutral ซึ่งในกรณีที่ Smile Voice + อารมณ์ Happy หรือ Neutral แสดงถึงเสียงยิ้มและอารมณ์ Happy ด้วยซึ่งเสียง และอารมณ์บนใบหน้าเป็นไปในแนวทางเดียวกัน ส่วนในกรณี Smile Voice + ไม่พบ อารมณ์ Happy และ Neutral ถือว่าเสียงและอารมณ์บนใบหน้าไม่ได้เป็นไปในแนวทางเดียวกัน ซึ่งผู้วิจัยพบว่า การที่ Smile Voice และอารมณ์บนใบหน้าจะมีความสอดคล้องกันหรือไม่ขึ้นอยู่กับปัจจัยอื่นด้วยนั่นคือ ประสบการณ์ หรือความเชี่ยวชาญในการใช้เสียง จะสังเกตได้ว่าส่วนใหญ่ที่ความสัมพันธ์ของ Smile Voice ไม่ได้เป็นไปในทางเดียวกับอารมณ์บนใบหน้าจะพบได้ใน คอลเซ็นเตอร์ที่มีประสบการการทำงานมากกว่า 2 ปีขึ้นไป และความสัมพันธ์ที่ไปในทางเดียวกันจะพบได้มากในคนที่มีประสบการณ์น้อยกว่า 2 ปี ซึ่งผู้วิจัยได้สัมภาษณ์เพิ่มเติมกับทั้ง 2 กลุ่มนี้ ในกลุ่มแรก คือ กลุ่มที่มีประสบการณ์มากว่า 2 ปี การทำเสียง Smile Voice เขาสามารถทำได้เป็นธรรมชาติ เพราะสามารถทำได้โดยใช้การปรับโทนเสียง และไม่จำเป็นต้องเน้นเสียงอะไร โดยที่อารมณ์บนใบหน้าจะเป็นอย่างไรก็ได้ ในส่วนของกลุ่มที่สอง การออกเสียงให้เป็น Smile Voice ถูกฝึกมาว่าถ้าหน้า

Happy หรือยิ้ม เสียงที่ออกมาจะเป็น Smile Voice ซึ่งเป็นวิธีการง่ายที่สุดที่สามารถสร้างเสียง Smile Voice ได้

เพราะฉะนั้นสามารถสรุปได้ว่าการออกเสียง Smile Voice นอกจากจะให้ผู้พูด พูดด้วยอารมณ์ของ Happy แล้วยังสามารถทำได้ด้วยผู้เชี่ยวชาญที่มีการใช้เสียงเป็นประจำ เพื่อสร้างเสียง Smile Voice โดยที่ไม่จำเป็นต้องมีความสัมพันธ์กับอารณ์บนใบหน้าได้

ในการต่อยอดงานวิจัยผู้วิจัยมองว่า สามารถนำโมเดล Smile Voice ไปประยุกต์ใช้เพื่อฝึกการพูดให้เป็น Smile Voice โดยเฉพาะใช้ในการฝึกอบรมของคอลเซ็นเตอร์ เพื่อช่วยแบ่งเบาภาระผู้ฝึกสอน ที่จะเข้ามานั่งฟังและวิเคราะห์เสียงในการฝึกอบรมว่าเป็นเสียง Smile Voice หรือไม่ และอาจนำไปต่อยอดเพื่อเป็นการมอนิเตอร์การใช้เสียงของคอลเซ็นเตอร์ในระหว่างวัน เพื่อดูว่า มีการใช้งานเสียงยิ้ม หรือ ไม่ยิ้มมากหรือน้อยเพียงใด เพื่อจะได้ปรับปรุงการใช้เสียงของคอลเซ็นต์เตอร์ให้ดีขึ้น และในส่วนของต่อยอดของการพัฒนาโมเดล Smile Voice ที่จะช่วยเพิ่มประสิทธิภาพให้ดีขึ้นได้ การใช้คลังข้อมูลของ Smile Voice ที่มีขนาดใหญ่ขึ้น โดยให้ผู้เชี่ยวชาญทางด้านการออกเสียง หรือคนที่มีประสบการณ์การใช้เสียงมาสร้างคลังข้อมูล Smile Voice ที่มีขนาดใหญ่ขึ้น ซึ่งจะช่วยในการเพิ่มประสิทธิภาพของโมเดล Smile Voice ได้ และนอกจากใช้คลังข้อมูล Smile Voice ที่มีขนาดใหญ่ขึ้นแล้ว ต้องมีการปรับเปลี่ยนพารามิเตอร์ในขั้นตอนของการสร้างโมเดลด้วย โดยจะเห็นได้ว่าการวิเคราะห์เสียง Smile จะมีประสิทธิภาพดีที่สุดถ้าใช้วิธีการสกัด Feature แบบ MFCC และใช้ 2DCNN ในการ Training Model ซึ่งนอกจากใช้คลังข้อมูลที่มีขนาดใหญ่ขึ้นแล้ว ต้องทำควบคู่ไปกับการปรับพารามิเตอร์ที่ใช้ใน Model เพื่อให้ได้ค่าที่ให้ประสิทธิภาพของ Model สูงสุด และนำโมเดลที่ได้ไปประยุกต์ใช้งานผ่านทางการสร้าง Application เพื่อใช้วิเคราะห์เสียงยิ้มต่อไปได้

