KKU SCIENCE JOURNAL
Journal Home Page : https://ph01.tci-thaijo.org/index.php/KKUSciJ
Published by the Faculty of Science, Khon Kaen University, Thailand

การจำแนกเสียงคนจริงและเสียงสังเคราะห์ปัญญาประดิษฐ์ด้วยโครงข่ายประสาทเทียมแบบคอนโวลูชัน

Real Human Voice and Artificial Intelligence Synthetic Voice Recognition with Convolutional Neural Networks

กฤติภูมิ ผลาจันทร์[1] สุพศิน วงศ์ลาภสุวรรณ[1] ธนพล รุ่มนุ่ม[1] สัจจาภรณ์ ไวจรรยา[2*] และ ณัฐโชติ พรหมฤทธิ์[2*]

Krittipoom Phalachun[1], Supasin Wonglapsuwan[1], Tanaphon Rumnum[1], Sajjaporn Waijanya[2*] and Nuttachot Promrit[2*]

[1]สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร วิทยาเขต พระราชวังสนามจันทร์ จังหวัดนครปฐม 73000

[2]ภาควิชาคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร วิทยาเขต พระราชวังสนามจันทร์ จังหวัดนครปฐม 73000

[1]Data Science Major, Faculty of Science, Silpakorn University, Nakhon Pathom, 73000, Thailand

[2]Department of Computing, Faculty of Science, Silpakorn University, Nakhon Pathom, 73000, Thailand

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อจำแนกเสียงคนจริงและเสียงสังเคราะห์ เพื่อป้องกันการเกิดอาชญากรรมอันเนื่องมาจากการปลอมแปลงเสียงด้วยเทคโนโลยี Deepfake Voice โดยมีกรณีที่บริษัทพลังงานที่ถูกหลอกให้โอนเงินประมาณ 200,000 ปอนด์ (260,000 ดอลลาร์) หลังจากที่มิจฉาชีพใช้ Deepfake Voice เทคโนโลยีเสียงเพื่อเลียนแบบเสียงของประธานบริษัทเพื่ออนุมัติการชำระเงิน ในงานวิจัยนี้สร้างชุดข้อมูลขึ้นมาเองจากตัวอย่างเสียงของคนที่มีชื่อเสียง 15 คน โดยแบ่งเป็นชุดข้อมูลสำหรับฝึกฝนโมเดล ชุดข้อมูลสำหรับตรวจสอบโมเดล และชุดข้อมูลสำหรับทดสอบโมเดล คิดเป็น อัตราส่วน 75:15:10 ใช้วิธีการสกัดคุณลักษณะของเสียงด้วยเทคนิคสัมประสิทธิ์เซปสตรัมบนสเกลเมล (MFCC) จากนั้นสร้างโมเดลโครงข่ายประสาทเทียมแบบคอนโวลูชันในการจำแนกเสียง และใช้วิธีการวัดประสิทธิภาพของโมเดลด้วย Confusion matrix ได้ค่าความถูกต้องเท่ากับ 97%

ABSTRACT

This research aims to distinguish between real human voices and synthesized voices in order to prevent crime resulting from voice impersonation using deepfake voice technology. There have been cases where energy companies were scammed out of nearly £200,000 (USD 260,000) after criminals used deepfake voice technology to imitate the CEO's voice and approve payment. For the dataset used in this research, 15 famous individuals' voices were recorded and divided into three sets: a training set, a validation set, and a testing set, with a ratio of 75:15:10. The Mel-frequency cepstral coefficients (MFCC) were extracted

*Corresponding Author, E-mail: waijanya_s@silpakorn.edu, promrit_n@silpakorn.edu

Received: date: 2 June 2023 | Revised date: 20 June 2023 | Accepted date: 7 July 2023

as the features of the voices and a convolutional neural network (CNN) was used to classify the voices. The performance of the model was evaluated using a confusion matrix, and the accuracy was found to be 97%.

คำสำคัญ: เสียงสังเคราะห์ จำแนกเสียง โครงข่ายประสาทเทียมแบบคอนโวลูชัน
Keywords: Voice Synthesis, Voice Classification, Convolutional Neural Network

บทนำ

การใช้รหัสผ่านอย่างเดียวในการยืนยันตัวตนสำหรับรับบริการต่างๆ นั้นยังไม่เพียงพอในเรื่องของความปลอดภัย เนื่องจากเทคนิควิธีการคาดเดาและถอดรหัสผ่านนั้นแพร่หลายขึ้นและมีประสิทธิภาพมากขึ้นเรื่อยๆ

ในปัจจุบันธนาคารบางแห่งในประเทศไทยในปัจจุบัน เช่น ธนาคารกรุงศรีและธนาคารกสิกรไทย และบริษัทประกันภัยอย่างเมืองไทยประกันชีวิต เริ่มหันมาให้ความสนใจการใช้เสียงในการยืนยันตัวตน การยืนยันตัวตนด้วยเสียงนั้นมีความปลอดภัยมากขึ้นเนื่องจากมีคุณสมบัติที่ไม่ซ้ำซ้อนกับคนอื่นอย่างเช่น สำเนียง ภาษาถิ่น ความเร็ว และระดับเสียง

