การรู้จำและการบ่งตัวตนของเสียงสภาพแวดล้อมและเสียงปืน-ปืนใหญ่ ด้วย MLP SVM และ DNN

นายชินวัฒน์ จัตุรัส

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมและเทคโนโลยีการป้องกันประเทศ ไม่สังกัดภาควิชา/เทียบเท่า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

CLASSIFICATION AND IDENTIFICATION OF ENVIRONMENTAL AND GUN-ARTILLERYSOUND USING MLP SVM AND DNN

Mr. Chinnavat Jatturas

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Defense Engineering and Technology
Common Course
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2020

หัวข้อวิทยานิพนธ์	การรู้จำและการบ่งตัวตนของเสียงสภาพแวดล้อมและเสียงปืน-ปืนใหญ่ ด้วย MLP SVM และ DNN
โดย	นายชินวัฒน์ จัตุรัส
สาขาวิชา	วิศวกรรมและเทคโนโลยีการป้องกันประเทศ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.วิทยากร อัศดรวิเศษ
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	พันเอก รองศาสตราจารย์ ดร.ผเดิม หนังสือ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์เจิดกุล โสภาวนิตย์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.วิทยากร อัศดรวิเศษ)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(พันเอก รองศาสตราจารย์ ดร.ผเดิม หนังสือ)

..... กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.ปริญญา สงวนสัตย์)

ชินวัฒน์ จัตุรัส : การรู้จำและการบ่งตัวตนของเสียงสภาพแวดล้อมและเสียงปืน-ปืนใหญ่ด้วย MLP SVM และ DNN. (CLASSIFICATION AND IDENTIFICATION OF ENVIRONMENTAL AND GUN-ARTILLERYSOUND USING MLP SVM AND DNN) อ.ที่ปรึกษาหลัก : ผศ. ดร.วิทยากร อัศดรวิเศษ, อ.ที่ปรึกษาร่วม : พ.อ.รศ. ดร.ผเดิม หนังสือ

วิทยานิพนธ์ฉบับนี้เสนอแนวทางการรู้จำและการบ่งตัวตนของเสียงสภาพแวดล้อมและเสียงปืน-ปืนใหญ่ โดยเสนอแบบจำลอง Support Vector Machine (SVM) Multi-Layer Perceptron (MLP) และ Deep Neural Networks (DNNs) อีกสองชนิด ได้แก่ Convolutional Neural Networks (CNNs) และ Recurrent Neural Networks (RNNs) วัตถุประสงค์หลักเพื่อศึกษาการรู้จำเสียงสภาพแวดล้อมและเสียงปืน-ปืนใหญ่ และขยายขอบเขตให้สามารถจำแนกระหว่างเสียงที่ไม่เป็นอันตรายและเสียงที่เป็นอันตราย ปัญหาหลักของการจำแนกเสียงเกิดจากสัญญาณเสียงมีคุณลักษณะที่ไม่คงที่ (Non-Stationary) และข้อมูลมีขนาดมิติทางเวลาสูง ด้วยเหตุนี้วิทยานิพนธ์นี้จึงเสนอแนวทางการแก้ปัญหาด้วยการประมวลผลก่อนหน้าด้วยผลการแปลงฟูเรียร์สั้น (Short-Time Fourier Transform, STFT) แล้วทำการสกัดคุณลักษณะด้วยการวิเคราะห์องค์ประกอบหลัก (Principal Components Analysis, PCA) และทำการจำแนกด้วย SVM และ MLP นอกจากนี้ด้วยสมมติฐาน เบื้องต้นที่ว่า STFT สามารถแปลงจากสัญญาณเสียงที่มีมิติขนาดหนึ่งมิติมาเป็นสัญญาณภาพ (image) ที่มีขนาดสองมิติได้ ทำให้เราสามารถนำ spectrogram ที่ได้จาก STFT มาประยุกต์ใช้กับการเรียนรู้ลึกชนิด CNN หรือ RNN ได้ในกรณีนี้ CNN และ RNN จะทำหน้าที่สกัดคุณลักษณะ และจำแนกไปพร้อมกับในระหว่างการเรียนรู้ ผลการทดลองวิทยานิพนธ์สรุปได้ว่าเครื่องมือที่สามารถทำนายเสียงสภาพแวดล้อมและเสียงปืน-ปืนใหญ่ ได้แม่นยำสุดคือ DNN ชนิด CNN

สาขาวิชา	วิศวกรรมและเทคโนโลยีการป้องกันประเทศ	ลายมือชื่อนิสิต .....
ปีการศึกษา	2563	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6070440621 : MAJOR DEFENSE ENGINEERING AND TECHNOLOGY
KEYWORD: Feature Extraction and Sound Classification

Chinnavat Jatturas : CLASSIFICATION AND IDENTIFICATION OF ENVIRONMENTAL AND GUN-ARTILLERYSOUND USING MLP SVM AND DNN . Advisor: Asst. Prof. WIDHYAKORN ASDORNWISED, Ph.D. Co-advisor: Assoc. Prof. Phaderm Nangsue, Ph.D.

This thesis proposes classification and identification of environmental and gun-artillery sound using Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) and two Deep Neural Networks (DNNs); Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The main objective is to study environment and gun-artillery sound classification and extend it to the classification between the harmful sound and environmental sound, issues in Sound Classification are from the non-stationary characteristic and high-dimensional temporal data. As a result, we proposes Short-Time Fourier Transform (STFT) for pre-processing. Next, the features will be extracted Principal Components Analysis (PCA) and then will be classified by SVM and MLP. In addition, according to assumption that, STFT is able to transform one dimensional speech signal to image which is two dimensional signal. Thus, we can use spectrogram from STFT with both of CNNs and RNNs approaches. In this case, CNNs and RNNs are able to extract the features in training process. The results conclude that, in case of environment and gun-artillery sound classification, CNNs of DNNs achieved the highest accuracy.

Field of Study:	Defense Engineering and Technology	Student's Signature .....
Academic Year:	2020	Advisor's Signature .....
		Co-advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของ ผู้ช่วยศาสตราจารย์ ดร. วิทยากร อัศดรวิเศษ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก และ พันเอก รองศาสตราจารย์ ดร.ผเดิม หนังสือ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ซึ่งได้ให้คำแนะนำ และให้การสนับสนุนการวิจัยเป็นอย่างดีตลอดมา ทำให้นิสิตมีความรู้ ความเข้าใจทฤษฎีและปฏิบัติมากยิ่งขึ้น ผู้วิจัยจึงขอขอบพระคุณไว้ ณ ที่นี้

ขอขอบคุณรองศาสตราจารย์เจิดกุล โสภาวนิตย์ ประธานกรรมการสอบวิทยานิพนธ์ รองศาสตราจารย์ ดร.ปริญญา สงวนสัตย์ กรรมการสอบวิทยานิพนธ์ ที่ได้สละเวลาตรวจสอบ ให้คำแนะนำและตอบคำถามทุกคำถาม เพื่อให้ผู้วิจัยเรียนรู้ถึงรายละเอียดของวิทยานิพนธ์ทุกประเด็นเพื่อทำให้วิทยานิพนธ์ฉบับนี้สมบูรณ์ยิ่งขึ้น และขอขอบคุณอาจารย์ทุกท่านที่ประสิทธิ์ประสาทวิชาความรู้ อบรมสั่งสอน จนทำให้ข้าพเจ้ามีความรู้ ความสามารถในการทำงาน และดำรงชีวิตในสังคมได้อย่างมีความสุข

ขอขอบคุณหลักสูตรวิศวกรรมและเทคโนโลยีการป้องกันประเทศและภาควิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยที่ให้ความรู้และประสบการณ์ดี ๆ ทั้งด้านวิชาการ ด้านสังคมและอื่น ๆ แก่ข้าพเจ้า

ชินวัฒน์ จัตุรัส

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของวิทยานิพนธ์

ปัจจุบันงานวิจัยด้านการเรียนรู้จำเสียงสภาพแวดล้อม (environmental sound recognition) เป็นที่ได้รับความสนใจอย่างแพร่หลาย และมีการนำไปประยุกต์ใช้งานในการสร้างแอพพลิเคชันต่าง ๆ เพื่ออำนวยความสะดวกให้กับผู้บริโภค อาทิเช่น ระบบสั่งการด้วยเสียงอัจฉริยะหรือที่เรียกว่า สิริ เป็นต้น อุปสรรคของงานด้านการเรียนรู้จำเสียงสภาพแวดล้อมมีปัจจัยหลักคือสัญญาณเสียงมีคุณลักษณะที่ไม่คงที่ (non-Stationary) และสัญญาณเสียงสภาพแวดล้อมมีขนาดมิติ (dimension) ใหญ่มาก ทำให้เครื่องมือที่จะมาจำแนก (classification) เสียงสภาพแวดล้อมเป็นเรื่องยาก

จากที่กล่าวมา หนึ่งในวิธีที่จะมาแก้ปัญหาของการเรียนรู้จำเสียงสภาพแวดล้อมได้แก่ การสกัดคุณลักษณะ ข้อดีของการสกัดคุณลักษณะช่วยแยกสัญญาณเสียงสภาพแวดล้อมและลดมิติข้อมูลที่มีขนาดใหญ่มาก ด้วยเหตุนี้ งานวิจัยเราได้เสนอวิธีการสกัดคุณลักษณะด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้น (Short-Time Fourier Transform, STFT) [1] เพื่อใช้สำหรับการแยกสัญญาณเสียงสภาพแวดล้อมและแก้ปัญหาสัญญาณที่ไม่คงที่ โดยวิธีการทำของผลการแปลงฟูเรียร์ช่วงเวลาสั้น ขั้นตอนแรกทำการสร้างฟังก์ชันหน้าต่าง (window function) เพื่อนำมาตัดเสียงแต่ละช่วงเวลา–ความถี่เท่า ๆ กัน จากนั้นนำสัญญาณมาวิเคราะห์ด้วยผลการแปลงฟูเรียร์แบบเร็ว (Fast Fourier Transform, FFT) แปลงจากสัญญาณโดเมนทางเวลา (time-domain) มาเป็นสัญญาณโดเมนทางความถี่ (frequency-domain) ในช่วงเวลาสั้น ทำให้ได้ค่าสเปกโตรแกรม (spectrograms) นำไปใช้ในงานจำแนกเสียงสภาพแวดล้อม นอกจากนี้ ผลการแปลงฟูเรียร์ช่วงเวลาสั้นสามารถนำมาประยุกต์ใช้กับโครงข่ายประสาทลึก (Deep Neural Networks, DNNs) ที่ปกติใช้ในการจำแนกข้อมูลด้วยภาพ (image)

หลังจากแก้ปัญหาการเรียนรู้จำเสียงสภาพแวดล้อมด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้น อีกปัญหาของงานวิจัยนี้ข้อมูลมีมิติขนาดใหญ่มาก ด้วยเหตุนี้ เราจึงต้องทำการลดขนาดมิติ (reduce dimension) เสียงสภาพแวดล้อมก่อนเข้าสู่เครื่องมือการจำแนก ซึ่งวิธีการลดมิติข้อมูลเสียงสภาพแวดล้อมมีหลากหลายวิธีขึ้นอยู่กับความเหมาะสมกับเครื่องมือการจำแนก

จากงานวิจัยเครื่องมือที่นำมาใช้ในการจำแนกเสียงได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine, SVM) เพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron, MLP) และโครงข่ายประสาทลึกโดยมีเครือย่อยได้แก่ โครงข่ายประสาทคอนโวลูชัน (Convolutional Neural Networks, CNNs) โครงข่ายประสาทเกิดซ้อน (Recurrent Neural Networks, RNNs) จากการทดลองเครื่องมือการจำแนกเสียงสภาพแวดล้อมมีวิธีการสกัดคุณลักษณะที่แตกต่างกัน ด้วยเหตุนี้ งานวิจัยเราจะแบ่งวิธีการสกัดคุณลักษณะกับเครื่องมือการจำแนกออกเป็น 2 วิธี วิธีแรกคือ ซัพพอร์ตเวกเตอร์แมชชีนกับเพอร์เซ็ปตรอนหลายชั้นจะใช้การสกัดคุณลักษณะด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้นและการวิเคราะห์องค์ประกอบหลัก (Principal Components Analysis, PCA) [2] สำหรับการลดมิติที่มีขนาดข้อมูลใหญ่มาก ส่วนวิธีที่สองเครื่องมือจำแนกโครงข่ายประสาทคอนโวลูชันใช้วิธีการสกัดคุณลักษณะจากผลการแปลงฟูเรียร์ช่วงเวลาสั้น ที่ซึ่งเราได้มาจากการแปลงข้อมูลสัญญาณเสียงมาเป็นข้อมูลภาพ นอกจากนี้ ข้อดีของโครงข่ายประสาทคอนโวลูชันภายในจะมี convolutional layer และ pooling layer ที่ลดมิติแทนการวิเคราะห์องค์ประกอบหลัก ส่วนเครื่องมือการจำแนกสุดท้ายคือโครงข่ายประสาทเกิดซ้อนจะใช้การสกัดคุณลักษณะโดยอ้อมเหมือนกับโครงข่ายประสาทคอนโวลูชัน ข้อดีของโครงข่ายประสาทเกิดซ้อนไม่ต้องลดมิติข้อมูลเพราะว่าโครงข่ายประสาทเกิดซ้อนวิเคราะห์ข้อมูลแบบเป็นลำดับ (sequence) ส่วนข้อเสียของโครงข่ายประสาทเกิดซ้อนไม่สามารถจดจำข้อมูลได้หลายลำดับ ซึ่งสามารถแก้ปัญหาได้ด้วยการใช้หน่วยความจำระยะสั้นระยะยาว (Long Short-Term Memory, LSTM)

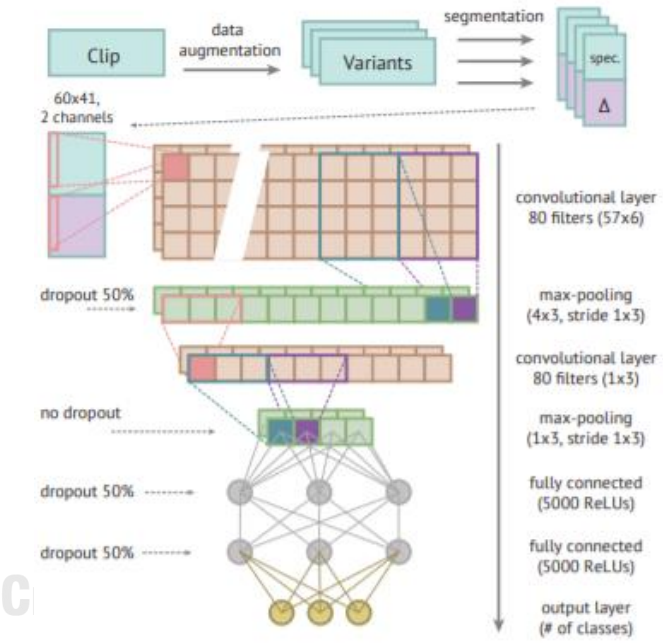
งานวิจัยนี้เกี่ยวกับการเรียนรู้จำเสียงสภาพแวดล้อม (Environmental Sound Recognition) แต่เพื่อให้สอดคล้องกับหลักสูตรป้องกันประเทศ เราจึงเพิ่มการเรียนรู้จำเสียงสภาพแวดล้อมเสียงปืนกับเสียงปืนใหญ่หรืออีกนัยหนึ่ง เพื่อต้องการแยกระหว่างเสียงปกติ (normal) กับเสียงอันตราย (harmful)

1.2 งานวิจัยที่เกี่ยวข้อง

C. Chang [3] ได้ทำเกี่ยวกับการจำแนกเสียงสภาพแวดล้อมและชุดข้อมูลเสียงสภาพแวดล้อมที่ใช้ทดสอบคือ Urbansound 8K ประกอบด้วยตัวอย่างการบันทึกจริง 8723 ตัวอย่าง มีทั้งหมด 10 ประเภทเสียงได้แก่ เสียงเครื่องปรับอากาศ เสียงแตรรถยนต์ เสียงเด็กเล่น เสียงสุนัขเห่า เสียงการเจาะของสว่าน เสียงการทำงานของรถยนต์เมื่อไม่เคลื่อนที่ เสียงยิงปืน เสียงเจาะจากเครื่องเจาะหิน เสียงไซเรน และเสียงดนตรีที่เล่นในสถานที่เปิด จากงานวิจัยของเขาขั้นตอนการทำจะต้องนำสัญญาณเสียงมาผ่านวิธีการสกัดคุณลักษณะด้วยสัมประสิทธิ์เซปสตรัมความถี่เมล (Mel Frequencies Cepstral Coefficients, MFCC) ซึ่งจะทำหน้าที่แปลงจากสัญญาณเสียงมาเป็นข้อมูลภาพเพื่อนำมาใช้กับเครื่องมือการจำแนก CNN

ตารางที่ 1.1 การเปรียบเทียบประสิทธิภาพของเครื่องมือการจำแนก

Model Architecture	Accuracy
1. CNN	73%
2. DeepNN	68%
3. Random Forest	61%
4. SVM	59%
5. RNN	56%
6. Naïve Bayes	23%



รูปที่ 1.1 แบบจำลองโครงข่ายคอนโวลูชัน

จากงานวิจัยของเขาเครื่องมือที่นำมาใช้ในการจำแนกเสียงสภาพแวดล้อม ได้แก่ Random Forest SVM DNN RNN และ CNN สรุปได้ว่าเครื่องมือการจำแนกที่สามารถเรียนรู้จำเสียงสภาพแวดล้อมได้ดีที่สุดคือ CNN แสดงดังตารางที่ 1.1

K. Piczak [4] ได้ทำเกี่ยวกับการเรียนรู้จำเสียงสภาพแวดล้อม จากงานวิจัยของเขาใช้การสกัดคุณลักษณะด้วยสัมประสิทธิ์เซปสตรัมบนสเกลเมลแปลงจากสัญญาณเสียงเป็นข้อมูลภาพที่เหมาะสำหรับการจำแนกเสียงด้วยวิธีโครงข่ายคอนโวลูชัน แสดงดังรูปที่ 1.1 จะเห็นได้ว่าข้อมูลภาพที่ใช้มีขนาด 64x41 และมี 2 channels เป็นอินพุต 2 นอกจากนี้ การจำแนกสำหรับโครงข่ายคอนโวลูชัน จะแบ่งออกเป็น 4 ส่วนหลัก ๆ ส่วนแรกคือการทำ convolutional layer โดยการลดขนาด

จาก 64x41 เหลือขนาด 57x6 และเพิ่มขนาดเป็น 80 คุณลักษณะ ส่วนที่สองคือการทำ max pooling ซึ่งมีขนาดตัวกรองเป็น 4x3 stride 1x3 ในการลดขนาดข้อมูลจะเลือกเฉพาะค่า max สำหรับการสร้างข้อมูลขึ้นมาใหม่ ส่วนที่สามคือการทำ fully connected พร้อมกับ relu ขั้นตอนนี้จะเป็นขั้นตอนสุดท้าย ก่อนเข้าสู่เครื่องมือการจำแนก นอกจากนี้ งานวิจัยเขาจะเพิ่มจำนวน dropout ที่ 50 เปอร์เซ็นต์เพิ่มตรงส่วน max pooling และตรงส่วน hidden layer คือ dropout จะช่วยลดการทำงานโดยเลือกสุ่ม node จาก hidden layer มาบางส่วนทำให้การฝึกข้อมูลเร็วขึ้น

เนื่องจากเราวิเคราะห์เสียงเป็นจำนวนมากจึงทำให้ข้อมูลมีขนาดมิติที่ใหญ่มาก ด้วยเหตุนี้การเรียนรู้จำเสียงสภาพแวดล้อมจึงต้องมีการสกัดคุณลักษณะของสัญญาณเสียงสภาพแวดล้อมก่อนการจำแนกเสียง การสกัดคุณลักษณะ [5] คือการแปลงข้อมูลต้นฉบับเป็นชุดข้อมูลใหม่ที่มีจำนวนตัวแปรลดลง ซึ่งจะทำให้การวิเคราะห์ข้อมูลเสียงง่ายขึ้น

C. Wang [6] ได้กล่าวว่าผลการแปลงฟูเรียร์ในกรณีที่มีสัญญาณหลายความถี่ การวิเคราะห์สัญญาณด้วยวิธีนี้นับว่ามีประโยชน์อย่างยิ่ง เนื่องจากให้ค่าความแม่นยำทางความถี่สูง โดยที่ว่าผลการแปลงฟูเรียร์เหมาะกับสัญญาณที่เป็นรายคาบ (stationary signal) มีความคงที่ของสัญญาณตลอดช่วงเวลา สำหรับกรณีที่สัญญาณเสียงไม่เป็นรายคาบ เช่น สัญญาณที่มีภาวะชั่วครู่และสัญญาณที่มีการเปลี่ยนแปลงแบบทันทีทันใด เป็นต้น การวิเคราะห์สัญญาณด้วยวิธีนี้อาจส่งผลให้เกิดความผิดพลาด

W.T. Cochran [7] ได้กล่าวว่าผลการแปลงฟูเรียร์แบบเร็วเป็นเครื่องมือคำนวณที่อำนวยความสะดวกในการวิเคราะห์สัญญาณ เช่น การวิเคราะห์สเปกตรัมโดยใช้คอมพิวเตอร์เป็นวิธีการคำนวณจากผลการแปลงฟูเรียร์แบบไม่ต่อเนื่อง หรือที่เรียกว่า อนุกรมเวลา (time series) ผลการแปลง ฟูเรียร์แบบเร็วเป็นที่นิยมใช้ในการแยกสัญญาณเสียงโดยการแปลงจากสัญญาณโดเมนทางเวลามาเป็นสัญญาณโดเมนทางเวลา-ความถี่

การวิเคราะห์ข้อมูลขนาดใหญ่เป็นที่รู้จักกันดีคือการวิเคราะห์ข้อมูล (data analytics) อาทิเช่น ชนิดของยา การเมือง และการขนส่ง แม้ว่าการวิเคราะห์ข้อมูลขนาดใหญ่จะใช้ในการปรับปรุงชีวิตมนุษย์ในหลายด้าน แต่ก็มาพร้อมกับปัญหาหนึ่งในนั้นคือคำสาปมิติ (curse of dimensionality) ซึ่งเป็นการเพิ่มขึ้นแบบเลขชี้กำลังของขนาดข้อมูลที่เกิดขึ้นจากมิติข้อมูลที่มีจำนวนมาก

เทคนิคการจำแนกประเภท (Classification) [8] คือเทคนิคที่สามารถสืบค้นหาฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Database, KDD) เป็นเทคนิคการจำแนกประเภทข้อมูลสำหรับกระบวนการสร้างแบบจำลองที่เรียกว่าข้อมูลสอนระบบ (training data) โดยแต่ละแถวของข้อมูลประกอบด้วยฟิลด์หรือแอทริบิวท์จำนวนมาก โดยแอทริบิวท์นี้อาจเป็นค่าต่อเนื่อง (continuous) หรือค่ากลุ่ม (categorical) โดยจะมีแอทริบิวท์แบ่ง (classifying attribute)

ซึ่งเป็นตัวบ่งชี้คลาสของข้อมูล จุดประสงค์ของการจำแนกประเภทข้อมูลคือการสร้างแบบจำลองนำไปประยุกต์ใช้ในหลายด้าน เช่น การจัดกลุ่มลูกค้าทางการตลาด การตรวจสอบความผิดปกติและการวิเคราะห์ทางการแพทย์ เป็นต้น

G. Guo [9] มีการนำเสนอเครื่องมือการจำแนก SVM เป็นอัลกอริทึมการเรียนรู้ใหม่สำหรับการจดจำรูปแบบต่าง ๆ บทความนี้พวกเขาได้กล่าวว่า SVM มีกลยุทธ์ต้นไม้ทวิภาค (binary tree) และถูกนำมาใช้เพื่อแก้ไขปัญหาการจำแนกเสียง โดยแสดงให้เห็นถึงศักยภาพของ SVM ในฐานข้อมูลเสียงทั่วไปซึ่งประกอบด้วย 409 เสียงจาก 16 คลาส ผลการจำแนกประเภทเปรียบเทียบ SVM กับแนวทางที่ได้รับความนิยมอื่น ๆ สำหรับการดึงข้อมูลเสียงได้มีการนำเสนอตัวชี้วัดใหม่ที่เรียกว่าระยะทางจากขอบเขต (Distance-From-Boundary, DFB)

1.3 วัตถุประสงค์

1. ศึกษาทฤษฎีที่เกี่ยวข้องสามารถนำมาประยุกต์ใช้กับงานด้านการเรียนรู้จำเสียงสภาพแวดล้อม
2. ศึกษาการสกัดคุณลักษณะ ได้แก่ ผลการแปลงฟูเรียร์แบบเร็ว ผลการแปลงฟูเรียร์ช่วงเวลาสั้น และการวิเคราะห์องค์ประกอบหลัก เพื่อใช้สำหรับการแยกคุณลักษณะเสียงสภาพแวดล้อมและลดขนาดมิติที่ใหญ่มาก
3. ศึกษาเครื่องมือการจำแนก ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน เพอร์เซ็ปตรอนหลายชั้น โครงข่ายประสาทคอนโวลูชัน โครงข่ายประสาทเกิดซ้อนและโครงข่ายประสาทลึก เพื่อใช้สำหรับการจำแนกเสียงสภาพแวดล้อมกับเสียงปืนใหญ่
4. สร้างแบบจำลองที่สามารถจำแนกเสียงสภาพแวดล้อมกับเสียงปืนใหญ่
5. ตีพิมพ์บทความทางวิชาการและจัดทำวิทยานิพนธ์

1.4 ขอบเขตวิทยานิพนธ์

1. สามารถรู้จำเสียงสภาพแวดล้อม
2. สร้างแบบจำลองการรู้จำเสียงสภาพแวดล้อมและเสียงปืนใหญ่
3. สามารถนำไปประยุกต์ต่อวิทยาการป้องกันประเทศ

บทที่ 2
หลักการและทฤษฎีทั่วไป

การศึกษาค้นคว้าหลักการและทฤษฎีทั่วไปสำหรับงานวิจัยด้านการเรียนรู้จำเสียงสภาพแวดล้อม เราจะแบ่งออกเป็น 2 หัวข้อ ได้แก่ การสกัดคุณลักษณะและเครื่องมือการจำแนกเสียงสภาพแวดล้อม

2.1 การสกัดคุณลักษณะ (feature extraction)

จากงานวิจัยนี้การสกัดคุณลักษณะใช้สำหรับการแยกสัญญาณเสียงสภาพแวดล้อมและลดมิติข้อมูลที่มีขนาดใหญ่มาก โดยวิธีการแยกเสียงเราได้เสนอผลการแปลงฟูเรียร์ช่วงเวลาสั้นในการแปลงจากสัญญาณเสียงโดเมนทางเวลามาเป็นโดเมนทางความถี่ ส่วนวิธีการการวิเคราะห์องค์ประกอบหลักเรานำมาใช้ลดมิติขนาดข้อมูลที่ใหญ่มาก จากที่กล่าวมาผลการแปลงฟูเรียร์ช่วงเวลาสั้นและการวิเคราะห์องค์ประกอบหลักใช้กับเครื่องมือการจำแนกเพอร์เซ็ปตรอนหลายชั้นและซัพพอร์ตเวกเตอร์แมชชีน

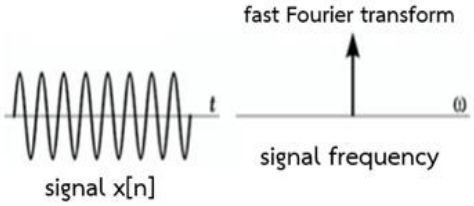
ส่วนวิธีการจำแนกเสียงสภาพแวดล้อมที่ใช้การสกัดคุณลักษณะด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้น ทำการประยุกต์จากข้อมูลสัญญาณเสียงแปลงเป็นข้อมูลภาพที่เหมาะสำหรับเครื่องมือการจำแนกโครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อน

2.1.1 ผลการแปลงฟูเรียร์แบบเร็ว

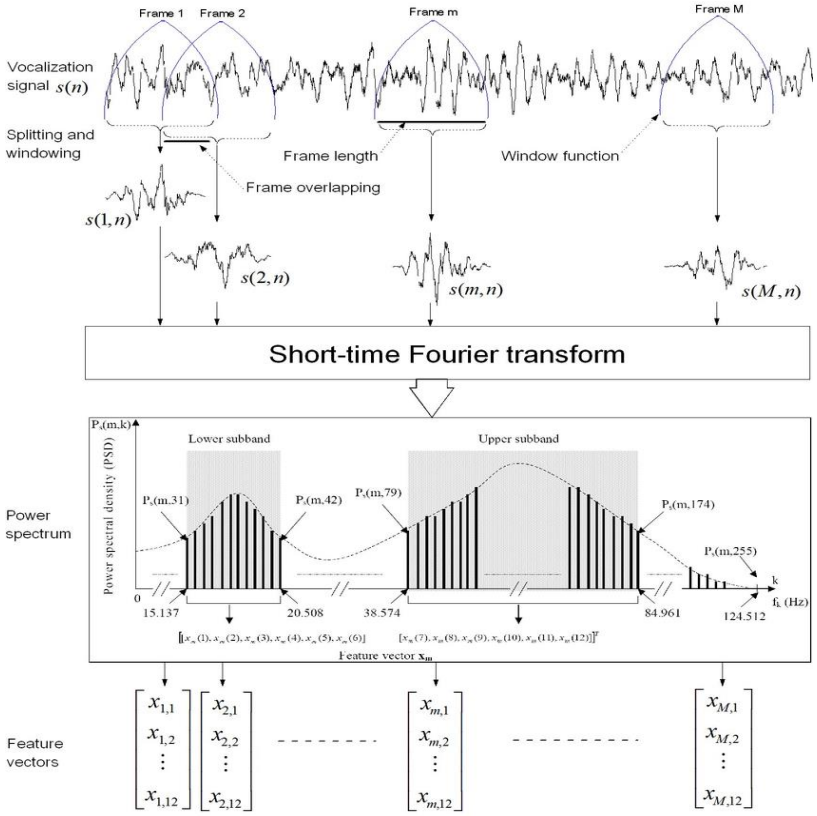
ผลการแปลงฟูเรียร์แบบเร็ว [7] เป็นเทคนิคที่พัฒนามาจากผลการแปลงฟูเรียร์แบบไม่ต่อเนื่อง (Discrete Fourier Transform, DFT) เพื่อช่วยในการประหยัดเวลาคำนวณหรือการประมวลผลทางคอมพิวเตอร์

$$y[k] = \sum_{n=0}^{N-1} x[n] e^{-2\pi jkn/N} \qquad (2\text{-}1)$$

ผลการแปลงฟูเรียร์แบบเร็วใช้สำหรับการแปลงสัญญาณจากโดเมนทางเวลาไปเป็นโดเมนทางความถี่หรือที่เรียกกันโดยทั่วไปว่าสเปกตรัม (spectrum) แสดงดังรูปที่ 2.2 และสมการที่ใช้สำหรับการแปลงฟูเรียร์แบบเร็ว ดังสมการที่ (2-1)



รูปที่ 2.2 ผลการแปลงฟูเรียร์แบบเร็ว



รูปที่ 2.3 ผลการแปลงฟูเรียร์ช่วงเวลาสั้น

2.1.2 ผลการแปลงฟูเรียร์ช่วงเวลาสั้น

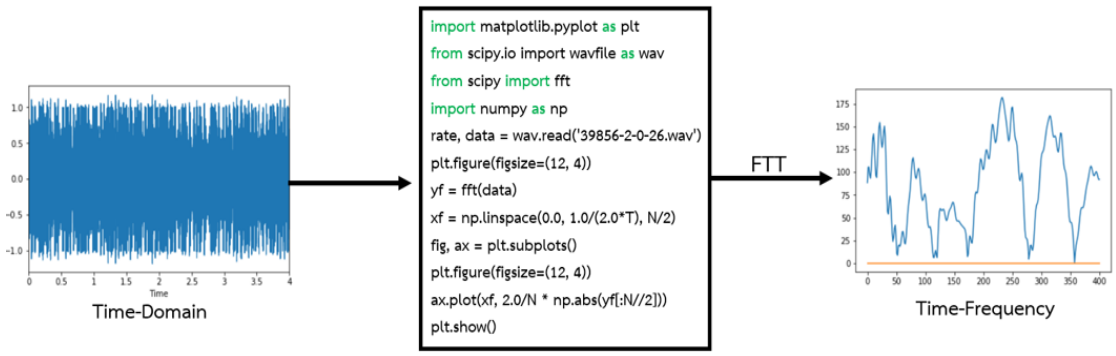
ผลการแปลงฟูเรียร์ช่วงเวลาสั้น [1] คือการสกัดคุณลักษณะประเภทหนึ่งที่นิยมใช้ในงานด้านการเรียนรู้จำเสียงสภาพแวดล้อมสามารถนำไปใช้สำหรับการแยกคุณลักษณะแต่ละสัญญาณเสียงได้ ขั้นตอนการทำงานจะสร้างฟังก์ชันหน้าต่าง (window function) ดังสมการที่ (2-2) ทำหน้าที่ตัดสัญญาณแต่ละช่วงเวลาที่ขนาดหน้าต่างเท่า ๆ กัน แล้วนำสัญญาณแต่ละช่วงเวลาไปหาค่าสเปกตรัมจากการแปลงด้วยผลการแปลงฟูเรียร์แบบเร็ว จากนั้นเราจะได้ค่าสเปกตรัมแต่ละช่วงเวลา แสดงดังรูปที่ 2.3

$$X(m,\omega) = \sum_{n}^{\infty} s[n]\, w[n-m] e^{-j\omega n} \qquad (2\text{-}2)$$

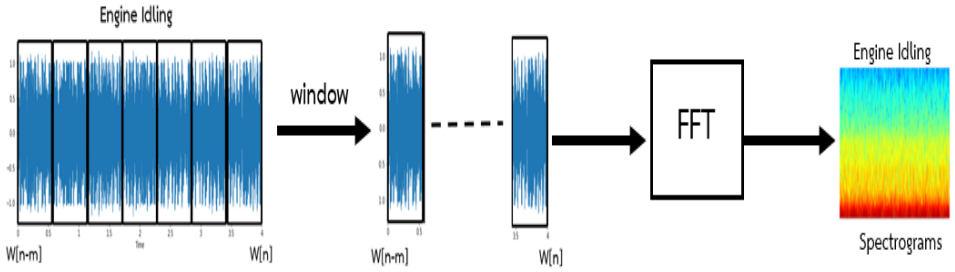
โดยที่ $X(m,\omega)$ คือ ค่าสเปกตรัมของเสียงสภาพแวดล้อม

$s[n]$ คือ สัญญาณเสียงสภาพแวดล้อม

$w[n]$ คือ ฟังก์ชันหน้าต่าง



รูปที่ 2.4 ตัวอย่างผลการแปลงฟูเรียร์แบบเร็ว

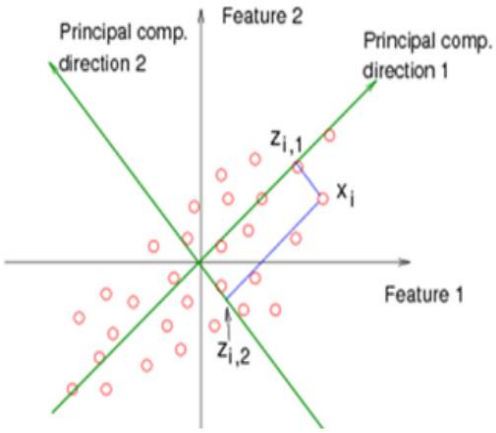


รูปที่ 2.5 ตัวอย่างผลการแปลงฟูเรียร์แบบเร็วและผล
การแปลงฟูเรียร์แบบสั้น

จากรูปที่ 2.4 เป็นตัวอย่างการแปลงสัญญาณเสียงสภาพแวดล้อมด้วยวิธีผลการแปลงฟูเรียร์แบบเร็ว จากสัญญาณโดเมนทางเวลามาเป็นสัญญาณโดเมนทางเวลา-ความถี่