รายการอ้างอิง

Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel Frequency Cepstral Coefficient and its Applications: A Review. *IEEE Access*, *10*, 122136-122158. https://doi.org/10.1109/ACCESS.2022.3223444

Alom, M. Z., Taha, T., Yakopcic, C., Westberg, S., Hasan, M., Esesn, B., Awwal, A., & Asari, V. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches.

Bonner, A. (2019). *What is Deep Learning and How Does it Work?* https://towardsdatascience.com/what-is-deep-learning-and-how-does-it-work-f7d02aa9d477

Brown, W. M., & Moore, C. (2002). Smile asymmetries and reputation as reliable indicators of likelihood to cooperate: An evolutionary analysis. In *Advances in psychology research, Vol. 11.* (pp. 19-36). Nova Science Publishers.

BrunelloN. (2021). *Deep neural network.* https://commons.wikimedia.org/wiki/File:Example_of_a_deep_neural_network.png

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335-359. https://doi.org/10.1007/s10579-008-9076-6

Deng, J., Dong, W., Socher, R., Li, L. J., Kai, L., & Li, F.-F. (2009, 20-25 June 2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition,

Dickson, B. (2020). *What are convolutional neural networks (CNN)?* https://bdtechtalks.com/2020/01/06/convolutional-neural-networks-cnn-convnets/

Drahota, A., Costall, A., & Reddy, V. (2008). The vocal communication of different kinds of smile. *Speech Communication*, *50*(4), 278-287. https://doi.org/https://doi.org/10.1016/j.specom.2007.10.001

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: investigator's guide*. Consulting Psychologists Press.

Ekman, P., & Friesen, W. V. (1982). Felt, false, and miserable smiles. *Journal of Nonverbal*

Behavior, *6*(4), 238-252. https://doi.org/10.1007/BF00987191

Emond, C., & Laforest, M. (2013). Prosodic correlates of smiled speech. *Proceedings of Meetings on Acoustics*, *19*(1), 060220. https://doi.org/10.1121/1.4799490

Fagel, S. (2010). Effects of Smiling on Articulation: Lips, Larynx and Acoustics. In A. Esposito, N. Campbell, C. Vogel, A. Hussain, & A. Nijholt (Eds.), *Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Dublin, Ireland, March 23-27, 2009, Revised Selected Papers* (pp. 294-303). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-12397-9_25

Fayek, H. (2016). *Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between*. https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

Forgas, J. P. (1987). The role of physical attractiveness in the interpretation of facial expression cues. *Personality and Social Psychology Bulletin*, *13*(4), 478-489. https://doi.org/10.1177/0146167287134005

Fukushima, K. (2014). Modeling Vision with the Neocognitron. In (pp. 765-782). https://doi.org/10.1007/978-3-642-30574-0_44

Haddad, K. E., Çakmak, H., Moinet, A., Dupont, S., & Dutoit, T. (2015, 7-10 Dec. 2015). An HMM approach for synthesizing amused speech with a controllable intensity of smile. 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT),

Hu, H., Xu, M. X., & Wu, W. (2007, 15-20 April 2007). GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07,

Huang, K. Y., Wu, C. H., Hong, Q. B., Su, M. H., & Zeng, Y. R. (2018, 26-29 Nov. 2018). Speech Emotion Recognition using Convolutional Neural Network with Audio Word-based Embedding. 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP),

Huang, Z. Y., Chiang, C. C., Chen, J. H., Chen, Y. C., Chung, H. L., Cai, Y. P., & Hsu, H. C. (2023). A study on computer vision for facial emotion recognition. *Sci Rep*, *13*(1), 8425. https://doi.org/10.1038/s41598-023-35446-4

Kohler, K. (2008). 'Speech-Smile', 'Speech-Laugh', 'Laughter' and Their Sequencing in Dialogic

Interaction. *Phonetica*, *65*, 1-18. https://doi.org/10.1159/000130013

LaFrance, M., & Hecht, M. A. (1995). Why smiles generate leniency. *Personality and Social Psychology Bulletin*, *21*(3), 207-214. https://doi.org/10.1177/0146167295213002