อย่างไรก็ตามการยืนยันตัวตนด้วยเสียงก็ยังไม่สมบูรณ์แบบดี นักวิจัยจากมหาวิทยาลัยอลาบามาในเบอร์มิงแฮม (Mukhopadhyay *et al*., 2015) ได้แสดงให้เห็นว่าเทคโนโลยีการจดจำเสียงมีความเสี่ยงที่จะถูกปลอมแปลงโดยใช้ตัวอย่างเสียงที่ลอกแบบมา เทคโนโลยีเสียงสังเคราะห์สามารถเก็บรวบรวมตัวอย่างเสียงปริมาณมากได้อย่างง่ายและรวดเร็วเช่นจากโซเชียลมีเดีย หรือการดักจับหรืออัดเสียงผ่านโทรศัพท์มือถือ จึงสามารถมองได้ว่าการยืนยันตัวตนด้วยเสียงสามารถปลอมแปลงได้ง่ายและมีความปลอดภัยน้อย

นอกจากนี้ในบทความของ Ring (2021) ได้กล่าวถึงการใช้เทคโนโลยี deepfake voice นำไปก่ออาชญากรรมไซเบอร์โดยมีกรณีหนึ่งที่เกี่ยวข้องกับการรายงานในสหราชอาณาจักรบริษัทพลังงานที่ถูกหลอกให้โอนเงินประมาณ 200,000 ปอนด์ (260,000 ดอลลาร์) ให้กับบัญชีธนาคารของชาวฮังการี หลังจากที่มิจฉาชีพใช้ deepfake เทคโนโลยีเสียงเพื่อเลียนแบบเสียงของ CEO ของบริษัทเพื่ออนุมัติการชำระเงิน และ ในงานวิจัยของ Amezaga และ Hajek (2022) ได้กล่าวถึงความง่ายต่อการเข้าถึงเทคโนโลยี deepfake voice ได้อย่างง่ายดายในปัจจุบันซึ่งอาจจะเพิ่มอัตราการเกิดอาชญากรไซเบอร์เพิ่มมากยิ่งขึ้น เพื่อป้องกันปัญหาเหล่านี้ผู้วิจัยจึงศึกษาแนวทางในการตรวจจับเสียงสังเคราะห์ปัญญาประดิษฐ์

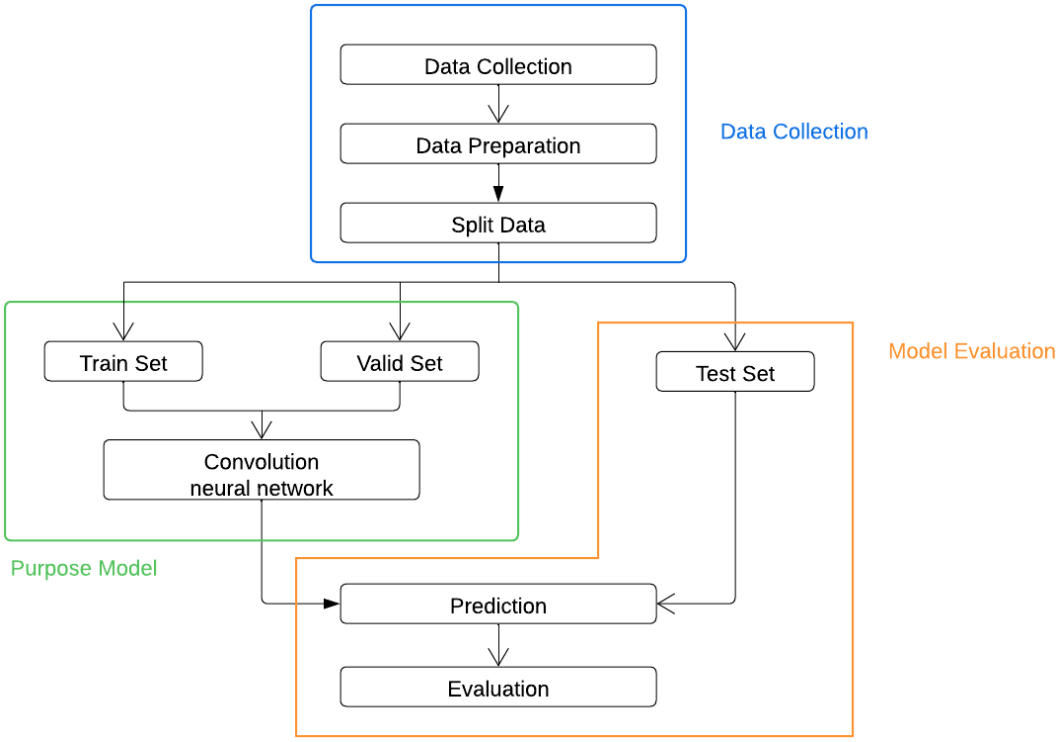
ผลจากการวิจัยที่นำเสนอในบทความนี้สามารถนำไปใช้เป็นวิธีการเบื้องต้นในการแยกเสียงของเสียงพูดมนุษย์และเสียงพูดสังเคราะห์จากปัญญาประดิษฐ์ เพื่อป้องกันไม่ให้เกิดการปลอมแปลงหรือสร้างความเสียหายอื่นๆ เช่น ปลุกปั่นสังคม สร้างหลักฐาน ปรักปรำผู้อื่น หรือบิดเบือนความจริง โดยผู้วิจัยจะได้อธิบายรายละเอียดของวิธีการดำเนินการวิจัย ผลการวิจัย และการอธิปรายผล รวมทั้งสรุปผลการวิจัยในแต่ละหัวข้อเป็นลำดับต่อไป