จากรูปที่ 2.5 คือการสกัดคุณลักษณะเสียงสภาพแวดล้อมจากสัญญาณโดเมนทางเวลาแปลงมาเป็นสัญญาณเวลา-ความถี่ โดยขั้นตอนแรกเราจะต้องสร้างฟังก์ชันหน้าต่างเพื่อนำมาตัดเสียงแต่ละช่วงเวลาที่ขนาดเท่า ๆ กัน และนำสัญญาณแต่ละช่วงเวลามาวิเคราะห์ด้วยวิธีผลการแปลงฟูเรียร์แบบเร็วเพื่อต้องการหาค่าแต่ละสเปกตรัม ส่วนขั้นตอนสุดท้ายคือนำค่าสเปกตรัมแต่ละช่วงเวลาทั้งหมดนำมาเรียงต่อกันทำให้ได้ค่าสเปกโตรแกรม

หลังจากทำการสกัดคุณลักษณะด้วยผลการแปลงฟูเรียร์แบบเร็วและผลการแปลงฟูเรียร์ช่วงเวลาสั้น ผลลัพธ์ที่ได้คือสเปกโตรแกรม ซึ่งสามารถนำไปใช้งานด้านการเรียนรู้จำเสียงสภาพแวดล้อม ทั้งนี้งานวิจัยนี้เราจะนำสเปกโตรแกรมไปใช้กับการวิเคราะห์องค์ประกอบหลักเพื่อลดขนาดมิติก่อนเข้าสู่เครื่องมือการจำแนกเพอร์เซ็ปตรอนหลายชั้นและซัพพอร์ตเวกเตอร์แมชชีน นอกจากนี้ ผลการแปลงฟูเรียร์ช่วงเวลาสั้นยังมีคุณสมบัติที่เหมาะสมต่อการนำไปประยุกต์ใช้กับเครื่องมือการจำแนกโครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อน



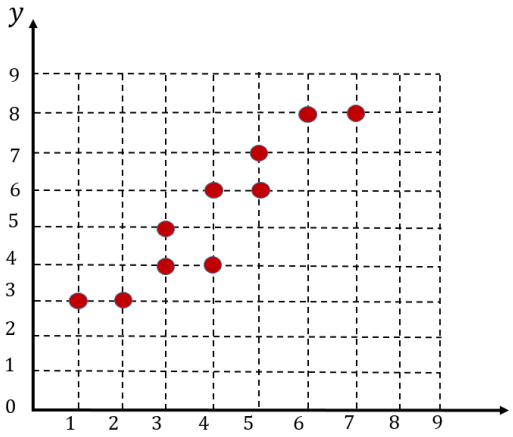
รูปที่ 2.6 การปรับแกนด้วยการวิเคราะห์องค์ประกอบหลัก

2.1.3 การวิเคราะห์องค์ประกอบหลัก

การวิเคราะห์องค์ประกอบหลัก [10] เป็นแปลงแบบพิกัดเชิงตั้งฉากของข้อมูล (orthogonal transformation) โดยค่าเส้นพิกัดใหม่ที่ได้จะถูกเรียกว่าองค์ประกอบหลัก (principal components) ประโยชน์ของการวิเคราะห์องค์ประกอบหลักคือช่วยให้เราสามารถอธิบายข้อมูลที่มีขนาดมิติใหญ่มาก (multi-dimensional data) ให้เหลือขนาดข้อมูลที่เล็กลง

การวิเคราะห์องค์ประกอบหลักจากรูปที่ 2.6 เป็นการปรับแกนเพื่อทำให้องค์ประกอบที่สัมพันธ์มาอยู่ด้วยกันทำให้เราสามารถลดมิติลงได้ โดยเราจะเลือกตัดแกนบางส่วนที่ไม่สำคัญทิ้งไป และทำการตั้งสมมุติฐานว่าแกนไหนที่มีความแปรปรวน (variant) ค่อนข้างมากและแกนไหนที่มีความแปรปรวนน้อย จากนั้นเราจะตัดแกนที่มีความแปรปรวนน้อย ๆ แล้วเลือกเฉพาะค่าที่มีความแปรปรวนสูง ๆ เก็บไว้

จากงานวิจัยข้อมูลเสียงสภาพแวดล้อมมีขนาดมิติใหญ่มาก เราจึงเสนอวิธีการวิเคราะห์องค์ประกอบหลักเป็นอีกทางเลือกหนึ่งที่นำมาใช้ในงานรู้จำเสียงสภาพแวดล้อม ซึ่งในส่วนของการวิเคราะห์องค์ประกอบหลักจะใช้กับเครื่องมือการจำแนกเพอร์เซ็ปตรอนหลายชั้นและซัพพอร์ตเวกเตอร์แมชชีน



x	y
1	3
2	3
3	4
3	5
4	4
4	6
5	6
5	7
6	8
7	8

$m_x = 4$ $m_y = 5.4$

รูปที่ 2.7 ตัวอย่างการวิเคราะห์องค์ประกอบหลัก 2 คุณลักษณะ

ตัวอย่างที่ X = [1 2 3 3 4 4 5 5 6 7]

Y = [3 3 4 5 4 6 6 7 8 8]

จากรูปที่ 2.7 เป็นตัวอย่างการลดมิติของคุณลักษณะ x และคุณลักษณะ y ด้วยการวิเคราะห์องค์ประกอบหลัก ซึ่งจากการทดลองแต่ละคุณลักษณะมีตัวอย่างทั้งหมด 10 ค่า โดยขั้นตอนการทำการวิเคราะห์องค์ประกอบหลัก อันดับแรกเราต้องนำค่าตัวอย่างทั้งหมดของคุณลักษณะ x และ y มาหาค่าเฉลี่ยดังสมการที่ (2-3)

$$m = \frac{1}{M}\sum_{i=1}^{M} x_i \qquad (2\text{-}3)$$

โดยที่ m คือค่าเฉลี่ยเวกเตอร์ของ x และ y

M คือจำนวนเวกเตอร์ตัวอย่างทั้งหมดของ x และ y

x_i คือค่าเวกเตอร์ตัวอย่างทั้งหมดของ x และ y (ขนาด 10x1)

จะได้ $m_x = \frac{1}{10}(1+2+3+3+4+4+5+5+6+7) = \frac{1}{10}(40)$ เพราะฉะนั้นค่าเฉลี่ย $m_x = 4$

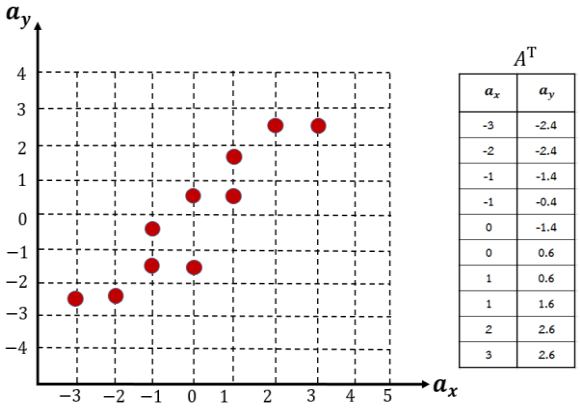
$m_y = \frac{1}{10}(3+3+4+5+4+6+6+7+8+8) = \frac{1}{10}(54)$ เพราะฉะนั้นค่าเฉลี่ย $m_y = 5.4$

หลังจากที่เราสามารถหาค่าเฉลี่ยเวกเตอร์ x และ y ขึ้นมาใหม่ด้วยสมการที่ (2-4)

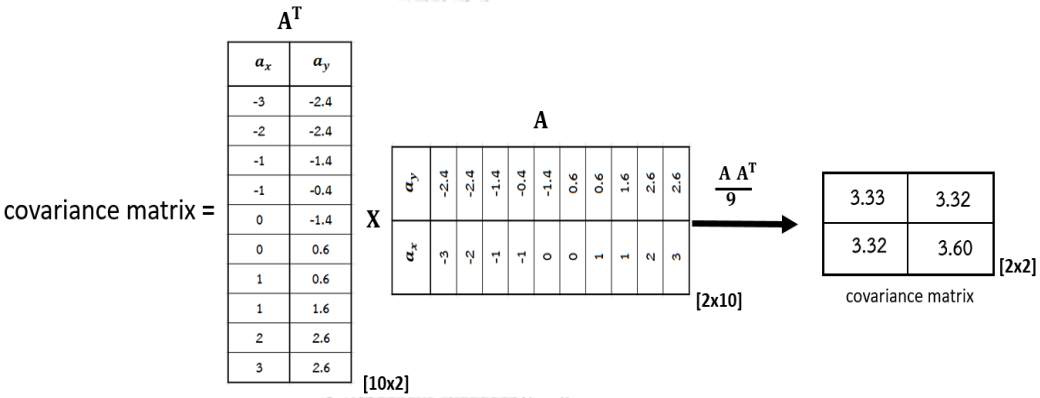
$$a_i = x_i - m \qquad (2\text{-}4)$$

จะได้ว่า $a_x = x_i - m_x$ เท่ากับ [-3 -2 -1 -1 0 0 1 1 2 3]

$a_y = x_i - m_y$ เท่ากับ [-2.4 -2.4 -1.4 -0.4 0.6 0.6 1.6 2.6 2.6]



รูปที่ 2.8 การสร้างคุณลักษณะของ x และ y



รูปที่ 2.9 การหาค่าเมทริกซ์โคแวเรียนซ์

ขั้นตอนต่อไปเป็นการหาค่าเมทริกซ์โคแวเรียนซ์ (covariance matrix) ดังสมการที่ (2-5) โดยวิธีการทำเราจะต้องนำเมทริกซ์ของคุณลักษณะ x และ y มาทำการทรานสโพสจากเมทริกซ์ [10x2] กลายเป็นเมทริกซ์ [2x10] แล้วนำระหว่างสองเมทริกซ์มาคูณกัน แสดงดังรูปที่ 2.9

$$C = \frac{1}{M-1} AA^{\mathrm{T}} = \frac{1}{M-1} \sum_{n=1}^{M} a_n a^{\mathrm{T}}{}_n \qquad (2\text{-}5)$$

โดยที่ C คือเมทริกซ์โคแวเรียนซ์ของ x และ y (ขนาด 2x2)

A คือเมทริกซ์ของ x และ y (ขนาด 2x10)

A^T คือทรานโพสเมทริกซ์ A (ขนาด 10x2)

ต่อมาเราจะนำค่าเมทริกซ์โคแวเรียนซ์ไปหาค่าไอเกนแวลลิว (Eigen- value) $\boldsymbol{\lambda}$ และค่าไอเกนเวกเตอร์ (Eigen vector) $\mathbf{v}$ สมการที่ (2-6)

$$Cv = \lambda v \qquad (2\text{-}6)$$

$$(C-\lambda I)V=0 \longrightarrow \begin{bmatrix} 3.33-\lambda & 3.22 \\ 3.22 & 3.60-\lambda \end{bmatrix} = 0 \quad [2x2]$$

รูปที่ 2.10 การหาค่าไอเกนเวกเตอร์และไอเกนแวลลิว

	V_1	V_2
	−0.722	−0.692
	0.692	−0.722

[2x2]

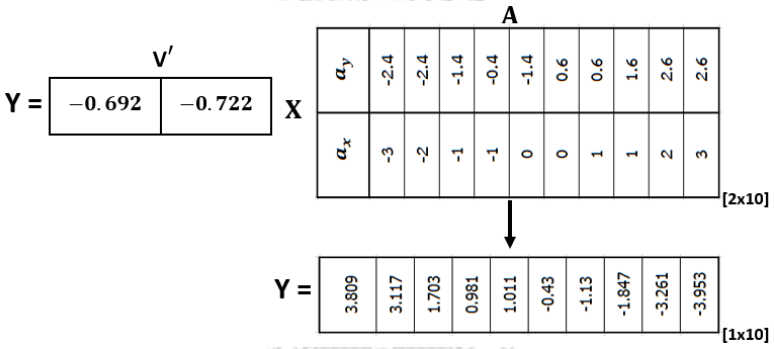
(ก)

λ_2	6.692
λ_1	0.242

[1x1]

(ข)

รูปที่ 2.11 ผลลัพธ์ของ (ก) ไอเกนเวกเตอร์และ (ข) ไอเกนแวลลิว



รูปที่ 2.12 การลดมิติด้วยการวิเคราะห์องค์ประกอบหลัก

เราสามารถหาค่าไอเกนเวกเตอร์กับไอเกนแวลลิว แสดงดังรูปที่ 2.10 โดยการหาดีเทอร์มิแนนต์ (determinant) เพื่อให้ได้ค่าไอเกนแวลลิวไปแทนในสมการ $(C-\lambda I)V=0$ เพื่อหาค่าไอเกนเวกเตอร์ แสดงดังรูปที่ 2.11 คือค่าไอเกนเวกเตอร์ $\mathbf{V_1}$ เท่ากับ [-0.722 0.692] จะมีความสัมพันธ์กับไอเกนแวลลิว $\boldsymbol{\lambda}_{\mathbf{1}}$ เท่ากับ 0.242 ส่วนค่าเวกเตอร์ของ $\mathbf{V_2}$ เท่ากับ [-0.692 -0.722] จะมีความสัมพันธ์กับ $\boldsymbol{\lambda}_{\mathbf{2}}$ เท่ากับ 6.692

หลังจากหาค่าไอเกนเวกเตอร์กับไอเกนแวลลิวได้แล้วนั้น เราจะนำค่าไอเกนเวกเตอร์มาทรานสโพสเป็น V' แล้วคูณกับเมทริกซ์ A แสดงดังสมการที่ (2-7) จะได้ขนาดเมทริกซ์ [1x10] แสดงดังรูปที่ 2.12

$$Y = V'A \tag{2-7}$$

โดยที่ Y คือจำนวนองค์ประกอบหลัก (ขนาด 1x10)

V′ คือค่าไอเกนเวกเตอร์ที่นำมาทรานสโพส

A คือเวกเตอร์ค่าเฉลี่ยของข้อมูล x และ y

A^{T}

a_x	a_y
-3	-2.4
-2	-2.4
-1	-1.4
-1	-0.4
0	-1.4
0	0.6
1	0.6
1	1.6
2	2.6
3	2.6

X

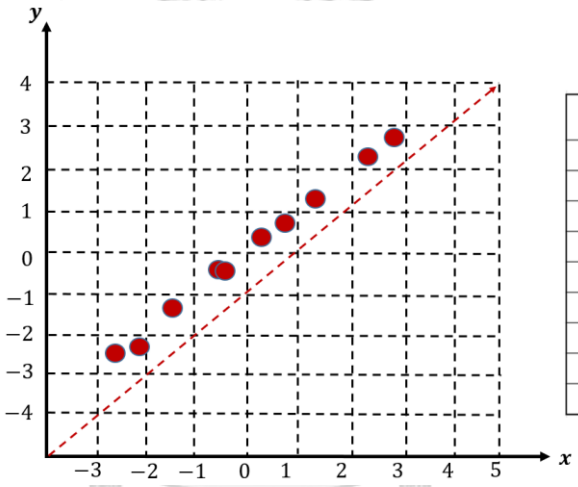
Y^{T}

3.809
3.117
1.703
0.981
1.011
-0.43
-1.13
-1.847
-3.261
-3.953

convert →

-2.635	-2.75
-2.156	-2.25
-1.178	-1.229
-0.678	-0.708
-0.699	-0.729
0.299	0.312
0.778	0.812
1.278	1.333
2.256	2.354
2.735	2.854

รูปที่ 2.13 การแปลงข้อมูลย้อนกลับ



y	x
-2.635	-2.75
-2.156	-2.25
-1.178	-1.229
-0.678	-0.708
-0.699	-0.729
0.299	0.312
0.778	0.812
1.278	1.333
2.256	2.354
2.735	2.854

PCA

รูปที่ 2.14 การวิเคราะห์องค์ประกอบหลัก

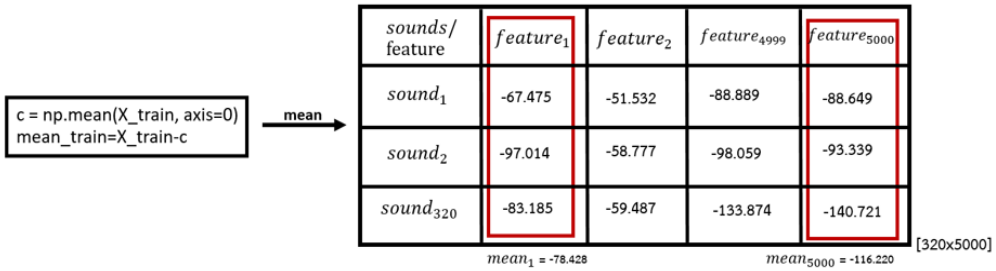
จากตัวอย่างเราสามารถแปลงข้อมูลย้อนกลับ โดยการนำเมทริกซ์เวกเตอร์ของ A^{T} คูณกับเมทริกซ์เวกเตอร์ Y^{T} แสดงดังรูปที่ 2.13 แล้วนำค่าเวกเตอร์ของคุณลักษณะ x และ y นำมาพล็อตจะเห็นได้ว่าข้อมูลที่ผ่านการทำการวิเคราะห์องค์ประกอบหลัก ทำให้ข้อมูลที่ได้มีการกระจัดกระจายน้อยกว่าข้อมูลเก่า แสดงดังรูปที่ 2.14

สรุปการทำการวิเคราะห์องค์ประกอบหลัก ขั้นตอนแรกนำแต่ละคุณลักษณะมาหาค่าเฉลี่ย แล้วนำค่าเฉลี่ยที่หามาได้นำไปลบกับค่าสมาชิกของคุณลักษณะเพื่อทำให้ข้อมูลอยู่ตรงกลาง ทำให้เราสามารถหาค่า covariance matrix ได้ โดยการนำเมทริกซ์ของคุณลักษณะมาทรานสโพส แล้วนำมาคูณกันระหว่างสองเมทริกซ์ทำให้ได้ค่า covariance matrix เพื่อนำไปหาค่าของไอเกนเวกเตอร์กับไอเกนแวลลิวใช้ในการสร้างข้อมูลขึ้นมาใหญ่ที่มีขนาดมิติเล็กลงจากเดิม

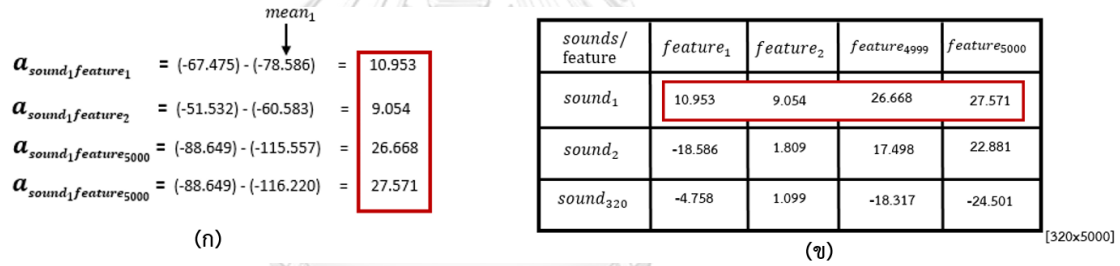
$sounds/$ feature	$feature_1$	$feature_2$	$feature_{4999}$	$feature_{5000}$
$sound_1$	-67.475	-51.532	-88.889	-88.649
$sound_2$	-97.014	-58.777	-98.059	-93.339
$sound_{320}$	-83.185	-59.487	-133.874	-140.721

[320x5000]

รูปที่ 2.15 ตัวอย่างเมทริกซ์เวกเตอร์ของเสียงสภาพแวดล้อม



รูปที่ 2.16 การหาค่าเฉลี่ยแต่ละเสียงสภาพแวดล้อม



รูปที่ 2.17 (ก) ตัวอย่างการลบระหว่างค่าเฉลี่ยและค่าสมาชิกของเสียงสภาพแวดล้อม
(ข) เมทริกซ์เสียงสภาพแวดล้อม

จากงานวิจัยเราได้เสนอวิธีการการวิเคราะห์องค์ประกอบหลักในการลดขนาดมิติเสียงสภาพแวดล้อม จากรูปที่ 2.15 เป็นตัวอย่างเสียงสภาพแวดล้อมทั้งหมด 320 เสียง และมีขนาดมิติเท่ากับ 5000 มิติ ขั้นตอนการลดมิติเสียงสภาพแวดล้อมมี 3 ขั้นตอนดังนี้

1. นำค่าสมาชิกของเสียงสภาพแวดล้อมทั้งหมดมาหาค่าเฉลี่ย แล้วนำมาลบกับค่าสมาชิกของเสียงสภาพแวดล้อมตัวเดิม เพื่อทำให้ข้อมูลย้ายมาอยู่ตรงกลาง
2. หา covariance matrix ของเสียงสภาพแวดล้อม
3. หาค่าไอเกนแวลลิวกับไอเกนเวกเตอร์ เพื่อนำไปลดมิติเสียงสภาพแวดล้อม

การวิเคราะห์องค์ประกอบหลัก อันดับแรกเราจะต้องหาค่าเฉลี่ยทั้งหมดของเสียงสภาพแวดล้อมก่อน แสดงดังรูปที่ 2.16 แล้วนำค่าเฉลี่ยที่หามาได้นำมาลบกับค่าสมาชิกของเสียงสภาพแวดล้อม แสดงดังรูปที่ 2.17

$sounds/$ feature	$sound_1$	$sound_2$	$sound_{320}$
$feature_1$	10.953	-18.586	-4.758
$feature_2$	9.054	1.809	1.099
$feature_{4999}$	26.668	17.498	-18.317
$feature_{5000}$	27.571	22.881	-24.501

[5000x320]

T

X

$sounds/$ feature	$feature_1$	$feature_2$	$feature_{4999}$	$feature_{5000}$
$sound_1$	10.953	9.054	26.668	27.571
$sound_2$	-18.586	1.809	17.498	22.881
$sound_{320}$	-4.758	1.099	-18.317	-24.501

[320x5000]

(ก)

covariance matrix	C_1	C_2	C_{5000}
C_1	14.716	7.915	3.242
C_2	7.915	11.822	4.889
C_{5000}	3.242	4.889	19.117

[5000x5000]

(ข)

รูปที่ 2.18 การหาค่าเมทริกซ์โคแวเรียนซ์ (ก) การคูณระหว่างเมทริกซ์เสียงสภาพแวดล้อมเสียงกับตัวทรานโพส (ข) ค่าเมทริกซ์โคแวเรียนซ์ [5000x5000]

```
eigvals_train, eigvecs_train = np.linalg.eig(cov_train.T)
print("eigenvalue = ", eigvals_trian.shape)
print("eigenvector =", eigvecs_train.shape)
```

	λ_1	λ_2	λ_3	λ_{5000}
Eigenvalue	5.206	1.693	0.639	0.001

[1x5000]

Eigenvector	V_1	V_2	V_3	V_{5000}
V_1	8.300	8.912	9.432	18.129
V_2	11.312	6.170	11.281	-18.213
V_{5000}	-2.632	-0.159	-0.532	-10.236

[5000x5000]

รูปที่ 2.19 การหาค่าไอเกนแวลลิวกับไอเกนเวกเตอร์

หลังจากที่เราสามารถย้ายข้อมูลเสียงสภาพแวดล้อมมาอยู่ตรงกลาง ทำให้เราสามารถหาค่า covariance matrix ได้ โดยการนำเมทริกซ์ของเสียงสภาพแวดล้อมมาทำทรานโพสเมทริกซ์แล้วนำเมทริกซ์ทั้งสองมาคูณกัน แสดงดังรูปที่ 2.18 (ก) ทำให้เราได้ค่า covariance matrix ที่มีขนาดเมทริกซ์ [5000x5000] แสดงดังรูปที่ 2.18 (ข) ค่าของ covariance matrix ที่ได้เราสามารถนำไปหาค่าไอเกนแวลลิวกับไอเกนเวกเตอร์ แสดงดังรูปที่ 2.19

A_n

$sounds/$ feature	$feature_1$	$feature_2$	$feature_{4999}$	$feature_{5000}$
$sound_1$	10.953	9.054	26.668	27.571
$sound_2$	-18.586	1.809	17.498	22.881
$sound_{320}$	-4.758	1.099	-18.317	-24.501

[320x5000]

Y = X

X_n

$Eigenvector$	V_1	V_2	V_3
V_1	8.300	8.912	9.432
V_2	11.312	6.170	11.281
V_{5000}	-2.632	-0.159	-0.532

[5000x3]

(ก)

Y=

$sounds/$ feature	$feature_1$	$feature_2$	$feature_3$
$sound_1$	-6.523	1.325	0.852
$sound_2$	-0.852	2.313	-4.895
$sound_{320}$	3.561	-0.632	5.532

[320x3]

(ข)

รูปที่ 2.20 (ก) ตัวอย่างการลดมิติเสียง (ข) ข้อมูลเสียงสภาพแวดล้อมถูกลดมิติ

$$Y = X_n A_n \tag{2-8}$$

โดยที่ Y คือจำนวนองค์ประกอบหลักของเสียงสภาพแวดล้อม (ขนาด 320*x3*)

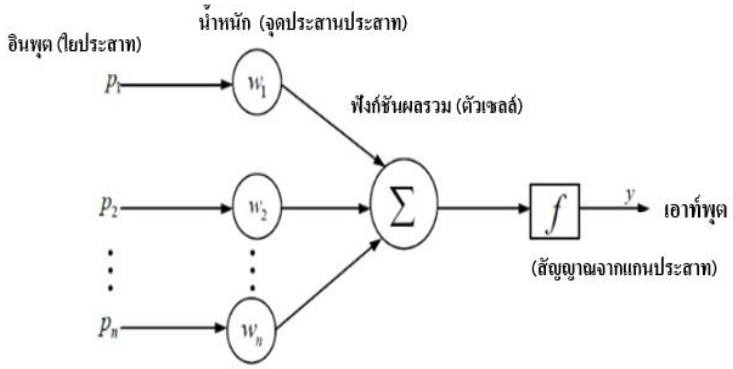
A_n คือเมทริกซ์เฉลี่ยของเสียงสภาพแวดล้อมทั้งหมด (ขนาด 320*x*5000)

X_n คือค่าไอเกนเวกเตอร์ของเสียงสภาพแวดล้อม (ขนาด 5000x3)

หลังจากที่หาค่าเมทริกซ์ของไอเกนเวกเตอร์กับไอเกนแวลลิว ทำให้เราสามารถลดมิติของเสียงสภาพแวดล้อมได้ โดยการลดมิติเสียงเราจะนำเมทริกซ์ของค่าไอเกนเวกเตอร์ขนาด [5000x5000] ลดมิติเหลือ [5000x3] ซึ่งเลือกมาจากค่าไอเกนแวลลิวที่มีลำดับสูงสุด ได้แก่ [5.206 1.693 และ 0.639] แล้วนำไปทำผลคูณภายใน (inner product) กับเสียงสภาพแวดล้อม A_n ที่มีขนาดเท่ากับ [320x5000] แสดงดังรูปที่ 2.20 (ก) จะทำให้ได้ข้อมูลเสียงสภาพแวดล้อมที่มีขนาดเมทริกซ์เท่ากับ [320x3] ดังตารางที่ 2.20 (ข) เราสามารถนำข้อมูลที่ลดมิติไปใช้กับเครื่องมือการจำแนกซัพพอร์ตเวกเตอร์แมชชีนและเพอร์เซ็ปตรอนหลายชั้น

2.2 เครื่องมือการจำแนก

หัวข้อนี้จะกล่าวถึงเครื่องมือการจำแนกเสียงสภาพแวดล้อมที่ผ่านกระบวนการสกัดคุณลักษณะของผลการแปลงฟูเรียร์ช่วงเวลาสั้นและการลดมิติของเสียงสภาพแวดล้อม ซึ่งจากงานวิจัยเราได้มีการทดสอบเปรียบเทียบประสิทธิภาพของเครื่องมือการจำแนกเสียงสภาพแวดล้อม 4 ประเภทได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน เพอร์เซ็ปตรอนหลายชั้น โครงข่ายคอนโวลูชัน และโครงข่ายประสาทเกิดซ้อน



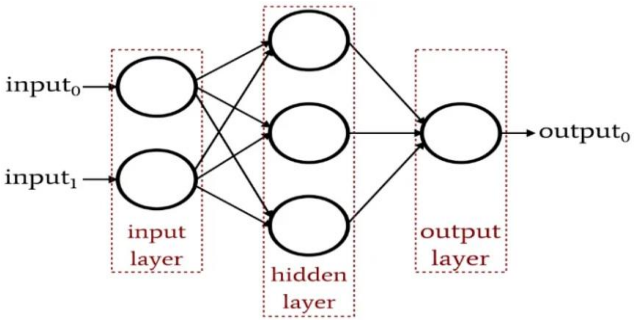
รูปที่ 2.21 โครงสร้างการทำงานของโครงข่ายประสาทเทียม

2.2.1 โครงข่ายประสาทเทียม (Artificial Neuron Network, ANN)

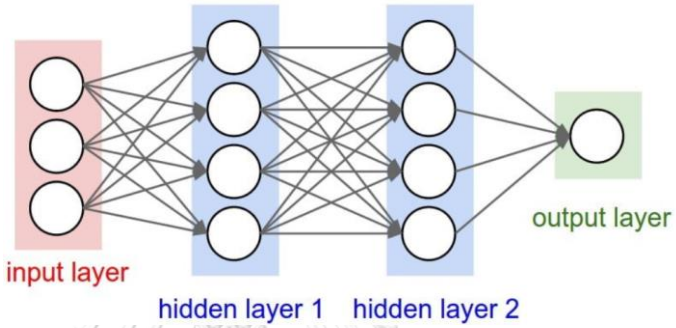
โครงข่ายประสาทเทียมทำหน้าที่คล้ายกับทางด้านปัญญาประดิษฐ์ (Artifitial Intelligence, AI) มีลักษณะของการทำงานเหมือนกับสมองมนุษย์ สามารถปรับเปลี่ยนตัวเองเมื่อมีการตอบสนองสิ่งที่รับเข้ามา เรียกว่า กฎของการเรียนรู้ (learning rule) โดยสมองมนุษย์ประกอบไปด้วยหน่วยประมวลผลที่เรียกว่า นิวรอน (เซลล์ประสาท หรือ neuron) จำนวนนิวรอนในสมองมนุษย์มีการเชื่อมต่อกันหลายจุด ซึ่งสมองมนุษย์เปรียบเสมือนคอมพิวเตอร์ที่มีการปรับตัวเอง (adaptive) ไม่เป็นเชิงเส้น (nonlinear) ทำงานแบบขนาน (parallel) ด้วยการดูแลจัดการการทำงานร่วมกันของนิวรอนในสมอง

โครงข่ายประสาทเทียมได้ถูกพัฒนาโดยหลักการทำงานของสมองมนุษย์ ซึ่งภายในสมองมนุษย์ประกอบด้วยนิวรอนจำนวนมากและมีจุดเชื่อมต่อเหมือนเซลล์ประสาทของมนุษย์ โดยโครงข่ายประสาทเทียมประกอบด้วย 3 ส่วนสำคัญ ดังรูปที่ 2.26 ได้แก่ ใยประสาท (dendrite) ตัวเซลล์ (soma) และแกนประสาท (axon) ในแต่ละโครงข่ายประสาทเทียมจะเชื่อมต่อกันโดยจุดประสานประสาท (synapse) ทำให้สามารถเปลี่ยนค่าความต้านทานได้ตามสัญญาณที่ส่งระหว่างกันของเซลล์ประสาท

การประมวลผลต่าง ๆ เกิดขึ้นในหน่วยประมวลผลย่อยที่เรียกว่าโหนด (node) คือการจำลองลักษณะการทำงานมาจากเซลล์การส่งสัญญาณระหว่างโหนดที่เชื่อมต่อกัน โดยการจำลองมาจากการเชื่อมต่อของใยประสาท ซึ่งภายในโหนดจะมีฟังก์ชันกำหนดสัญญาณส่งออกที่เรียกว่าฟังก์ชันการแปลง (transfer function) โครงข่ายประสาทเทียมประกอบด้วย 5 องค์ประกอบ 1. ข้อมูลอินพุต (input) 2. ข้อมูลเอาต์พุต (output) 3. ค่าน้ำหนัก (weights) 4. ฟังก์ชันผลรวม (Summation function: S) และ 5. ฟังก์ชันการแปลง (transfer function)



รูปที่ 2.22 โครงข่ายประสาทแบบชั้นเดียว



รูปที่ 2.23 โครงข่ายประสาทหลายชั้น

2.2.2 เพอร์เซ็ปตรอนหลายชั้น

เพอร์เซ็ปตรอนหลายชั้น (MLP) เป็นศาสตร์แขนงหนึ่งทางด้านปัญญาประดิษฐ์ที่สามารถนำไปประยุกต์ใช้กับงานหลายด้านได้อย่างมีประสิทธิภาพ เช่น การจำแนกรูปแบบ การทำนาย การควบคุม การหาความเหมาะสม และการจัดกลุ่ม เป็นต้น

หลักการสำคัญของโครงข่ายประสาทคือความพยายามลอกเลียนแบบการทำงานของเซลล์ประสาทในสมองมนุษย์ โดยลักษณะทั่วไปของโครงข่ายประสาทคือการที่โหนดต่าง ๆ จำลองมาจากไซแนปของเซลล์ประสาทระหว่างเดนไดร์ทและแอกซอน โดยมีฟังก์ชันเป็นตัวกำหนดสัญญาณส่งออกนั่นเอง ลักษณะของโครงข่ายประสาทสามารถแบ่งได้ 2 แบบ โครงข่ายประสาทแบบชั้นเดียว (single layer) จะมีเพียงชั้นสัญญาณประสาทแค่เพียงขาเข้ากับชั้นสัญญาณประสาทขาออกเท่านั้น ส่วนของโครงข่ายประสาทแบบหลายชั้นจะมีลักษณะเช่นเดียวกับโครงข่ายประสาทแบบชั้นเดียวแต่จะมีชั้นแอบแฝง hidden เพิ่มขึ้นอยู่ระหว่างชั้นนำข้อมูลป้อนเข้าและชั้นส่งข้อมูลออก

2.2.2.1 เพอร์เซ็ปตรอนแบบชั้นเดียว

เพอร์เซ็ปตรอนแบบชั้นเดียวเป็นโครงข่ายประสาทที่มีเพียงชั้นรับข้อมูลกับส่งข้อมูล โดยชั้นรับข้อมูลทำหน้าที่รับข้อมูลเข้า (input value) แล้วส่งข้อมูลไปยังเส้นเชื่อมโยงให้กับโหนดอื่น ๆ ส่วนชั้นส่งข้อมูลออกปริมาณข้อมูลจะขึ้นอยู่กับค่าน้ำหนักที่อยู่บนเส้นเชื่อมโยงในโหนดชั้นส่งข้อมูลออก จากนั้นนำข้อมูลที่ได้รับมาคำนวณโดยใช้ฟังก์ชันทางคณิตศาสตร์ที่เรียกว่าฟังก์ชันการแปลง (transfer function) ที่เหมาะสมกับปัญหาแล้วส่งผลลัพธ์ที่ได้เป็นข้อมูลส่งออก เช่น โครงข่ายแบบชั้นเดียวแบบเพอร์เซ็ปตรอนอย่างง่าย และโครงข่ายโฮปฟิลด์ (hopfield networks) ลักษณะโครงข่ายแบบชั้นเดียวแสดงดังรูปที่ 2.27

2.2.2.2 เพอร์เซ็ปตรอนแบบหลายชั้น

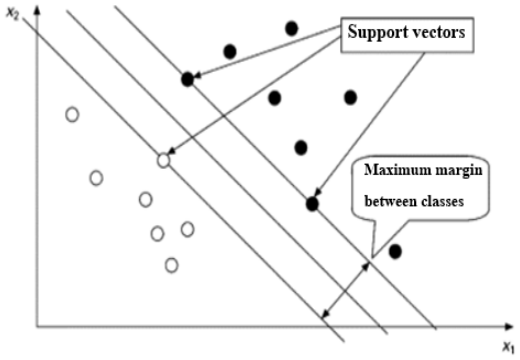
โครงข่ายแบบหลายชั้นเป็นโครงข่ายที่มีชั้นแอบแฝงตั้งแต่ 1 ชั้นขึ้นไปโครงข่ายแบบหลายชั้นจะใช้สำหรับกรณีที่มีความซับซ้อน ซึ่งโครงข่ายแบบชั้นเดียวไม่สามารถแก้ปัญหาได้ จึงเพิ่มจำนวนโหนดที่มีการคำนวณหรือชั้นแอบแฝงให้กับโครงข่าย จากตัวอย่างของโครงข่ายแบบหลายชั้น เช่น การแพร่ย้อนกลับ (back propagation) โครงข่ายโคโฮเนน (Kohonen networks) และการเผยแผ่แบบเคาน์เตอร์ (counter propagation) เป็นต้น ลักษณะโครงสร้างโครงข่ายแบบหลายชั้นแสดงดังรูปที่ 2.28

