การวิเคราะห์ตัวแทนของเพลงเพื่อระบุข้อมูลเพลง

ศักรินทร์ นุ้ยพิน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2563

Content-based Representations for Retrieve Song Information

Sakkarin Nuipin

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

Department of Big Data Engineering,

College of Innovative Technology and Engineering,

Dhurakij Pundit University

2020

ใบรับรองงานวิทยานิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

หัวข้อวิทยานิพนธ์ การวิเคราะห์ตัวแทนของเพลงเพื่อระบุข้อมูลเพลง

เสนอโดย ศักรินทร์ นุ้ยพิน

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบวิทยานิพนธ์แล้ว

.....ประธานกรรมการ

(ดร.สรรพฤทธิ์ มฤคทัต)

.....กรรมการและอาจารย์ที่ปรึกษาหลัก

(ผศ.ดร.ดวงใจ จิตคงชื่น)

.....กรรมการ

(ดร.ธนภัทร ฆังคะจิตร)

.....กรรมการ

(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา)

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์รับรองแล้ว

.....

(ผู้ช่วยศาสตราจารย์ ดร.ณรงค์เดช กีรติพรานนท์)

คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 27 เดือน เม.ย. พ.ศ. 2563

หัวข้อวิทยานิพนธ์	การวิเคราะห์ตัวแทนของเพลงเพื่อระบุข้อมูลเพลง
ชื่อผู้เขียน	ศักรินทร์ นุ้ยพิน
อาจารย์ที่ปรึกษา	ผศ. ดร. ดวงใจ จิตคงชื่น
สาขาวิชา	วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา	2562

บทคัดย่อ

ในอุตสาหกรรมเพลงผู้บริหารหรือเจ้าของผลงานเพลงมีความจำเป็นที่จะต้องตรวจสอบว่าเพลงแต่ละเพลงถูกเปิดไปแล้วกี่ครั้ง เพื่อนำมาใช้ในตัดสินใจทางธุรกิจ แต่ในปัจจุบันการตรวจสอบเพลงที่เปิดผ่านสถานีวิทยุนั้นเป็นไปได้ยาก ดังนั้นงานวิจัยนี้มีวัตถุประสงค์เพื่อแก้ปัญหาการค้นหาและระบุข้อมูลเพลงจากสัญญาณเสียงวิทยุ โดยประยุกต์ใช้งานโครงข่ายทริปเล็ตเพื่อสร้างตัวแทนของคุณลักษณะสำคัญของเสียง ซึ่งในการทดลองสร้างโมเดลจากปัจจัยที่แตกต่างกันคือ กระบวนการแยกคุณลักษณะสำคัญของเสียง, ขนาดของข้อมูลที่ใช้เป็นอินพุต (Input), ลักษณะโครงสร้างย่อยภายในโครงข่ายทริปเล็ต และ ขนาดของคุณลักษณะเวกเตอร์ที่จะใช้เป็นเอาท์พุต (Output) หลังจากได้โมเดลสำหรับสร้างตัวแทนของเสียงแล้ว จึงทดลองจำลองการค้นหาข้อมูลของเสียงเพลงด้วยวิธีการค้นหาเพื่อนบ้านใกล้สุด (Nearest Neighbor Search) และนอกจากนี้ยังทดลองเพิ่มสัญญาณรบกวน (Noise) ของสัญญาณ เสียงที่จะใช้เป็นข้อมูลสำหรับค้นหาด้วย จากผลการทดลองโดยใช้ข้อมูลเพลงจำนวน 100 เพลงพบว่าโมเดลโครงข่ายทริปเล็ตสามารถทำนายข้อมูลเพลงได้ถูกต้อง โดยให้ค่าความแม่นยำ (Accuracy) 0.86 นอกจากนี้ยังแสดงให้เห็นว่าตัวแทนของเสียงที่สร้างจากโมเดลโครงข่ายทริปเล็ตยังทนทานต่อสัญญาณรบกวนพอสมควรเพราะมีความผิดพลาดไปเพียง 3%

Thesis Title	Content-based Representations for Retrieve Song Information
Author	Sakkarin Nuipin
Thesis Advisor	Asst.Prof. Dr. Duangjai Jitkongchuen
Department	Big Data Engineering
Academic Year	2019

ABSTRACT

The executive, in the music industry, or music owner is necessary to know how many times each song has been played for used to business decisions. But the process to detect songs played via radio stations is difficult. Thus, the purpose of this study attempt to solve the problem of finding and identifying the song information from the radio signals by applying Triplet Networks to representation of song. Four factors were considered: feature extraction, size of input data, architecture of the sub-networks within Triplet Networks, and size of feature vector used as output. After model Triplet Networks trained. We implement to search song information by audio with Nearest Neighbor Search. In addition, we add noise to audio data (query data). The result base on 100 songs, show that model can accurately identification song. The accuracy rate is 0.86 also shows model can resistant to noise because there is mistakes 3%.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยการให้ความช่วยเหลือแนะนำของ ผศ. ดร. ดวงใจ จิตคงชื่น ซึ่งเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ที่ได้กรุณาที่ให้คำแนะนำข้อคิดเห็นตรวจสอบ และแก้ไขร่างวิทยานิพนธ์มาโดยตลอด ผู้เขียนจึงขอกราบขอบพระคุณไว้ ณ โอกาสนี้