วิธีการดำเนินการวิจัย

ภาพรวมของกระบวนการดำเนินงานและการสร้างโมเดลแสดงดังรูปที่ 1 ซึ่งประกอบด้วยส่วนการรวบรวมข้อมูล (Data Collection) ส่วนการฝึกฝนโมเดลที่เป็นเป้าหมาย (Purpose Model) และส่วนการประเมินประสิทธิภาพโมเดล ซึ่งจะอธิบายรายละเอียดของแต่ละส่วน ในลำดับถัดไป

1. การรวบรวมข้อมูล

งานวิจัยนี้สร้างชุดข้อมูลขึ้นมาเองโดยรวบรวมเสียงสังเคราะห์จากเว็บไซต์ Fakeyou และ Speechify และเสียงของคนจากเว็บไซต์ Youtube มีรายละเอียดตามตารางที่ 1



รูปที่ 1 ภาพรวมของกระบวนการดำเนินงานและการสร้างโมเดล

ตารางที่ 1 รายละเอียดชุดข้อมูล

ประเภท	ชื่อ	ความยาวเสียง (วินาที)	ประเภทไฟล์	จำนวนไฟล์
fake	Adam Driver	114	wav	12
fake	Al Michaels	95	wav	11
fake	Alan Rickman	89	wav	10
fake	Gwyneth Paltrow	92	mp3	1
fake	Jeff Goldblum	98	wav	14
fake	Jim Ross	85	wav	8
fake	John F. Kennedy	86	wav	9
fake	Mitch Mcconnell	58	wav	6
fake	Obama	74	mp3	1
fake	Richard Nixon	241	wav	9
fake	Ronald Reagan	110	wav	12
fake	Snoop Dogg	118	mp3	1
fake	Stan Lee	74	wav	7
fake	Tobey Maguire	63	wav	7
fake	Tom Hanks	103	wav	10
real	Adam Driver	126	mp3	1

ตารางที่ 1 รายละเอียดชุดข้อมูล (ต่อ)

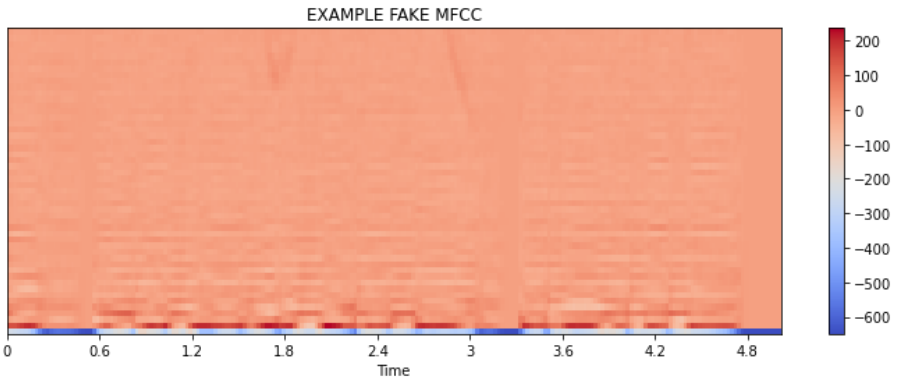
ประเภท	ชื่อ	ความยาวเสียง (วินาที)	ประเภทไฟล์	จำนวนไฟล์
real	Al Michaels	137	mp3	1
real	Alan Rickman	129	mp3	1
real	Gwyneth Paltrow	128	mp3	1
real	Jeff Goldblum	120	mp3	1
real	Jim Ross	126	mp3	1
real	John F. Kennedy	128	mp3	1
real	Mitch Mcconnell	122	mp3	1
real	Obama	126	mp3	1
real	Richard Nixon	130	mp3	1
real	Ronald Reagan	142	mp3	1
real	Snoop Dogg	125	mp3	1
real	Stan Lee	119	mp3	1
real	Tobey Maguire	108	mp3	1
real	Tom Hanks	123	mp3	1