2.2.2.3 การเรียนรู้แบบมีผู้สอน (supervised learning)

การเรียนรู้แบบมีผู้สอนคือการทำให้ข้อมูลที่ได้เป็นรูปแบบเดียวกัน โดยวิธีการเริ่มจากนำข้อมูลป้อนเข้าของโครงข่ายจะกำหนดค่าผลลัพธ์ที่ให้กับข้อมูลป้อนเข้าแต่ละโครงข่ายจะนำค่าผิดพลาดระหว่างค่าเป้าหมายกับค่าผลลัพธ์ที่ได้ มาใช้สำหรับการปรับค่าน้ำหนัก เพื่อให้ค่าผลลัพธ์ที่ใกล้เคียงกับเป้าหมายมากที่สุด ถ้าหากเปรียบเทียบจะเหมือนกับการสอนนักเรียน โดยมีครูผู้สอนคอยให้คำชี้แนะ ตัวอย่างแบบจำลองนี้ได้แก่ การแพร่ย้อนกลับและเพอร์เซ็ปตรอน เป็นต้น

2.2.2.4 การเรียนรู้แบบไม่มีผู้สอน (unsupervised learning)

การเรียนรู้แบบไม่มีผู้สอนเป็นโครงข่ายที่มีข้อมูลป้อนเข้าอย่างต่อเนื่อง โดยไม่มีการส่งค่าผลลัพธ์เป้าหมายให้กับข้อมูลป้อนเข้าแต่ละตัว โดยการปรับน้ำหนักจะใช้ข้อมูลที่นำมาสอนเป็นตัวปรับค่า โดยค่าน้ำหนักจะปรับตามกลุ่มที่ข้อมูลป้อนเข้าที่มีรูปแบบคล้ายคลึงกัน



รูปที่ 2.24 การแบ่งกลุ่มของซัพพอร์ตเวกเตอร์แมชชีน

2.2.3 ซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีน [3] เป็นเครื่องมือการจำแนกแบ่งกลุ่มโดยใช้เทคนิคการสร้างไฮเปอร์เพลน (hyperplane) นอกจากนี้อัลกอริทึมของซัพพอร์ตเวกเตอร์แมชชีนสามารถแก้ปัญหาความซับซ้อนจากวิธีเชิงเส้น (linear) โดยการประยุกต์ใช้วิธีของเคอร์เนล (kernal methods)

แนวคิดของซัพพอร์ตเวกเตอร์แมชชีนสำหรับการแบ่งกลุ่มข้อมูล จากรูปที่ 2.21 เป็นตัวอย่างของการแบ่งข้อมูล 2 กลุ่ม สำหรับวิธีการแบ่งกลุ่มจะต้องสร้างไฮเปอร์เพลนเส้นตรงแบ่งระหว่างสองกลุ่มและต้องทำให้ระยะห่างทั้งสองข้อมูลมีระยะห่างมากที่สุด ทำให้สามารถแบ่งข้อมูลสำหรับการเรียนรู้ใหม่ได้ดีขึ้น นอกจากนี้ ข้อมูลที่อยู่ใกล้กับเส้นแบ่งของซัพพอร์ตเวกเตอร์แมชชีนมากที่สุดจะถูกเลือกนำมาใช้งานแล้วตัดส่วนที่เหลือทิ้ง

การแบ่งกลุ่มของซัพพอร์ตเวกเตอร์แมชชีนนั้นสามารถอธิบายโดยสมการดังต่อไปนี้ กำหนดให้ D คือชุดข้อมูล, $\mathsf{D} = \{(x_1, y_1),\ldots,(x_n, y_n)\}$, $\mathrm{x} \in \mathrm{R}^{\mathrm{n}}, \mathrm{y} \in \{-1, 1\}$

โดยที่ x = ข้อมูลอินพุตของเสียงสภาพแวดล้อม

y = ค่าข้อมูลสำหรับการแบ่งเป็น 2 กลุ่ม

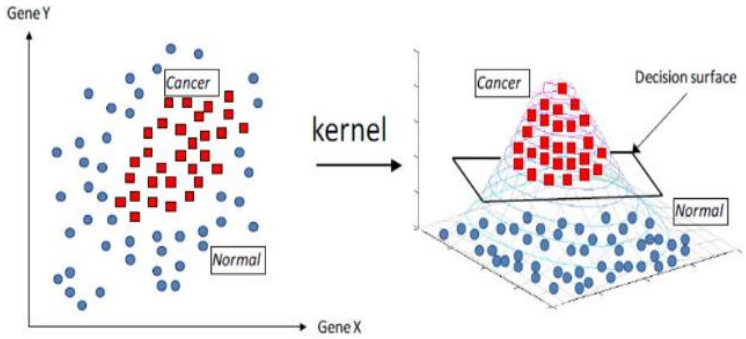
จะหาไฮเปอร์เพลนที่สามารถแบ่งกลุ่มออกเป็น 2 กลุ่มโดยที่มีระยะห่างมากที่สุด โดยการใช้เส้นขอบเขต (boundary lines) ดังสมการที่ (2-7)

$$(w, x) + b = 0 \qquad (2\text{-}7)$$

โดยที่ w คือ ขอบเขต

x คือ ค่าอินพุตเวกเตอร์

b คือ ค่าเริ่มต้น (Scalar Threshold)



รูปที่ 2.25 การแบ่งข้อมูลด้วยวิธีซัพพอร์ตไม่เป็นเชิงเส้น

จะได้คู่ของค่า (w, x) ที่ทำให้ไฮเปอร์เพลนไม่ซ้ำกัน โดยสมการรูปแบบปกติจะกำหนดได้จากค่า (w, x) ดังนี้

$$(w, x) + b = 1, \quad (w, x) + b = -1 \tag{2-8}$$

จะได้ฟังก์ชั่นการตัดสินใจคือ

$$y = \text{sign}((w, x) + b) \tag{2-9}$$

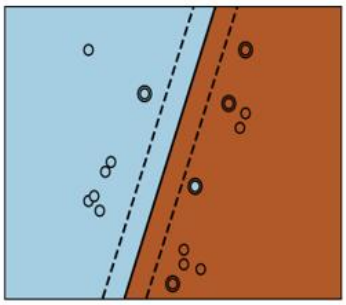
ดังนั้น จะได้ไฮเปอร์เพลนที่แยกจากกันในรูปแบบปกติดังนี้

$$[(w, x) + b] \geq 1, i = 1, \ldots., l \tag{2-10}$$

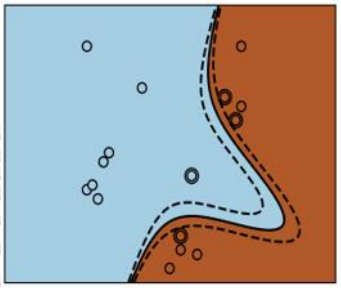
โดยที่ l คือ จำนวนของชุดข้อมูลสำหรับเรียนรู้

เราสามารถสร้างไฮเปอร์เพลนหลาย ๆ อันในการแบ่งชุดข้อมูลได้ แต่จะมีเพียง 1 ไฮเปอร์เพลนที่สามารถแบ่งข้อมูลที่มีค่าความผิดพลาดน้อยที่สุด โดยหลักการสร้างไฮเปอร์เพลนจะต้องหาระยะห่าง (margin) ของข้อมูลที่มีความกว้างมากที่สุด นอกจากนี้ การสร้างไฮเปอร์เพลนเหมาะสำหรับการแบ่งกลุ่มเพียง 2 คลาส ด้วยเหตุนี้เราจึงแก้ปัญหาด้วยวิธีของซัพพอร์ตเวกเตอร์แมชชีนแบบไม่เป็นเชิงเส้น (non–linear)

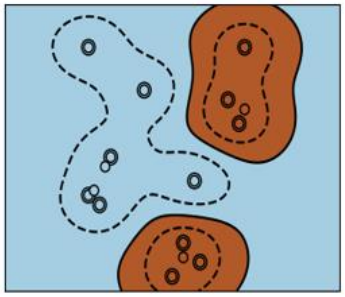
เนื่องจากวิธีของซัพพอร์ตเวกเตอร์แมชชีนแบบเชิงเส้นไม่สามารถแบ่งหลาย ๆ ข้อมูลได้ เราจึงได้เสนอวิธีการของซัพพอร์ตเวกเตอร์แมชชีนแบบไม่เป็นเชิงเส้น (non–linear) แสดงดังรูปที่ 2.22 เป็นการสร้างไฮเปอร์เพลนในรูปแบบ 3 มิติ ทำให้เราสามารถสร้างไฮเปอร์เพลนในการแบ่งหลาย ๆ ข้อมูลได้



รูปที่ 2.26 เคอร์เนลแบบเชิงเส้น



รูปที่ 2.27 เคอร์เนลแบบโพลีโนเมียล



รูปที่ 2.28 เคอร์เนลแบบเรเดียลเบสิสฟังก์ชั่น

2.2.3.1 เคอร์เนลฟังก์ชั่น (Kernal Function)

เราประยุกต์วิธีเคอร์เนลใช้กับเครื่องมือการจำแนกซัพอร์ตเวกเตอร์แมชชีนในการแบ่งข้อมูลมากกว่า 2 ขึ้นไป โดยวิธีการทำเราจะสร้างไฮเปอร์เพลนสำหรับแบ่ง 2 กลุ่ม แต่เราจะสร้างหลาย ๆ ไฮเปอร์เพลน จากนั้นนำไฮเปอร์เพลนทั้งหมดมารวมกัน (union)

การจำแนกประเภทข้อมูลด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีนจะต้องมีการเลือกใช้เคอร์เนลฟังก์ชั่น (Kernal Fuction) [11] ที่เหมาะสมกับงาน ได้แก่ เคอร์เนลแบบเชิงเส้น (Linear) เคอร์เนลแบบโพลีโนเมียล (Polynomial) และเรเดียลเบสิสฟังก์ชั่น (Radial Basis Function) แสดงดังรูปที่ 2.26 2.27 และ 2.28 เป็นต้น

ตัวอย่างพื้นฐานของเคอร์เนล สมมติว่าข้อมูลที่จะวิเคราะห์เป็นเวกเตอร์ของจริงนั่นคือ $x = \mathrm{R}^{\mathrm{p}}$ วัตถุใด ๆ จะถูกเขียนเป็น $x = (x_1, \ldots, x_p)^{\mathrm{T}}$ หนึ่งในการเปรียบเทียบเวกเตอร์ดังกล่าวโดยใช้สำหรับ inner product คือ $\mathrm{x}, \mathrm{x}' \in \mathrm{R}^{\mathrm{p}}$

$$k_L(\mathbf{x}, \mathbf{x}') := \mathbf{x}^\top \mathbf{x}' = \sum_{i=1}^{p} x_i x_i'. \tag{2-11}$$

ฟังก์ชั่นนี้เป็นเคอร์เนล $x^T\, x' = x'^T\, x$ เพราะวิธีนี้ผลลัพธ์ที่ได้เป็นผลบวกจากการคำนวณอย่างง่าย ๆ ดังต่อไปนี้ใช้ได้กับทุก ๆ ที่ n > 0 และ $x_1, \ldots, x_n \in \mathrm{R}^{\mathrm{p}}$ ฟังก์ชั่นของ $c_1, \ldots, c_n \in \mathrm{R}^{\mathrm{p}}$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k_L(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \mathbf{x}_i^\top \mathbf{x}_j = ||\sum_{i=1}^{n} c_i \mathbf{x}_i||^2 \geq 0 \tag{2 -12}$$

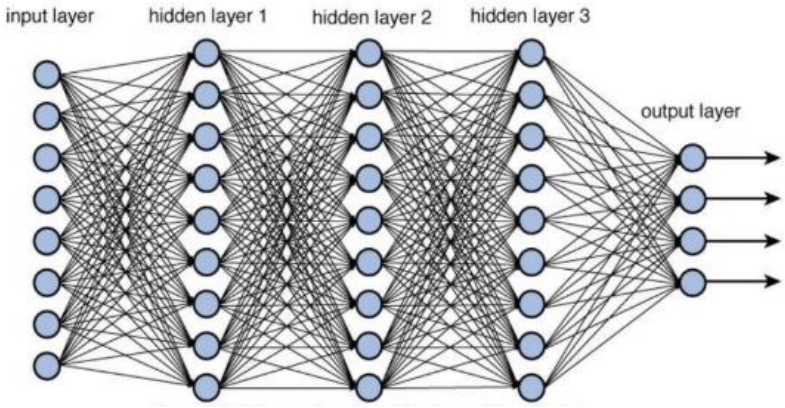
ผลิตภัณฑ์ชั้นในระหว่างเวกเตอร์เป็นเคอร์เนลตัวแรกที่เรามักพบจะเรียกว่าเคอร์เนลเชิงเส้น โดยข้อจำกัดที่ชัดเจนของเคอร์เนลจะถูกกำหนดไว้เฉพาะเมื่อข้อมูลเป็นพาหะ สำหรับวัตถุทั่วไปเพิ่มเติมของ $\mathrm{x} \in \mathrm{X}$ เช่นเวกเตอร์ $\phi(\mathrm{x}) \in \mathrm{R}^{\mathrm{p}}$ จากนั้นกำหนดค่าเคอร์เนลสำหรับใด ๆ $x, x' \in \mathrm{X}$ โดย

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}'). \tag{2-13}$$

หลังจากการคำนวณสมการที่ (2-12) ผู้อ่านสามารถตรวจสอบได้อย่างง่ายว่าฟังก์ชัน k ที่กำหนดใน (2-13) เป็นเคอร์เนลที่ถูกต้องบน space X ซึ่งไม่จำเป็นต้องเป็นพื้นที่เวกเตอร์

การแมพใด ๆ $\phi : \mathrm{X} \rightarrow \in \mathrm{R}^{\mathrm{p}}$ สำหรับ $p \geq 0$ ส่งผลให้เคอร์เนลที่ถูกต้องผ่าน (2-13) ในทางกลับกันอาจสงสัยว่ามีข้อมูลทั่วไปมากกว่านี้หรือไม่ ดังที่ต่อไปนี้ของ *Aronszajn (*1950) แสดงให้เห็นว่าคำตอบนั้นเป็นค่าลบ

จากงานวิจัยการเรียนรู้จำเสียงสภาพแวดล้อมได้เสนอวิธีการจำแนกด้วยซัพพอร์ตเวกเตอร์แมชชีนที่นำไปประยุกต์ใช้กับวิธีเคอร์เนลเรเดียลเบสิสฟังก์ชั่น สำหรับการจำแนกข้อมูลเสียงสภาพแวดล้อม



รูปที่ 2.29 โครงข่ายประสาทลึก

2.2.4 โครงข่ายประสาทลึก

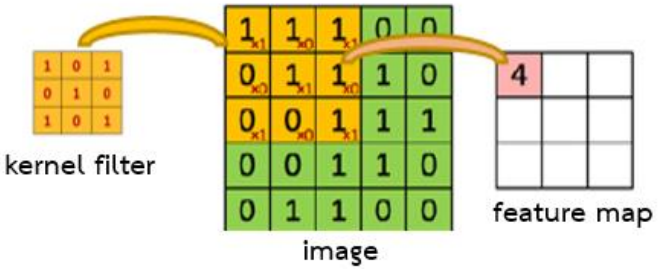
การเรียนรู้เชิงลึก (deep learning) เป็นสาขาของปัญญาประดิษฐ์ที่พัฒนามาจากเครื่องมือการจำแนกอื่น ๆ พื้นฐานของการเรียนรู้เชิงลึกคืออัลกอริทึมที่พยายามจะสร้างแบบจำลองที่ประกอบไปด้วยโครงสร้างย่อย ๆ หลายอัน แสดงดังรูปที่ 2.29

โครงข่ายประสาทลึกเป็นโครงข่ายประสาทที่มีหลาย ๆ ชั้นต่อกัน นอกจากนี้ โครงข่ายประสาทลึกสามารถแบ่งออกได้ดังนี้ เช่น โครงข่ายประสาทเกิดซ้อน (Recurrent Neural Network , RNNs) [3] หน่วยความจำระยะสั้นแบบยาว (Long-Shot Term Memory, LSTM) และ โครงข่ายประสาทคอนโวลูชัน (Convolutional Neural Networks, CNNs) เป็นต้น

2.2.4.1 โครงข่ายประสาทคอนโวลูชัน

โครงข่ายประสาทคอนโวลูชัน (Convolutional Neural Networks) [13] เป็นโครงข่ายประสาทลึกประเภทหนึ่ง โดยเริ่มต้นมาจากงานวิจัยทางด้านการรู้จำตัวอักษรที่มักจะใช้ข้อมูลเมทริกซ์ที่ได้มาจากการแปลงข้อมูลภาพ ส่วนของโครงสร้างโครงข่ายประสาทคอนโวลูชันประกอบได้ดังนี้

- คอนโวลูชัน (convolutional Layer)
- พลูลิง (pooling หรือ subsampling layer)
- การเชื่อมโยงเต็มรูปแบบ (fully connected layer)



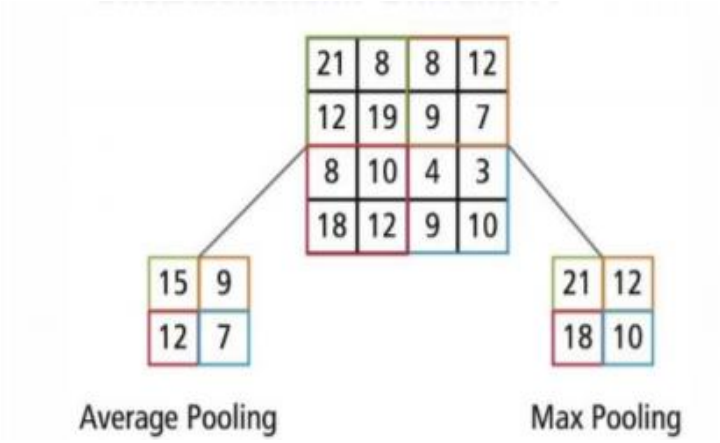
รูปที่ 2.30 การทำคอนโวลูชันภาพนำเข้า

1. ชั้นคอนโวลูชัน

คอนโวลูชันประกอบด้วยเคอร์เนลฟิลเตอร์ (kernel filter) โดยแต่ละเคอร์เนลจะถูกกำหนดจากการสุ่มของค่าการเรียนรู้เริ่มต้นที่ได้จากการปรับค่าการแพร่กระจายย้อนกลับ ทำให้ผลลัพธ์ที่ได้ถูกที่เรียกว่า แมพคุณลักษณะ (feature map) นอกจากนี้ คอนโวลูชันมักจะตามด้วยฟังก์ชันกระตุ้น ซึ่งเป็นฟังก์ชันแบบไม่เชิงเส้น (non-linear Function) ขั้นตอนการทำคอนโวลูชัน แสดงดังรูปที่ 2.30

- **ขนาดของตัวกรอง (filter size)**
  คือความกว้างและความสูงของตัวกรองที่จะนำมาใช้สำหรับการทำคอนโวลูชัน
- **การทำคอนโวลูชันแบบแคบ (narrow convolution)**
  นิยมใช้สำหรับการทำคอนโวลูชันทั่วไป โดยการทำคอนโวลูชันตัวกรองที่นำไปดอทกับเมทริกซ์จะไม่มีการกระทำเลยขอบของเมทริกซ์ กล่าวได้ว่าการทำคอนโวลูชันที่มีข้อมูลอินพุตขนาด NxN กับตัวกรองขนาด mxm จะทำให้เราได้ขนาดเมทริกซ์เอาท์พุตได้จากสูตร (N-m+1)x(N-m+1)

นอกจากนี้ยังมีการทำคอนโวลูชันแบบกว้าง (wide Convolution) จะมีการกระทำเลยขอบของเมทริกซ์อินพุตออกไป โดยส่วนที่เกินออกไปนั้นจะมีการแทนค่าของข้อมูลช่องนั้น ๆ ด้วย 0 เรียกว่า การเสริมด้วยศูนย์ (zero padding) กล่าวคือ การทำคอนโวลูชันแบบกว้างที่มีข้อมูลรับขนาด NxN กับตัวกรองขนาด mxm จะได้เมทริกซ์ขนาด (N-m+1)x(N-m+1) การทำคอนโวลูชันแบบกว้างนี้มีจุดประสงค์เพื่อป้องกันการสูญเสียข้อมูลตรงบริเวณขอบของข้อมูลอินพุต



รูปที่ 2.31 การหาพลูลิงแบบหาค่าเฉลี่ยและ
หาค่ามากที่สุด

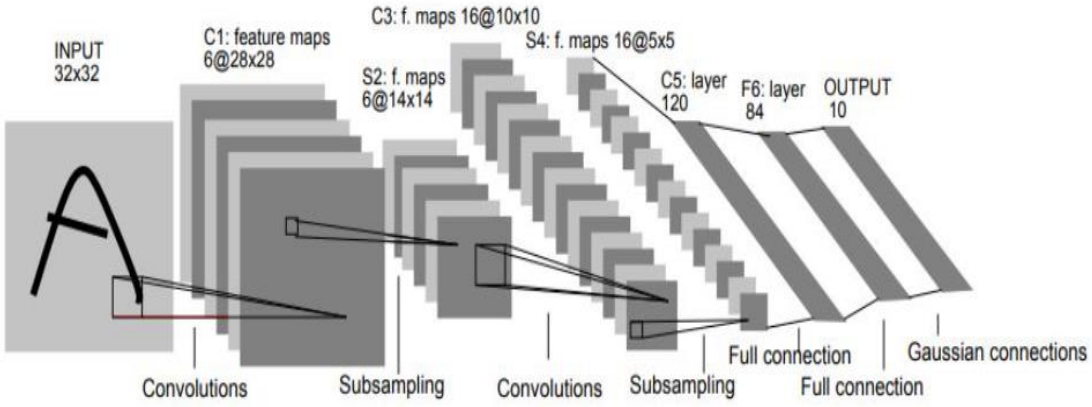
1. ชั้นพลูลิง

มีจุดประสงค์เพื่อทำการลดขนาดของข้อมูลที่ผ่านการทำคอนโวลูชัน โดยนิยมนำมาต่อกับชั้นคอนโวลูชัน แต่ก็อาจไม่จำเป็นต้องนำมาต่อกันเสมอไป ดังนั้นการออกแบบของพลูลิงที่นิยมมีสองวิธีคือ

- **การทำพลูลิงแบบหาค่ามากที่สุด (Max Pooling)**
  เป็นวิธีที่นิยมอย่างแพร่หลายมากในงานวิจัยด้านคอนโวลูชันแบบโครงข่ายประสาทในปัจจุบัน
- **การทำพลูลิงแบบหาค่าเฉลี่ย (Average Pooling)**
  เป็นการเลือกข้อมูลทั้งหมดแล้วนำมาหาค่าเฉลี่ย แสดงดังรูปที่ 2.31 จะแตกต่างจากการทำพลูลิงแบบหาค่ามากที่สุด

2. ชั้นการเชื่อมโยงเต็มรูปแบบ

เป็นชั้นสุดท้ายของนิวรอลของโครงข่ายประสาทคอนโวลูชัน ทำหน้าที่การจำแนกข้อมูลที่ถูกป้อนเข้ามา โครงข่ายประสาทคอนโวลูชันจะเป็นการเชื่อมโยงเต็มรูปแบบ หลังจากการประกอบกันของชั้นคอนโวลูชันและชั้นพลูลิงจะมีเพอร์เซ็ปตรอนอยู่จำนวนหนึ่ง ซึ่งเพอร์เซ็ปตรอนแต่ละตัวจะมีเส้นเชื่อมกับเพอร์เซ็ปตรอนทุกตัวในชั้นก่อนหน้า นอกจากนี้ เพอร์เซ็ปตรอนทุกตัวถัดไปจะมีการคำนวณแบบป้อนไปข้างหน้าและการแพร่กระจายย้อนกลับ

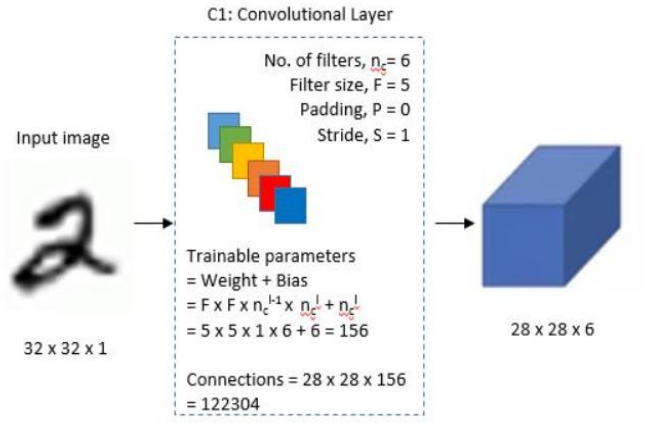


รูปที่ 2.32 แบบจำลองของ Lenet–5

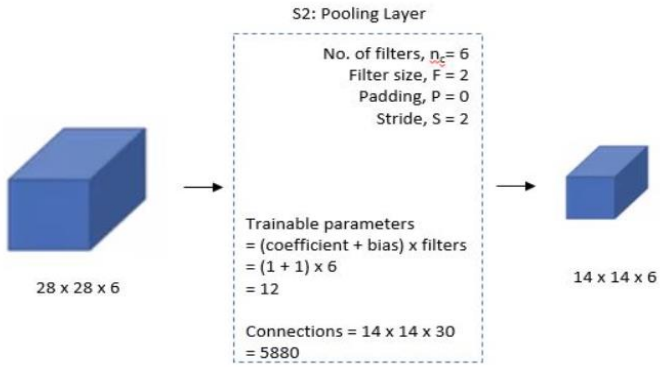
Y. LeCun [12] ได้เสนอสถาปัตยกรรมโครงข่ายประสาทเทียมสำหรับการรู้จำอักขระที่เขียนด้วยลายมือและพิมพ์ด้วยเครื่องจักรในปี 1990 ซึ่งเรียกว่า LeNet-5 แสดงดังรูปที่ 2.32 โดยภายใน Lenet–5 ประกอบไปด้วยเลเยอร์ convolutional pooling และ fully connected ในการสร้างแบบจำลองรู้จำอักขระ

จากงานวิจัยเราได้เสนอแบบจำลองของ LeNet–5 ในการจำแนกเสียงสภาพแวดล้อมโดยการใช้แบบจำลอง LeNet–5 ข้อมูลที่ป้อนเข้าแบบจำลองต้องเป็นข้อมูลภาพ เนื่องจากงานวิจัยเราข้อมูลเป็นสัญญาณเสียง เราจึงประยุกต์จากข้อมูลเสียงแปลงมาเป็นข้อมูลภาพด้วยวิธีผลการแปลงฟูเรียร์ช่วงเวลาสั้น ตัวอย่างการเรียนรู้จำอักขระของแบบจำลอง LeNet–5 มีดังต่อไป

1. ขนาดข้อมูลตัวอักขระมีขนาดรูปภาพไซต์เท่ากับ 32x32
2. ขั้นตอนการฝึกฝนเครื่องมือภายใน LeNet–5 จะประกอบไปด้วย
   - convolutional layer จะประกอบไปด้วย 2 conv ในขั้นตอนการทำงาน
   - pooling layer จะประกอบไปด้วย 2 pool ในขั้นตอนการทำงาน
   - fully connected จะประกอบไปด้วย 2 Fully ในขั้นตอนการทำงาน



รูปที่ 2.33 ตัวอย่างการรู้จำอักขระของแบบจำลอง Lenet–5
สำหรับการทำคอนโวลูชันเลเยอร์แรก



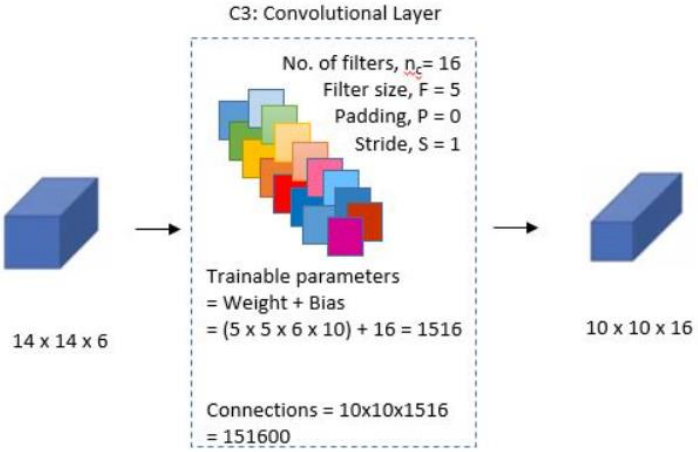
รูปที่ 2.34 ตัวอย่างการรู้จำอักขระของแบบจำลอง Lenet–5
สำหรับการทำพลูลิงเลเยอร์แรก

- **ขั้นตอนแรกคอนโวลูชันเลเยอร์แรก**

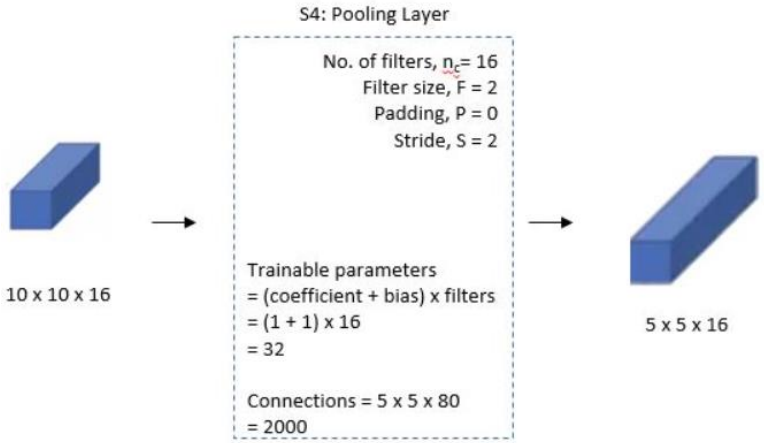
  อินพุตของ LeNet-5 เป็นข้อมูลภาพระดับสีเทาขนาด 32×32 ขั้นตอนแรกนำข้อมูลภาพมาทำคอนโวลูชัน แสดงดังรูปที่ 2.33 จะมีตัวกรองที่มีขนาด 5×5 ทำการลดขนาดมิติข้อมูลภาพจากขนาด 32x32x1 เหลือเพียง 28x28 และมีขนาด 6 คุณลักษณะ ในการเรียนรู้จำอักขระ

- **ขั้นตอนสองพลูลิงเลเยอร์แรก**

  การทำพลูลิงเลเยอร์จะเป็นชั้นที่ต่อจากเลเยอร์ของคอนโวลูชัน วิธีการทำคือนำข้อมูลมาผ่านตัวกรองขนาด 2x2 ทำให้ลดขนาดมิติจากเดิม 28x28x6 เหลือเพียง 14x14x6 แสดงดังรูปที่ 2.34



รูปที่ 2.35 ตัวอย่างการรู้จำอักขระของแบบจำลอง Lenet–5 สำหรับการทำคอนโวลูชันเลเยอร์สอง



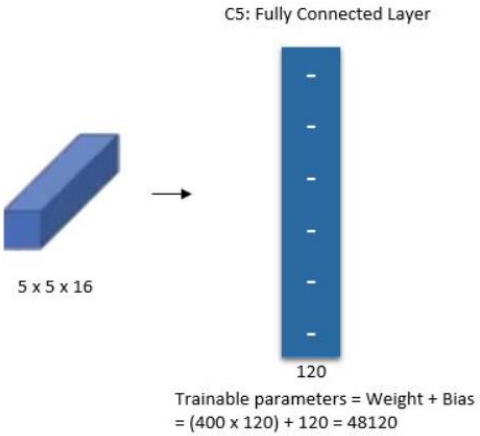
รูปที่ 2.36 ตัวอย่างการรู้จำอักขระของแบบจำลอง Lenet–5 สำหรับการทำพลูลิงเลเยอร์สอง

- **ชั้นตอนสามคอนโวลูชันเลเยอร์สอง**

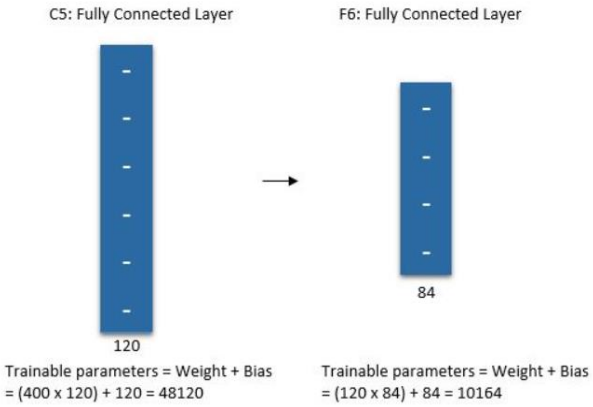
  การทำคอนโวลูชันเลเยอร์สอง ขนาดตัวกรองใช้เหมือนกับเลเยอร์แรกคือขนาด 5 × 5 สำหรับการลดขนาดมิติภาพจาก 14x14x6 เหลือเพียง 10x10 และได้ 16 คุณลักษณะ โดยค่าพารามิเตอร์ที่ใช้ฝึกฝน (trainable parameters) เท่ากับ 151600 มิติ แสดงดังรูปที่ 2.35

- **ชั้นตอนสองพลูลิงเลเยอร์สอง**

  นำคอนโวลูชันเลเยอร์สองมาผ่านการทำพลูลิงอีกรอบจากขนาด 10x10x16 ลดเหลือ 5x5x16 เหลือค่าพารามิเตอร์ที่ฝึกฝนเท่ากับ 2000 มิติ แสดงดังรูปที่ 2.36



รูปที่ 2.37 การทำการเชื่อมโยงเต็มรูปเลเยอร์แรก



รูปที่ 2.38 การทำการเชื่อมโยงเต็มรูปเลเยอร์สอง

- **ขั้นตอนห้าทำการเชื่อมโยงเต็มรูปเลเยอร์แรก**

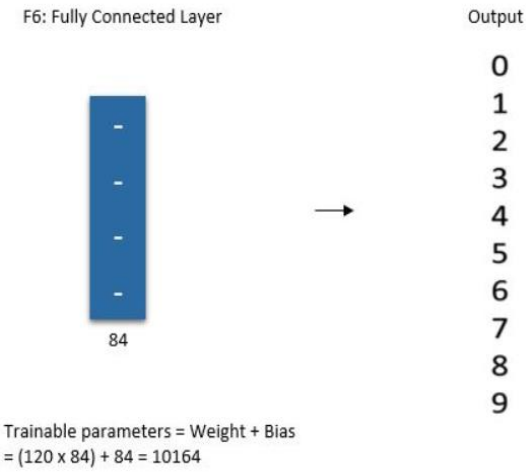
  ขั้นตอนต่อมาคือการรวมข้อมูลที่ได้จากการทำคอนโวลูชันและพลูลิง เข้าสู่กระบวนการทำการเชื่อมโยงเต็มรูปเลเยอร์ทำหน้าที่จำแนกข้อมูลที่ถูกป้อนเข้ามาการเชื่อมโยงเต็มรูปแบบมีขนาด 5x5x16 หลังจากนั้นข้อมูลจะถูกเพิ่มขนาดคุณลักษณะเป็น 120 ทำให้ขนาดการเชื่อมโยงเต็มรูปเลเยอร์แรกมีขนาดค่าพารามิเตอร์ที่ฝึกฝนเท่ากับ 48120 มิติ แสดงดังรูปที่ 2.37

- **ขั้นตอนห้าทำการเชื่อมโยงเต็มรูปเลเยอร์สอง**

  ก่อนหน้านี้เราได้มีการเพิ่มการเชื่อมโยงเต็มรูปแบบ 120 คุณลักษณะ หลังจากนั้นได้มีการปรับลดเหลือ 84 คุณลักษณะ ทำให้เหลือค่าพารามิเตอร์ที่ฝึกฝนเท่ากับ 10164 มิติ แสดงดังรูปที่ 2.38