Li, Y., Zhao, T., & Kawahara, T. (2019). Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. Interspeech,

Ma, E. (2019). *Data Augmentation for Audio*. https://medium.com/@makcedward/data-augmentation-for-audio-76912b01fdf6

Mehu, M., & Dunbar, R. (2008). Relationship between Smiling and Laughter in Humans (Homo sapiens): Testing the Power Asymmetry Hypothesis. *Folia primatologica; international journal of primatology*, *79*, 269-280. https://doi.org/10.1159/000126928

Mehu, M., Grammer, K., & Dunbar, R. I. M. (2007). Smiles when sharing. *Evolution and Human Behavior*, *28*(6), 415-422. https://doi.org/https://doi.org/10.1016/j.evolhumbehav.2007.05.010

Prombut, N., Waijanya, S., & Promrit, N. (2021). *Feature Extraction Technique Based on Conv1D and Conv2D Network for Thai Speech Emotion Recognition*. https://doi.org/10.1145/3508230.3508238

Promrit, N. (2020). *Visualizing Kernels and Feature Maps in Deep Learning Model (CNN)*. https://blog.pjjop.org/visualizing-filters-and-feature-maps-in-deep-learning-cnn/

Rath, S. R. (2019). *Deep Learning: An Introduction to Convolutional Neural Networks*. https://debuggercafe.com/deep-learning-an-introduction-to-convolutional-neural-networks/

Roberts, L. (2020). *Understanding the Mel Spectrogram*. https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

Satt, A., Rozenberg, S., & Hoory, R. (2017). *Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms*. https://doi.org/10.21437/Interspeech.2017-200

Serengil, S. I., & Ozpinar, A. (2020, 15-17 Oct. 2020). LightFace: A Hybrid Deep Face Recognition Framework. 2020 Innovations in Intelligent Systems and Applications Conference (ASYU),

Shor, R. E. (1978). The Production and Judgment of Smile Magnitude. *Journal of General Psychology*, *98*, 79-96.

Singh, N., Khan, R. A., & Shree, R. (2012). MFCC and Prosodic Feature Extraction Techniques: A Comparative Study. *International Journal of Computer Applications*, *54*, 9-13.

Suk, M., & Prabhakaran, B. (2014, 23-28 June 2014). Real-Time Mobile Facial Expression Recognition System -- A Case Study. 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops,

Tartter, V. C., & Braun, D. (1994). Hearing smiles and frowns in normal and whisper registers. *The Journal of the Acoustical Society of America*, *96 4*, 2101-2107.

Team, I. D. a. A. (2023). *AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the difference?* https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/

Tensorflow.org. (2015). *Introduction to TensorFlow*. https://www.tensorflow.org/

Torre, I. (2013). *Production and Perception of Smiling Voice*.

Torre, I., Goslin, J., & White, L. (2020). If your device could smile: People trust happy-sounding artificial agents more. *Computers in Human Behavior*, *105*, 106215. https://doi.org/https://doi.org/10.1016/j.chb.2019.106215

van Hooff, J. (1972). A comparative approach to the phylogeny of laughter and smiling.

VISTEC. (2021). *Thai Speech Emotion Dataset*. VISTEC. https://github.com/vistec-AI/dataset-releases/releases

Wikipedia. (2023). *Machine learning*. https://en.wikipedia.org/wiki/Machine_learning

Wikipedia, S. (2023). *Sound*. https://en.wikipedia.org/wiki/Sound

Yalçın, O. G. (2021). *The Brief History of Convolutional Neural Networks*. https://towardsdatascience.com/the-brief-history-of-convolutional-neural-networks-45afa1046f7f

Zendesk. (2023). *What is a call center? Definition, types, and how they work*. https://www.zendesk.com/th/blog/ultimate-guide-call-centers/

มหาวิทยาลัยศิลปากร

ประวัติผู้เขียน

ชื่อ-สกุล	นริศร์ พรหมบุตร
วุฒิการศึกษา	ป.โท มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ป.ตรี มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ผลงานตีพิมพ์	Waijanya, S., Promrit, N., Prombut, N. (2021), Feature Extraction Technique Based on Conv1D and Conv2D Network for Thai Speech Emotion Recognition. NLPIR 2021: 2021 5th International Conference on Natural Language Processing and Information Retrieval
รางวัลที่ได้รับ	BEST PRESENTATION AWARD ผลงาน Feature Extraction Technique based on Conv1D and Conv2D Network for Thai Speech Emotion Recognition ในงานประชุมวิชาการระดับนานาชาติ 2021 5th International Conference on Natural Language Processing and Information (NLPIR 2021)