จากตารางที่ 1 จะเห็นว่ามีไฟล์ที่เป็นเสียงสังเคราะห์ 118 ไฟล์และเสียงคน 15 ไฟล์ ที่มาจากเสียงของบุคคลที่มีชื่อเสียง 15 คนและใช้ประโยคในการพูดเหมือนกัน

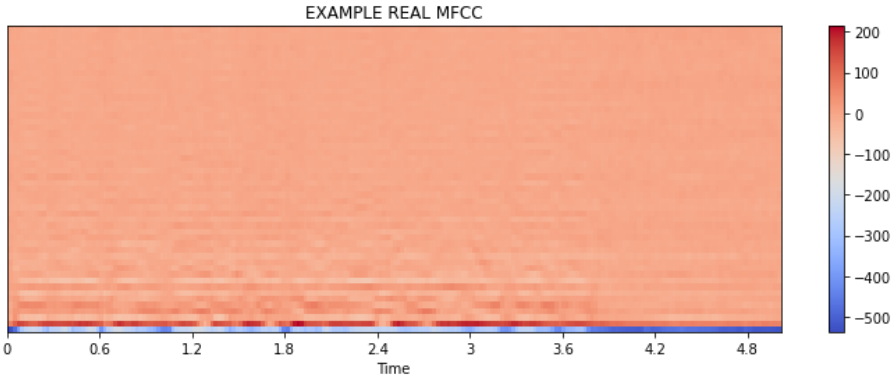
2. การเตรียมข้อมูล

งานวิจัยนี้จะนำเสียงสังเคราะห์และเสียงคนมาตัด โดยตัดแบ่งออกมา 5 วินาทีโดยข้อมูลที่ถูกตัดออกมาแล้วจะมีข้อมูลที่ไม่ถึง 5 วินาที จึง padding เพื่อทำให้ข้อมูลมีความยาวเท่ากันก็คือ 5 วินาที สุดท้ายจะได้ข้อมูลเสียงออกมาทั้งหมด 564 เสียง

จากนั้น สกัดคุณลักษณะโดยใช้วิธี MFCC ซึ่งตัวอย่างผลลัพธ์คุณลักษณะจาก MFCC ที่ได้จากตัวอย่างเสียงสังเคราะห์และเสียงคน แสดงได้ดังรูปที่ 2 และ 3



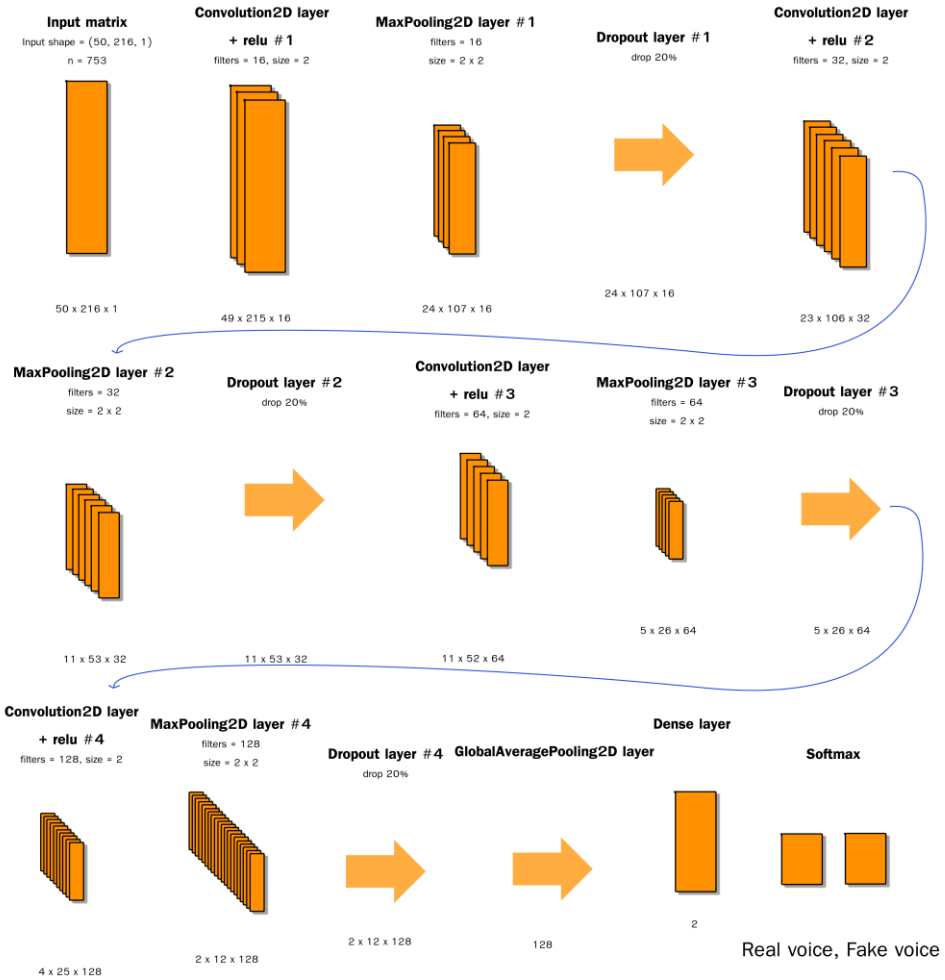
รูปที่ 2 คุณลักษณะจาก MFCC ที่ได้จากตัวอย่างเสียงสังเคราะห์