ตารางที่ 2.2 สถาปัตยกรรมของแบบจำลองของ LeNet-5

Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	32x32	-	-	-
1	Convolution	6	28x28	5x5	1	tanh
2	Average Pooling	6	14x14	2x2	2	tanh
3	Convolution	16	10x10	5x5	1	tanh
4	Average Pooling	16	5x5	2x2	2	tanh
5	Convolution	120	1x1	5x5	1	tanh
6	FC	-	84	-	-	tanh
Output	FC	-	10	-	-	softmax



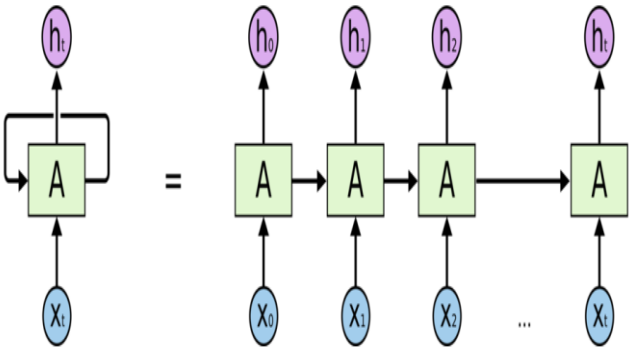
รูปที่ 2.39 เอาท์พุตการเชื่อมโยงเต็มรูปเลเยอร์

- **ขั้นตอนสุดท้ายการเชื่อมโยงเต็มรูปเลเยอร์**

  ขั้นตอนนี้เป็นการจำแนกทั้งหมด 10 อักขระออกมา เริ่มตั้งแต่อักขระ 0 ถึง 9 ที่ได้มาจากการทำคอนโวลูชัน พลูลิง และการเชื่อมโยงเต็มรูปแบบ แสดงดังรูปที่ 2.39

จากงานวิจัยของ LeNet–5 การออกแบบจำลองสำหรับการรู้จำตัวอักขระใช้สถาปัตยกรรมทั้งหมด แสดงดังตารางที่ 1.2 คือ คอนโวลูชันทั้งหมด 2 เลเยอร์ ส่วนของพลูลิงเลเยอร์ใช้ไปทั้งหมด 2 เลเยอร์ และส่วนสุดท้ายของการเชื่อมโยงเต็มรูปเลเยอร์ใช้ทั้งหมด 2 เลเยอร์

จากที่เราได้ศึกษาแบบจำลองของ LeNet-5 เราสามารถนำมาประยุกต์ใช้กับงานด้านการเรียนรู้จำเสียงสภาพแวดล้อมกับเสียงปืนใหญ่ได้



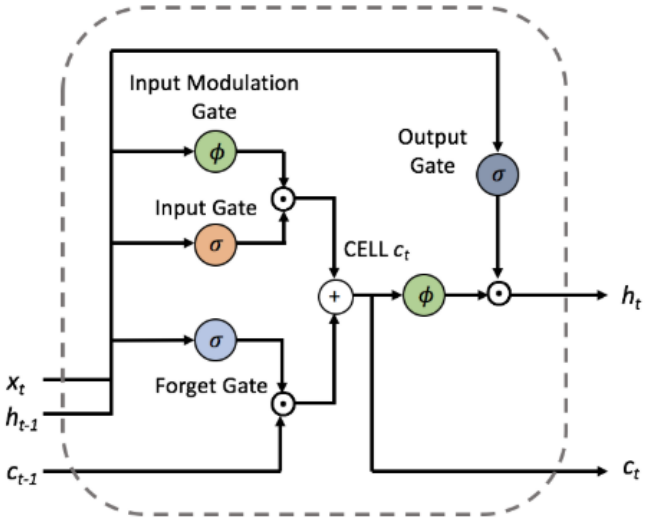
รูปที่ 2.40 โครงข่ายประสาทเกิดซ้อน

2.2.4.2 โครงข่ายประสาทเกิดซ้อน

โครงข่ายประสาทเกิดซ้อน [13] คือโครงข่ายที่มีการเชื่อมต่อระหว่างโหนดในรูปที่ 2.40 ตามลำดับ โครงข่ายประสาทเกิดซ้อนสามารถใช้สถานะภายในหน่วยความจำ เพื่อการประมวลผลลำดับของอินพุตทำให้สามารถใช้กับงานต่าง ๆ เช่น การจำแนกลายมือหรือการรู้จำเสียงพูดหรือแม้แต่รูปภาพเองก็ตาม โดยแต่ละโหนดของโครงข่ายประสาทเกิดซ้อนจะมีข้อมูลขาเข้าสองอย่าง ได้แก่ ข้อมูลอินพุต ณ โหนดนั้น ๆ ผลลัพธ์ที่ได้จากการคำนวณโหนดก่อนหน้า ทั้งสองข้อมูลจะถูกนำมารวมเข้าด้วยกันและออกผลลัพธ์มาเป็นสองทางคือผลลัพธ์ที่ออก ณ โหนดนั้น ๆ เพื่อนำข้อมูลขาเข้าในโหนดถัดไป

ข้อดีของโครงข่ายประสาทเกิดซ้อนคือการใช้ข้อมูลก่อนหน้าในการทำนายข้อมูลอนาคต ทำให้ลดปัญหาการสิ้นเปลืองทรัพยากรในการทำนายข้อมูล ด้วยเหตุนี้โครงข่ายประสาทเกิดซ้อนจึงไม่ต้องมีการลดมิติด้วยการวิเคราะห์องค์ประกอบหลัก

ข้อเสียของโครงข่ายประสาทเกิดซ้อนที่ต้องพบเจอคือโครงข่ายประสาทเกิดซ้อนสามารถดูย้อนกลับได้แค่เพียงในช่วงระยะเวลาสั้น ๆ เท่านั้น โดยปัญหาหลักของโครงข่ายประสาทเกิดซ้อนคือค่าเกรเดียน (gradient) เริ่มน้อยลงมาจากขนาดข้อมูลที่มีความยาวมากขึ้น ทำให้ไม่สามารถเห็นความเปลี่ยนแปลงของเกรเดียนได้เลย ซึ่งปัญหานี้ถูกเรียกว่า vanishing gradient problem



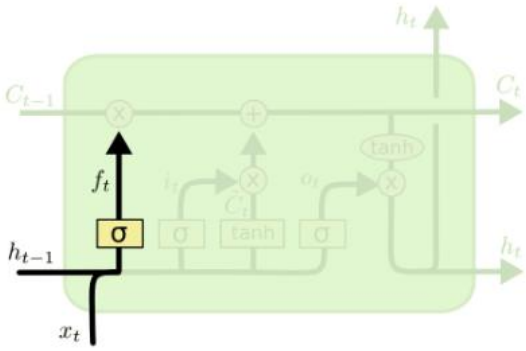
รูปที่ 2.41 หน่วยความจำระยะสั้นระยะยาว

2.2.4.2.1 หน่วยความจำระยะสั้นระยะยาว

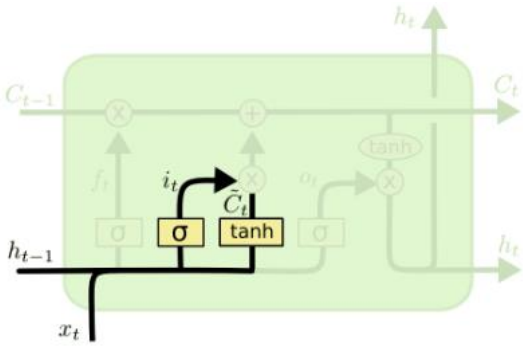
หน่วยความจำระยะสั้นระยะยาว หรือเรียกว่า LSTM แสดงดังรูปที่ 2.41 เป็นโครงข่ายประสาทเกิดซ้อนชนิดพิเศษที่สามารถเรียนรู้ข้อมูลที่มีระยะยาวได้ มาตรฐานของหน่วยความจำระยะสั้นระยะยาวมีการเชื่อมต่อข้อเสนอแนะสามารถประมวลผลจุดข้อมูลเดียวคือข้อมูลภาพโดยสามารถเรียงลำดับข้อมูลทั้งหมด เช่น คำพูดหรือวิดีโอ ตัวอย่างเช่น หน่วยความจำระยะสั้นระยะยาวสามารถใช้ได้กับงานต่าง ๆ เช่น unsegmented การรู้จำลายมือที่เชื่อมต่อการรู้จำเสียงและการตรวจจับความผิดปกติในการรับส่งข้อมูลโครงข่าย

หน้าที่ของหน่วยความจำระยะสั้นระยะยาวเพื่อจะนำมาแก้ปัญหาของโครงข่ายประสาทเกิดซ้อนที่ไม่สามารถจดจำลำดับข้อมูลยาว ๆ ได้ ด้วยเหตุนี้เราจึงได้นำเสนอหน่วยความจำระยะสั้นระยะยาวนำมาใช้กับสำหรับการจำเสียงสภาพแวดล้อม หลักการทำงานของหน่วยความจำระยะสั้นระยะยาวมีดังต่อไปนี้

1. การทำ forget gate layer
2. การทำ input gate layer
3. การทำ update cell memory
4. การทำ output cell update



รูปที่ 2.42 forget gate layer



รูปที่ 2.43 input gate layer

2.2.4.2.2 หลักการทำงานการของหน่วยความจำระยะสั้นระยะยาว

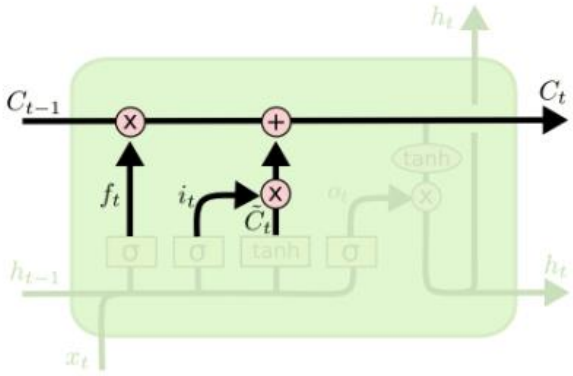
ขั้นตอนแรกของหน่วยความจำระยะสั้นระยะยาวคือการตัดสินใจว่าข้อมูลใดควรจะถูกตัดทิ้งจากสถานะเซลล์ โดยการตัดสินใจจะขึ้นอยู่กับ sigmoid ที่เรียกว่า "forget gate layer" แสดงดังรูปที่ 2.42 จะเห็นได้ว่า $\boldsymbol{h_{t-1}}$ ทำหน้าที่ส่งข้อมูลออกระหว่างตัวเลข 0 ถึง 1 แสดงดังสมการที่ (2-11) จะเห็นได้ว่าแต่ละหมายเลขในสถานะเซลล์ $\boldsymbol{C_{t-1}}$ หมายถึงเก็บสิ่งนี้ไว้อย่างสมบูรณ์ ในขณะที่ 0 หมายถึงกำจัดสิ่งนี้อย่างสมบูรณ์

$$f_t = \sigma\left(W_f * [h_{t-1}, x_t] + b_f\right) \qquad (2 - 11)$$

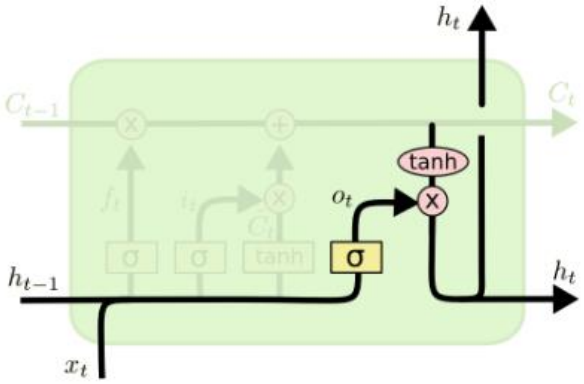
ส่วนที่สองคือการตัดสินใจว่าควรจะเลือกเก็บข้อมูลใหม่ไว้ในสถานะของเซลล์ ในส่วนของเลเยอร์ sigmoid ที่เรียกว่า “input gate layer” แสดงดังรูปที่ 2.43 คือทำหน้าที่ตัดสินใจว่าเราจะอัปเดตค่าใดส่งถัดไปให้กับเลเยอร์ tanh โดยเราจะสร้างค่าเวกเตอร์ตัวเลือกใหม่คือ $\boldsymbol{C'_t}$ แสดงดังสมการที่ (2-12) และ (2-13) ที่สามารถเพิ่มเข้าไปในสถานะขั้นตอนถัดไป โดยเราจะรวมสองสิ่งนี้เพื่อสร้างการอัปเดตให้เป็นสถานะขึ้นมาใหม่

$$i_t = \sigma\left(W_i * [h_{t-1}, x_t] + b_i\right) \qquad (2 - 12)$$

$$C'_t = tanh\left(W_c * [h_{t-1}, x_t] + b_c\right) \qquad (2 - 13)$$



รูปที่ 2.44 update cell memory



รูปที่ 2.45 output cell update

ขั้นตอนที่สามคือการอัปเดตสถานะเซลล์เก่าคือ C_{t-1} แสดงดังรูปที่ 2.44 เข้าสู่สถานะเซลล์ใหม่ โดยขั้นตอนก่อนหน้านี้จะต้องตัดสินใจว่าจะทำอย่างไรกับข้อมูลก่อนหน้า

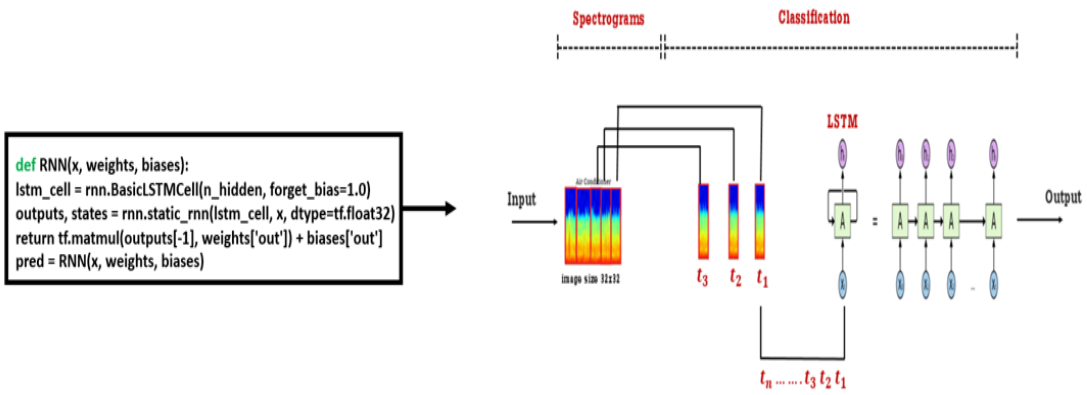
เราจะเพิ่มความทวีคูณของสถานะเก่าด้วยการลืมสิ่งที่เราตัดสินใจไปก่อนหน้า จากนั้นเราจะเพิ่ม $i_t * C'_t$ แสดงดังสมการที่ (2-14) คือค่าตัวเลือกใหม่สำหรับปรับขนาดตามจำนวนที่เราตัดสินใจที่จะอัปเดตค่าสถานะของแต่ละค่า

$$C_t = f_c * C_{t-1} + i_t * C'_t \quad (2 - 14)$$

ส่วนสุดท้ายเราจะต้องตัดสินใจว่าจะส่งข้อมูลอะไรออกไป ผลลัพธ์ที่ได้ขึ้นอยู่กับสถานะเซลล์สำหรับเวอร์ชันที่ถูกกรองมาก่อนหน้านี้ ก่อนอื่นเราจะต้องเรียกใช้ sigmoid layer เพื่อทำการตัดสินใจว่าส่วนใดของสถานะเซลล์ที่เราจะเลือกส่งข้อมูลออกไป จากนั้นเราจะใส่ผ่านสถานะของเซลล์ tanh (เพื่อดันค่าให้อยู่ระหว่าง −1 ถึง 1) และคูณมันด้วยเอาต์พุตของ sigmoid gate เพื่อให้เราส่งออกเฉพาะส่วนที่เราตัดสินใจเท่านั้น

$$O_t = \sigma\left(W_0 * [h_{t-1}, x_t] + b_0\right) \quad (2 - 15)$$

$$h_t = O_t * \tanh(C_t) \quad (2 - 16)$$



รูปที่ 2.46 ตัวอย่างการทำของโครงข่ายประสาทเกิดซ้อนกับเสียงสภาพแวดล้อม

2.2.4.2.3 หลักการทำงานของโครงข่ายประสาทเกิดซ้อนและหน่วยความจำระยะสั้นระยะยาว

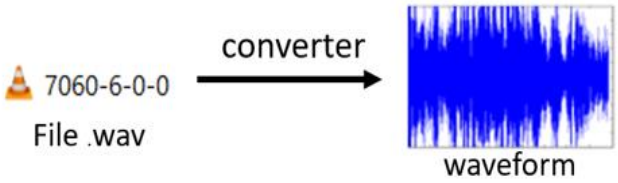
การจำแนกเสียงสภาพแวดล้อมด้วยโครงข่ายประสาทเกิดซ้อนและหน่วยความจำระยะสั้นระยะยาว จากตัวอย่างรูปที่ 2.46 เราจะเห็นได้ว่าข้อมูลอินพุตจะต้องเป็นข้อมูลภาพ ด้วยเหตุนี้เราจึงต้องแปลงจากสัญญาณเสียงสภาพแวดล้อมให้ได้คุณสมบัติภาพ ด้วยวิธีของผลการแปลงฟูเรียร์ช่วงเวลาสั้น

ขั้นตอนการทำโครงข่ายประสาทเกิดซ้อนกับเสียงสภาพแวดล้อมจะแตกต่างจากเครื่องมือการจำแนกอื่นที่วิเคราะห์ข้อมูลทั้งก้อน แต่สำหรับโครงข่ายประสาทเกิดซ้อนนั้นจะวิเคราะห์ข้อมูลเป็นที่ละ step หรือเป็นลำดับ แสดงดังรูปที่ 2.46 จะเห็นได้ว่า $t_1\ t_2\ t_3$ จนถึง t_n นั้นเป็นการลำดับข้อมูลแต่ละช่วงเวลา-ความถี่ ให้กับโครงข่ายประสาทเกิดซ้อน แต่เนื่องจากโครงข่ายประสาทเกิดซ้อนไม่สามารถจดจำข้อมูลได้หลาย ๆ ลำดับ เราจึงเสนอวิธีหน่วยความจำระยะสั้นระยะยาว หรือเรียกตัวย่อว่า LSTM ที่สามารถแก้ปัญหาการจดจำข้อมูลที่มีขนาดยาวได้ ทำให้ประสิทธิภาพของการจำแนกโครงข่ายประสาทเกิดซ้อนเกิดความแม่นยำเพิ่มขึ้น จากตัวอย่างวิธีการสร้างแบบจำลองของโครงข่ายประสาทเกิดซ้อนและหน่วยความจำระยะสั้นระยะยาว โดยเราจะใช้คำสั่ง **lstm_cell = rnn.BasicLSTMCell(n_hidden, forget_bias=1.0)** และ **outputs, states = rnn.static_rnn(lstm_cell, x, dtype = tf.float32)**

สรุปการทำโครงข่ายประสาทเกิดซ้อนกับเสียงสภาพแวดล้อม โดยเราได้เสนอวิธีหน่วยความจำระยะสั้นระยะยาวนำมาแก้ปัญหาของโครงข่ายประสาทเกิดซ้อนที่ไม่สามารถจดจำข้อมูลได้หลาย ๆ ลำดับ

```
import soundfile as sf
[data,rate] = sf.read(file_name)
```

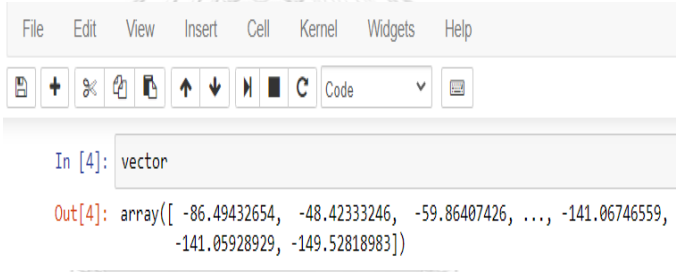
รูปที่ 2.47 ตัวอย่างโปรแกรมแปลงสัญญาณไฟล์เสียง .wav



รูปที่ 2.48 ตัวอย่างการแปลงไฟล์ .wav มาเป็นสัญญาณโดมเมนทางเวลา

```
import numpy as np
```

รูปที่ 2.49 โปรแกรม import numpy as np



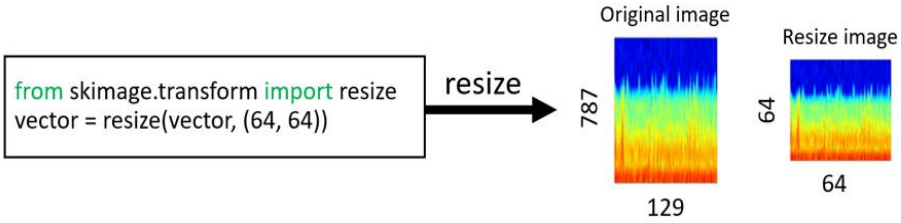
รูปที่ 2.50 ตัวอย่างโปรแกรม numpy เก็บค่าเวกเตอร์ใน array

2.3 คำสั่งพื้นฐาน

จากงานวิจัยเราใช้ข้อมูลเสียงสภาพแวดล้อมของ Urbansound 8K มีไฟล์เสียงทั้งหมด 8732 เสียง และเป็นไฟล์ .wav ทั้งหมด ทำให้เราต้องแปลงสัญญาณเสียงอยู่ในรูปแบบ waveform หรือสัญญาณโดเมนทางเวลา โดยเราใช้คำสั่ง import soundfile as sf และ [data, rate] = sf.read(file_name) แสดงดังรูปที่ 2.47 ทำหน้าที่แปลงไฟล์เสียง .wav ให้เป็นสัญญาณโดเมนทางเวลา แสดงดังรูปที่ 2.48 จึงทำให้นำข้อมูลมาใช้ในงานเรียนรู้จำเสียงสภาพแวดล้อมได้

จากงานวิจัยการเรียนรู้จำเสียงสภาพแวดล้อม ข้อมูลส่วนใหญ่เป็นค่าเวกเตอร์ที่ใช้สำหรับการเรียนรู้จำเสียง เราจึงใช้คำสั่ง import numpy as np แสดงดังรูปที่ 2.49 ทำหน้าที่สำหรับเก็บค่าเวกเตอร์ไว้ใน array แสดงดังรูปที่ 2.50

รูปที่ 2.51 Scikit-Image



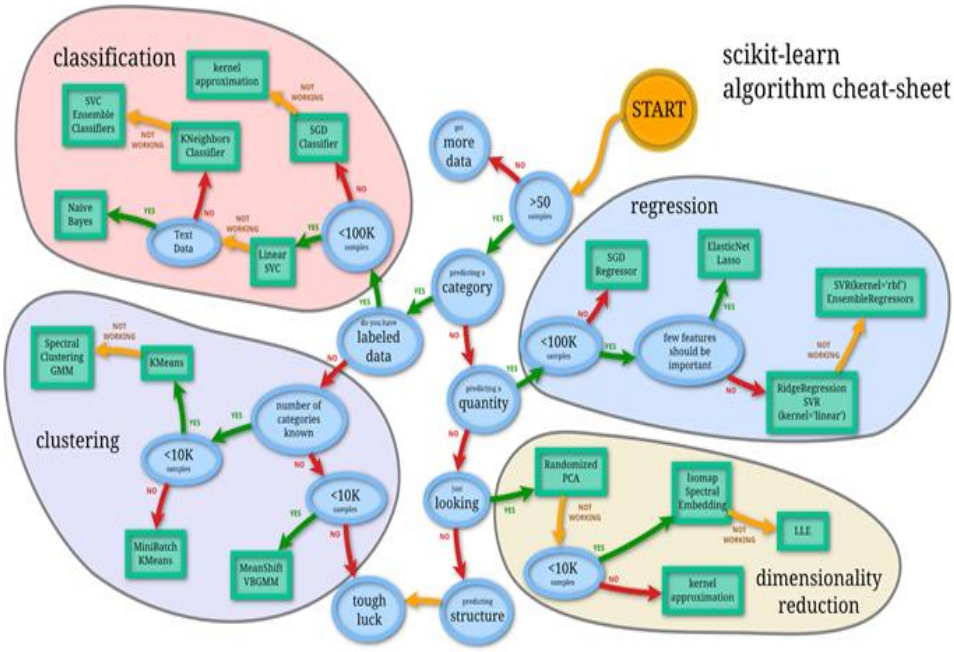
รูปที่ 2.52 การปรับขนาดภาพด้วยคำสั่ง resize

2.4 ฟังก์ชัน Scikit-image

scikit-image [13] คือชุดของอัลกอริทึมสำหรับการประมวลผลทางภาพจากรูปที่ 2.51 ข้อดีของฟังก์ชัน scikit-image สามารถปรับขนาดภาพได้ตามความเหมาะสมสำหรับผู้ใช้งานและสามารถปรับแต่งพื้นที่สีและอื่น ๆ นอกจากนี้ ตัวฟังก์ชันยังถูกออกแบบมาเพื่อใช้งานร่วมกันกับ Python NumPy และ SciPy ตัวอย่างคำสั่งของ scikit-image มีดังต่อไปนี้

1. **from skimage import data**
2. **from skimage import color**
3. **from skimage import img_as_float**
4. **from skimage.transform import resize**

จากหัวข้อนี้เราจะพูดถึง scikit-image ที่นำมาใช้การปรับขนาดภาพของเสียงสภาพแวดล้อมให้แต่ละด้านมีขนาดเท่า ๆ กัน เนื่องจากเครื่องมือการจำแนกของโครงข่ายคอนโวลูชันข้อมูลภาพในการทดสอบต้องมีขนาดแต่ละด้านที่เท่ากัน จากตัวอย่างรูปที่ 2.52 จะเห็นได้ว่าข้อมูลสัญญาณเสียงสภาพแวดล้อมที่ถูกแปลงด้วยผลการแปลงฟูเรียร์ ทำให้ได้ขนาดภาพ 787x129 เราจึงใช้คำสั่ง **from skimage.transform import resize** ในการปรับขนาดภาพของเสียงสภาพแวดล้อมให้ลดลงเหลือแต่ละด้าน 64x64



รูปที่ 2.53 โมดูล Scikit-learn

2.5 ฟังก์ชัน Scikit-learn

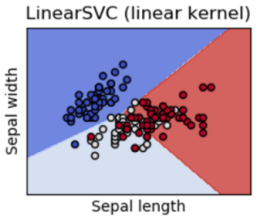
scikit-learn [14] เป็นซอฟต์แวร์สำหรับการเขียนโปรแกรมภาษาไพธอน (python) สามารถทำเกี่ยวกับ classification regression clustering และ dimensionality reduction แสดงดังรูปที่ 2.53 โดยการเรียกใช้ scikit-learn ใน python จะต้องใช้คำสั่ง import เข้ามาอาทิ เช่น **from sklearn import svm** เป็นต้น

จากหัวข้อนี้เราจะพูดถึง scikit-learn ที่นำมาใช้สำหรับงานวิจัยการเรียนรู้จำเสียงสภาพแวดล้อม อาทิเช่น การลดขนาดมิติเสียงสภาพแวดล้อมด้วยวิธีการการวิเคราะห์องค์ประกอบ คำสั่งของ scikit-learn ที่ใช้สำหรับงานวิจัยหลัก ๆ มีดังต่อไปนี้

1. **from sklearn.svm import SVC**
2. **from sklearn.svm import LinearSVC**
3. **from sklearn.neural_network import MLPClassifier**

```
from sklearn.svm import LinearSVC
linear_svc = svm.SVC(kernel='linear')
linear_svc.kernel
```

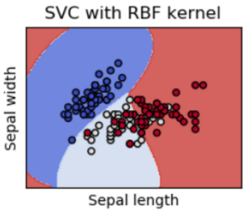
linear



รูปที่ 2.54 การสร้างแบบจำลอง linear kernel

```
from sklearn.svm import SVC
rbf_svc = svm.SVC(kernel='rbf')
rbf_svc.kernel
```

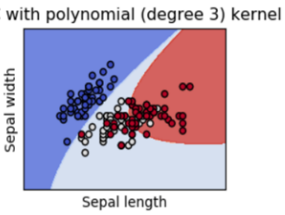
RBF



รูปที่ 2.55 การสร้างแบบจำลอง RBF kernel

```
from sklearn.svm import SVC
poly_svc = svm.SVC(kernel='poly')
poly_svc.kernel
```

poly



รูปที่ 2.56 การสร้างแบบจำลอง poly kernel

2.5.1 การสร้างแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนด้วยวิธี linear kernel

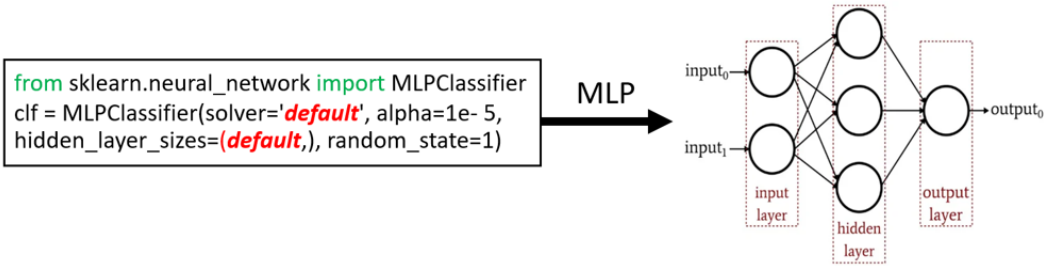
การสร้างแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนด้วยวิธี linear kernel โดยเราจะใช้คำสั่ง **from sklearn.svm import LinearSVC** จากนั้นใช้คำสั่ง **(kernel='linear')** แสดงดังรูปที่ 2.54 ทำให้เราสามารถสร้างแบบจำลอง linear kernel ได้

2.5.2 การสร้างแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนด้วยวิธี RBF kernel

การสร้างแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนด้วยวิธี RBF kernel โดยเราจะใช้คำสั่ง **from sklearn.svm import SVC** จากนั้นใช้คำสั่ง **(kernel='RBF')** แสดงดังรูปที่ 2.55 ทำให้เราสามารถสร้างแบบจำลอง RBF kernel ได้

2.5.3 การสร้างแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนด้วยวิธี poly kernel

การสร้างแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนด้วยวิธี poly kernel โดยเราจะใช้คำสั่ง **from sklearn.svm import SVC** จากนั้นใช้คำสั่ง **(kernel='poly')** แสดงดังรูปที่ 2.56 ทำให้เราสามารถสร้างแบบจำลอง poly kernel ได้

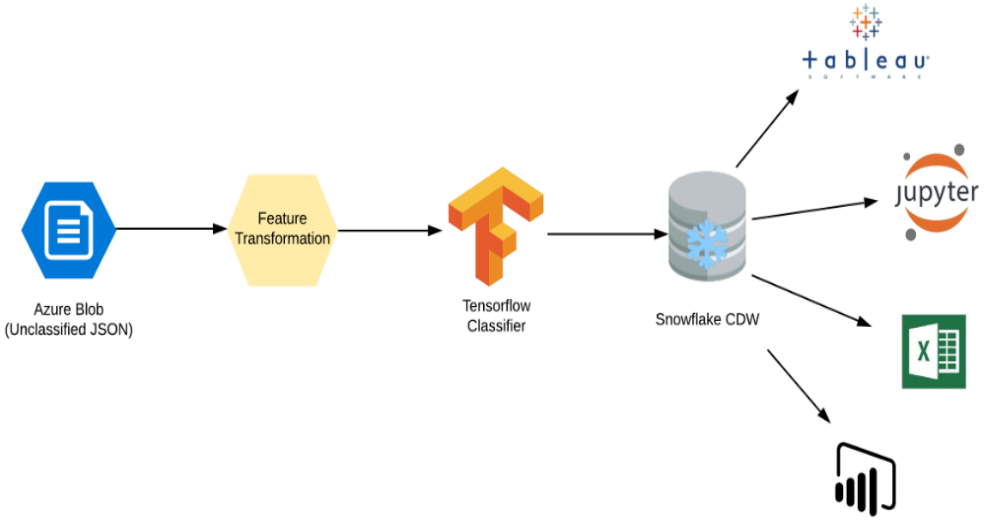


รูปที่ 2.57 การสร้างแบบจำลองเพอร์เซ็ปตรอน

2.5.4 การสร้างแบบจำลองเพอร์เซ็ปตรอน

การสร้างแบบจำลองเพอร์เซ็ปตรอนเราจะ import โมดูลจาก scikit-learn ด้วยคำสั่ง **from sklearn.neural_network import MLPClassifier** เราสามารถกำหนดค่าพารามิเตอร์ตามวัตถุประสงค์ที่เราใช้งานวิจัยเสียงสภาพแวดล้อม แสดงดังรูปที่ 2.57 จะเห็นได้ว่าเรากำหนดค่าพารามิเตอร์ **solver = 'default'** ซึ่งการตั้งค่าเป็น **default** จะทำให้โปรแกรมบ่งชี้กำหนดค่าพารามิเตอร์มาให้เป็น **solver = 'adam'** หน้าที่ของ **adam** คือการเพิ่มประสิทธิภาพการไล่ระดับสี นอกจากนี้ คำสั่งของ **solver** สามารถกำหนดได้ 3 ค่าดังต่อไปนี้ **adam lbfgs** และ **sgd** โดยที่ **lbfgs** เป็นเครื่องมือเพิ่มประสิทธิภาพสำหรับ **quasi-Newton** ส่วนของ **sgd** เป็นการไล่ระดับสีแบบสุ่ม

ต่อมาเป็นการกำหนดค่า hidden layer ถ้าเรากำหนดค่าเป็น **default** โปรแกรมจะให้มา 100 hidden layer แต่ความเป็นจริงเราสามารถกำหนดค่าให้เหมาะกับวัตถุประสงค์ของงานวิจัยได้ ส่วนสุดท้าย **random_ state = 1** คือการสร้างหมายเลขสุ่มสำหรับค่าน้ำหนัก (weights)



รูปที่ 2.58 หลักการทำงานของ Tensorflow

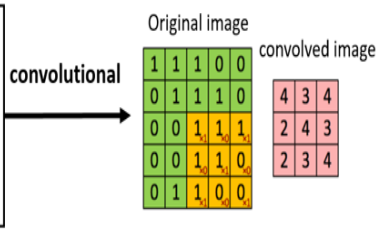
2.6 ฟังก์ชัน Tensorflow

tensorflow [15] เป็นซอฟต์แวร์สำหรับการเขียนโปรแกรมภาษาไพธอน (python) แสดงดังรูปที่ 2.58 สามารถทำเกี่ยวกับ deep learning และ machine learning ส่วนของแพลตฟอร์ม (platform) ที่สามารถติดตั้งใช้สำหรับ tensorflow ได้แก่ jupyter notebook colab spyder และ visual studio เป็นต้น

จากหัวข้อนี้เราจะพูดถึง tensorflow ที่นำมาใช้ในการสร้างแบบจำลองโครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อน สำหรับการจำแนกเสียงสภาพแวดล้อม ตัวอย่างคำสั่งของ tensorflow สำหรับการสร้างแบบจำลองมีดังต่อไปนี้

1. **import tensorflow as tf**
2. **from tensorflow.contrib.layers import flatten**
3. **tf.nn.rnn_cell.BasicRNNCell**
4. **from tensorflow.keras.layers import Dense, Flatten, Conv2D**
5. **from tensorflow.keras import Model**

```
import tensorflow as tf
from tensorflow.contrib.layers import flatten
conv1_W = tf.Variable(tf.truncated_normal(shape=(3, 3, 1, 1), mean = mu, stddev = sigma))
conv1_b = tf.Variable(tf.zeros(20))
conv1   = tf.nn.conv2d(x, conv1_W, strides=[1, 1, 1, 1], padding='VALID') + conv1_b
conv1 = tf.nn.relu(conv1)
```



รูปที่ 2.59 การทำคอนโวลูชันเลเยอร์

```
import tensorflow as tf
from tensorflow.contrib.layers import flatten
conv1 = tf.nn.max_pool(conv1, ksize=[1, 2, 2, 1], strides=[1, 2, 2, 1], padding='VALID')
```