ผู้เขียนขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.สรรพฤทธิ์ มฤคทัต ที่กรุณาให้เกียรติเป็นประธานโดยมี ดร.เอกสิทธิ์ พัชรวงศ์ศักดา และ ดร.ธนภัทร ฆังคะจิตร เป็นกรรมการในการสอบวิทยานิพนธ์ ซึ่งได้กรุณาตรวจแก้ไขวิทยานิพนธ์ฉบับนี้ให้ถูกต้องสมบูรณ์ยิ่งขึ้น ตลอดจน นางสาวกุลธิดา รอดบุญ รวมถึงเจ้าหน้าที่บัณฑิตมหาวิทยาลัยธุรกิจบัณฑิตย์ ทุกท่านที่ให้ความสะดวกด้านอำนวยการ และประสานงาน ในการทำวิทยานิพนธ์ให้ผู้เขียนตลอดมาตลอดจนค้นคว้าหาข้อมูลในการจัดทำวิทยานิพนธ์ของผู้เขียนครั้งนี้สำเร็จลุล่วงไปด้วยดี

ท้ายนี้ผู้เขียนขอน้อมรำลึกถึงอำนาจบารมีของคุณพระศรีรัตนตรัย และสิ่งศักดิ์สิทธิ์ทั้งหลายที่อยู่ในสากลโลก อันเป็นที่พึ่งให้ผู้เขียนมีสติปัญญาในการจัดทำวิทยานิพนธ์ให้สำเร็จลุล่วงไปด้วยดี ผู้เขียนขอให้เป็นกตเวทิตาแด่บิดา มารดา ครอบครัวของผู้เขียน ตลอดจนผู้เขียนหนังสือ และบทความต่าง ๆ ที่ให้ความรู้แก่ผู้เขียนจนสามารถให้วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยดี

ศักรินทร์ นุ้ยพิน

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

สารบัญภาพ (ต่อ)

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในอุตสาหกรรมเพลงสิ่งหนึ่งที่ผู้บริหารหรือเจ้าของผลงานเพลงต้องการทราบคือผลงานเพลงที่เผยแพร่ออกไปนั้นมีผู้รับฟังมากน้อยเพียงใด ในปัจจุบันมีหลายช่องทางที่ผู้ผลิตเพลงสามารถเผยแพร่ผลงานออกไปสู่ผู้บริโภคหรือผู้ฟัง และยังสามารถตรวจสอบได้ด้วยว่าเพลงเหล่านั้นถูกเปิดฟังไปแล้วกี่ครั้ง ช่องทางเหล่านั้น อาทิ Apple Music, JOOX, Spotify, YouTube เป็นต้น แต่หากย้อนไปในอดีตช่องทางหนึ่งที่ผู้ผลิตเพลงสามารถเผยแพร่ผลงานออกไปยังผู้ฟังได้คือสถานีวิทยุ ซึ่งปัจจุบันก็คงยังเป็นช่องทางหนึ่งที่ยังนิยมใช้กันอยู่ แต่การที่จะตรวจสอบว่าเพลงที่เผยแพร่ไปนั้นถูกเปิดไปแล้วกี่ครั้งนั้นไม่สามารถตรวจสอบได้

หากต้องการตรวจสอบผลงานเพลงที่เผยแพร่ผ่านทางสถานีวิทยุ วิธีการทั่วไปใช้วิธีการสุ่มเลือกช่วงเวลาที่สนใจเพื่อตรวจสอบโดยการฟังของพนักงานในอุตสาหกรรมนั้น แต่เนื่องจากมีข้อจำกัดอยู่หลายอย่าง อาทิ ความยากในการวิเคราะห์ข้อมูลเสียงที่มีระยะเวลานาน โดยอาศัยเพียงแค่การฟังของคนเพียงอย่างเดียว ประสบการณ์ของผู้ฟังมีผลต่อความถูกต้องของการวิเคราะห์ข้อมูลโดยตรง ความคล้ายคลึงกันของเพลงก็เป็นปัจจัยหนึ่งที่มีผลต่อการวิเคราะห์ข้อมูล เป็นต้น ดังนั้นจึงมีการนำวิธีการสืบค้นข้อมูลดนตรี (Music information retrieval) [3] มาช่วยในการตรวจสอบข้อมูลเพลงแทนการฟัง