รูปที่ 3 คุณลักษณะจาก MFCC ที่ได้จากตัวอย่างเสียงคน

3. การสร้างโมเดล

ในการนิยามโมเดลแสดงดังรูปที่ 4 มีการกำหนด Input Shape = (50,216,1) ตามขนาดสูงสุดของ tensor ของข้อมูล โดยในเลเยอร์แรกของโมเดลคือ Convolution2D layer ที่มีขนาดของ filters เท่ากับ 2 และชั้นของ filters เท่ากับ 16 ชั้น และใช้ ReLU activation function ชั้นต่อมาเป็น max pooling2D layer ที่มีขนาด 2 x 2 และมี Dropout layer ในการสุ่มปิดโหนด เพื่อลดปัญหา Overfitting โดยกำหนดค่าเท่ากับ 0.2 ทำซ้ำทั้ง 3 เลเยอร์ เป็นจำนวน 4 รอบ โดยค่าทุกอย่างคงเดิม แต่ปรับชั้นของ filters เป็น 32 64 และ 128 ตามลำดับ หลังจากนั้นเป็น GlobalAveragePooling2D layer และชั้นสุดท้ายคือ Dense layers ที่มีโหนดเท่ากับจำนวน label และใช้ Softmax Activation Function



รูปที่ 4 Convolutional Neural Network Model ของการจำแนกเสียงคนจริงและเสียงสังเคราะห์

โมเดลในงานวิจัยนี้เรียนรู้ด้วย Learning rate เท่ากับ 0.001 Momentum เท่ากับ 0.9 และจำนวน Epoch เท่ากับ 100 รอบ

การวัดประสิทธิภาพของ Convolutional Neural Network Model จะแสดงผลออกมาในรูปตาราง Confusion matrix โดยประกอบด้วยค่าที่สำคัญดังนี้

Accuracy เป็นการวัดความแม่นยำในการทำนายของโมเดล โดยพิจารณารวมทุกประเภทของ output

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision เป็นการวัดความเที่ยงของข้อมูล โดยพิจารณาแยกทีละประเภท

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall เป็นการวัดความถูกต้องของโมเดล โดยพิจารณาแยกทีละประเภท