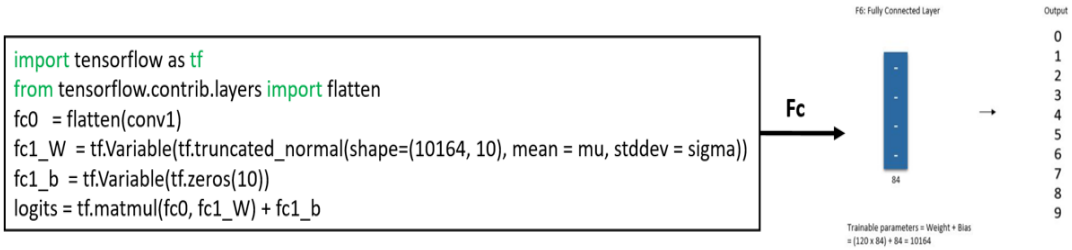
รูปที่ 2.60 การทำพลูลิง

2.6.1 การใช้ฟังก์ชัน Tensorflow สร้างคอนโวลูชันเลเยอร์

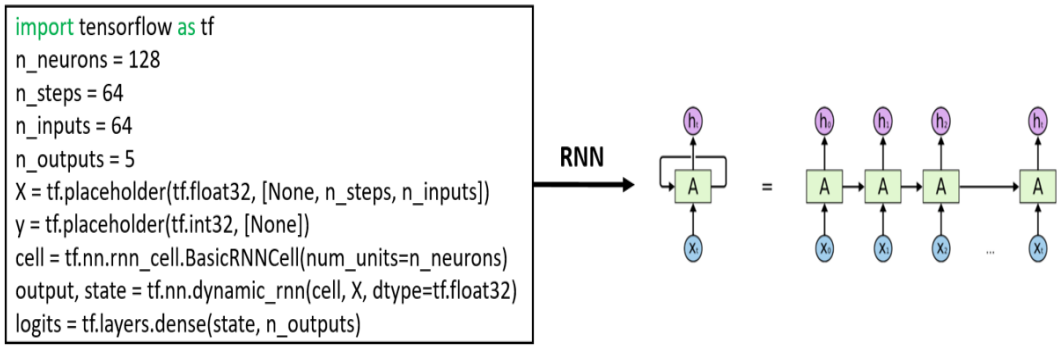
ขั้นตอนการสร้างแบบจำลองโครงข่ายคอนโวลูชัน อันดับแรกเริ่มจากการสร้างคอนโวลูชัน คือการลดมิติข้อมูลที่ใหญ่มากให้มีขนาดมิติเล็กลง โดยใช้คำสั่ง **import tensorflow as tf** และ **from tensorflow.contrib.layers import flatten** จากตัวอย่างเราทำการปรับขนาด **original image** ที่มีขนาดอินพุตเท่ากับ 5x5 ให้เหลือขนาดเอาท์พุตเท่ากับ 3x3 วิธีการทำคอนโวลูชัน เราจะลดมิติด้วยคำสั่ง **shape = (3,3,1,1)** เป็นการสร้างตัวกรองขนาด 3x3 นำมา padding กับข้อมูลภาพของ original image ทำให้ขนาดข้อมูลเอาท์พุตที่ได้คือ 3x3 แสดงดังรูปที่ 2.59 หลังจากข้อมูลที่ได้จาก **conv1** โดยเราจะนำมาผ่านกระบวนการของ **Relu** ด้วยคำสั่ง **conv1 = tf.nn.relu(conv1)** เพราะเราต้องการให้ slope เป็น 1 จึงทำให้ค่าของเกเดียน (Gradient) ไม่หาย จึงช่วยลดปัญหา Vanishing Gradient ทำให้เราสามารถเทรนแบบจำลองได้อย่างรวดเร็ว

2.6.2 การใช้ Tensorflow ทำพลูลิงเลเยอร์

การทำพลูลิงส่วนใหญ่มักจะทำหลังจากคอนโวลูชันเลเยอร์ หน้าที่ของพลูลิงคล้ายกับคอนโวลูชันเลเยอร์คือการลดมิติข้อมูล จากตัวอย่าง original image มีขนาดอินพุตเท่ากับ 4x4 โดยเราจะสร้างพลูลิงลดขนาดมิติเหลือเพียง 2x2 โดยใช้คำสั่ง **ksize=[1, 2, 2, 1], strides=[1, 2, 2, 1], padding='VALID'** แสดงดังรูปที่ 2.60



รูปที่ 2.61 การทำ Fully Connected



รูปที่ 2.62 การสร้างแบบจำลองโครงข่ายประสาทเกิดซ้อน

2.6.3 การใช้ Tensorflow ทำ Fully Connected

การทำ fully connected จะทำเป็นขั้นตอนสุดท้ายหลังจากการทำ convolutional relu และ pooling โดยขั้นตอนนี้จะเป็นการจำแนกจำนวนคลาสข้อมูลออกมาทั้งหมด ซึ่งความแม่นยำของการจำแนกข้อมูลทั้งหมดขึ้นอยู่กับการทำ convolutional relu และ pooling คำสั่งของ fully connected ที่เราใช้คือ **shape=(10164, 10)** แสดงดังรูปที่ 2.61

2.6.4 การใช้ Tensorflow สร้างแบบจำลองโครงข่ายประสาทเกิดซ้อน

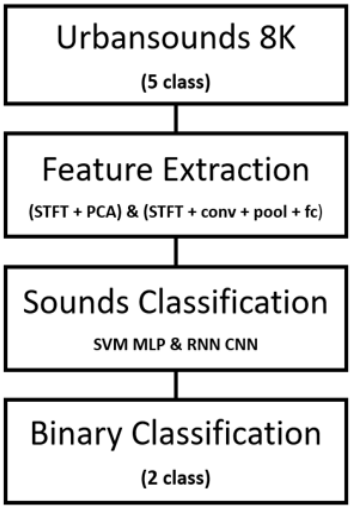
การสร้างแบบจำลองโครงข่ายประสาทเกิดซ้อนเราจะต้องติดตั้ง Tensorflow ก่อน ด้วยคำสั่ง **import tensorflow as tf** ทำให้เราสามารถสร้างแบบจำลองโครงข่ายประสาทเกิดซ้อนได้ คำสั่งที่ใช้คือ **tf.nn.rnn_cell.BasicRNNCell** จากตัวอย่างเราได้กำหนดค่าพารามิเตอร์ n_neurons เท่ากับ 128 หมายถึงจำนวน 128 neurons และ n_steps เท่ากับ 64 หมายถึงเราจะลับดับข้อมูล 64 ครั้ง และ n_output เท่ากับ 5 คือจำนวนข้อมูลที่เราป้อนเข้ามา จากนั้นคำสั่งที่เราสร้าง RNN คือ **tf.nn.rnn_cell.BasicRNNCell** แสดงดังรูปที่ 2.62

บทที่ 3
แนวทางที่เสนอ

3.1 ภาพรวมขั้นตอนการทำงาน

จากงานวิจัยการเรียนรู้จำเสียงจะแบ่งออกเป็น 2 ส่วนหลัก ๆ ส่วนวิธีแรกคือการจำแนกสัญญาณเสียงสภาพแวดล้อม ได้แก่ เสียงเครื่องปรับอากาศ เสียงเด็กเล่น เสียงของรถยนต์ เสียงไซเรน และเสียงดนตรีที่เล่นสถานที่เปิด ส่วนวิธีที่สองคือการจำแนกเสียงสภาพแวดล้อมกับเสียงปืนใหญ่ โดยวัตถุประสงค์ของวิจัยเราต้องการที่จะแยกระหว่างเสียงที่ไม่เป็นอันตรายและเสียงที่เป็นอันตราย

จากงานวิจัยการเรียนรู้จำเสียงสภาพแวดล้อมวิธีดำเนินงาน แสดงดังรูปที่ 3.63 ขั้นตอนแรกนำชุดข้อมูล Urbansounds 8K มาทำการสกัดคุณลักษณะเสียงสภาพแวดล้อมจะแบ่งออกเป็น 2 วิธี วิธีแรกคือการแยกคุณลักษณะด้วยผลการแปลงฟูเรียร์ และวิธีที่สองคือการลดมิติข้อมูลเสียงที่มีขนาดใหญ่มาก ส่วนขั้นตอนที่สองคือเครื่องมือการจำแนก สำหรับงานวิจัยเราได้ใช้เครื่องมือการจำแนกในทดสอบการเปรียบเทียบสมรรถนะการรู้จำเสียงสภาพแวดล้อม ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน เพอร์เซ็ปตรอนหลายชั้น โครงข่ายประสาทเกิดซ้อน และโครงข่ายคอนโวลูชัน ขั้นตอนสุดท้ายคือการทำ binary classification คือการจำแนกระหว่างเสียงที่ไม่เป็นอันตรายและเสียงที่เป็นอันตราย



รูปที่ 3.63 โฟลว์ชาร์ตกระบวนการทำงานของการเรียนรู้จำเสียงสภาพแวดล้อม

ตารางที่ 3.3 รายละเอียดของจำนวนเสียงแต่ละชนิดในฐานข้อมูลเสียง UrbanSound 8K

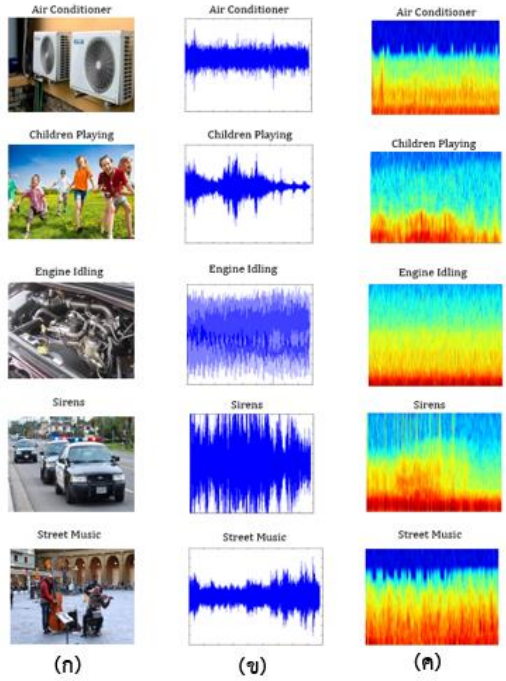
ลำดับ	ชนิดเสียง	จำนวน
1	เสียงเครื่องปรับอากาศ	1000
2	เสียงแตรรถยนต์	429
3	เสียงเด็กเล่น	100
4	เสียงสุนัขเห่า	1000
5	เสียงการเจาะของสว่าน	1000
6	เสียงการทำงานของรถยนต์เมื่อไม่เคลื่อนที่	1000
7	เสียงยิงปืน	374
8	เสียงเจาะจากเครื่องเจาะหิน(Jackhammer)	1000
9	เสียงไซเรน	929
10	เสียงดนตรีที่เล่นในสถานที่เปิด (Street music)	1000
	รวม	8732

3.1.1 การเตรียมฐานข้อมูลสำหรับการฝึกฝนของเสียงสภาพแวดล้อม

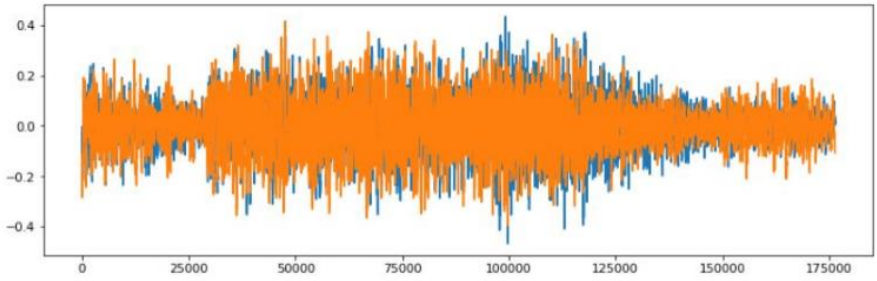
จากงานวิจัยเราใช้ฐานข้อมูล UrbanSound 8K ซึ่งประกอบด้วยไฟล์เสียงสภาพแวดล้อมทั้งหมด 10 คลาสมีทั้งหมด 8,732 เสียง ดังตารางที่ 3.3 คือเสียงเครื่องปรับอากาศ เสียงแตรรถยนต์ เสียงเด็กเล่น เสียงสุนัขเห่า เสียงการเจาะของสว่าน เสียงการทำงานของรถยนต์เมื่อไม่เคลื่อนที่ เสียงยิงปืน เสียงเจาะจากเครื่องเจาะหิน เสียงไซเรน และเสียงดนตรีที่เล่นในสถานที่เปิด โดยในแต่ละชนิดจะมีชุดไฟล์เสียงอยู่อีกหลายชุด ซึ่งมีความยาวเฉลี่ยสูงสุด 4 วินาทีต่อ 1 เสียง แต่ก็มีบางเสียงที่มีความยาว 2 วินานที และหนึ่งในนั้นคือเสียงปืน

จากงานวิจัยเราได้เลือกเสียงสภาพแวดล้อมจากทั้งหมด 10 คลาสเลือกมาเพียง 5 คลาส ได้แก่ เสียงเครื่องปรับอากาศ เสียงเด็กเล่น เสียงการทำงานของรถยนต์เมื่อไม่เคลื่อนที่ เสียงไซเรน และเสียงดนตรีที่เล่นในสถานที่เปิด

หลังจากการจัดเตรียมชุดข้อมูลเสียงสภาพแวดล้อม เราจะนำสัญญาณเสียงมาทำการสกัดคุณลักษณะและการจำแนกเสียง



รูปที่ 3.64 คุณลักษณะของเสียงสภาพแวดล้อม 5 ประเภท (ก) ชนิดของเสียง (ข) สัญญาณโดเมนทางเวลา และ (ค) สัญญาณโดเมนทางเวลา-ความถี่



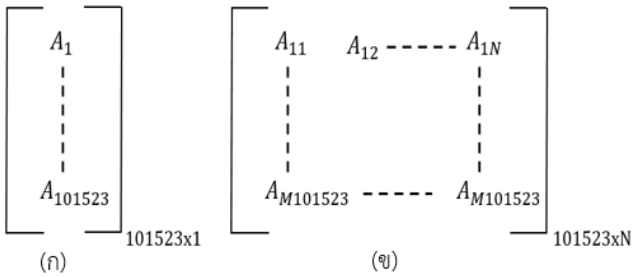
รูปที่ 3.65 ตัวอย่างของเสียงสภาพแวดล้อมจำนวน 1 เสียงที่มีขนาดความยาว 4 วินาที

3.1.2 การเลือกประเภทของเสียงที่จะนำมาใช้สำหรับการฝึกฝนและทดสอบ

จากงานวิจัยการเรียนรู้จำเสียงสภาพแวดล้อม เราได้เลือกเสียงที่นำมาใช้ในการฝึกฝนและทดสอบจำนวนทั้งหมด 5 ประเภท แสดงดังรูปที่ 3.64 จะเห็นได้ว่าแต่ละเสียงสภาพแวดล้อมมีขนาดเท่ากับ 4 วินาที [1วินาที เท่ากับ 44100 มิติ] เพราะฉะนั้นข้อมูลเสียงสภาพแวดล้อมจำนวน 1 เสียงมีขนาดเท่ากับ 176,400 มิติ แสดงดังรูปที่ 3.65

ตารางที่ 3. 4 การฝึกฝนและทดสอบของเสียงสภาพแวดล้อม

Class	Training Set	Testing set	Sounds
air conditioner	64	28	92
children playing	64	28	92
engine dling	64	28	92
siren	64	28	92
street music	64	28	92
sum	320	140	460



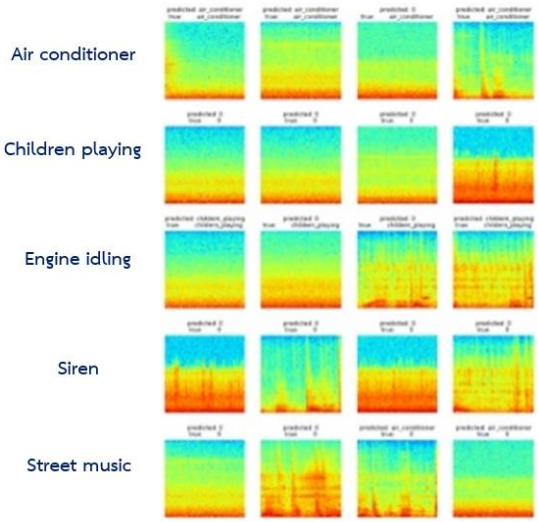
รูปที่ 3.66 เมทริกซ์ของเสียงสภาพแวดล้อม (ก) เมทริกซ์ขนาด 101523x1 (ข) เมทริกซ์ขนาด 101523xN

3.1.3 การแบ่งฐานข้อมูลเสียงเพื่อใช้ในการฝึกฝนและทดสอบ

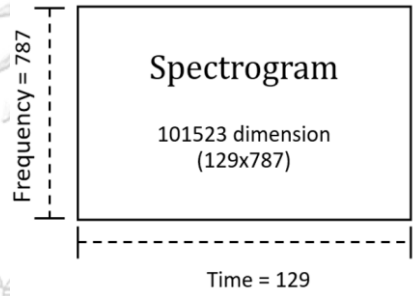
การแบ่งข้อมูลสำหรับการเรียนรู้จำเสียงสภาพแวดล้อม เราจะใช้หลักการสุ่มแบ่งข้อมูลออกเป็น 2 ชุด คือชุดที่หนึ่งใช้สำหรับการฝึกฝน ส่วนชุดที่สองใช้สำหรับการทดสอบ เพื่อให้เกิดการกระจายตัวของฐานข้อมูลเสียงทั้งในส่วนของการฝึกฝนและทดสอบ โดยทั่วไปจะทำการแบ่งอัตราส่วนข้อมูลชุดฝึกฝนต่อข้อมูลชุดทดสอบที่ 70:30 หรือ 80:20 ตามความต้องการ โดยในวิจัยนี้เราเลือกที่จะแบ่งข้อมูลของการฝึกฝนและการทดสอบเป็นอัตราส่วน 70:30 แสดงดังตารางที่ 3.4

3.1.4 การจัดเรียงฐานข้อมูลเสียงสภาพแวดล้อม

หลังจากที่เราสามารถแบ่งจำนวนเสียงสภาพแวดล้อมในการเรียนรู้จำเสียงสภาพแวดล้อมได้แล้ว ขั้นตอนนี้จะเป็นการบ่งบอกถึงการจัดเรียงข้อมูลก่อนเข้าสู่วิธีการจำแนกเสียง จากรูปที่ 3.66 (ก) จะเห็นได้ว่าจำนวนเสียง 1 สัญญาณข้อมูลจะถูกจัดเรียงให้อยู่ในลักษณะแถว (row) ส่วนรูปที่ 3.66 (ข) คือข้อมูลวิเคราะห์หลายสัญญาณการเรียงข้อมูลจะนำมาเรียงต่อกันเป็นคอลัมน์ (column)



รูปที่ 3.67 ตัวอย่างสเปกโตรแกรมของสัญญาณเสียงสภาพแวดล้อม



รูปที่ 3.68 ขนาดของสเปกโตรแกรมของเสียงสภาพแวดล้อม

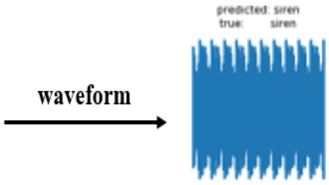
3.2 การสกัดคุณลักษณะ (Feature Extraction)

จากงานวิจัยการเรียนรู้จำเสียงสภาพแวดล้อมสิ่งที่ช่วยทำให้การวิเคราะห์ข้อมูลนั้นง่ายขึ้น คือการสกัดคุณลักษณะ จากงานวิจัยการสกัดคุณลักษณะแบ่งออกเป็น 2 วิธี วิธีแรกคือการแยกคุณลักษณะแต่ละประเภทเสียงด้วยวิธีการผลการแปลงฟูเรียร์ช่วงเวลาสั้น ส่วนวิธีที่สองคือการลดมิติของสัญญาณเสียงที่มีขนาดใหญ่มาก

3.2.1 การหาผลการแปลงฟูเรียร์ช่วงเวลาสั้น

จากงานวิจัยการสกัดคุณลักษณะด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้น เราใช้สำหรับการแยกคุณลักษณะแต่ละประเภทของเสียงสภาพแวดล้อม แสดงดังรูปที่ 3.67 นอกจากนี้ เรายังนำมาประยุกต์แปลงจากข้อมูลเสียงมาเป็นข้อมูลภาพใช้สำหรับเครื่องมือการจำแนกเครื่องข่ายประสาทลึก วิธีการทำของผลการแปลงฟูเรียร์ช่วงเวลาสั้น เราจะนำสัญญาณเสียงสภาพแวดล้อมมาแปลงจากสัญญาณโดเมนทางเวลามามาเป็นสัญญาณทางเวลา-ความถี่ ทำให้ได้คุณสมบัติสเปกโตรแกรมที่มีขนาด 101523 มิติ แสดงดังรูปที่ 3.68

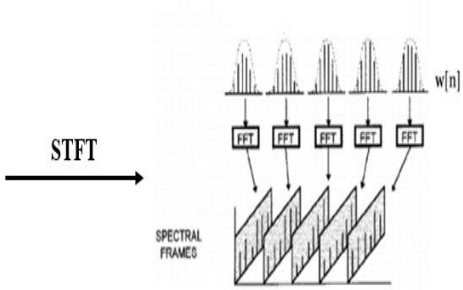
```
for file_name in raw_train:
    file_detail = file_name.split('-')
    types = int(file_detail[1])
    [data,rate] = sf.read(file_name)
    dim = data.shape
    all_sample = dim[0]
    duration = dim[0]/rate
```



program: waveform

รูปที่ 3.69 ตัวอย่างโปรแกรมแปลงเสียงสภาพแวดล้อม

```
    data_left = data
    f, t, Sxx = signal.spectrogram(data_left,rate)
    vector = (np.transpose(10*np.log10(np.abs(Sxx)))).flatten()
    train_bef_vec.append(Sxx)
    train_vec_list.append(vector)
    train_label_list.append(types)
X_train = np.array(train_vec_list)
y_train = np.array(train_label_list)
```



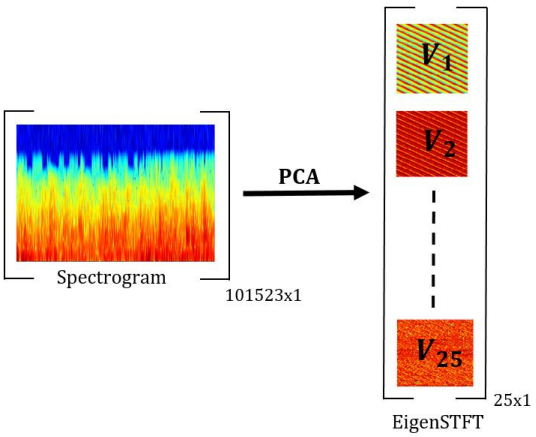
program: STFT

รูปที่ 3.70 ตัวอย่างโปรแกรมผลการแปลงฟูเรียร์ช่วงเวลาสั้นของเสียงสภาพแวดล้อม

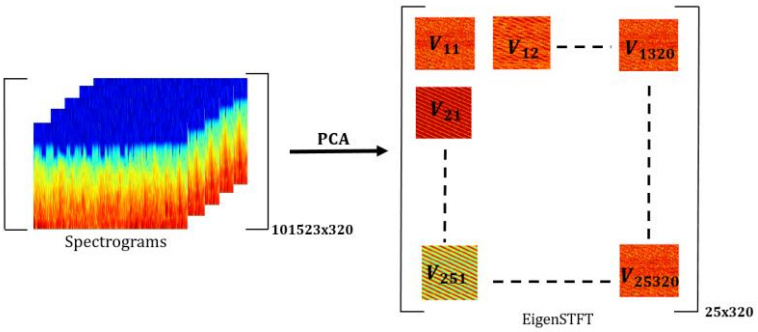
ตัวอย่างโปรแกรมการแปลงสัญญาณเสียงสภาพแวดล้อม จากไฟล์เสียง .wav ถูกแปลงเป็นสัญญาณ waveform หรือสัญญาณโดเมนทางเวลา แสดงดังรูปที่ 3.69 ด้วยคำสั่ง **[data, rate] = sf.read(file_name)** ก่อนจะเข้าสู่กระบวนการสกัดคุณลักษณะและเครื่องมือการจำแนกเสียงสภาพแวดล้อม

หลังจากที่เราแปลงไฟล์สัญญาณเสียง .wav ให้อยู่ในรูปแบบสัญญาณโดเมนทางเวลาได้แล้ว เราจะนำสัญญาณเข้าสู่กระบวนการสกัดคุณลักษณะของผลการแปลงฟูเรียร์ช่วงเวลาสั้น โดยคำสั่ง **f, t, Sxx = signal.spectrogram(data_left,rate)** แสดงดังรูปที่ 3.70 จะเห็นได้ว่าสัญญาณจะแปลงจากสัญญาณโดเมนทางมาเป็นสัญญาณโดเมนทางเวลา-ความถี่ ผลลัทธ์เอาท์พุตที่ได้คือ สเปกโตรแกรม ที่สามารถนำไปใช้กับเครื่องมือการจำแนกเสียงสภาพแวดล้อม

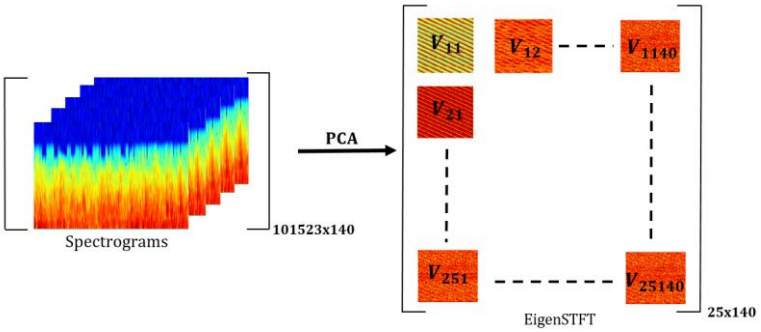
สรุปวิธีการเตรียมการข้อมูลคือเราต้องแปลงไฟล์สัญญาณเสียง .wav ให้อยู่ในรูปแบบสัญญาณโดเมนทางเวลา ทำให้เราสามารถนำไปใช้กับการสกัดคุณลักษณะและการจำแนกเสียงสภาพแวดล้อมได้



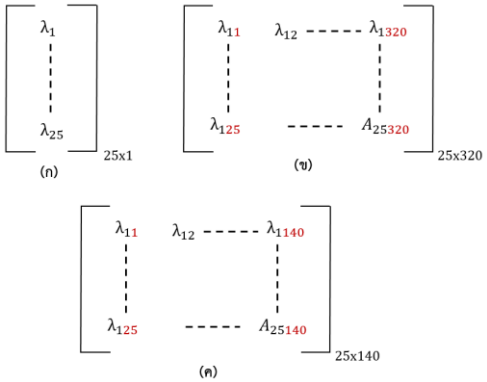
รูปที่ 3.71 ตัวอย่างการวิเคราะห์องค์ประกอบหลัก 1 สัญญาณเสียง ได้ค่าไอเกน 25 EigenSTFT



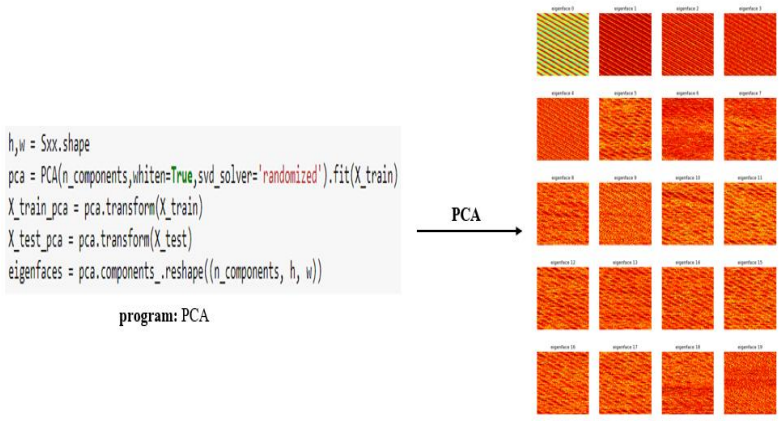
รูปที่ 3.72 การสร้างชุดฝึกฝนเสียงสภาพแวดล้อมจำนวน 320 เสียง สำหรับการลดมิติด้วยการแปลง PCA ที่จำนวน (25 EigenSTFT) x (320 Sounds)



รูปที่ 3.73 การสร้างชุดฝึกฝนเสียงสภาพแวดล้อมจำนวน 320 เสียง สำหรับการลดมิติด้วยการแปลง PCA ที่จำนวน (25 EigenSTFT) x (140 Sounds)



รูปที่ 3.74 ค่าไอเกนเวกเตอร์ของเสียงสภาพแวดล้อม (ก) ค่าไอเกนเวกเตอร์จำนวนเสียงสภาพแวดล้อม 1 เสียง (ข) ไอเกนเวกเตอร์ชุดฝึกฝน (ค) ไอเกนเวกเตอร์ชุดทดสอบ

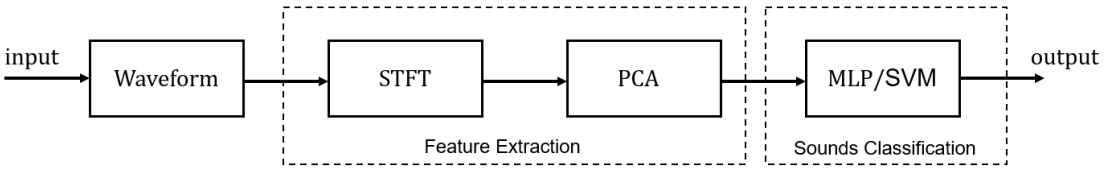


รูปที่ 3.75 ตัวอย่างโปรแกรมทำการวิเคราะห์องค์ประกอบหลัก

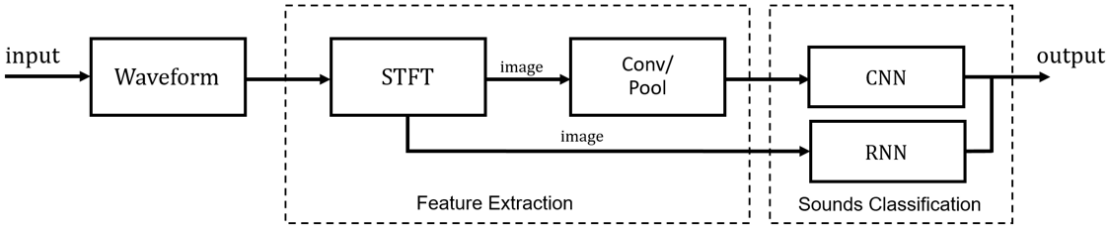
3.2.2 หลักการการวิเคราะห์องค์ประกอบหลัก

หลังจากเราแยกข้อมูลเสียงสภาพแวดล้อมด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้น ข้อมูลเสียงมีขนาดมิติใหญ่มาก ด้วยเหตุนี้เราจึงเสนอการวิเคราะห์องค์ประกอบหลัก วิธีการทำคือตัดตัวแปรที่ไม่สำคัญทิ้ง จากนั้นสร้างข้อมูลขึ้นมาใหม่ด้วยไอเกนผลการแปลงฟูเรียร์ แสดงดังรูปที่ 3.74 (ก) จะเห็นได้ว่าเราลดมิติจากสัญญาณ 1 เสียงที่มีจำนวน 101523 มิติ ให้เหลือเพียง 25 eigenSTFT ใช้สำหรับการเรียนรู้จำเสียงสภาพแวดล้อม และสำหรับค่าไอเกน STFT ที่ใช้ในการชุดฝึกฝนและชุดทดสอบของเครื่องมือการจำแนกเสียงสภาพแวดล้อม แสดงดังรูปที่ 3.72 (ข) และ 3.72 (ค)

จากรูปที่ 3.75 เป็นตัวอย่างการสร้าง eigenSTFT สำหรับการเรียนรู้จำเสียง ด้วยคำสั่ง **pca = PCA(n_components,whiten=True,svd_solver='randomize').fit(X_train)** และ **eigenfaces = pca.components_.reshape((n_components, h, w))** ทำให้ได้ค่าไอเกนผลการแปลงฟูเรียร์ แสดงดังรูปที่ 3.75



รูปที่ 3.76 บล็อกไดอะแกรมของเครื่องมือการจำแนกซัพพอร์ตเวกเตอร์แมชชีน และเพอร์เซ็ปตรอนหลายชั้นที่นำเสนอ



รูปที่ 3.77 บล็อกไดอะแกรมของเครื่องมือการจำแนกโครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อนที่นำเสนอ

3.3 ขั้นตอนการทำงานของเครื่องมือการจำแนก

เครื่องมือที่ใช้สำหรับการจำแนกเสียงสภาพแวดล้อมได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน เพอร์เซ็ปตรอนหลายชั้น โครงข่ายคอนโวลูชัน และโครงข่ายประสาทเกิดซ้อน จากการทดลองวิธีการการสกัดคุณลักษณะในการจำแนกจะแบ่งได้ 2 วิธี วิธีแรกคือการสกัดคุณลักษณะด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้นและการวิเคราะห์องค์ประกอบหลัก ใช้กับเครื่องมือการจำแนกของซัพพอร์ตเวกเตอร์แมชชีนและเพอร์เซ็ปตรอนหลายชั้น แสดงดังรูปที่ 3.76 ส่วนวิธีที่สองเราใช้การสกัดคุณลักษณะของผลการแปลงฟูเรียร์ช่วงเวลาสั้นในการแปลงจากข้อมูลเสียงมาเป็นข้อมูลภาพประยุกต์ใช้งานกับเครื่องการมือจำแนกของโครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อน แสดงดังรูปที่ 3.77

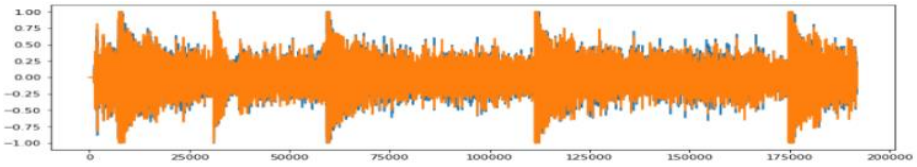
หลังจากที่เราสามารถจำแนกเสียงสภาพแวดล้อมทั้ง 5 คลาสได้ ซึ่งในงานวิจัยเราวัตถุประสงค์หลักคือการจำแนกเสียงสภาพแวดล้อมกับเสียงปืนใหญ่ โดยการทำ binary classification คือการแยกระหว่างเสียงปกติกับเสียงอันตราย

รูปที่ 3.78 รูปแบบคุณลักษณะของเสียง 7 ประเภท (ก) ชนิดของเสียง
(ข) สัญญาณโดเมนทางเวลา

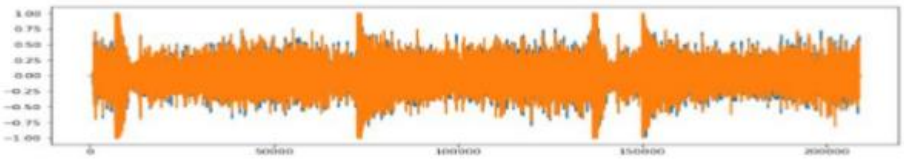
3.4 การจัดเตรียมข้อมูลเสียงปืนใหญ่

จุดประสงค์หลักของงานวิจัยคือต้องการแยกระหว่างเสียงอันตรายกับเสียงปกติ จากงานวิจัยก่อนหน้านี้เราจำแนกเสียงปกติคือเสียงสภาพแวดล้อมทั้งหมด 5 คลาส ได้แก่ เสียงเครื่องปรับอากาศ เสียงเด็กเล่น เสียงของรถยนต์ เสียงไซเรน และเสียงดนตรีที่เล่นสถานที่เปิด ด้วยเหตุนี้ เราจึงเพิ่มเสียงปืนกับเสียงปืนใหญ่เข้ามาในการทดลอง เพื่อเราต้องการทำ binary classification โดยกำหนดให้เสียงสภาพแวดล้อมเป็นคลาส 0 คือเสียงปกติ ส่วนเสียงปืนและปืนใหญ่เป็นคลาส 1 คือเสียงอันตราย