การสืบค้นข้อมูลด้วยเสียงเป็นการนำเอาความรู้จากหลายวิชามาประยุกต์ อาทิ การประมวลผลสัญญาณดิจิทัล (Digital signal processing) การจดจำรูปแบบ (Pattern recognition) ทฤษฎีทางดนตรี (Music theory) เป็นต้น เทคนิคการรู้จำดนตรี (Music recognition technique) [4] ถูกนำมาประยุกต์เพื่อใช้ดึงคุณลักษณะเฉพาะเพื่อใช้เป็นตัวแทนเนื้อหาของเสียงและเปรียบเทียบเพื่อตรวจสอบความคล้ายคลึงการของเสียง คุณลักษณะเฉพาะที่ถูกเลือกเป็นตัวแทนเนื้อหาของเสียงมีหลายหลายรูปแบบทั้งนี้ขึ้นอยู่กับวัตถุประสงค์การใช้งาน หนึ่งในคุณลักษณะเฉพาะที่ถูกนำมาใช้ในการตรวจสอบความคล้ายคลึงกันของเสียงคือ ลายนิ้วมือทางเสียง (Audio fingerprint) [5]

ลายนิ้วมือของเสียงคือคุณลักษณะเฉพาะของข้อมูลที่สามารถระบุถึงเนื้อหาของเสียงเช่นเดียวกับลายนิ้วมือของคนที่ต่างกันไปในแต่ละบุคคล จึงทำให้สามารถระบุตัวตนของคนนั้นได้

ข้อมูลเสียงที่อินพุต (Input) จะถูกแปลงเป็นลายนิ้วมือเสียงเพื่อนำไปเปรียบเทียบกับลายนิ้วมือของเสียงที่อยู่ในฐานข้อมูลอ้างอิง หากพบว่าเนื้อหาตรงกันก็จะสามารถระบุได้ว่าเสียงนั้นเป็นเสียงเดียวกัน ซึ่งเทคนิคนี้ถูกนำไปประยุกต์ใช้ในแอพพลิเคชั่นต่างๆ อาทิ ระบบกรองไฟล์เสียงในระบบแชร์ไฟล์ [6] ระบบห้องสมุดเสียงอัตโนมัติ (Automatic Music Library organization) [7] แอพค้นหาเสียงหรือเพลง [4], [8] เป็นต้น

งานวิจัยนี้มุ่งเน้นที่จะแก้ปัญหาการค้นหาข้อมูลเพลงจากสัญญาณเสียงวิทยุ ซึ่งดูแล้วสามารถประยุกต์ใช้ลายนิ้วมือของเสียง [4] เพื่อใช้ในการแก้ปัญหานี้ได้ แต่เนื่องจากลายนิ้วมือของเสียงยังมีปัญหาในการแยกเสียงเพลงที่มีเนื้อเสียง (Tone) ระดับเสียง (Pitch) จังหวะดนตรี (Rhythm) และ เสียงร้อง ที่มีความคล้ายคลึงกันค่อนข้างสูง ในงานวิจัยฉบับนี้จึงนำเสนอวิธีการสกัดคุณลักษณะเฉพาะเพื่อนำมาใช้เป็นตัวแทนของเนื้อหาเสียงโดยใช้โมเดลโครงข่ายทริปเล็ต เนื่องจากมีวิธีการที่ใช้ในการตรวจจับอินสแตนซ์ที่คล้ายกัน ซึ่งก่อนหน้านี้มีหลายงานวิจัยที่นำเสนอการใช้โมเดลโครงข่ายสยามเพื่อสร้างตัวแทนและทำความเข้าใจเกี่ยวกับภาพและเสียง [8], [9], [10] ดังนั้นนี่จึงเป็นแนวทางในการนำโครงข่ายทริปเล็ตไปประยุกต์ใช้เพื่อสกัดคุณลักษณะของเสียงและนำคุณสมบัติที่ได้มาใช้ในการตรวจสอบความคล้ายคลึงของเสียงเพลง เพื่อที่จะสามารถระบุถึงข้อมูลเพลงของเสียงเพลงนั้นได้

1.2 วัตถุประสงค์ของงานวิจัย

1.2.1 วิเคราะห์คุณลักษณะสำคัญของเสียง (Audio features) ที่สัมพันธ์ระหว่างเสียงมนุษย์และเสียงดนตรี อาทิ จังหวะเสียงดนตรี (Rhythm หรือ Tempo), เสียงประสาน (Harmonic), ค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล (Mel Frequency Cepstral Coefficient), ค่าลอการิทึมของเสียงสเปคตรัมบนสเกลเมล (Log scaled Mel-spectrogram)

1.2.2 สร้างคุณลักษณะสำคัญเฉพาะของเสียงเพื่อใช้เป็นตัวแทนเนื้อหาของเสียงเพลง

1.3 ขอบเขตงานวิจัย

งานวิจัยนี้มุ่งเน้นศึกษาแนวทางการวิเคราะห์คุณลักษณะของเสียงเพื่อหาคุณลักษณะสำคัญเฉพาะของเสียงเพื่อใช้เป็นตัวแทนเนื้อหาของเสียงโดยใช้เทคนิคการเรียนรู้ด้วยตัวอย่างจำนวนน้อย และจะใช้ตัวอย่างเพลงจำนวน 300 เพลง โดยตัดเป็นคลิปเสียงความยาวคลิปละ 2 นาที

1.4 สมมติฐานของงานวิจัย

1.4.1 ค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลเป็นวิธีการแยกคุณลักษณะสำคัญของเสียงที่ส่งผลกับเสียงของมนุษย์อย่างมีนัยสำคัญ