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

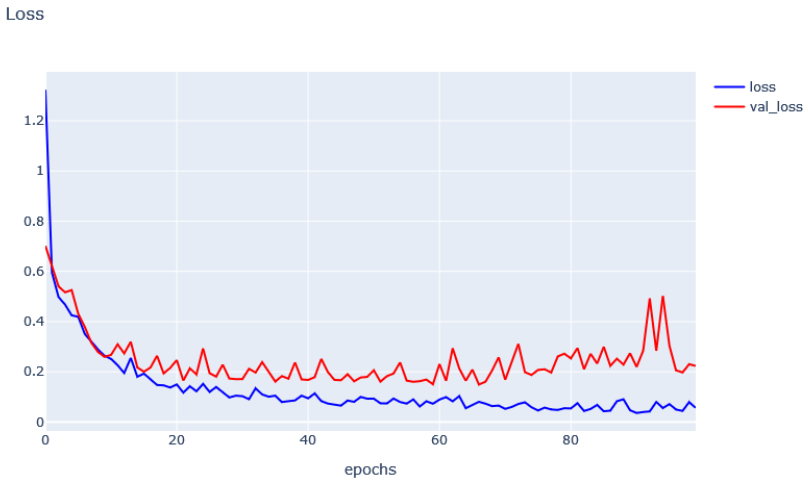
F1-Score เป็นการวัดความสามารถของโมเดล

$$F1 = 2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

ผลการวิจัยและวิจารณ์ผล

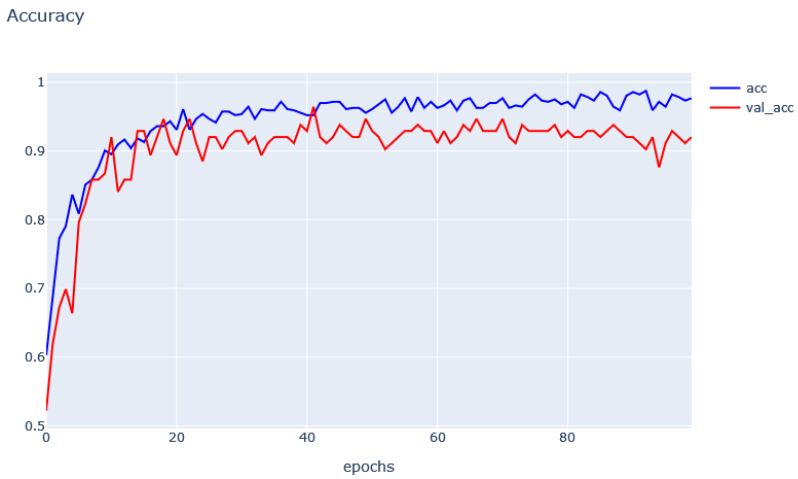
การจำแนกเสียงพูดของคนและเสียงสังเคราะห์ของปัญญาประดิษฐ์ใช้ข้อมูลสำหรับฝึกฝน 564 เสียง ชุดข้อมูลสำหรับตรวจสอบโมเดล 113 เสียง และชุดข้อมูลสำหรับทดสอบโมเดล 76 เสียง และฝึกฝนโมเดล 100 รอบ ด้วย Convolutional Neural Network Model ดังรูปที่ 4 ได้ผลลัพธ์โมเดล ดังต่อไปนี้

1. การทดสอบวัดประสิทธิภาพโมเดล



รูปที่ 5 Training and Validation loss

จากการสังเกตกราฟในรูปที่ 5 ที่ฝึกฝนโมเดล 100 รอบ กราฟนี้ค่า training loss กับ validation loss พบว่า ค่า loss สำหรับชุดข้อมูลฝึกฝนโมเดลลดลงอย่างต่อเนื่องในช่วงประมาณรอบที่ 1 ถึงรอบที่ 20 จากนั้นค่าคงที่ ที่ค่า loss เกือบเท่ากับ 0 และสำหรับชุดข้อมูลตรวจสอบโมเดลมีค่า loss ลดลงอย่างต่อเนื่องในช่วงประมาณรอบที่ 1 ถึงรอบที่ 20 และคงที่จนถึงรอบที่ 70 มีค่า loss เพิ่มขึ้นและลดลงมา



รูปที่ 6 ค่าความถูกต้องของ Training and Validation

จากการสังเกตรูปที่ 6 ที่ฝึกฝนโมเดล 100 รอบ กราฟค่าความถูกต้องของ training กับ validation พบว่าค่าความถูกต้องสำหรับชุดข้อมูลฝึกฝน โมเดลเพิ่มขึ้นอย่างต่อเนื่องในช่วงประมาณรอบที่ 1 ถึงรอบที่ 20 จากนั้นคงที่ ที่ค่าความถูกต้องเกือบเท่ากับ 1.0 และสำหรับชุดข้อมูล ตรวจสอบโมเดลค่าความถูกต้องเพิ่มขึ้นอย่างต่อเนื่องในช่วง ประมาณรอบที่ 1 ถึงรอบที่ 20 จากนั้น ค่าความถูกต้องมีการแกว่งอยู่ในช่วง 0.857 ถึง 0.96 ตั้งแต่รอบที่ 20 ขึ้นไป

ทุกรอบของการฝึกฝนในการสร้างโมเดลจะบันทึกโมเดลที่มีค่าความถูกต้องของ validation สูงสุดเอาไว้เป็นไฟล์ประเภท h5 เพื่อให้สามารถโหลดมาใช้ในการทดสอบสรุปผลการดำเนินการวิจัย เพื่อให้ได้โมเดลที่มีประสิทธิภาพมากที่สุด จึงโหลดโมเดลรอบที่ 41 หรือโมเดลที่บันทึกก่อนรอบที่ 42 ซึ่งเริ่มเกิดการ overfit มาใช้ในการทดสอบประสิทธิภาพ

ตารางที่ 2 Confusion matrix จาก CNN Model

		Actual	
		Real	Fake
Predict	Real	35	2
	Fake	0	39

จากตารางที่ 2 Confusion matrix อธิบายได้ว่าโมเดลสามารถทำนายเสียงคนจริงว่าเป็นเสียงคนจริงได้ถูกต้องทั้งหมดในชุดข้อมูลทดสอบ หมายความว่าไม่มีเสียงคนจริงที่โมเดลทำนายผิดว่าเป็นเสียงสังเคราะห์ แต่การทำนายในส่วนของเสียงสังเคราะห์มีการทำนายผิดว่าเป็นเสียงคนจริงอยู่ 2 จุดเท่านั้น คิดเป็น 0.026%

ค่า Confusion matrix ข้างต้น เมื่อนำมาเทียบค่าคำนวณความถูกต้องและความแม่นยำ จะได้ค่า Accuracy Precision Recall และ F1-score ดังตารางที่ 3

ตารางที่ 3 Precision Recall และ F1 Score จาก CNN Model

	Precision	Recall	F1-Score	Support	Accuracy
Fake	1	0.95	0.97	37	0.97
Real	0.95	1	0.97	39	0.97