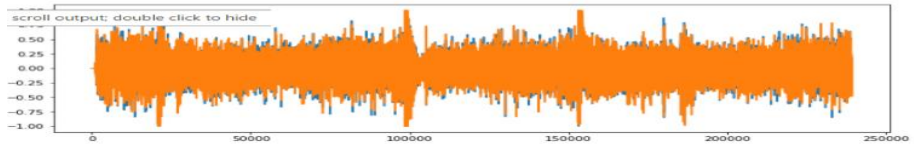
การจัดเตรียมเสียงปืนใหญ่และเสียงปืน สำหรับงานวิจัยเราในส่วนของเสียงปืนเราได้ใช้ข้อมูลของ Urbansound 8K แต่ก็ยังมีปัญหาเนื่องจากข้อมูลเสียงปืนความยาว 4 วินาทีที่เราต้องการนั้นมีเพียง 14 เสียง เราจึงนำเสียงปืนที่มีความยาว 1 วินาที 2 วินาที และ 3 วินาทีมาสร้างสัญญาณโดยการซุ่มให้ได้ครบ 78 เสียง ส่วนเสียงปืนใหญ่เราได้ใช้เสียงจากเกมปืนใหญ่ โดยเราตัดความยาวของเสียงปืนใหญ่เพียง 500000 มิติ จากนั้น เราก็สร้างด้วยวิธีการสุ่มมค่าให้ได้ 92 เสียง ใช้สำหรับการทำไบนารี



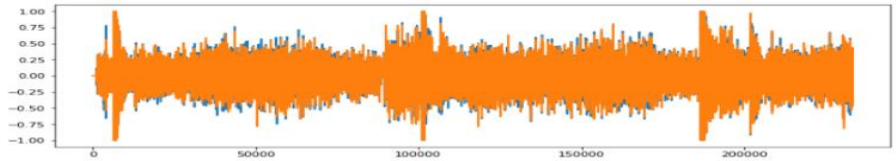
รูปที่ 3.79 สัญญาณเสียงปืนขนาด 1 วินาที 2 วินาที และ 2 วินาที



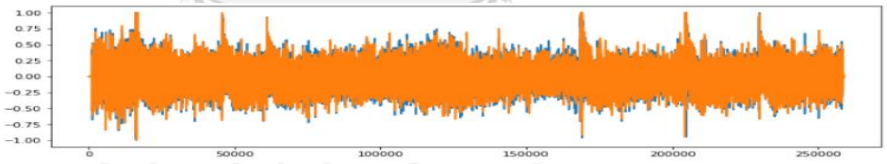
รูปที่ 3.80 สัญญาณเสียงปืนขนาด 1 วินาที 2 วินาที และ 2 วินาที



รูปที่ 3.81 สัญญาณเสียงปืนขนาด 2 วินาที 1 วินาที และ 2 วินาที

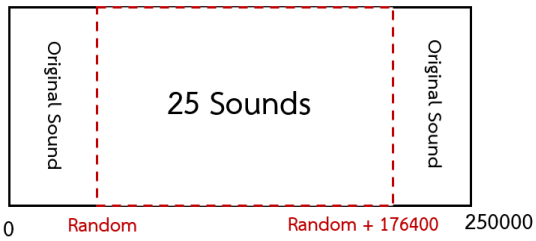


รูปที่ 3.82 สัญญาณเสียงปืนขนาด 2 วินาที 2 วินาที และ 1 วินาที

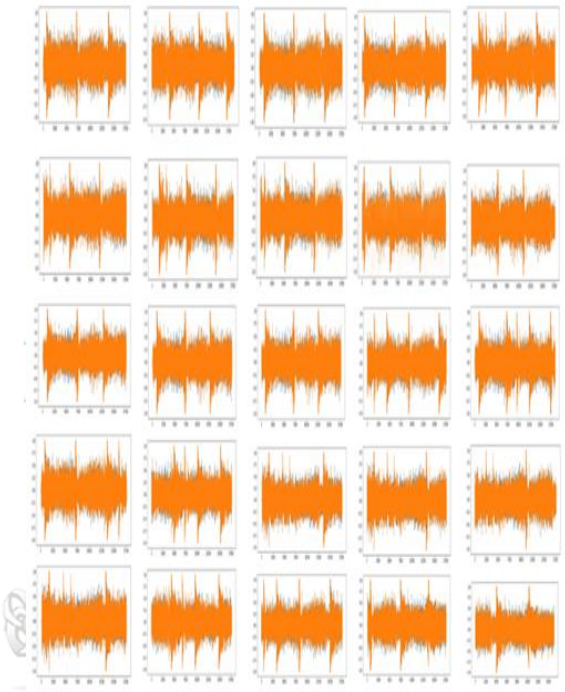


รูปที่ 3.83 สัญญาณเสียงปืนขนาด 3 วินาที และ 2 วินาที

จากงานวิจัยเราต้องการสร้างจำนวนเสียงปืนให้ได้จำนวน 78 เสียง โดยแต่ละเสียงจะมีความยาว 4 วินาที ด้วยเหตุนี้เราจะนำเสียงปืน 1 วินาที 2 วินาที และ 3 วินาที นำมาสร้างให้ได้ 5 สัญญาณที่ไม่ซ้ำกัน แสดงดังรูปที่ 3.79 ถึง 3.83 แต่ละสัญญาณจะมีความยาวขนาด 5 วินาที (เสียง 1 วินาที มีจำนวนมิติ 44,100 มิติ) จากงานวิจัยความยาวของเสียงที่ใช้ในการเรียนรู้จำเสียงมีขนาด 1 เสียงเท่ากับ 176400 มิติ ดังนั้น สัญญาณเสียงทั้งหมด 5 สัญญาณ โดยแต่ละสัญญาณมีขนาดข้อมูลเสียง 250000 มิติ ซึ่งเราสามารถนำ 1 สัญญาณเสียงมาสร้างโดยการสุ่มค่าทำ overlap สามารถสร้างจำนวนเสียงได้ 25 สัญญาณ



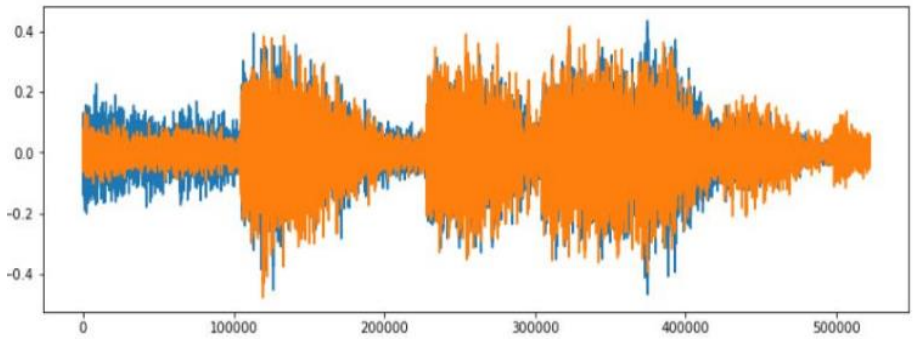
รูปที่ 3.84 การสร้างสัญญาณเสียงปืน



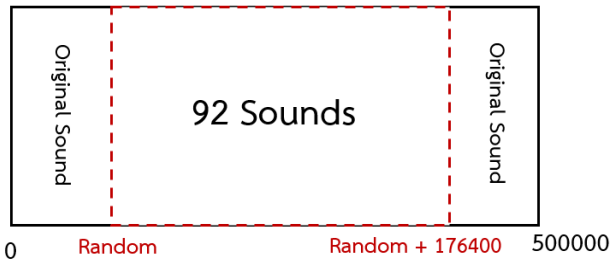
รูปที่ 3.85 ตัวอย่างการสุ่มเสียงปืน 25 สัญญาณ

ขั้นตอนการสร้างสัญญาณเสียงปืน แสดงดังรูปภาพที่ 3.84 เราจะนำสัญญาณเสียงขนาด 250000 มิติมาสร้างให้ได้ 25 สัญญาณ โดยการสุ่มค่ามาทั้งหมด 25 ค่า แล้วนำแต่ละค่าที่สุ่มมาได้ไปบวกด้วย 176400 มิติ ทำให้ได้สัญญาณเสียงที่ไม่ซ้ำกัน แสดงดังรูปที่ 3.85

หลังจากที่เราสามารถสร้างสัญญาณเสียงปืนรวมกันแล้วได้ 92 เสียง เราสามารถนำสัญญาณเข้าสู่กระบวนการสกัดคุณลักษณะด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้นในการแยกคุณลักษณะเสียงและนำไปประยุกต์จากข้อมูลเสียงมาเป็นข้อมูลภาพ เพื่อนำไปใช้งานกับเครื่องมือการจำแนกเสียงของ ซัพพอร์ตเวกเตอร์แมชีน เพอร์เซ็ปตรอนหลายชั้น โครงข่ายประสาทเกิดซ้อนและโครงข่ายคอนโวลูชัน ส่วนสุดท้ายคือการทำไบนารีแยกระหว่างเสียงปกติกับเสียงอันตราย



รูปที่ 3.86 ตัวอย่างเสียงปืนใหญ่ขนาด 500,000 มิติ ใช้ในการสร้างเสียงชุดฝึกฝนกับชุดทดลองจำนวน 92 เสียง



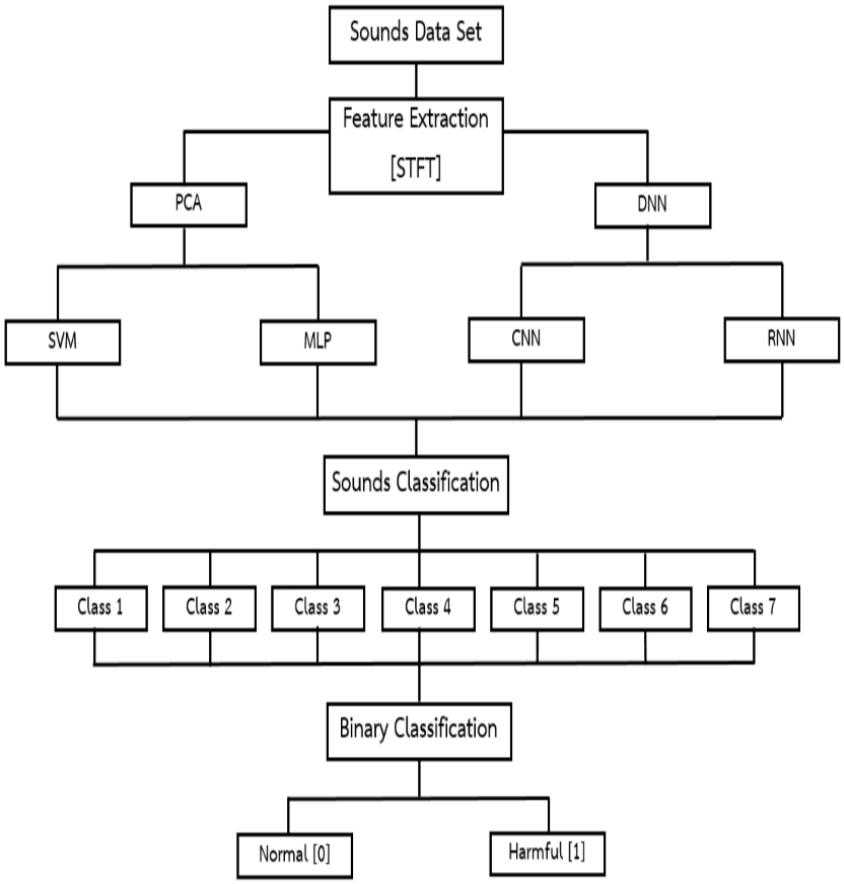
รูปที่ 3.87 ตัวอย่างการสร้างสัญญาณเสียงปืนใหญ่

การสร้างสัญญาณเสียงปืนใหญ่ เนื่องจากเรามีขนาดของเสียงปืนใหญ่เท่ากับ 500000 มิติ แสดงดังรูปที่ 3.86 ที่ใช้สำหรับการสร้างเสียงปืนใหญ่ให้ได้จำนวน 92 เสียง วิธีการเราจะสุ่มค่าให้ได้ทั้งหมด 92 ค่า แล้วนำแต่ละค่าที่สุ่มมาไปบวกด้วย 176400 มิติ แสดงดังรูปที่ 3.87 ทำให้ได้สัญญาณเสียงปืนใหญ่ที่ไม่ซ้ำกัน

หลังจากที่เราสามารถสร้างสัญญาณเสียงปืนใหญ่ได้ 92 เสียง เราสามารถนำสัญญาณเข้าสู่กระบวนการสกัดคุณลักษณะด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้นในการแยกคุณลักษณะเสียงและนำไปประยุกต์จากข้อมูลเสียงมาเป็นข้อมูลภาพ เพื่อนำไปใช้งานกับเครื่องมือการจำแนกเสียงของซัพพอร์ตเวกเตอร์แมชีน เพอร์เซ็ปตรอนหลายชั้น โครงข่ายประสาทเกิดซ้อนและโครงข่ายคอนโวลูชัน ส่วนขั้นตอนสุดท้ายคือการทำไบนารีแยกระหว่างเสียงปกติกับเสียงอันตราย

จากงานวิจัยหัวข้อนี้เราก็สามารถสร้างเสียงปืนใหญ่และเสียงปืน เพื่อนำไปใช้กับการเรียนรู้จำเสียงสภาพแวดล้อม โดยวิธีการทำเราจะทำการจำแนกเสียงสภาพแวดล้อมกับเสียงปืนใหญ่และปืน จากนั้น เราจะทำการแยกระหว่างเสียงปกติกับเสียงอันตราย

บทที่ 4
การทดสอบและผลลัพธ์

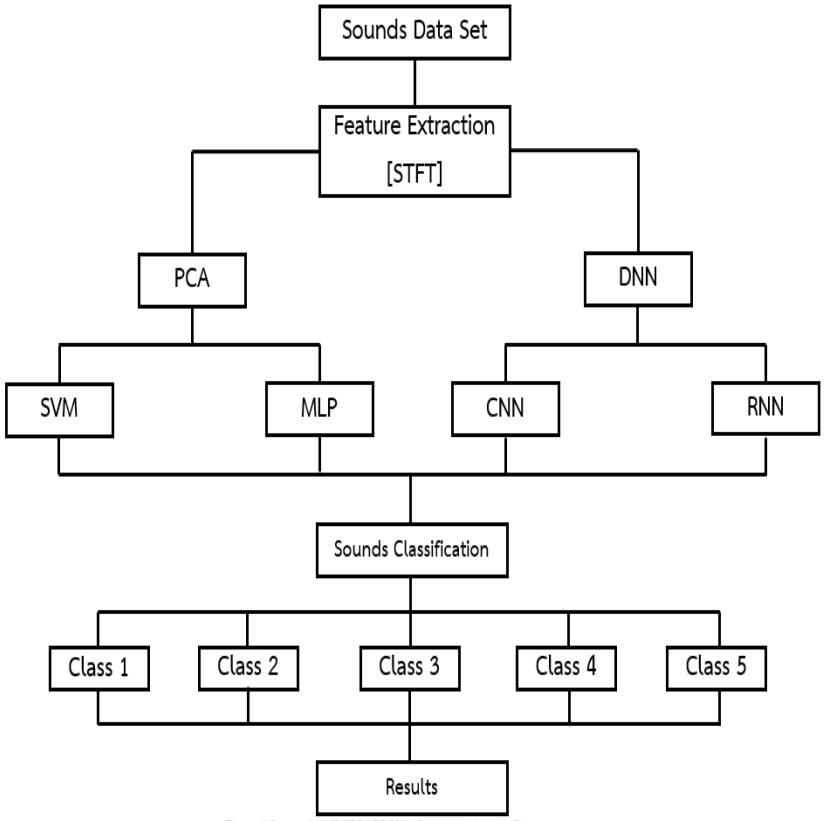


รูปที่ 4.88 โฟลว์ชาร์ตกระบวนการทดสอบสมรรถนะการจำแนกและไบนารี

4.1 ขั้นตอนการทดสอบ

จากการทดสอบการเปรียบเทียบสมรรถนะเครื่องมือการจำแนกเสียงสภาพแวดล้อมจะแบ่งออกเป็น 2 ส่วน โดยส่วนแรกคือการสกัดคุณลักษณะด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้นและการลดมิติด้วยการวิเคราะห์องค์ประกอบหลัก จะใช้กับเครื่องมือการจำแนกเพอร์เซ็ปตรอนหลายชั้นและซัพพอร์ตเวกเตอร์แมชชีน ส่วนวิธีที่สองใช้การสกัดคุณลักษณะของผลการแปลงฟูเรียร์ช่วงเวลาสั้นแปลงข้อมูลเสียงให้ได้ข้อมูลภาพ นำมาใช้กับเครื่องมือการจำแนกของโครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อน

หลังจากที่เราทำการทดสอบการจำแนกเสียงสภาพแวดล้อม ขั้นตอนต่อไปคือการจำแนกเสียงสภาพแวดล้อมระหว่างเสียงปืนกับปืนใหญ่ จากนั้น ทำไบนารีแยกระหว่างเสียงปกติกับเสียงอันตราย แสดงดังรูปที่ 4.88



รูปที่ 4.89 โฟลว์ชาร์ตกระบวนการทดสอบสมรรถนะการทำงานของการจำแนกเสียงสภาพแวดล้อม

4.2 การเปรียบเทียบสมรรถนะของเครื่องมือการจำแนกเสียงสภาพแวดล้อม

จากงานวิจัยการทดสอบสมรรถนะของการรู้จำเสียงสภาพแวดล้อมจะแบ่งวิธีการสกัดคุณลักษณะและการเครื่องมือการจำแนกออกเป็น 2 วิธี วิธีแรกคือการสกัดคุณลักษณะด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้นและการลดมิติเสียงที่มีขนาดใหญ่มากด้วยการวิเคราะห์องค์ประกอบหลัก สำหรับวิธีนี้ใช้กับเครื่องมือการจำแนกเพอร์เซ็ปตรอนหลายชั้นและซัพพอร์ตเวกเตอร์แมชชีน

ส่วนวิธีที่สองใช้การสกัดคุณลักษณะของผลการแปลงฟูเรียร์ช่วงเวลาสั้นในการแปลงจากข้อมูลเสียงให้ได้ข้อมูลภาพนำมาใช้กับเครื่องมือการจำแนกโครงข่ายประสาทเกิดซ้อนและโครงข่ายคอนโวลูชัน แสดงดังรูปที่ 4.89 แต่สำหรับวิธีนี้ทำไมไม่ใช้การวิเคราะห์องค์ประกอบหลักในการลดมิติ เพราะว่าโครงข่ายคอนโวลูชัน ภายในมีคอนโวลูชันเลเยอร์และพลูลิงเลเยอร์ทำหน้าที่ลดมิติอยู่แล้ว ส่วนโครงข่ายประสาทเกิดซ้อนข้อมูลอินพุตที่ลำดับเข้าไปรู้จำก็มีขนาดมิติน้อยมาก

ตารางที่ 4.5 เมทริกซ์ความสับสนของการจำแนกเสียงสภาพแวดล้อมด้วย SVM

	air conditioner	children playing	engine idling	siren	street music
air conditioner	28	0	0	0	0
children playing	0	26	0	0	2
engine idling	0	0	23	0	5
siren	0	1	0	27	0
street music	1	4	2	0	21

ตารางที่ 4.6 ประสิทธิภาพการจำแนกเสียงสภาพแวดล้อมของ SVM

	Prediction	Recall	F1-score	Support
Air conditioner	0.97	1.00	0.98	28
children playing	0.81	0.93	0.87	28
engine dling	0.92	0.82	0.81	28
siren	1.00	0.96	0.98	28
street music	0.74	0.71	0.73	28
average	0.89	0.89	0.89	140

4.2.1 การเปรียบเทียบสมรรถนะของเครื่องมือการจำแนกของเสียงสภาพแวดล้อมด้วยซัพพอร์ตเวกเตอร์แมชชีนและเพอร์เซ็ปตรอนหลายชั้น

จากงานวิจัยส่วนนี้คือการเปรียบเทียบสมรรถนะของเครื่องมือการจำแนกของเสียงสภาพแวดล้อมของซัพพอร์ตเวกเตอร์แมชชีนและเพอร์เซ็ปตรอนหลายชั้น จากการทดลองเราได้จัดเตรียมเสียงสภาพแวดล้อมทั้งหมด 5 คลาส ได้แก่ air conditioner children playing engine idling siren และ street music จากนั้น เราได้ทำการแบ่งข้อมูลสำหรับชุดฝึกฝนทั้งหมด 320 เสียง และชุดทดสอบ 140 เสียง ส่วนวิธีการสกัดคุณลักษณะทางเราได้ใช้วิธีของผลการแปลงฟูเรียร์ช่วงเวลาสั้นและการวิเคราะห์องค์ประกอบหลัก ในส่วนของการวิเคราะห์องค์ประกอบหลักในการใช้ลดมิติข้อมูลเสียงสภาพแวดล้อม เครื่องมือการจำแนกระหว่างซัพพอร์ตเวกเตอร์แมชชีนและเพอร์เซ็ปตรอนหลายชั้นเลือกค่า eigenSTFT ที่ดีที่สุดคือจำนวน 25 components

จากการทดสอบสมรรถนะของเครื่องมือการจำแนกของเสียงสภาพแวดล้อมด้วยซัพพอร์ตเวกเตอร์แมชชีน จากตารางที่ 4.5 จะเห็นได้ว่าเครื่องมือการจำแนกสามารถทำนายได้ถูกต้องหมดคือเสียงของ air conditioner ส่วนในการจำแนกเสียงสภาพแวดล้อมที่แย่สุดคือ street music ที่มีการทำนายผิดพลาดเป็นเสียง children playing ถึง 4 เสียงด้วยกัน และประสิทธิภาพของการทำนายทั้งหมดคือ 89% [16] แสดงดังตารางที่ 4.6

ตารางที่ 4.7 เมทริกซ์ความสับสนของการจำแนกเสียงสภาพแวดล้อมด้วย MLP

	air conditioner	children playing	engine idling	siren	street music
air conditioner	27	0	1	0	0
children playing	0	27	0	0	1
engine idling	2	0	21	0	5
siren	0	2	0	24	2
street music	1	7	0	3	17

ตารางที่ 4.8 ประสิทธิภาพการจำแนกเสียงสภาพแวดล้อมของ MLP

	Prediction	Recall	F1-score	Support
air conditioner	0.90	0.96	0.93	28
children playing	0.75	0.96	0.84	28
engine idling	0.95	0.75	0.84	28
siren	0.89	0.86	0.87	28
street music	0.68	0.61	0.64	28
average	0.83	0.83	0.83	140

หลังจากที่ได้ทำการทดสอบสมรรถนะของเครื่องมือการจำแนกซัพพอร์ตเวกเตอร์แมชชีนเป็นที่เรียบร้อย ต่อมาเป็นการเปรียบเทียบสมรรถนะของเครื่องมือการจำแนกเพอร์เซ็ปตรอนหลายชั้น ซึ่งในการทดลองทั้งสองเครื่องมือการจำแนกเราได้ใช้ข้อมูลเสียงชุดฝึกฝนและชุดทดสอบเดียวกัน และส่วนวิธีการสกัดคุณลักษณะใช้ผลการแปลงฟูเรียร์ช่วงเวลาสั้นและการวิเคราะห์องค์ประกอบหลักจำนวน 25 components เท่ากัน

จากการทดสอบสมรรถนะของเครื่องมือการจำแนกของเสียงสภาพแวดล้อมด้วยเพอรีเซ็ปตรอนหลายชั้น จากตารางที่ 4.7 จะเห็นได้ว่าเครื่องมือการจำแนกสามารถทำนายเสียงได้แม่นยำสุดคือเสียงของ air conditioner ส่วนในการจำแนกเสียงสภาพแวดล้อมที่แย่สุดคือ street music ที่มีการทำนายผิดพลาดเป็นเสียง children playing ถึง 7 เสียงด้วยกัน และประสิทธิภาพของการทำนายทั้งหมดคือ 83% [15] แสดงดังตารางที่ 4.8

สรุปผลการทดลองการเปรียบเทียบสมรรถนะระหว่างการจำแนกเสียงสภาพแวดล้อมของซัพพอร์ตเวกเตอร์แมชชีนและเพอร์เซ็ปตรอนหลายชั้น จากการทดลองเครื่องมือการจำแนกของซัพพอร์ตเวกเตอร์แมชชีนสามารถทำนายได้แม่นยำกว่า แต่จากการทดลองเราก็ได้เห็นว่าทั้งสองวิธีสามารถทำนายเสียงได้เหมือนกัน เช่น ทำนายเสียง air conditioner คือเสียงที่ทำนายแม่นยำสุด และเสียงที่ทำนายแย่สุดคือ street music เหมือนกัน

ตารางที่ 4.9 ประสิทธิภาพการจำแนกเสียงสภาพแวดล้อมของเครื่องมือการจำแนก LeNet-5

	Air conditioner	children playing	engine idling	siren	street music	Recall
air conditioner	25	0	0	0	3	89.28%
children playing	0	24	1	0	3	85.71%
engine idling	2	0	23	0	3	82.14%
siren	0	2	0	22	4	78.57%
street music	1	1	4	0	22	78.57%
average						82.85%

รูปที่ 4.90 แบบจำลองของ LeNet-5

4.2.2 การเปรียบเทียบสมรรถนะของแบบจำลอง LeNet-5 และ Original CNN

จากกรณีที่ศึกษาค้นคว้าทำให้เรารู้ว่าสัญญาณเสียงสามารถนำไปประยุกต์ใช้กับเครื่องมือการจำแนกข้อมูลภาพได้ โดยใช้วิธีผลการแปลงฟูเรียร์แปลงจากสัญญาณโดเมนทางเวลา 1 มิติ มาเป็นสัญญาณโดเมนทางเวลา-ความถี่ ได้ขนาดเป็น 2 มิติ ทำให้ได้ข้อมูลภาพเสียงสภาพแวดล้อม ด้วยเหตุนี้ จากงานวิจัยเราได้เสนอแบบจำลองของ original CNN และ LeNet-5 ใช้สำหรับการเรียนรู้จำเสียงสภาพแวดล้อม

จากการทดลองแบบจำลอง LeNet-5 [12] ในการจำแนกเสียงสภาพแวดล้อม จากรูปที่ 4.90จะเห็นได้ว่าแบบจำลองของ LeNet-5 วิธีการจำแนกได้ใช้แบบจำลองโครงข่ายคอนโวลูชันที่มีขนาดคอนโวลูชันเลเยอร์ 2 เลเยอร์ และมีขนาดตัวกรองเท่ากับ 5x5 ส่วนการทำพลูลิงมี 2 เลเยอร์ และชั้นสุดท้ายคือชั้นการเชื่อมโยงเต็มรูปแบบมี 4 เลเยอร์ จากการทดลองสมรรถนะของแบบจำลองของ LeNet-5 ทำได้ดีที่สุดคือ 82.85% [17] แสดงดังตารางที่ 4.9

ตารางที่ 4.10 ประสิทธิภาพเสียงสภาพแวดล้อมของเครื่องมือการจำแนก original CNN

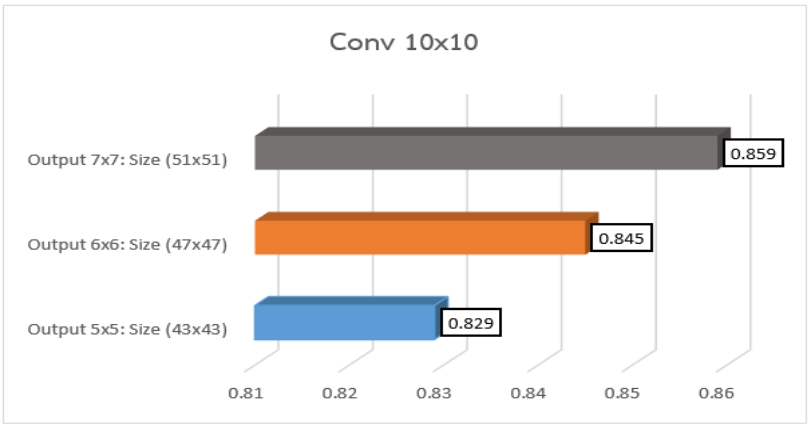
	Air conditioner	children playing	engine idling	siren	street music	Recall
air conditioner	21	4	0	3	3	75.00%
children playing	3	23	0	2	0	82.14%
engine idling	5	4	17	1	0	60.71%
siren	2	3	0	23	4	82.14%
street music	3	7	3	1	14	50.00%
average						69.99%

รูปที่ 4.91 แบบจำลองของ original CNN

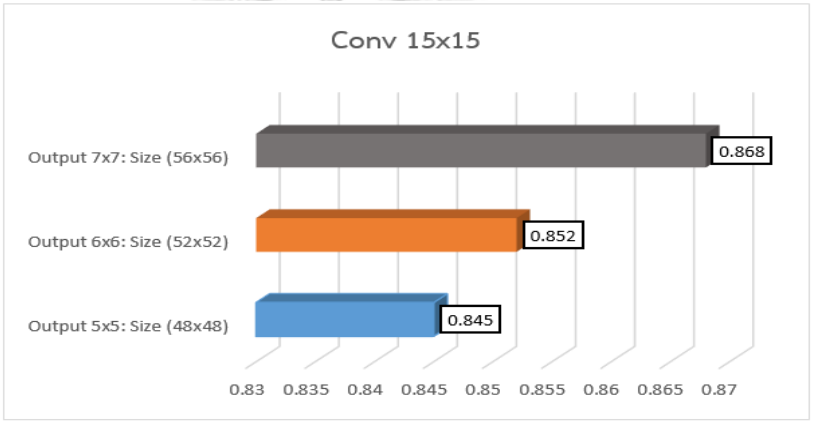
จากงานวิจัยเราได้ใช้แบบจำลองของ original CNN นำมาประยุกต์ใช้กับการเรียนรู้จำเสียงสภาพแวดล้อม สำหรับวิธีการจำแนกได้ใช้เครื่องมือการจำแนกโครงข่ายคอนโวลูชันที่มีเพียง 1 คอนโวลูชันเลเยอร์ และใช้ขนาดตัวกรองเท่ากับ 30x30 ส่วนการทำพลูลิงก็มีเพียง 1 เลเยอร์ ส่วนชั้นสุดท้ายคือชั้นการเชื่อมโยงเต็มรูปแบบมี 2 เลเยอร์ แสดงดังรูปที่ 4.91 จากการทดลองการจำแนกเสียงสภาพแวดล้อมความแม่นยำที่ทำได้ดีที่สุดคือ 69.99% แสดงดังตารางที่ 4.10

จากการทดลองจะเห็นได้ว่าทั้งสองแบบจำลองใช้โครงข่ายคอนโวลูชันในการจำแนกเหมือนกัน แต่สิ่งที่แตกต่างกันก็คือจำนวนของ คอนโวลูชันเลเยอร์ พลูลิงเลเยอร์ และการเชื่อมโยงเต็มรูปแบบ ซึ่งแบบจำลองของ LeNet-5 นั้นมีจำนวนทำแต่ละเลเยอร์มากกว่า original CNN ด้วยเหตุนี้การทดลองการเปรียบเทียบสมรรถนะของแบบจำลอง LeNet-5 และ original CNN จากการทดลองสรุปได้ว่าแบบจำลองของ LeNet-5 ให้ความแม่นยำในการทำนายเสียงสภาพแวดล้อมมากกว่าตัว original CNN

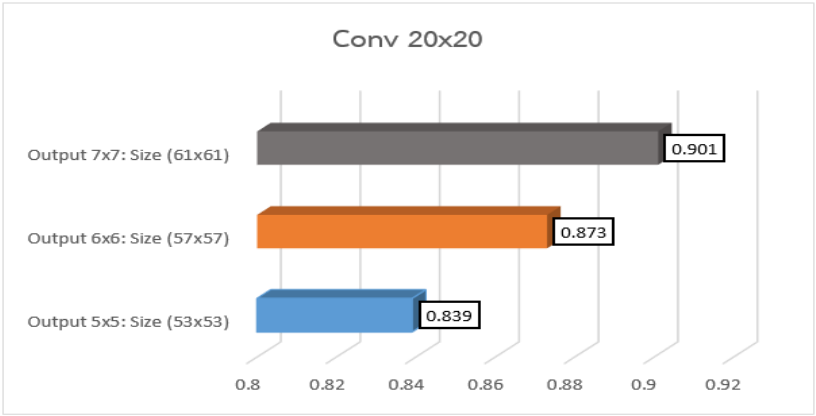
จากการทดลองทดสอบสมรรถนะของเครื่องมือการจำแนกที่จะนำมาใช้สำหรับการเรียนรู้จำเสียงสภาพแวดล้อมคือแบบจำลองของ LeNet-5



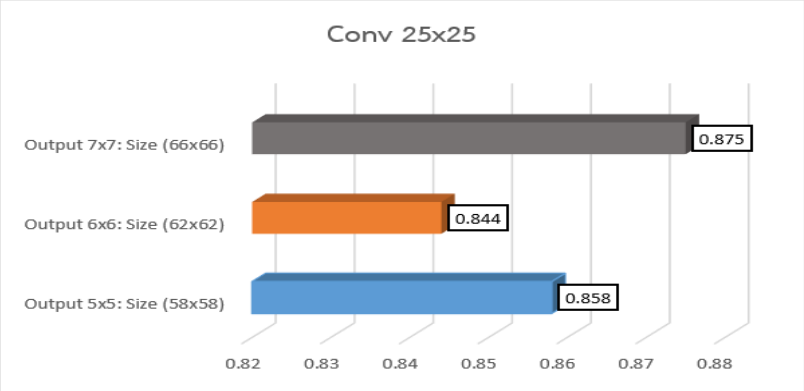
รูปที่ 4.92 การเปรียบเทียบประสิทธิภาพของโครงข่ายคอนโวลูชันขนาดเอาท์พุตเท่ากับ 5x5 size 43x43 6x6 size 47x47 และ 7x7 size 51x51



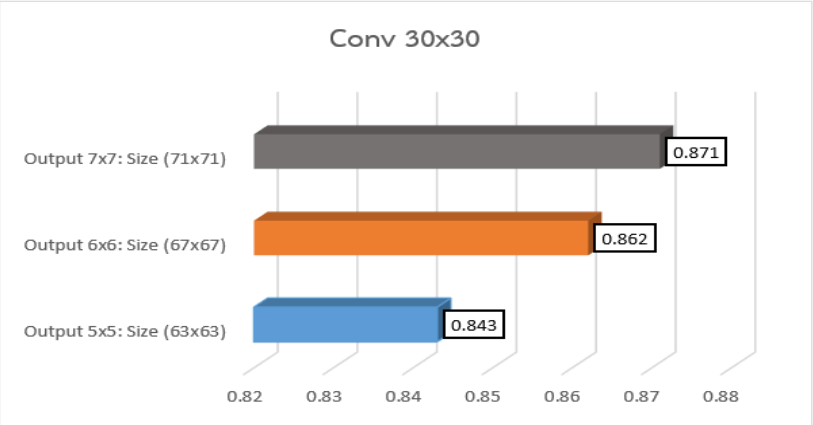
รูปที่ 4.93 การเปรียบเทียบประสิทธิภาพของโครงข่ายคอนโวลูชันขนาดเอาท์พุตเท่ากับ 5x5 size 48x48 6x6 size 52x52 และ 7x7 size 56x56



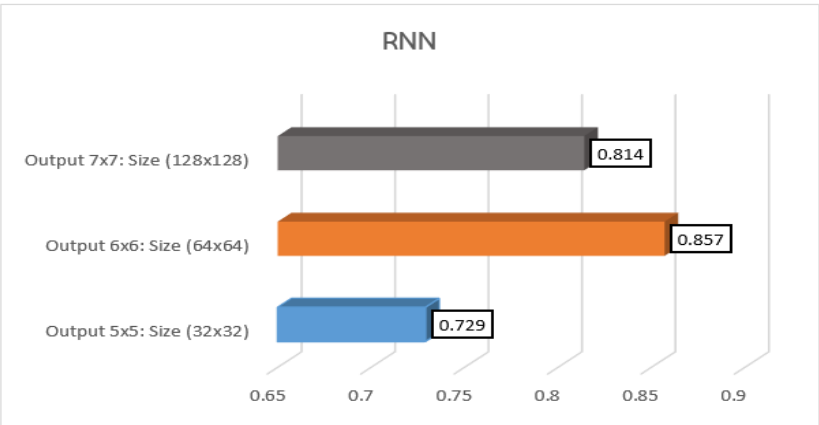
รูปที่ 4.94 การเปรียบเทียบประสิทธิภาพของโครงข่ายคอนโวลูชันขนาดเอาท์พุตเท่ากับ 5x5 size 57x57 6x6 size 61x61 และ 7x7 size 65x65