1.4.2 เทคนิคการเรียนรู้ด้วยตัวอย่างจำนวนน้อยสามารถสร้างคุณลักษณะสำคัญเฉพาะของเสียงเพื่อใช้เป็นตัวแทนเนื้อหาของเสียงเพลงได้

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 ได้วิธีการเลือกและวิเคราะห์คุณลักษณะสำคัญที่ส่งผลต่อการสร้างคุณลักษณะสำคัญเฉพาะของเสียง

1.5.2 โมเดลที่ใช้แยกคุณลักษณะสำคัญเฉพาะที่ได้สามารถนำไปใช้เพื่อพัฒนาระบบที่สามารถค้นหาข้อมูลเพลงได้

1.6 นิยามศัพท์

1.6.1 เพลง หรือ Song ถ้อยคำที่นักประพันธ์เรียงร้อยหรือเรียบเรียงขึ้น ซึ่งประกอบด้วยเนื้อร้อง ทำนอง จังหวะ

1.6.2 ข้อมูลเพง ให้ความหมายถึงชื่อเพลง, ศิลปิน, และแนวเพลง

1.6.3 คุณลักษณะสำคัญของเสียง หรือ Audio features เป็นคุณลักษณะสำคัญของเสียงนั้น เช่น จังหวะดนตรีและเสียงประสาน

1.6.4 คุณลักษณะสำคัญเฉพาะของเสียง หรือ Local features เป็นคุณลักษณะสำคัญเฉพาะ ที่สามารถแสดงถึงเนื้อหาของเสียงนั้นได้

1.6.5 การเรียนรู้ด้วยตัวอย่างจำนวนน้อย หรือ Few-shot learning เป็นเรียนรู้โดยอาศัยข้อมูลจำนวนน้อยในแต่ละคลาส (classes) ในชุดข้อมูล ตัวอย่างเช่น หากต้องการจำแนกภาพสุนัขหรือแมว หากในชุดข้อมูลมีภาพสุนัขเพียง 10 ภาพ และแมวเพียง 10 ภาพ เป็นต้น

1.6.6 โครงข่ายสยาม หรือ Siamese network เป็นสถาปัตยกรรมโครงข่ายที่ประกอบด้วยโครงข่ายย่อยสองโครงข่ายที่เหมือนกันทั้งโครงสร้างและน้ำหนัก มักใช้ในงานที่ต้องการหาความสัมพันธ์ของชุดข้อมูลสองชุด

1.6.7 โครงข่ายทรีพเล็ท หรือ Triplet network เป็นสถาปัตยกรรมโครงข่ายที่ประกอดด้วยโครงข่ายย่อยสามโครงข่ายที่เหมือนกันทั้งโครงสร้างและนำ้หนัก มักใช้ในงานที่ต้องการหาความสัมพันธ์ของชุดข้อมูลสามชุด

บทที่ 2
แนวคิด ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีดนตรี

ดนตรี (Music) เป็นสิ่งที่เกิดจากความคิดและจินตนาการของผู้แต่ง (Composers) ที่ใช้เครื่องมือหรืออุปกรณ์ต่างๆ มาประกอบเป็นดนตรี โดยนำเอาระดับเสียงที่ต่างกัน มาจัดเป็นกลุ่มตามกฎเกณฑ์ทางดนตรี เพื่อนำมาเรียบเรียงเป็นผลงานที่เรียกว่า บทเพลง หรือเพลง ในการผลิตหรือสร้างผลงานทางดนตรีนั้นมีองค์ประกอบที่สำคัญประกอบไปด้วยปัจจัยดังนี้

2.1.1 เสียง (Tone)

เรื่องเกี่ยวกับเสียงเป็นสาขาหนึ่งทางวิทยาศาสตร์เรียกสวนศาสตร์ (Acoustics) ซึ่งเป็นวิชาที่ศึกษาเกี่ยวกับเสียง และเนื่องจากดนตรีมีความสัมพันธ์กับเสียง ดังนั้นจึงมีความจำเป็นต้องเข้าใจเกี่ยวกับเสียงในเบื้องต้นก่อน เสียงนั้นเกิดจากการสั่นสะเทือนของวัตถุ ซึ่งมีการอัดและขยาวตัวของคลื่น โดยเสียงที่ได้ยินนั้นจะมีลักษณะเป็นอย่างไรจะขึ้นอยู่กับแหล่งกำเนิดเสียง และจำนวนรอบต่อวินาทีของการสั่นสะเทือนที่ทำให้เกิดเสียงนั้น คุณสมบัติของเสียงประกอบด้วย

2.1.1.1 ระดับเสียง (Pitch) คือความสูงต่ำของเสียงในเชิงกายภาพ หากความถี่ของการสั่นสะเทือนเป็นไปอย่างรวดเร็ว จะทำให้เกิดเสียงสูง ถ้าความถี่ของการสั่นสะเทือนเป็นลักษณะช้าก็จะทำให้เกิดเสียงต่ำ ซึ่งหูของมนุษย์สามารถแยกเสียงได้ตั้งแต่ระดับความถี่ของการสั่นสะเทือน 16 ครั้งต่อวินาที จนถึง 20,000 ครั้งต่อวินาที