จากตารางที่ 3 จะเห็นได้ว่า ค่าความถูกต้องของโมเดลโครงข่ายประสาทเทียมแบบคอนโวลูชันจะได้อยู่ที่ 0.97 การใช้โครงข่ายประสาทเทียมแบบคอนโวลูชันเข้ามาช่วยในการสร้างโมเดล ถือว่าเป็นปัจจัยที่สำคัญที่ส่งผลให้การจำแนกเสียงมีประสิทธิภาพ สามารถจำแนกเสียงจริงและเสียงสังเคราะห์ได้

2. ผลการดำเนินงานเปรียบเทียบกับงานวิจัยที่เกี่ยวข้อง

นอกจากงานวิจัยบทความนี้ที่มีการจำแนกเสียงแล้ว ยังมีงานวิจัยอื่นๆ ที่ได้ทดลองในลักษณะเดียวกัน ซึ่งได้สรุปรายละเอียด แสดงดังตารางที่ 4

ตารางที่ 4 เปรียบเทียบผลการดำเนินงานกับงานวิจัยที่เกี่ยวข้อง

ผู้จัดทำวิจัย	เทคนิค	ชุดข้อมูล	ผลสรุป
Narasimhan *et al*. (2017)	CNN กับ encoder-decoder architecture	ข้อมูลสำหรับแยกเสียงนก 2 ชุด	acc = 98.8%
Changwei *et al*. (2020)	CNN	ข้อมูลสำหรับแยกเสียงปกติอยู่ 53 เสียง และมีเสียงผิดปกติ 133 เสียง	acc = 96.51%
Hireš *et al*. (2022)	CNN	ข้อมูลสำหรับแยกคนสุขภาพดี 50 และคนเป็นโรคพาร์กินสัน 50 เสียง	acc = 99.0%
Kao *et al*. (2021)	CNN	ข้อมูล 5 ชุด สำหรับแยกอารมณ์ของมนุษย์จากเสียง พูด	acc = 80%
Ballesteros *et al*. (2021)	CNN	ข้อมูล 2092 สำหรับแยกเสียงจริงและเสียงปลอม	Precision = 0.985 Recall = 0.944
Khochare *et al*. (2021)	TCN, STN	Fake or Real (FoR) dataset สำหรับแยกเสียงจริงและเสียงปลอม	acc = 92%
Hamza *et al*. (2022)	MFCC, SVM	Fake or Real (FoR) dataset สำหรับแยกเสียงจริงและเสียงปลอม	acc = 67%
Reimao *et al*. (2019)	SVM	Fake or Real (FoR) dataset สำหรับแยกเสียงจริงและเสียงปลอม	acc = 73.46%
งานวิจัยนี้	MFCC และ CNN	รวบรวมและสร้างขึ้นมา Youtube และ Fakeyou	acc = 97.7% loss = 7.9%

จากตารางที่ 4 พบว่า งานวิจัยการจำแนกประเภทเสียงส่วนใหญ่ใช้เทคนิค CNN ถัดมาคือเทคนิค SVM และการสกัดคุณลักษณะด้วยเทคนิค MFCC นอกจากนี้สังเกตได้ว่างานวิจัยการจำแนกประเภทเสียงที่สรุปในตารางมีประสิทธิภาพการทำนายค่อนข้างดี โดยที่ค่าความถูกต้องต่ำสุดอยู่ที่งานวิจัยที่ใช้เทคนิค SVM และ MFCC กับชุดข้อมูล Fake or Real เท่ากับ 67%

สรุปผลการวิจัยและอภิปรายผล

บทความนี้เสนอการศึกษาและวิเคราะห์การจำแนกเสียงพูดสังเคราะห์และเสียงพูดจริง โดยในการสร้างโมเดล รวบรวมข้อมูลด้วยโปรแกรมสังเคราะห์เสียงจากเว็บไซต์ Fakeyou และ Speechify และเสียงของคนจากเว็บไซต์ Youtube ได้ชุดข้อมูลประกอบด้วยไฟล์ที่เป็นเสียงสังเคราะห์ 118 ไฟล์และเสียงคน 15 ไฟล์ ที่มาจากเสียงของบุคคลที่มีชื่อเสียง 15 คน สกัดคุณสมบัติจากเสียงเป็นสัมประสิทธิ์เซปสตรัมบนสเกลเมล (Mel Frequency Cepstral Coefficients: MFCC) เก็บไว้อยู่ในรูปแบบของ vector จากนั้นนำข้อมูลคุณสมบัติที่สกัดได้และผลเฉลยทั้งหมดแบ่งเป็น Training set Validation set และ Testing set อัตราส่วน 75:15:10 นำไปสร้างโมเดลโครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neural Network: CNN) ซึ่งนำไปสร้างเป็น Confusion matrix ได้ค่าความถูกต้องเท่ากับ 0.97 แสดงให้เห็นว่าโมเดลโครงข่ายประสาทเทียมแบบคอนโวลูชันสามารถจำแนกเสียงจริงและเสียงสังเคราะห์ได้อย่างมีประสิทธิภาพ ซึ่งจากผลลัพธ์ที่ได้ การใช้ cnn layers เข้ามาช่วยในการสร้างโมเดล deep learning ถือว่าเป็นปัจจัยที่สำคัญที่ส่งผลให้การจำแนกเสียงมีประสิทธิภาพ ซึ่งสอดคล้องกับงานวิจัยของ Ballesteros *et al*. (2021) ที่ได้ค่า precision 0.997 ในการจำแนกเสียงการกล่าวสุนทรพจน์ปลอม ดังนั้นในงานบางประเภทที่มีความจำเป็นต้องทราบว่าเสียงที่ถูกป้อนเข้ามาเป็นเสียงพูดของมนุษย์จริงหรือไม่ เช่น การใช้เสียงระบุตัวตน หรือการพิสูจน์เสียงพูดของบุคคลที่อาจใช้เทคโนโลยีสังเคราะห์เสียงโดยมีจุดประสงค์ร้ายสามารถใช้โมเดลงานวิจัยชิ้นนี้ในการทดสอบได้