รูปที่ 4.95 การเปรียบเทียบประสิทธิภาพของโครงข่ายคอนโวลูชันขนาดเอาท์พุตเท่ากับ 5x5 size 58x58 6x6 size 62x62 และ 7x7 size 66x66



รูปที่ 4.96 การเปรียบเทียบประสิทธิภาพของโครงข่ายคอนโวลูชันขนาดเอาท์พุตเท่ากับ 5x5 size 63x63 6x6 size 67x67 และ 7x7 size 71x71



รูปที่ 4.97 การเปรียบเทียบประสิทธิภาพของโครงข่ายประสาทเกิดซ้อนขนาดภาพ 32x32 64x64 และ 128x128

ตารางที่ 4.11 เมทริกซ์ความสับสนของการจำแนกเสียงสภาพแวดล้อมด้วยโครงข่ายคอนโวลูชัน

	air conditioner	children playing	engine idling	siren	street music
air conditioner	140	0	1	0	0
children playing	0	124	3	3	8
engine idling	0	0	123	6	11
siren	0	6	5	122	7
street music	0	16	2	0	122

ตางรางที่ 4.12 ประสิทธิภาพการจำแนกเสียงสภาพแวดล้อมของโครงข่ายคอนโวลูชัน

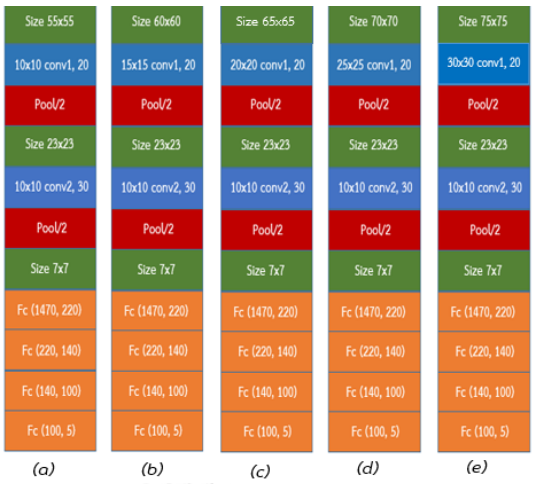
	Prediction	Recall	F1-score	Support
air conditioner	0.92	1.00	0.93	140
children playing	0.89	0.88	0.88	140
engine idling	0.89	0.88	0.87	140
siren	0.88	0.87	0.87	140
street music	0.89	0.87	0.92	140
average	0.90	0.90	0.90	700

4.2.3 การเปรียบเทียบสมรรถนะของเครื่องมือการจำแนกเสียงสภาพแวดล้อมด้วยโครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อน

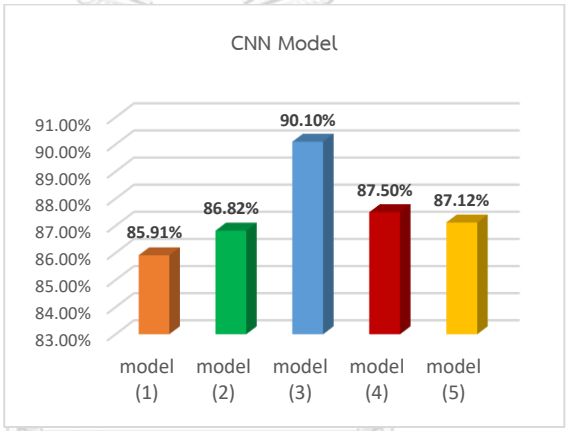
จากงานวิจัยส่วนนี้คือการเปรียบเทียบสมรรถนะของเครื่องมือการจำแนกเสียงสภาพแวดล้อมของโครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อน จากการทดลองเราได้จัดเตรียมเสียงสภาพแวดล้อมทั้งหมด 5 คลาส ได้แก่ air conditioner children playing engine idling siren และ street music จากนั้น เราได้ทำการแบ่งข้อมูลสำหรับชุดฝึกฝนทั้งหมด 320 เสียง และชุดทดสอบ 140 เสียง ส่วนวิธีการสกัดคุณลักษณะทั้งสองเครื่องมือการจำแนกใช้ผลการแปลงฟูเรียร์ช่วงเวลาสั้น ในการแปลงข้อมูลเสียงมาเป็นข้อมูลภาพ

จากผลการทดลองของโครงข่ายคอนโวลูชันจากรูป (4.92 4.93 4.94 4.95 และ 4.96) จะเห็นได้ว่าการทำคอนโวลูชันเลเยอร์ขนาด 20x20 ได้เอาท์พุตเท่ากับ 7x7 ขนาดภาพที่ได้ 61x61 ให้ความแม่นยำสูงสุดคือ 90% [16] แสดงดังตารางที่ 4.12 ส่วนเครื่องมือการจำแนกของโครงข่ายประสาทเกิดซ้อน เราได้ทำการทดลองสร้างแต่ละขนาดภาพ ได้แก่ 32x32 64x64 และ 128x128 จากการทดลองขนาดอินพุต 64x64 ให้ความแม่นยำสูงสุดถึง 85.7% [18] แสดงดังรูปที่ 4.97

สรุปจาการทดลองโครงข่ายคอนโวลูชันสามารถจำแนกได้ดีกว่าโครงข่ายประสาทเกิดซ้อนมากถึง 0.05%



รูปที่ 4.98 แบบจำลองโครงข่ายประสาทคอนโวลูชัน



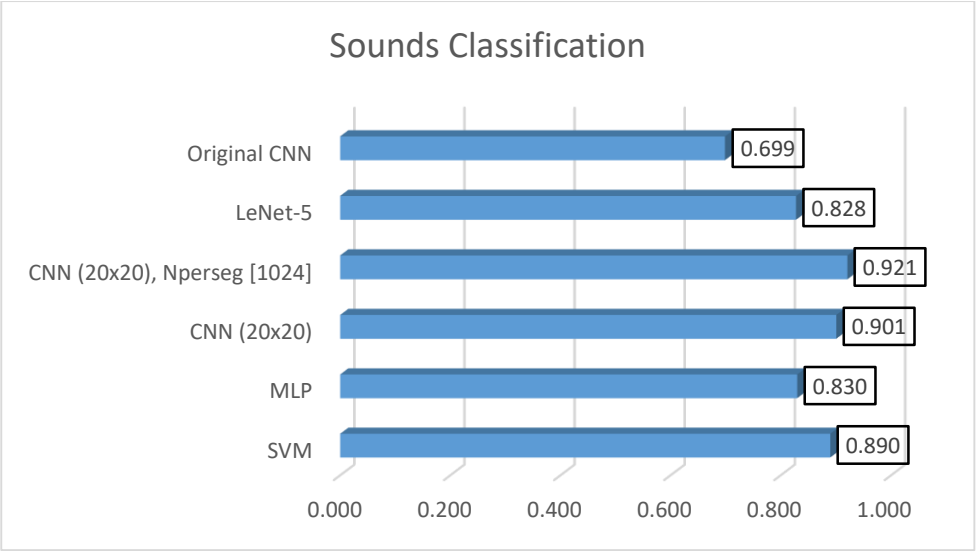
รูปที่ 4.99 การเปรียบเทียบประสิทธิภาพแบบจำลองของโครงข่ายประสาทคอนโวลูชัน

4.2.4 สรุปผลการทดลอง

จากการทดลองของแบบจำลอง CNN ทั้ง 5 รูปแบบที่มีขนาดภาพอินพุตและคอนโวลูชันที่แตกต่างกัน แสดงดังรูปที่ 4.98 และจากรูปที่ 4.99 ผลการทดลองที่สามารถทำนายเสียงสภาพแวดล้อมได้แม่นยำสุดคือ model (3) ที่คอนโวลูชันเท่ากับ 20x20

จากผลการทดลองจะเห็นได้ว่าการปรับคอนโวลูชัน model (1) ขนาด 10x10 และขนาด 15x15 ของ model (2) ซึ่งในการทดลองมันเกิด under-fitting คือแบบจำลองมีคุณลักษณะหรือมิติน้อยเกินไป ทำให้ความแม่นยำเริ่มลดลงหรือน้อยกว่าคอนโวลูชัน model (3) ขนาด 20x20 และจากการทดลองเพิ่มขนาดของคอนโวลูชันมาเป็น 25x25 และ 30x30 มันทำให้เกิด over-fitting คือแบบจำลองมีคุณลักษณะหรือมิติมากเกินไป ทำให้ความแม่นยำเริ่มลดลง แสดงดังรูปที่ 4.99

จากผลการทดลองของ CNN จะคล้ายกับวิธี SVM และ MLP ที่ 25 PCA มีความแม่นยำสูงสุด หลังจากที่เรามีการ ลด-เพิ่ม ขนาดของ PCA ความแม่นยำในการทำนายเสียงก็เริ่มลดลง



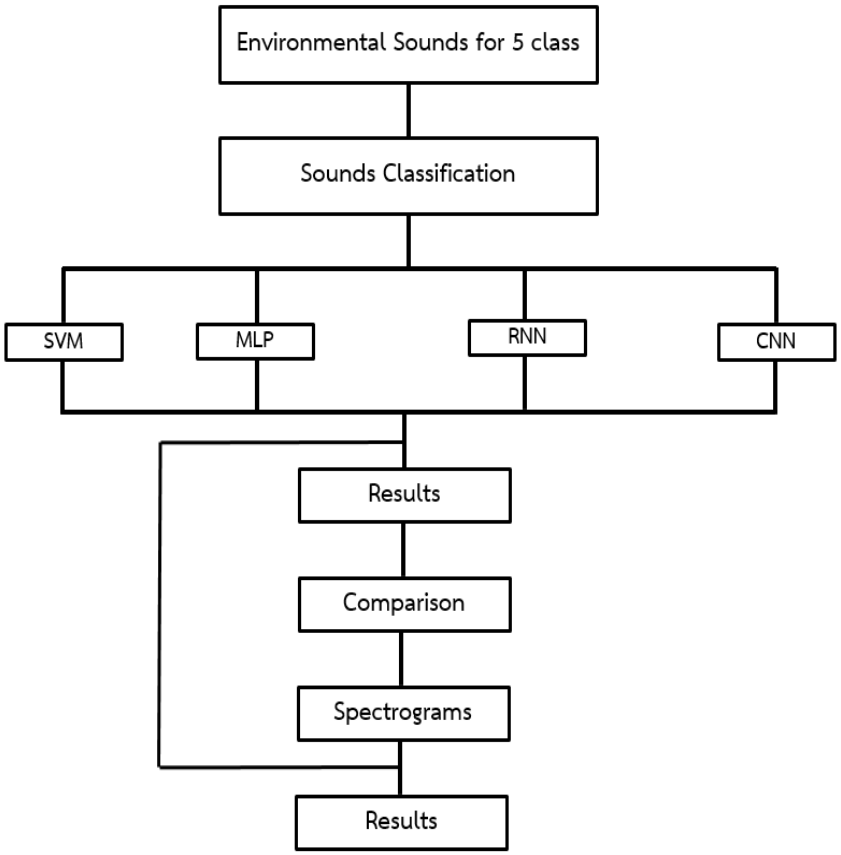
รูปที่ 4.100 การเปรียบเทียบสมรรถนะของเครื่องมือการจำแนกเสียงสภาพแวดล้อม

จากงานวิจัยการเรียนรู้จำเสียงสภาพแวดล้อม เราได้ทำการเปรียบเทียบสมรรถนะของเครื่องมือการจำแนกแบ่งออกเป็น 2 วิธี วิธีแรกเครื่องมือการจำแนกของซัพพอร์ตเวกเตอร์แมชชีนและเพอร์เซ็ปตรอนหลายชั้น ส่วนวิธีที่สองคือโครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อน

จากการทดสอบเครื่องมือการจำแนกของซัพพอร์ตเวกเตอร์แมชชีนและเพอร์เซ็ปตรอนหลายชั้น วิธีการทำงานของทั้งสองเครื่องมือการจำแนกใช้วิธีการสกัดคุณลักษณะด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้น และส่วนการลดมิติใช้การวิเคราะห์องค์ประกอบหลัก โดยการทำการวิเคราะห์องค์ประกอบหลักค่า eigenSTFT ที่ลดมิติได้ดีที่สุดคือ 25 components จากการทดลองเครื่องมือการจำแนกด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนสามารถทำนายได้แม่นยำกว่าเพอร์เซ็ปตรอนหลายชั้น

ส่วนการทดสอบเครื่องมือการจำแนกของโครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อน วิธีการทำงานเราจะต้องประยุกต์จากข้อมูลเสียงแปลงมาเป็นข้อมูลภาพด้วยวิธีของผลการแปลงฟูเรียร์ช่วงเวลาสั้น จากการทดลองสรุปได้ว่าเครื่องมือการจำแนกที่ดีที่สุดคือโครงข่ายคอนโวลูชันสามารถทำนายได้แม่นยำถึง 90.10%

จากการทดสอบสมรรถนะของเครื่องมือการจำแนก ซัพพอร์ตเวกเตอร์แมชชีน เพอร์เซ็ปตรอนหลายชั้น โครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อน เครื่องมือที่จำแนกได้แม่นยำสุดคือโครงข่ายคอนโวลูชัน แสดงดังรูปที่ 4.100 ด้วยเหตุนี้ เราจึงเสนอวิธีของโครงข่ายคอนโวลูชันนำไปประยุกต์ใช้ในงานเรียนรู้จำเสียงที่ปกติและเสียงที่อันตราย

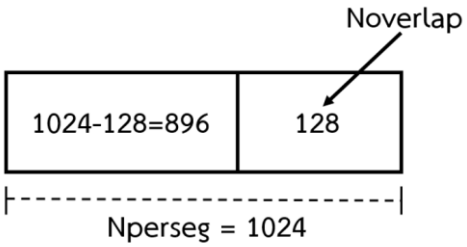


รูปที่ 4.101 โฟลว์ชาร์ตกระบวนการทดสอบสมรรถนะการทำงาน
ของการจำแนกเสียงสภาพแวดล้อม

4.3 ขนาดหน้าต่างของผลการแปลงฟูเรียร์ต่อสมรรถนะของการจำแนกเสียงสภาพแวดล้อม

จากการทดสอบสมรรถนะของการเรียนรู้จำเสียงสภาพแวดล้อมจากหัวข้อที่ 4.2 ผลการทดลองสรุปได้ว่าเครื่องมือที่สามารถทำนายได้แม่นยำสุดคือโครงข่ายคอนโวลูชันสามารถทำนายได้แม่นยำถึง 90.10% แต่จากการทดสอบสมรรถนะของการรู้จำเสียงสภาพแวดล้อมในครั้งนั้นเราไม่ได้มีการปรับขนาดฟังก์ชันหน้าต่างของผลการแปลงฟูเรียร์ ทำให้ได้ขนาดสเปกโตรแกรมเท่ากับ 101523 มิติ

จากหัวข้อนี้เราจะทำการทดสอบเปรียบเทียบระหว่างการปรับขนาดฟังก์ชันหน้าต่างของผลการแปลงฟูเรียร์และการที่ไม่ได้ปรับฟังก์ชันหน้าต่างนั้น จะส่งผลทำให้การทำนายเสียงสภาพแวดล้อมแม่นยำขึ้นหรือไม่ โดยขั้นตอนการทดสอบสมรรถนะของการปรับฟังก์ชันหน้าต่างแสดงดังรูปที่ 4.101



รูปที่ 4.102 ตัวอย่างการปรับขนาดของฟังก์ชันหน้าต่าง โดยใช้คำสั่ง Nperseg เท่ากับ 1024 และมีขนาด Noverlap เท่ากับ 128

4.3.1 การปรับขนาดฟังก์ชันหน้าต่างของผลการแปลงฟูเรียร์ช่วงเวลาสั้น

จากการทดลองปรับขนาดฟังก์ชันหน้าต่างของผลการแปงฟูเรียร์ ส่งผลทำให้ความละเอียดทางด้านเวลาและความถี่เปลี่ยนไป จึงทำให้ขนาดของสเปกโตรแกรมเปลี่ยนตามค่าฟังก์ชันหน้าต่างที่เรากำหนด ด้วยเหตุนี้ การทดลองปรับขนาดฟังก์ชันหน้าต่าง อาจทำให้ความแม่นของเครื่องมือการจำแนกเสียงเพิ่มขึ้น

การทดลองปรับขนาดฟังก์ชันหน้าของผลการแปลงฟูเรียร์ช่วงเวลาสั้น แสดงดังรูปที่ 4.102 จะเห็นได้ว่าเราใช้คำสั่ง nperseg เท่ากับ 1024 เป็นตัวกำหนดขนาดฟังก์ชันหน้าต่างเท่ากับ 896 และมีขนาด noverlap เท่ากับ 128 ทำให้ได้ขนาดของสเปกโตรแกรมเท่ากับ 100548 มิติ

จากการทดลอง Nperseg = 1024 เราจะได้ขนาดของ **Noverlap =** $\frac{\text{Nperseg}}{8}$ **=** $\frac{1024}{8}$ **= 128** จากนั้นนำ 1024-128 ทำให้ได้ขนาดฟังก์หน้าต่างเท่ากับ 896

ต่อมา หาขนาดและจำนวนของ window segments ในการทำฟังก์ชันหน้าต่าง ทำให้ได้ขนาดของ **f** ความละเอียดทางด้านความถี่และขนาดของ **t** ความละเอียดทางด้านเวลา หรือ จำนวนครั้งในการทำแต่ละ window segments

เพราะฉะนั้น t จะเท่ากับขนาดของเสียงสภาพแวดล้อมเท่ากับ 176400 มิติ แล้วถูกหารด้วยฟังก์ชันหน้าต่างคือ 896 ดังนั้น ค่าของ t = $\frac{176400}{896}$ = 196 ครั้ง

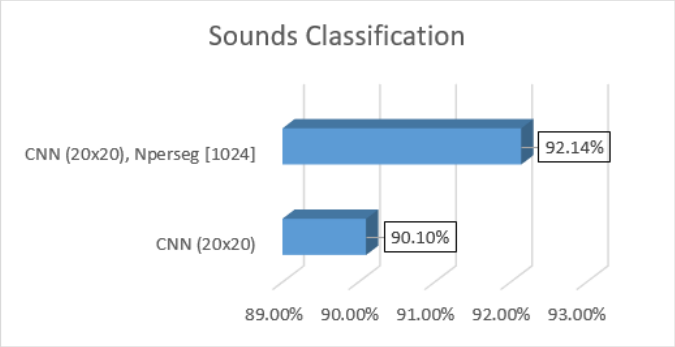
และส่วน f หรือขนาดของความถี่จะเท่ากับ $\frac{\text{Nperseg}}{2}+1$ = $\frac{1024}{2}+1$ = 513 [Hz]

ดังนั้น window segments มีขนาดเท่ากับ [f=513, t=196] ทำให้ได้ขนาดของสเปกโตรแกรมเท่ากับ [fxt] = 513x196 = 100548 มิติ

สำคัญ- ถ้าเราไม่ได้มีการกำหนดค่าของ nperseg โดยกำหนดให้เป็นค่า default จะทำให้เราได้ค่าเริ่มต้นของ nperseg คือ 256

ตารางที่ 4.13 การเปรียบเทียบประสิทธิภาพการจำแนกเสียงสภาพแวดล้อม โดยการปรับขนาดหน้าต่างของผลการแปลงฟูเรียร์ช่วงเวลาสั้น

Nperseg	Noverlap	Window_Segments	Spectrogram [fx t]	ite.1	ite.2	ite.3	ite.4	ite.5	Averages
256	32	f = 128, t = 787	101253	87.86%	89.23%	91.43%	92.86%	87.86%	89.85%
512	64	f = 257, t = 393	101001	82.14%	89.99%	88.57%	89.99%	90.00%	90.14%
768	96	f = 385, t = 262	100870	91.43%	88.57%	91.43%	85.71%	91.43%	89.71%
1024	128	f = 513, t = 196	100548	92.86%	91.43%	95.00%	92.14%	89.28%	92.14%



รูปที่ 4.103 การเปรียบสมรรถนะของการปรับฟังก์ชันหน้าต่างด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้น

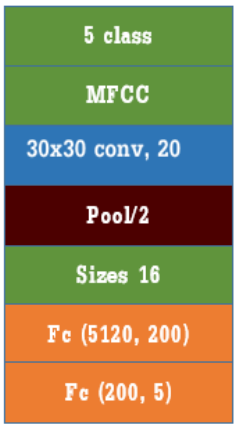
จากงานวิจัยการเรียนรู้จำเสียงสภาพแวดล้อมก่อนหน้านี้ไม่ได้มีการปรับขนาดฟังก์ชันหน้าต่างของผลการแปลงฟูเรียร์ เราจึงทำการทดลองปรับขนาดฟังก์ชันหน้าต่างทั้งหมด 4 ค่าคือ [256 512 768 และ 1024] แสดงดังตารางที่ 4.13

จากการทดลองจะเห็นได้ว่าขนาดของฟังก์ชันหน้าต่างที่เพิ่มขึ้น ทำให้ประสิทธิภาพของเครื่องมือการจำแนกเสียงสภาพแวดล้อมสามารถทำนายได้แม่นยำขึ้น เนื่องจากการปรับเพิ่มขนาดของฟังก์ชันหน้าต่างส่งผลทำให้ความละเอียดทางด้านความถี่มากกว่าความละเอียดทางด้านเวลา

4.3.2 สรุปผลการทดลอง

จากการเปรียบเทียบสมรรถนะระหว่างการปรับขนาดฟังก์ชันหน้าต่างด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้นและงานจากหัวข้อที่ 4.3.1 ที่ไม่ได้มีการปรับขนาดฟังก์ชันหน้าต่าง จากการทดลองสรุปได้ว่างานวิจัยที่มีการปรับขนาดฟังก์ชันหน้าต่าง สามารถทำให้เครื่องมือการจำแนกทำนายได้แม่นยำกว่า แสดงดังรูปที่ 4.103

จากการทดลองปรับขนาดฟังก์ชันหน้าต่าง สรุปได้ว่าการปรับขนาดของความละเอียดทางด้านความถี่มากกว่าความละเอียดทางด้านเวลา ส่งผลทำให้การทำนายเสียงสภาพแวดล้อมแม่นยำขึ้น



รูปที่ 4.104 แบบจำลอง Original CNN ใช้คุณลักษณะของสัมประสิทธิ์เซปสตรัมความถี่เมล

รูปที่ 4.105 แบบจำลอง Original CNN ใช้คุณลักษณะของผลการแปลงฟูเรียร์ช่วงเวลาสั้น

4.4 การเปรียบเทียบสมรรถนะระหว่างการสกัดคุณลักษณะ MFCC กับ STFT-PCA

จากหัวข้อนี้จะเป็นการทดสอบสมรรถนะระหว่างสัมประสิทธิ์เซปสตรัมความถี่เมลและผลการแปลงฟูเรียร์ช่วงเวลาสั้น โดยเราจะทำการทดสอบว่าคุณลักษณะประเภทใดที่สามารถให้ความแม่นยำสูงสุด จากการทดลองแบบจำลองของสัมประสิทธิ์เซปสตรัมความถี่เมลแสดงดังรูปที่ 4.104 และแบบจำลองของผลการแปลงฟูเรียร์ช่วงเวลาสั้น แสดงดังรูปที่ 4.105 จะเห็นได้ว่าแบบจำลองทั้งสองใช้สัญญาณเสียงและโครงข่ายคอนโวลูชันในการจำแนกเหมือนกัน แตกต่างตรงการสกัดคุณลักษณะ

ตารางที่ 4.14 การเปรียบเทียบแบบจำลองที่มีการสกัดคุณลักษณะที่แตกต่างกัน

Architecture Model	Feature Extraction	Convolution	Pooling	Fully Connected	Class	Sounds	Precsion
Original (CNN)	MFCC	1	1	2	10	8732	73.12%
Original (CNN)	MFCC	1	1	2	5	460	62.05%
Original (CNN)	STFT	1	1	2	5	460	69.99%
LeNet - 5	STFT	2	2	4	5	460	82.80%
CNN (20x20), Nperseg [1024]	STFT	2	2	4	5	460	92.14%

จากตารางที่ 4.14 เป็นการเปรียบเทียบแบบจำลองและคุณสมบัติที่แตกต่างกัน โดยเริ่มแรกของงานวิจัยเราได้นำงานวิจัยของ Original (CNNs) มาศึกษาในการทดลองการเรียนรู้จำเสียงสภาพแวดล้อม จากงานวิจัยของเขาได้ทดลองเสียง 10 ประเภท ได้แก่ dog bark children playing car horn air conditioner street music siren engine idling jckhammer gun shot และ drilling และมีทั้งหมด 8732 เสียง ส่วนวิธีการสกัดคุณลักษณะใช้สัมประสิทธิ์เซปสตรัมความถี่เมลทำหน้าที่แปลงข้อมูลเสียงมาเป็นข้อมูลภาพ และวิธีการจำแนกเสียงสภาพแวดล้อมใช้แบบจำลองของโครงข่ายคอนโวลูชัน จากงานวิจัยของเขาผลการทดลองประสิทธิภาพในการจำแนกได้ 73.12%

หลังจากที่เราได้ศึกษางาน Original (CNNs) เราได้ตั้งสมมุติฐานว่าถ้าเราลดจำนวนประเภทของเสียงให้น้อยลง เหลือเพียงเพียง 5 ประเภท ได้แก่ [Children Playing Air Conditioner Street Music Siren Engine idling และ Siren จะทำให้ประสิทธิภาพของการจำแนกเสียงอาจดีขึ้นหรือใกล้เคียง จากการทดลองเราสามารถทำนายได้ถึง 62.05% แสดงดังตารางที่ 4.14 ซึ่งไม่ได้ทำให้ความแม่นยำเพิ่มขึ้น เราเลยตั้งสมมุติฐานทดลองปรับเปลี่ยนการสกัดคุณลักษณะมาเป็นผลการแปลงฟูเรียร์ช่วงเวลาสั้น ทำให้ประสิทธิในการทำนายเสียงถึง 69.99% สิ่งที่เราตั้งสมมุติฐานก็เป็นจริง แต่ประสิทธิภาพที่ได้ยังไม่ดีหนัก เราจึงทำการศึกษาแบบจำลองของ LeNet-5 จากการทดลองเราจะเห็นได้ว่าโครงข่ายคอนโวลูชันของ Original (CNNs) มีความแตกตรงที่ LeNet-5 มีการเพิ่มเลเยอร์ของ convolutional pooling และ fully แสดงดังตารางที่ 4.14

จากการทดลองการเรียนรู้จำเสียงสภาพแวดล้อมด้วย LeNet-5 สามารถทำนายได้แม่นยำถึง 82.80% ซึ่งมากกว่าแบบจำลอง Original (CNNs) ทำให้เราสรุปจากการทดลองได้ว่าวิธีการสกัดคุณลักษณะไม่ใช่เหตุผลหลักที่ทำให้ความแม่นยำในการจำแนกเสียงที่เพิ่มขึ้น แต่เป็นเพราะการเพิ่มขึ้นของชั้นเลเยอร์ convolutional pooling และ fully ต่างหาก

ตารางที่ 4.15 การเปรียบเทียบความแตกต่างระหว่างเวลาในการปรับขนาด spcetrogram ด้วยขนาดของฟังก์ชันหน้าต่างเท่ากับ [256 512 768 และ 1024]

Nperseg	Noverlap	Window_Segments	Spectrogram [fxt]	Iter.1	Iter.2	Iter.3	Iter.4	Iter.5	Average (s)
256	32	f = 128, t = 787	101253	13.56	13.95	13.58	14.02	13.65	13.75
512	64	f = 257, t = 393	101001	13.25	13.12	13.56	13.68	13.12	13.35
768	92	f = 385, t = 262	100870	12.98	12.75	13.01	13.05	12.98	12.95
1024	128	f = 513, t = 196	100548	12.69	12.75	12.97	13.01	12.95	12.87

ตารางที่ 4.16 การเปรียบเทียบความแตกต่างความซับซ้อนในการคำนวณของ image เท่ากับ Size [(57,57) (61, 61) และ (65, 65)]

Input_image	total_memory	total_params
size (57x57)	253.39K	430.30K
size (61x61)	336.12K	503.40K
size (65x65)	433.17K	588.70K

4.5 การเปรียบเทียบความซับซ้อน (complexity) และเวลาในการคำนวณการจำแนกเสียงสภาพแวดล้อม

ความซับซ้อนในการคำนวณเป็นสาขาหนึ่งของทฤษฎีการคำนวณที่มุ่งเน้นไปในการวิเคราะห์ทางเวลาและทรัพยากร เพื่อแก้ปัญหาความซับซ้อนที่คอมพิวเตอร์ก็ไม่สามารถคำนวณได้ นอกจากนี้ ประโยชน์ของความซับซ้อนยังสามารถช่วยลดมิติข้อมูลที่มีขนาดใหญ่มากได้

จากงานวิจัยหัวข้อนี้ เราจะมีการคำนวณความซับซ้อนเพื่อหาที่มาของปัญหา โดยการทดลองจะเห็นได้ว่าความละเอียดทางด้านเวลาที่มีจำนวนมาก ทำให้ขนาดของสเปกโตรแกรมมีขนาดมิติที่ใหญ่กว่าความละเอียดทางด้านความถี่ ด้วยเหตุนี้ จากการทดลองทำให้การคำนวณทรัพยากรของความละเอียดทางด้านเวลาที่มีขนาดใหญ่มาก ส่งผลทำให้คอมพิวเตอร์ใช้เวลามากกว่าความละเอียดทางด้านความถี่ แสดงดังตารางที่ 4.15

จากการทดลองเราจะหาความซับซ้อนในการคำนวณของแบบจำลองโครงข่ายคอนโวลูชันโดยมีการปรับขนาดข้อมูลภาพทั้งหมด 3 ขนาด ได้แก่ [(57, 57) (61, 61) และ (65, 65)] จากการทดลองจะเห็นได้ว่าขนาดภาพที่เพิ่มขึ้น ทำให้ค่าความซับซ้อนในการคำนวณ total memory และ params เพิ่มขึ้น แสดงดังตารางที่ 4.16

Input_size (65, 65)
20x20 conv, 20
pool/2
10x10 conv, 30
pool/2
Fc (1470, 220)
Fc (220, 140)
Fc (140, 100)
Fc (100, 5)

รูปที่ 4.106 รูปแบบจำลองโครงข่ายคอนโวลูชัน
ที่มีขนาดข้อมูลภาพเท่ากับ (65, 65)

4.5.1 การคำนวณความซับซ้อนของแบบจำลองโครงข่ายคอนโวลูชัน

จากการทดลองส่วนนี้เป็นการคำนวณหาความซับซ้อนของแบบจำลองโครงข่ายคอนโวลูชัน จากตัวอย่างขนาดข้อมูลภาพเท่ากับ (65, 65) และมี 2 คอนโวลูชันเลเยอร์ 2 พลูลิงเลเยอร์ และ 4 การเชื่อมโยงเต็มรูปแบบ

วิธีการคำนวณหาความซับซ้อน ของแบบจำลองโครงข่ายคอนโวลูชัน โดยมีขนาดอินพุตที่ให้มาเท่ากับ 65x65 จากตัวอย่างนี้เราจะหาค่าหน่วยความจำ (memory) และจำนวนพารามิเตอร์ (parameter) ทั้งหมด

INPUT: [65x65x20] memory: 65x65 = 4.22K params; 0
CONV20-20: [65x65x20] memory: 46x46x20 = 42.32K params; (20x20x20)x20 = 160K
POOL/2: [23x23x20] memory: 23x23x20 = 10.58K params; 0
CONV10-30: [23x23x30] memory: 14x14x30 = 5.88K params; (10x10x20)x30 = 60K
POOL/2: [7x7x30] memory: 7x7x30 = 1.47K params; 0
FC1: [1x1x220] memory: 220 params: 7x7x30x220 = 323.4K
FC2: [1x1x140] memory: 140 params: 220x140 = 30.8K
FC3: [1x1x100] memory: 100 params: 140x100 = 14K
FC4: [1x1x5] memory: 5 params: 5x100 = 0.5K

Total Memory = (4.22+42.32+10.58+5.88+1.47+0.22+0.14+0.10+0.005)K = 64.93K
= 64.93x4 bytes = 259.72K byte / image
Total Params = (160+60+323.4+30.8+14+0.5)K = 588.7K

ตารางที่ 4.17 การเปรียบเทียบประสิทธิภาพของการจำแนกเสียงสภาพแวดล้อมระหว่างการปรับขนาดของ nperseg [256 512 768 และ 1024] และขนาด image [57x57 61x61 และ 65x65]

Nperseg	Noverlap	image size	total_memory	total_params	training (s)	testing (s)	precision
256	32	size(57, 57)	174.20K	430.30K	523.45	1.70	86.57%
256	32	size(61, 61)	214.72K	503.40K	654.57	2.12	85.56%
256	32	size(65, 65)	259.72K	588.70K	815.26	2.52	89.85%
512	64	size(57, 57)	253.39K	430.30K	523.36	1.73	88.42%
512	64	size(61, 61)	336.12K	503.40K	656.25	2.32	87.99%
512	64	size(65, 65)	433.17K	588.70K	815.56	2.62	90.14%
768	92	size(57, 57)	174.20K	430.30K	526.22	1.75	88.14%
768	92	size(61, 61)	214.72K	503.40K	657.95	2.56	89.85%
768	92	size(65, 65)	259.72K	588.70K	815.96	2.65	89.71%
1024	128	size(57, 57)	174.20K	430.30K	526.22	1.78	87.85%
1024	128	size(61, 61)	214.72K	503.40K	659.56	2.59	90.00%
1024	128	size(65, 65)	259.72K	588.70K	816.12	2.69	92.14%

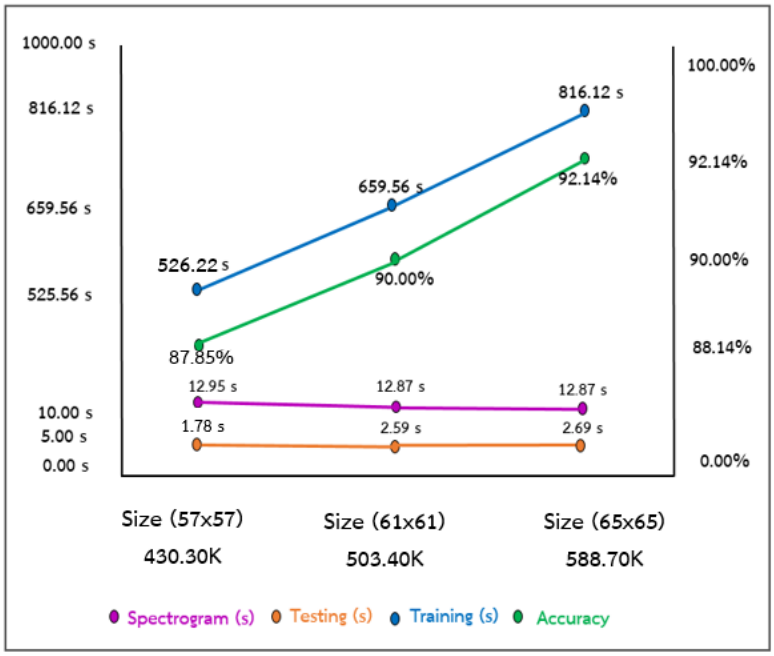
4.5.2 การคำนวณความซับซ้อนของแบบจำลองโครงข่ายคอนโวลูชัน

อนึ่งความซับซ้อนในการคำนวณโดยจำนวนค่าของหน่วยความจำและพารามิเตอร์ที่เพิ่มขึ้น จะทำให้ใช้เวลาในการฝึกฝนและทดสอบมากกว่าความซับซ้อนในการคำนวณจำนวนที่น้อยกว่า