2.1.1.2 ความสั้น-ยาวของเสียง (Duration) คือความแตกต่างกันในเรื่องของความสั้นยาวของเสียง บางครั้งจะได้ยินในลักษณะของการลากเสียงยาวๆ หรือไม่ก็เป็นลักษณะห้วนๆ สั้นๆ

2.1.1.3 ความดัง-เบาของเสียง (Dynamics) คือความแตกต่างกันในเรื่องของความดังเบาของเสียง ตัวอย่างเช่น บางครั้งจะได้ยินการบรรเลงเพลงที่มีความดัง อึกทึกครึกโครม หรือในบางครั้งก็จะได้ยินเสียงดนตรีที่นุ่มนวลหรือแผ่วเบา

2.1.1.4 สีสันของเสียง (Tone color หรือ Timbre) คือความแตกต่างที่เกิดจากแหล่งกำเนิดเสียงที่แตกต่างกัน เช่น เสียงเครื่องดนตรีชนิดต่างๆ และรวมไปถึงเสียงร้องของมนุษย์ด้วย ตัวอย่างเช่น ในเพลงหนึ่งๆ หากขับร้องโดยผู้ชายก็จะได้ความรู้สึกที่แตกต่างจากการขับร้องโดยผู้หญิง หรือในการบรรเลงดนตรี หากเป็นการบรรเลงเดี่ยวก็จะมีความแตกต่างไปจากการบรรเลงเป็นวงหรือบรรเลงโดยเครื่องดนตรีที่ต่างชนิดกัน

2.1.2 จังหวะ (Rhythm)

ความสั้นยาวของเสียงที่ทำให้เกิดห่วงทำนองที่สามารถสะท้อนอารมณ์ความรู้สึกได้หลากหลาย และอาจจะหมายถึงจังหวะของเครื่องดนตรีประเภทเครื่องตีอย่างกลอง หรือเครื่องเคาะต่างๆ ที่ถูกนำมาใช้เป็นเครื่องช่วยเน้นย้ำจังหวะที่ถูกสร้างขึ้นเพิ่มเติมให้มีความน่าสนใจ โดยการรับรู้ด้านจังหวะ แบ่งออกได้ 3 ลักษณะที่ตรงกันข้ามกัน

2.1.2.1 จังหวะที่ปกติสม่ำเสมอ (Regular) ให้อารมณ์ที่เรียบง่าย ตรงข้ามกับจังหวะที่ไม่ปกติสม่ำเสมอ (Irregular) ให้อารมณ์ที่อึดอัด คับข้อง ซึ่งจังหวะในลักษณะนี้อาจจะเพิ่มสีสันในบทเพลงได้ หากผู้แต่งหรือผู้บรรเลงมีศิลปะในการสร้างดนตรีให้ความคับข้องกลายเป็นเสน่ห์ของเพลงเช่น ดนตรีแจ๊ส ที่มีลักษณะของเครื่องดนตรีหยอกล้อกับเสียงกลอง เมื่อบรรเลงเดี่ยว

2.1.2.2 จังหวะหนัก (Strong) ให้อารมณ์ที่หนักแน่น มั่นคง สง่างาม กับจังหวะเบา (Weak) ที่ให้ความรู้สึกอ่อนไหว ไม่มั่นคง

2.1.2.3 จังหวะยาว (Long) ให้ความรู้สึกร่าเริง สดใส

2.1.3 ทำนอง (Melody)

การเรียบเรียงชองเสียงที่มีความแตกต่างกันของระดับเสียงและความยาวของเสียง โดยทั่วไปแล้วดนตรีประกอบด้วยทำนอง ซึ่งเป็นองค์ประกอบที่ง่ายต่อการจดจำ มีหลายลักษณะที่แตกต่างกันออกไป มีองค์ประกอบได้แก่

2.1.3.1 จังหวะของทำนอง (Melodic Rhythm) ความสั้นยาวของระดับเสียง แต่ละเสียงที่ประกอบกันเป็นทำนอง

2.1.3.2 มิติของทำนอง (Melodic Dimension) มีด้วยกัน 2 ส่วนคือความยาว และช่วงกว้าง

2.1.3.3 ความยาว (Length) บางครั้งทำนองอาจจะสั้นๆ เป็นส่วน ซึ่งส่วนที่เล็กที่สุดหรือสั้นที่สุด หรือบางครั้งเป็นทำนองที่ยาวมาก

2.1.3.4 ช่วงกว้าง (Range) ระยะห่างระหว่างระดับเสียงต่ำสุดจนถึงระดับเสียงสูงสุด

2.1.3.5 ช่วงเสียงของทำนอง (Register) ทำนองเพลงอาจจะอยู่ในช่วงเสียงหนึ่งเช่น ช่วงเสียงต่ำ กลาง หรือสูง บางครั้งทำนองอาจจะเคลื่อนที่จากช่วงเสียงหนึ่งไปยังอีกช่วงเสียงหนึ่งก็ได้