ข้อเสนอแนะ

จากตารางที่ 4 พบว่านอกเหนือจากการแยกประเภทเสียงด้วยเทคนิค CNN แล้ว ยังมีเทคนิคอื่นที่เห็นได้ว่ามีประสิทธิภาพการทำนายที่ดี เช่น TCN STN SVM และเทคนิคอื่นๆ ที่ไม่ได้นำเสนอในบทความ ซึ่งไม่ได้ทดลองในงานวิจัยนี้ ดังนั้น จึงอาจสามารถต่อยอดงานวิจัยนี้ด้วยการทดลองโดยใช้เทคนิคอื่นๆ

เอกสารอ้างอิง

Amezaga, N. and Hajek, J. (2022). Availability of Voice Deepfake Technology and its Impact for Good and Evil. In: SIGITE'22: Proceedings of the 23rd Annual Conference on Information Technology Education. Association for Computing Machinery, New York. 23 - 28.

Ballesteros, D.M., Rodriguez-Ortega, Y., Renza, D., and Arce, G. (2021). Deep4SNet: Deep learning for fake speech classification. Expert Systems with Applications 184: 115465. doi: 10.1016/j.eswa.2021.115465.

Changwei, Z., Lili, Z., Xiaojun, Z., Yuanbo, W., Di, W. and Zhi, T. (2020). Classification of normal and pathological voices using convolutional neural network. In: 2020 International Conference on Sensing, Measurement & amp; Data Analytics in the Era of Artificial Intelligence (ICSMD). Xi'an Jiaotong University, Xi'an, China. 325-329. doi: 10.1109/icsmd50554.2020.9261730.

Hamza, A., Javed, A.R.R., Iqbal, F., Kryvinska, N., Almadhor, A.S., Jalil, Z. and Borghol, R. (2022). Deepfake audio detection via MFCC features using machine learning. IEEE Access 10: 134018 – 134028. doi: 10.1109/access.2022.3231480.

Hireš, M., Gazda, M., Drotár, P., Pah, N.D., Motin, M.A. and Kumar, D.K. (2022). Convolutional neural network ensemble for Parkinson's disease detection from voice recordings. Computers in Biology and Medicine 141: 105021. doi: 10.1016/j.compbiomed.2021.105021.

Kao, Y.C., Li, C.T., Tai, T.C. and Wang, J.C. (2021). Emotional speech analysis based on convolutional neural networks. In: 2021 9th International Conference on Orange Technology (ICOT). CMICSD Laboratory, National Cheng Kung University, Tainan, Taiwan. 1 - 4. doi: 10.1109/icot54518.2021.968 0651.

Khochare, J., Joshi, C., Yenarkar, B., Suratkar, S. and Kazi, F. (2021). A deep learning framework for audio deepfake detection. Arabian Journal for Science and Engineering 47(3): 3447 – 3458. doi: 10.1007/s13369-021-06297-w.

Mukhopadhyay, D., Shirvanian, M. and Saxena, N. (2015). All Your Voices are Belong to Us: Stealing Voices to Fool Humans and Machines. In: Computer Security -- ESORICS 2015. Lecture Notes in Computer Science, Vienna. 599 - 621.

Narasimhan, R., Fern, X.Z. and Raich, R. (2017). Simultaneous segmentation and classification of bird song using CNN. In: 2017 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP). IEEE Signal Processing Society, New Orleans, USA. 146 - 150. doi: 10.1109/icassp.2017.7952135.

Reimao, R. and Tzerpos, V. (2019). FoR: A dataset for synthetic speech detection. In: 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). Telecommunications and Information Technology, Politehnica University of Bucharest, Timisoara, Romania. 1 - 10. doi: 10.1109/SPED.2019.8906599.

Ring, T. (2021). Europol: the AI hacker threat to biometrics. Biometric Technology Today 2021(2): 9 - 11. doi: 10.1016/S0969-4765(21)00023-0.

❑❑❑❑❑