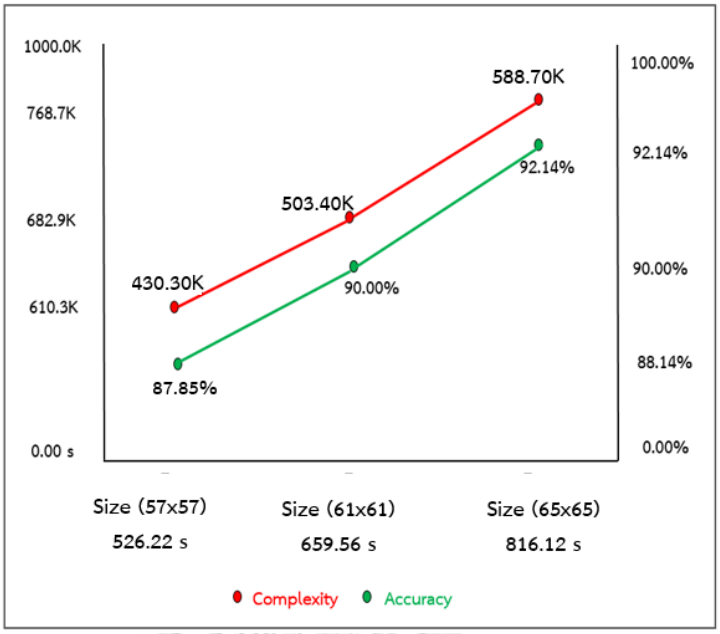
จากการทดลองจะเห็นได้ว่าเวลาในการฝึกฝนกับเวลาทดสอบที่เพิ่มขึ้น มาจากขนาดแต่ละข้อมูลภาพที่มีขนาดไม่เท่ากัน ได้แก่ [(57x57) (61x61) และ (65x65)] จะส่งผลทำให้ได้ขนาดของ total memory และ params มีขนาดเพิ่มขึ้น ตามลำดับ จากผลการทดลองหัวข้อนี้สรุปได้ว่าข้อมูลภาพที่มีขนาดเพิ่มขึ้นจะทำให้ขนาดมิติใหญ่มาก ส่งผลทำให้ใช้เวลาในการคำนวณระหว่างฝึกฝนและทดสอบนานกว่าขนาดข้อมูลภาพที่เล็กกว่า

จากผลการทดลองจะได้ว่าเมื่อความซับซ้อนในการคำนวณเพิ่มขึ้น [430.30k 503.40k และ 588.70k] ตามลำดับ จะทำให้การทำนายเสียงสภาพแวดล้อมได้แม่นยำมากกว่าโดยมีความถูกต้อง 87.85% 90.00% และ 92.14% ดังตารางที่ 4.17

รูปที่ 4.107 การเปรียบเทียบประสิทธิภาพของการจำแนกเสียงสภาพแวดล้อมเวลาของชุดฝึกฝนที่มีขนาด image เท่ากับ (57x57 61x61 และ 65x65)



รูปที่ 4.108 การเปรียบเทียบประสิทธิภาพของการจำแนกเสียงสภาพแวดล้อมเวลาระหว่างขนาดสเปกโตรแกรมชุดฝึกฝนและชุดทดสอบที่ขนาด image เท่ากับ (57x57 61x61 และ 65x65)



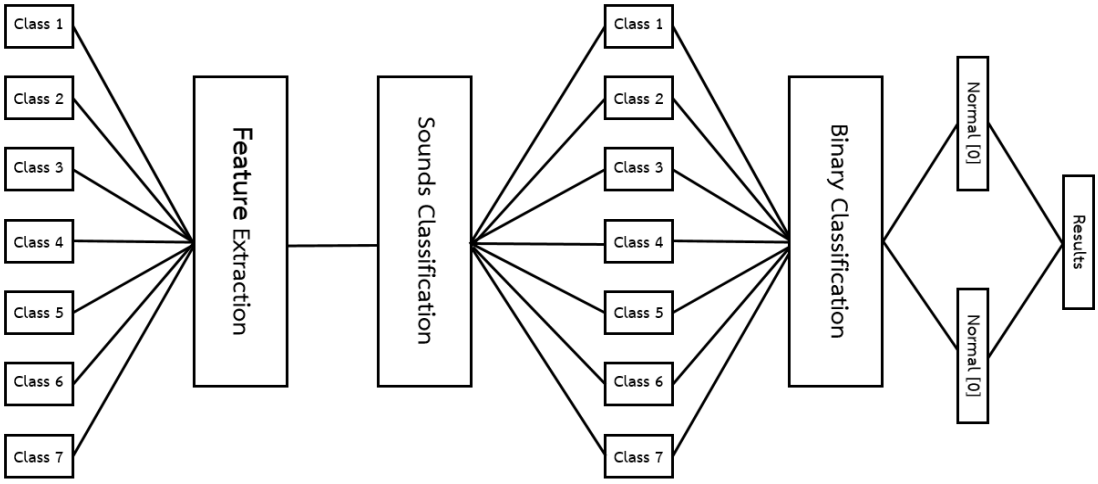
รูปที่ 4.109 กราฟการเปรียบเทียบประสิทธิภาพของการจำแนกเสียงสภาพแวดล้อมระหว่างความซับซ้อนในการคำนวณกับขนาด image ได้แก่ (57x57 61x61 และ 65x65)

4.5.3 สรุปผลการทดลอง

จากการทดลองจะเห็นได้ว่าขนาดของ image ที่เพิ่มขึ้น [(57x57) (61x61) และ (65x65)] ทำให้ค่าของความซับซ้อนของพารามิเตอร์การเรียนรู้จำเสียงมีจำนวนเพิ่มขึ้น [(430.30K) (503.40K) และ (588.70K)] ด้วยเหตุนี้ ค่าความซับซ้อนและเวลาที่ใช้ในการฝึกฝนที่มีขนาดสูงสุด ส่งผลทำให้การทำนายเสียงสภาพแวดล้อมแม่นยำกว่าค่าความซับซ้อนและเวลาที่ฝึกฝนจำนวนน้อย ๆ แสดงดังรูปที่ 4.107

จากรูปที่ 4.108 คือการคำนวณค่าความซับซ้อนทั้งหมด โดยเริ่มตั้งแต่การคำนวณเวลาของขนาดสเปกโตรแกรมที่ได้จากผลการแปลงฟูเรียร์ การทำคอนโวลูชันเลเยอร์ พลูลิง และการเชื่อมโยงเต็มรูปแบบ และการคำนวณเวลาของการฝึกฝนและทดสอบข้อมูล จากการทดลองสรุปได้ว่าความซับซ้อนและเวลาในการคำนวณเพิ่มขึ้น ทำให้ประสิทธิภาพของการทำนายเสียงแม่นยำขึ้น แสดงดังรูปที่ 4.109

หลังจากที่เราสามารถจำแนกเสียงสภาพแวดล้อม ปรับขนาดฟังก์ชันหน้าต่างด้วยผลการแปลงฟูเรียร์ช่วงเวลาสั้น และการหาความซับซ้อนและเวลาในการคำนวณ ทำให้เราสามารถหาวิธีการสกัดคุณลักษณะและแบบจำลองที่ดีที่สุดใช้ในงานวิจัยการเรียนรู้จำเสียงสภาพแวดล้อมและปืนใหญ่



รูปที่ 4.110 โฟลว์ชาร์ตขั้นตอนการทำงานของการจำแนกและการทำไบนารี
ของเสียงสภาพแวดล้อมและปืนใหญ่

4.6 การทดสอบการจำแนกเสียงไม่อันตรายกับเสียงอันตราย

จากงานวิจัยเราจะทำการทดสอบแบ่งออกเป็น 2 ประเภท คือการจำแนกและการทำไบนารีของเสียงสภาพแวดล้อมกับเสียงปืนใหญ่ แสดงดังรูปที่ 4.110 โดยเริ่มจากประเภทแรกคือการจำแนกเสียงสภาพแวดล้อมและปืนใหญ่ ได้แก่ เสียงเครื่องปรับอากาศ เสียงเด็กเล่น เสียงของรถยนต์ เสียงไซเรน เสียงดนตรีที่เล่นสถานที่เปิด เสียงปืน และเสียงปืนใหญ่ ส่วนวิธีการทำจะแบ่งออกเป็น 2 วิธี วิธีแรกคือการสกัดคุณลักษณะด้วยผลการแปลงฟูเรียร์และการวิเคราะห์องค์ประกอบหลัก ส่วนเครื่องมือการจำแนกเสียงใช้ซัพพอร์ตเวกเตอร์แมชชีนและเพอร์เซ็ปตรอนหลายชั้น ส่วนวิธีที่สองเราใช้การสกัดคุณลักษณะของผลการแปลงฟูเรียร์ช่วงเวลาสั้นในการประยุกต์จากข้อมูลเสียงแปลงมาเป็นข้อมูลภาพ ทำให้สามารถใช้กับเครื่องมือการจำแนกของโครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อน

หลังจากการจำแนกเสียงสภาพแวดล้อมและปืนใหญ่ได้ วัตถุประสงค์ของงานวิจัยเราต้องการแยกระหว่างกลุ่มเสียงที่ปกติกับเสียงอันตราย โดยการทำไบนารีแยกระหว่างเสียงสภาพแวดล้อมและเสียงปืนใหญ่ วิธีการทำไบนารีเราจะต้องแบ่งสัญญาณเสียง โดยเริ่มจาก 5 คลาสแรก ได้แก่ เสียงเครื่องปรับอากาศ เสียงเด็กเล่น เสียงของรถยนต์ เสียงไซเรน และเสียงดนตรีที่เล่นสถานที่เปิด จัดให้อยู่ในคลาส 0 เป็นเสียงปกติ และส่วน 2 คลาสที่เหลือคือเสียงปืนกับเสียงปืนใหญ่จะถูกแทนเป็นคลาส 1 ซึ่งเป็นคลาสอันตราย ขั้นตอนสุดท้ายเราจะเปรียบระหว่างเครื่องมือการจำแนกของซัพพอร์ตเวกเตอร์แมชชีนและโครงข่ายคอนโวลูชันอันไหนสามารถแยกระหว่างเสียงปกติและเสียงอันตรายได้ดีกว่ากัน

ตารางที่ 4.18 เมทริกซ์ความสับสนของการจำแนกเสียงสภาพแวดล้อมและปืนใหญ่ด้วย SVM

	Air Conditioner	Childern Playing	Engine Idling	Siren	Street Music	Gun Shot	Artillery
Air Conditioner	28	0	0	0	0	0	0
Childern Playing	0	23	0	1	4	0	0
Engine Idling	1	0	18	0	1	9	0
Siren	0	0	1	27	0	0	0
Street Music	2	4	1	1	20	0	0
Gun Shot	0	0	0	0	0	28	0
Artillery	0	0	1	0	0	0	27

ตารางที่ 4.19 เมทริกซ์ความสับสนการจำแนกเชิงไบนารีด้วย SVM

	Normal	Harmful
Normal	135	5
Harmful	1	55

เสียงปกติ คลาส [0] ได้แก่ เสียงเครื่องปรับอากาศ เสียงเด็กเล่น เสียงของรถยนต์ เสียงไซเรน และเสียงดนตรีที่เล่นสถานที่เปิด

เสียงอันตราย คลาส [1] ได้แก่ เสียงปืนและเสียงปืนใหญ่

ตารางที่ 4.20 เมทริกซ์ความสับสนการจำแนกเชิงไบนารีด้วย SVM

	Normal	Harmful
Normal	80	4
Harmful	2	54

เสียงปกติ คลาส [0] ได้แก่ เสียงเด็กเล่น เสียงของรถยนต์ และเสียงดนตรีที่เล่นสถานที่เปิด

เสียงอันตราย คลาส [1] ได้แก่ เสียงปืนและเสียงปืนใหญ่

4.6.1 การจำแนกเชิงไบนารีเสียงสภาพแวดล้อมกับปืนใหญ่ด้วยซัพพอร์ตเวกเตอร์แมชชีน

หัวข้อนี้เป็นการทดสอบการรู้จำและการจำแนกเชิงไบนารีระหว่างเสียงสภาพแวดล้อมกับเสียงปืน-ปืนใหญ่ จากตารางที่ 4.18 จะเห็นได้ว่าแบบจำลองของ SVM ทำนายเสียงของรถยนต์ผิดพลาดไปเป็นเสียงปืนถึง 9 เสียง ซึ่งในความเป็นจริงมนุษย์เราสามารถรับฟังเสียงปืนกับเสียงรถยนต์แล้วสามารถแยกออกจากกันได้ จากสมมติฐาน เสียงปืนในบางกรณีมีสัญญาณเสียงที่เกิดซ้ำ ๆ กัน ทำให้ไปคล้ายคลึงกับเสียงรถยนต์

การทำไบนารีแยกเสียงปกติกับเสียงอันตรายจากตารางที่ 4.19 และ 4.20 เราได้ทำ 2 วิธี โดยวิธีแรกทำการทดสอบข้อมูลเสียงสภาพแวดล้อมจำนวน 5 คลาส และวิธีที่สองทำการทดสอบเสียงสภาพแวดล้อม 3 คลาส ตามลำดับ ซึ่งในการทดลองได้ตั้งสมมุติฐานว่าคลาสเสียงสภาพแวดล้อมที่มีจำนวนมากกว่าสามารถทำนายเสียงได้แม่นยำกว่าคลาสเสียงสภาพแวดล้อมจำนวนน้อยกว่า

ตารางที่ 4.21 เมทริกซ์ความสับสนของการจำแนกเสียงสภาพแวดล้อม
และปืนใหญ่ด้วยโครงข่ายคอนโวลูชัน

	Air Conditioner	Childern Playing	Engine Idling	Siren	Street Music	Gun Shot	Artillery
Air Conditioner	28	0	0	0	0	0	0
Childern Playing	0	21	0	1	6	0	0
Engine Idling	1	0	26	0	1	0	0
Siren	0	1	1	25	1	0	0
Street Music	0	3	0	1	23	1	0
Gun Shot	0	0	0	0	0	28	0
Artillery	0	0	0	0	0	0	28

ตารางที่ 4.22 เมทริกซ์ความสับสนการจำแนกเชิงไบนารีระหว่างเสียงปกติ
และเสียงอันตรายด้วยโครงข่ายคอนโวลูชัน

	Normal	Harmful
Normal	139	1
Harmful	0	56

เสียงปกติ คลาส [0] ได้แก่ เสียงเครื่องปรับอากาศ เสียงเด็กเล่น เสียงของรถยนต์ เสียงไซเรน และเสียงดนตรีที่เล่นสถานที่เปิด
เสียงอันตราย คลาส [1] ได้แก่ เสียงปืนและเสียงปืนใหญ่

ตารางที่ 4.23 เมทริกซ์ความสับสนการจำแนกเชิงไบนารีระหว่างเสียงปกติ
และเสียงอันตรายด้วยโครงข่ายคอนโวลูชัน

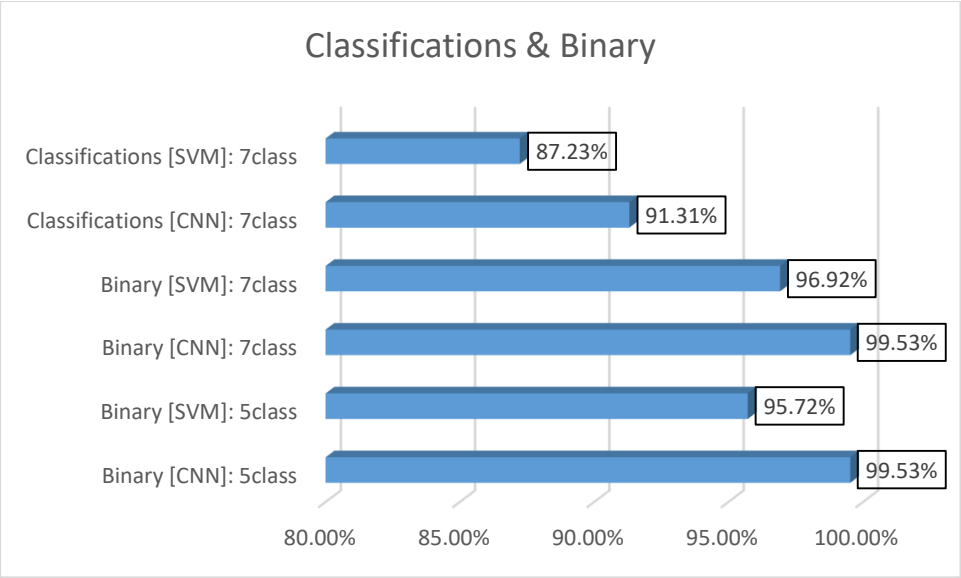
	Normal	Harmful
Normal	83	1
Harmful	0	56

เสียงปกติ คลาส [0] ได้แก่ เสียงเด็กเล่น เสียงของรถยนต์ และเสียงดนตรีที่เล่นสถานที่เปิด
เสียงอันตราย คลาส [1] ได้แก่ เสียงปืนและเสียงปืนใหญ่

4.6.2 การจำแนกเชิงไบนารีเสียงสภาพแวดล้อมกับปืนใหญ่ด้วยโครงข่ายคอนโวลูชัน

จากหัวข้อนี้เป็นการทดสอบการจำแนกและการทำไบนารีเสียงสภาพแวดล้อมกับเสียงปืนใหญ่ด้วยเครื่องมือการจำแนกของโครงข่ายคอนโวลูชัน จากการทดลองจะเห็นได้ว่าเสียงเด็กเล่นทำนายไปเป็นเสียงดนตรีถึง 6 เสียง แสดงดังตารางที่ 4.21 แต่จากผลการทดลองการจำแนกวิธีของโครงข่ายคอนโวลูชันสามารถทำนายได้แม่นยำกว่าวิธีของซัพพอร์ตเวกเตอร์แมชชีน

การทำไบนารีของโครงข่ายคอนโวลูชันเหมือนกับวิธีของซัพพอร์ตเวกเตอร์แมชชีน จากผลการทดการแยกเสียงปกติและเสียงอันตราย แสดงดังตารางที่ 4.22 และ 4.23 จะเห็นได้ว่าทั้ง 2 วิธีทำนายเสียงปกติผิดไปเป็นเสียงอันตราย 1 เสียง



รูปที่ 4.111 การเปรียบเทียบสมรรถนะของเครื่องมือการจำแนกระหว่างซัพพอร์ตเวกเตอร์แมชชีนและโครงข่ายประสาทเกิดซ้อน

4.6.3 สรุปผลการทดลอง

จากงานวิจัยหัวข้อนี้เราได้ทำการทดลองการจำแนกเสียงสภาพแวดล้อมกับเสียงปืนใหญ่และทำไบนารีแยกระหว่างเสียงปกติกับเสียงอันตราย โดยเราจะทำการทดสอบสมรรถนะของเครื่องมือการจำแนกของซัพพอร์ตเวกเตอร์แมชชีนและโครงข่ายคอนโวลูชัน

จากการทดลองการจำแนกเสียงทั้งหมด 7 คลาส ได้แก่ เสียงเครื่องปรับอากาศ เสียงเด็กเล่น เสียงของรถยนต์ เสียงไซเรน เสียงดนตรีที่เล่นสถานที่เปิด เสียงปืน และเสียงปืนใหญ่ โดยการทดสอบสมรรถนะของเครื่องมือการจำแนก ผลการทดลองโครงข่ายคอนโวลูชันสามารถจำแนกได้แม่นยำกว่าซัพพอร์ตเวกเตอร์แมชชีน แสดงดังรูปที่ 4.111 จากที่เราได้ทำการจำแนกเสียงทั้งหมด 7 คลาส เราจะทำการทดลองต่อไปคือไบนารีแยกเสียงปกติกับเสียงอันตราย ซึ่งการทำไบนารีเราจะลดให้เหลือเพียงคลาส [0 กับ 1] โดยที่คลาส 0 คือเสียงปกติหรือเสียงสภาพแวดล้อม ได้แก่ เสียงเครื่องปรับอากาศ เสียงเด็กเล่น เสียงของรถยนต์ เสียงไซเรน และเสียงดนตรีที่เล่นสถานที่เปิด และส่วนคลาส 1 คือเสียงอันตราย ได้แก่ เสียงปืนและเสียงปืนใหญ่ จากผลการทดลองเครื่องมือที่สามารถแยกระหว่างเสียงปกติกับเสียงอันตรายได้ดีสุดคือโครงข่ายคอนโวลูชันคือ 99.53%

บทที่ 5
สรุปผลการทดลอง

5.1 บทสรุป

วิทยานิพนธ์ฉบับนี้ทำเกี่ยวกับการเรียนรู้จำเสียงสภาพแวดล้อมและปืนใหญ่ โดยวัตถุประสงค์หลักคือต้องแยกระหว่างเสียงปกติกับเสียงอันตราย จากการทดลองเราจะแบ่งวิธีการทำงานเป็น 2 วิธีคือ sounds classification กับ binary classification

การทำ sounds classification คือการจำแนกเสียงสภาพแวดล้อมและเสียงปืนใหญ่ โดยขั้นตอนการทำงานจะแบ่งออกเป็น 2 ส่วนหลัก ๆ คือการสกัดคุณลักษณะเสียงสภาพแวดล้อมและการจำแนกเสียงสภาพแวดล้อม โดยวิธีการสกัดคุณลักษณะจะแบ่งออกเป็น 2 วิธีคือผลการแปลงฟูเรียร์ช่วงเวลาสั้นทำหน้าที่แยกคุณลักษณะแต่ละสัญญาณเสียงสภาพแวดล้อม และการลดมิติข้อมูลขนาดใหญ่มากใช้วิธีการวิเคราะห์องค์ประกอบหลัก สำหรับวิธีที่กล่าวมาใช้กับเครื่องการจำแนกของซัพพอร์ตเวกเตอร์แมชชีนและเพอร์เซ็ปตรอนหลายชั้น ส่วนวิธีที่สองเราใช้ผลการแปลงฟูเรียร์ช่วงเวลาสั้นนำมาประยุกต์แปลงจากข้อมูลเสียงมาเป็นข้อมูลภาพ ทำให้สามารถใช้กับเครื่องมือการจำแนกของโครงข่ายคอนโวลูชันและโครงข่ายประสาทเกิดซ้อน

การทำ binary classification คือการจำแนกระหว่างเสียงปกติกับเสียงอันตรายหรือการทำให้เหลือเพียง 2 คลาส โดยวิธีการทำเราจะกำหนดให้เสียงสภาพแวดล้อม ได้แก่ เสียงเครื่องปรับอากาศ เสียงเด็กเล่น เสียงของรถยนต์ เสียงไซเรน และเสียงดนตรีที่เล่นสถานที่เปิด เป็นคลาส 0 คือกลุ่มเสียงปกติ ส่วนเสียงปืนและปืนใหญ่คือเสียงอันตราย ซึ่งในการทดลองเราจะทำการเปรียบเทียบสมรรถนะของเครื่องมือการจำแนกที่สามารถแยกระหว่างเสียงปกติกับเสียงอันตรายได้แม่นยำที่สุด

จากบทที่ 3 เราจะพูดถึงขั้นตอนการเตรียมงานเริ่มจากหัวข้อ 3.1 เราจะเตรียมฐานข้อมูลเสียงจาก Urbansounds 8k ที่มีจำนวน 10 คลาส โดยเราเลือกมาเพียง 5 คลาส ได้แก่ เสียงเครื่องปรับอากาศ เสียงเด็กเล่น เสียงของรถยนต์ เสียงไซเรน และเสียงดนตรีที่เล่นสถานที่เปิด เพื่อใช้สำหรับการเรียนรู้จำเสียงสภาพแวดล้อม แล้วนำสัญญาณที่ได้มาผ่านกระบวนการสกัดคุณลักษณะจากหัวข้อที่ 3.2 เราจะแบ่งออกเป็น 2 วิธี วิธีแรกผลการแปลงฟูเรียร์ช่วงเวลาสั้น สำหรับวิธีการทำเราจะนำสัญญาณเสียงมาแปลงจากสัญญาณโดเมนทางเวลามาเป็นโดเมนทางเวลา-ความถี่ เพื่อต้องการแยกประเภทของแต่ละสัญญาณเสียงและยังสามารถประยุกต์มาเป็นข้อมูลภาพ ต่อมาหัวข้อ 3.2.2 คือการสกัดคุณลักษณะด้วยการวิเคราะห์องค์ประกอบหลักทำหน้าที่ลดมิติข้อมูลที่มีขนาดใหญ่มาก โดยวิธีการทำจะต้องหาค่าของเมทริกซ์โคแวเรียนซ์ของเสียงสภาพแวดล้อม เราถึงจะได้ค่า eigenSTFT และไอเกนแวลลิว นำมาลดมิติของเสียงสภาพแวดล้อมได้ แต่ในงานวิจัยเราการลดมิติไม่

จำเป็นต้องใช้การวิเคราะห์องค์ประกอบหลักอย่างเดียว อาทิเช่น โครงข่ายคอนโวลูชันภายในประกอบด้วยคอนโวลูชันเลเยอร์ และพลูลิงเลเยอร์ทำหน้าที่ลดมิติแทนการวิเคราะห์องค์ประกอบหลักได้ จากที่เราได้กล่าวมาเป็นขั้นตอนการสกัดคุณลักษณะ

การทดสอบหัวข้อที่ 4.2 เราได้เสนอวิธีการเรียนรู้จำเสียงสภาพแวดล้อม โดยเราจะแบ่งการทดสอบออกเป็น 2 วิธี วิธีแรกคือใช้การสกัดคุณลักษณะด้วย STFT กับ PCA ใช้เครื่องมือการจำแนกได้แก่ SVM และ MLP ส่วนวิธีที่สองเราจะใช้ STFT ในการประยุกต์จากสัญญาณเสียงให้กลายมาเป็นข้อมูลภาพทำให้สามารถใช้กับเครื่องมือการจำแนก CNN และ RNN จากผลการทดการเรียนรู้จำเสียงสภาพแวดล้อมนั้นเครื่องมือที่สามารถจำแนกได้แม่นยำสุดคือ CNN (20x20)

การทดสอบหัวข้อที่ 4.3 จะเป็นการเปรียบเทียบระหว่างการปรับขนาดฟังก์ชันหน้าต่างของผลการแปลงฟูเรียร์ช่วงเวลาสั้น จากหัวข้อ 4.2 ไม่ได้มีการปรับขนาดฟังก์ชันหน้าต่าง จากการทดสอบเราได้มีการปรับขนาดฟังก์ชันหน้าต่างทั้งหมด 4 ค่า ได้แก่ 256 512 768 และ 1024 จากผลการทดลองขนาดของฟังก์ชันหน้าต่างเท่ากับ 1024 ให้ความละเอียดทางด้านความถี่มากกว่า 256 512 และ 768 จึงทำให้การทำนายเสียงสภาพแวดล้อมแม่นยำมากกว่า สรุปการปรับขนาดฟังก์ชันหน้าต่างให้ความแม่นยำกว่าการที่ไม่ได้ปรับขนาดหน้าต่าง

การทดสอบหัวข้อที่ 4.4 จะเป็นการเปรียบเทียบระหว่างงาน original กับงานของวิจัยเราที่ใช้การสกัดคุณลักษณะแตกต่างกัน อีกสิ่งหนึ่งที่เหมือนกันระหว่างงานวิจัยเรากับวิจัยของ original คือใช้แบบจำลอง CNN แต่ก็ยังแตกต่างกันตรงที่เราใช้ 2 convolutional 2 pooling และ 4 fully connected ส่วนงานวิจัย original ใช้แค่ 1 convolutional 1 pooling และ 2 fully connected จากการทดลองงานวิจัยของ original จำแนกเสียงทั้งหมด 10 คลาส แต่สำหรับของเราจำแนกแค่ 5 คลาส เราจึงเอาชุดข้อมูล 5 คลาสไปทดสอบแบบจำลองของงานวิจัยของ original เพื่อจะทำการเปรียบเทียบเครื่องมือการจำแนก จากผลการทดลองสรุปได้ว่างานวิจัยเราสามารถทำนายได้แม่นยำกว่า อาจเป็นเพราะว่าเรามีการเพิ่ม convolutional pooling และ fully connected

การทดสอบหัวข้อที่ 4.5 จะเป็นการเปรียบเทียบความซับซ้อนและเวลาในการคำนวณการจำแนกเสียงสภาพแวดล้อม ซึ่งในหัวข้อนี้เราจะทำการเปรียบเทียบระหว่างการปรับพารามิเตอร์ต่าง ๆ นั้นส่งผลให้ใช้เวลาในการฝึกฝนข้อมูลนานขึ้นและประสิทธิภาพของความแม่นยำจะเพิ่มขึ้นหรือไหม จากการทดลองสรุปได้ว่าการปรับพารามิเตอร์ขนาดของฟังก์หน้าต่างเท่ากับ 1024 ทำให้ใช้เวลา น้อยกว่าการปรับขนาดเท่ากับ 256 512 และ 768 เพราะว่าขนาดหน้าต่างเท่ากับ 1024 จะไปลดความละเอียดทางด้านเวลาแต่จะไปเพิ่มความละเอียดทางด้านความถี่แทนนั่นเอง จากการทดลองสรุปได้ว่าค่า complexity จำนวนที่เพิ่มขึ้นส่งผลทำให้การเรียนรู้จำเสียงสามารถจำแนกได้แม่นยำกว่า complexity จำนวนน้อย ๆ

การทดสอบหัวข้อที่ 4.6 เราได้เสนอวิธีการทดสอบการจำแนกเสียงไม่อันตรายและเสียงอันตราย โดยเราทำการประยุกต์มาจากการเรียนรู้จำเสียงสภาพแวดล้อมที่มีการเพิ่มจำนวนเสียงปืนใหญ่กับปืน จากการทดสอบจะแบ่งออกเป็น 2 ขั้นตอน ขั้นตอนแรกการทำ sounds classification คือการจำแนกระหว่างเสียงสภาพแวดล้อมกับเสียงปืนใหญ่ โดยเครื่องมือที่จะนำมาทดสอบจะแบ่งออกเป็น 2 วิธีได้แก่ SVM กับ CNN ส่วนวิธีที่สองคือการทำ binary classification จะมีแค่ 2 คลาส คือคลาส [0] และคลาส [1] โดยที่คลาส [0] จะเป็นเสียงไม่อันตรายก็คือเสียงสภาพแวดล้อมและส่วนคลาส [1] จะเป็นเสียงอันตรายได้แก่เสียงปืนใหญ่และเสียงปืน จากการทดลองทั้งสองวิธีเครื่องมือที่สามารถทำนายได้แม่นยำสุดคือ CNN

5.2 ข้อเสนอแนะ

1. ทำการทดสอบโดยการเพิ่มจำนวนเสียงที่หลากหลายประเภท เพื่อดูว่าแบบจำลองที่ใช้ในการทดสอบยังสามารถทำนายได้แม่นยำเหมือนเดิมหรือไหม
2. ทำการทดสอบเปลี่ยนการสกัดคุณลักษณะด้วยวิธีผลการแปลงฟูเรียร์ช่วงเวลาสั้น มาเป็นวิธีการสกัดคุณลักษณะด้วยการแปลงเวฟเล็ต (wavelet transform)
3. ทำการทดสอบเสียงสภาพแวดล้อมมากกว่า 1 เสียงกับเสียงปืนใหญ่และปืน
4. ทำการทดสอบการจำแนกเสียงปืนใหญ่ใน background จากสถานที่จริง
5. ทำการลดขนาดของคอนโวลูชันจาก 20x20 เป็น 3x3 แล้วทำการเพิ่มความลึกของ convolutional Layer และ fully connected layer จากเดิม 6 เลเยอร์ เพิ่มมาเป็น 16 เลเยอร์

บรรณานุกรม

[1] A. Elbir, H. O. İlhan, G. Serbes, and N. Aydın, "Short Time Fourier Transform Based Music Genre Classification," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 2018, pp. 1-4.

[2] J. Novakovic and S. Rankov, "Classification Performance Using Principal Component Analysis and Different Value of the Ratio R," in *Computer Science, Int. J. Comput. Commun. Control*, 2011, vol. 6, pp. 317-327.

[3] C. Chang and B. Doran, "Urban Sound Classification: With Random Forest SVM DNN RNN and CNN Classifiers," in *CSCI E-81 Machine Learning and Data Mining Final Project Fall 2016*: Harvard University Cambridge, 2016.

[4] K. J. Piczak, "Environmental Sound Classification with Convolutional Neural Networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015, pp. 1-6.

[5] I. Y. Kim, S. Lee, H. Yeo, W. Han, and S. Hong, "Feature extraction for Heart Sound Recognition Based on Time-Frequency Analysis," in *Proceedings of the First Joint BMES/EMBS Conference. 1999 IEEE Engineering in Medicine and Biology 21st Annual Conference and the 1999 Annual Fall Meeting of the Biomedical Engineering Society*, 1999, vol. 2, pp. 960-967.

[6] C.-C. Wang and Y. Kang, "Feature Extraction Techniques of Non-Stationary Signals for Fault Diagnosis in Machinery Systems," in *Journal of Signal and Information Processing, Department of Mechanical Engineering, Chung Yuan Christian University*, 2012, vol. 3, pp. 16-25.

[7] W. T. Cochran *et al.*, "What is the fast Fourier transform?," in *Proceedings of the IEEE* 1967, vol. 55, no. 10, pp. 1664-1674.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2012.

[9] G. Guo and S. Z. Li, "Content-Based Audio Classification and Retrieval by Support Vector Machines," in *IEEE Transactions on Neural Networks* 2003, vol. 14, no. 1, pp. 209-215.

[10] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-Voice Separation from Monaural Recordings using Robust Principal Component Analysis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 57-60.

[11] G. Baudat and F. Anouar, "Kernel-Based Methods and Function Approximation," in *IJCNN'01. International Joint Conference on Neural Networks. Proceedings*, 2001, vol. 2, pp. 1244-1249.

[12] Y. LeCun, "LeNet-5, Convolutional Neural Networks," in http://yannlecun.com, 2015, vol. 20, no. 5, p. 14.

[13] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, and T. Yu, "Scikit-Image: Image Processing in Python," vol. 2, pp. 453-475, 2014.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, and V. Dubourg, "Scikit-learn: Machine learning in Python," vol. 12, pp. 2825-2830, 2011.

[15] J. V. Dillon, I. Langmore, D. Tran, E. Brevdo, and R. A. Saurous, "Tensorflow Distributions," 2017.

[16] C. Jatturas, P. D. N. Ayudhya, S. Pankaew, and W. Asdornwised, "Performance Comparison of Environmental Sound Classification using Scikit-learn," in *International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, 2018.

[17] C. Jatturas, S. Chokkoedsakul, P. D. N. Avudhva, S. Pankaew, C. Sopavanit, and W. Asdornwised, "Feature-Based and Deep Learning-Based Classification of Environmental Sound," in *2019 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 2019, pp. 126-130.

[18] C. Jatturas, S. Chokkoedsakul, P. D. N. Ayudhya, S. Pankaew, C. Sopavanit, and W. Asdornwised, "Recurrent Neural Networks for Environmental Sound Recognition using Scikit-learn and Tensorflow," in *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2019, pp. 806-809.

ประวัติผู้เขียน

ชื่อ-สกุล	ชินวัฒน์ จัตุรัส
วัน เดือน ปี เกิด	20 มีนาคม 2538
สถานที่เกิด	โรงพยาบาลสิงห์บุรี
วุฒิการศึกษา	ปริญญาบัณฑิต คณะวิศวกรรมศาสตร์ วิศวกรรมอิเล็กทรอนิกส์และโทรคมนาคม มหาวิทยาลัยพระจอมเกล้าพระนครเหนือ
ที่อยู่ปัจจุบัน	41/1 หมู่ 4 ต.ม่วงหมู่ อ.เมือง จ.สิงห์บุรี 16000
ผลงานตีพิมพ์	C. Jatturas, P. D. N. Ayudhya, S. Pankaew, and W. Asdornwised, "Performance Comparison of Environmental Sound Classification using Scikit-learn," in International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), 2018. C. Jatturas, S. Chokkoedsakul, P. D. N. Avudhva, S. Pankaew, C. Sopavanit, and W. Asdornwised, "Feature-Based and Deep Learning-Based Classification of Environmental Sound," in 2019 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), 2019, pp. 126-130. C. Jatturas, S. Chokkoedsakul, P. D. N. Ayudhya, S. Pankaew, C. Sopavanit, and W. Asdornwised, "Recurrent Neural Networks for Environmental Sound Recognition using Scikit-learn and Tensorflow," in International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2019, pp. 806-809.