2.1.3.6 ทิศทางของทำนอง (Direction) ทิศทางการเคลื่อนที่ของทำนอง กล่างคือทำนองอาจจะเคลื่อนที่ไปในหลายทิศทาง เช่น เคลื่อนที่ขึ้น เคลื่อนที่ลง หรืออยู่กับที่ โดยปกติทำนองมักจะเคลื่อนที่ขึ้นจุดสูงสุด เมื่อเนื้อหาของเพลงดำเนินไปถึงจุดสำคัญที่สุด โดยการเคลื่อนที่ของทำนองอาจจะเป็นลักษณะกระโดด (Disjunct Progression) หรือเรียงกันไป (Conjunct Progression) ซึ่งบทเพลงจะน่าสนใจ น่าฟัง ขึ้นอยู่กับผลรวมของคุณสมบัติต่างๆ ของทำนอง โดยทั่วไปทำนองที่เป็นหลักในบทเพลงหนึ่งจะเรียกว่าทำนองหลัก (Main Theme) ในแต่ละบทเพลงอาจจะมีทำนองหลักได้มากกว่า 1 ทำนอง

2.1.4 เสียงประสาน (Harmony)

องค์ประกอบของดนตรีที่เกิดขึ้นจากการผสมผสานของเสียงมากกว่าหนึ่งแนวเสียง เสียงประสานเป็นองค์ประกอบดนตรีที่สลับซับซ้อนกว่าจังหวะและทำนองแสดงถึงความประณีตในการประพันธ์ อย่างไรก็ตามในบางวัฒนธรรมอาจจะไม่พบการประสานเสียงของดนตรีเลย เช่น ดนตรีพื้นเมืองหรือดนตรีพื้นบ้านที่มีความเรียบง่ายของการประพันธ์ ซึ่งเป็นดนตรีที่แสดงถึงเอกลักษณ์ของตนเอง การประสานเสียงนั้นมี 2 ลักษณะคือ การประสานเสียงที่มีลักษณะของเสียงที่กลมกลืนกันและไม่กลมกลืนกัน

2.1.4.1 เสียงประสานที่กลมกลืน (Consonance) เสียงประสานในลักษณะนี้เมื่อฟังแล้วจะทำให้เกิดความรู้สึกกลมกล่อมสบายหูสามารถพบได้ในหลายวัฒนธรรมดนตรี

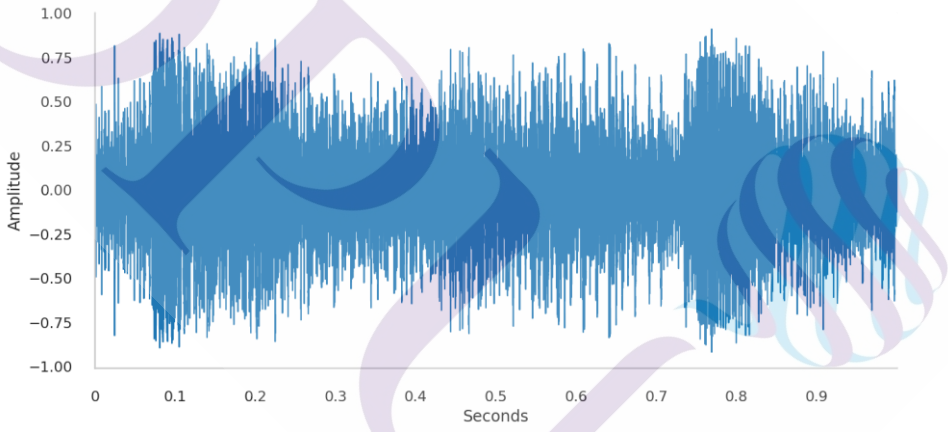
2.1.4.2 เสียงประสานที่ไม่กลมกลืน (Dissconance) เสียงประสานในลักษณะนี้เมื่อฟังแล้วจะทำให้เกิดความรู้สึกขัดหู ดึงเครียด พบในวัฒนธรรมดนตรีตะวันตกเพียงเท่านั้น

2.2 การประมวลผลเสียง (Acoustic Process)

การประมวลผลเสียงถือเป็นพื้นฐานในการวิเคราะห์ข้อมูลทางเสียง มีขั้นตอนหลักๆ อันได้แก่ การแทนค่าข้อมูลเสียงจากข้อมูลในลักษณะคลื่นเสียงไปเป็นสัญญาณดิจิทัลเพื่อใช้ในการประมวลผล, การดึงคุณลักษณะการกระจายตัวของความถี่จากคลื่นเสียง, และการแสดงภาพของเสียง

2.2.1 คลื่นเสียง (Sound Wave)

การรับรู้หรือได้ยินเสียงนั้นเกิดจากการถ่ายทอดพลังงานจากการสั่นสะเทือนของแหล่งกำเนิดเสียงผ่านโมเลกุลของตัวกลางไปยังผู้รับ อย่างเช่นหูของมนุษย์ที่สามารถรับรู้การสั่นสะเทือนของโมเลกุลเหล่านี้และได้ทำการเปลี่ยนแปลงออกมาในรูปแบบต่างๆ ซึ่งในการแทนค่าหรือแสดงค่าของคลื่นเสียงนั้นจะแสดงให้เห็นการเปลี่ยนแปลงของความดันอากาศในช่วงเวลาต่างๆ ดังตัวอย่างในภาพที่ 2.1



ภาพที่ 2.1 ตัวอย่างสัญญาณเสียงในรูปแบบคลื่นไซน์ความยาว 1 วินาที

2.2.2 การทำค่าเฉลี่ยเป็นศูนย์ (Zero Mean)

เนื่องจากสัญญาณเสียงที่ได้ในแต่ละครั้งจากการบันทึก จะมีค่าตรงแกนกลางที่สูงหรือต่ำกว่าศูนย์ ดังนั้นจึงทำให้ในการวิเคราะห์ข้อมูลเป็นไปได้ยาก เพื่อให้ง่ายต่อการวิเคราะห์และประมวลผลสัญญาณ จึงต้องมีการปรับสัญญาณที่อยู่นอกแกนศูนย์กลับเข้าสู่แกนศูนย์ โดยใช้สมการที่ (2-1)

$$signal = signal - mean(signal) \quad (2\text{-}1)$$

2.2.3 การลดทอนสัญญาณรบกวน (Noise Reduction)

การลดทอนสัญญาณรบกวนเป็นกระบวนการปรับแต่งความถี่ของสัญญาณให้มีลักษณะตามที่ต้องการโดยต้องการให้มีเฉพาะความถี่ต่ำ ความถี่สูง ช่วงความถี่บางช่วง หรือบางช่วงความถี่ที่ไม่ต้องการให้แสดงผลออกไป สามารถทำได้ดังนี้

2.2.3.1 การเน้นล่วงหน้า (Pre-emphasis) เป็นกระบวนการทำให้อัตราส่วนของสัญญาณต่อสัญญาณรบกวนมีค่าคงที่ ตลอดทุกช่วงความถี่ โดยนำสัญญาณผ่านวงจรกรองดิจิทัลอับดับที่หนึ่ง (First order digital filter) มีผลทำให้ลดผลกระทบจากสัญญาณรบกวนความถี่ต่ำ และทำให้มีอัตราส่วนสัญญาณเสียงต่อสัญญาณรบกวน (Signal to noise ratio: SNR) สูงขึ้น ดังสมการที่ (2-3) โดยกำหนดให้ค่า a จะมีค่าอยู่ระหว่าง 0.9 ถึง 1

$$H(z) = 1 - aZ^{-1} \quad (2\text{-}2)$$

$$\tilde{s}(n) = s(n) - as(n-1) \quad (2\text{-}3)$$

เมื่อ a เป็นค่าสัมประสิทธิ์ของวงจรกรอง

$\tilde{s}(n)$ เป็นค่าของสัญญาณเสียงเอาต์พุต (Output) ผ่านกรรมวิธีการเน้นล่วงหน้า

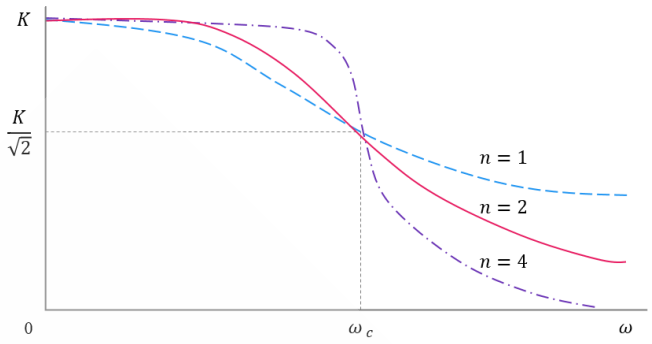
$s(n)$ เป็นค่าของสัญญาณเสียงอินพุต (Input) ที่ n

และ $s(n-1)$ เป็นค่าของสัญญาณเสียงอินพุตที่ $n-1$

2.2.3.2 วงจรกรองความถี่ต่ำแบบบัตเตอร์เวิร์ธ (Low Pass Butterworth Filter) เป็นวงจรกรองความพี่ที่มีคุณลักษณะเฉพาะที่ใกล้เคียงกับวงจรกรองความถี่ต่ำทางอุดมคติ โดยยอมให้ช่วงความถี่ที่ผ่านได้ มีขนาดของการเปลี่ยนแปลงเท่าเทียมตลอดทั้งย่านความถี่ที่ยอมให้ผ่านได้ โดยที่การตอบสนองเชิงขนาดของสัญญาณมีค่าตามสมการที่ (2-4)

$$H(j\omega) = \frac{K}{\sqrt{1 + (\omega/\omega_c)^{2n}}} \quad (2\text{-}4)$$

เมื่อ n เป็นค่าอันดับ (Order) ของวงจรกรองความถี่



ภาพที่ 2.2 ภาพแสดงการตอบสนองของวงจรกรองความถี่ต่ำแบบบัตเตอร์เวิร์ธ เมื่อเลือกใช้ n ที่ต่างกัน

2.2.4 การวิเคราะห์แบบหน้าต่าง (Windowing)

เนื่องด้วยสัญญาณเสียงมีลักษณะที่ไม่คงที่และมีการเปลี่ยนแปลงอย่างช้าๆ ไปตามเวลา ดังนั้นในการวิเคราะห์จะมีการกำหนดกรอบเพื่อแบ่งสัญญาณออกเป็นช่วงสั้นๆ ในการกำหนดขนาดกรอบของสัญญาณเสียงต้องมีความเหมาะสมไม่ควรสั้นเกินกว่าช่วงหนึ่งคาบของสัญญาณเสียงในช่วงที่สนใจ ซึ่งเงื่อนไขนี้จะมีผลต่อค่าเฟรมเรท (Frame Rate) ซึ่งเป็นจำนวนครั้งต่อวินาทีที่ทำการวิเคราะห์สัญญาณเสียง และในการขยับกรอบสัญญาณไปเป็นคาบๆ ตามแกนเวลา ตามปกติเฟรมเรทจะมีความเหลื่อมกันประมาณ 1/2 ถึง 1/3 ของขนาดกรอบสัญญาณ เพื่อให้ข้อมูลที่วิเคราะห์มีความต่อเนื่องกัน ฟังก์ชันหน้าต่างที่นิยมใช้กันตัวอย่างเช่น

2.2.4.1 ฟังก์ชันหน้าต่างสี่เหลี่ยม (Rectangular Window) ดังสมการที่ (2-5)

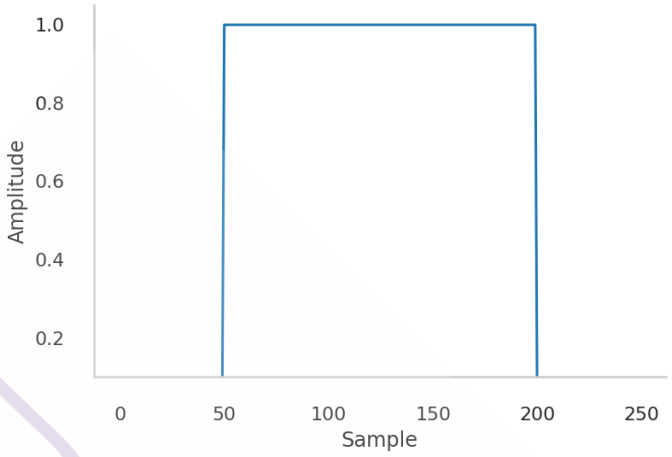
$$w(n) = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & Otherwise \end{cases} \qquad (2\text{-}5)$$

เมื่อ $w(n)$ เป็นผลลัพธ์ของฟังก์ชันกรอบตำแหน่งที่ n

N เป็นความกว้างของหน้าต่าง

และ n เป็นข้อมูลในหน้าต่าง มีค่าตั้งแต่ 0 จนถึง $\mathrm{N}-1$

ซึ่งมีลักษณะดังภาพที่ 2.3



ภาพที่ 2.3 ฟังก์ชันกรอบสัญญาณแบบหน้าต่างสี่เหลี่ยม

2.2.4.2 ฟังก์ชันหน้าต่างแฮมมิ่ง (Hamming Window) ดังสมการที่ (2-6)

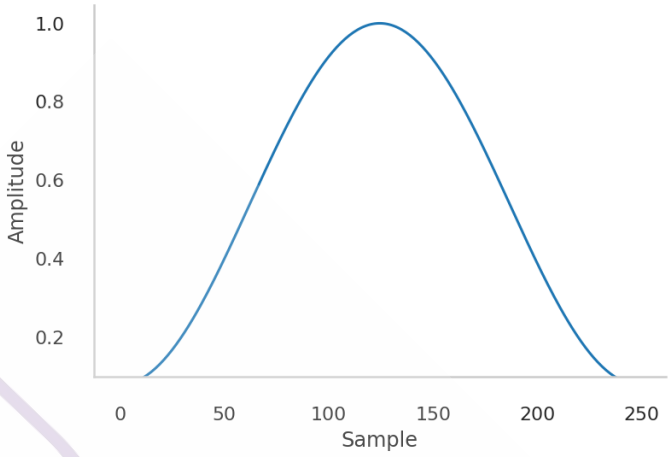
$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), & 0 \le n \le N-1 \\ 0, & Otherwise \end{cases} \quad (2\text{-}6)$$

เมื่อ $w(n)$ เป็นผลลัพธ์ของฟังก์ชันกรอบตำแหน่งที่ n

N เป็นความกว้างของหน้าต่าง

และ n เป็นข้อมูลในหน้าต่าง มีค่าตั้งแต่ 0 จนถึง $N-1$

ซึ่งมีลักษณะดังภาพที่ 2.4



ภาพที่ 2.4 ฟังก์ชันกรอบสัญญาณแบบหน้าต่างแฮมมิ่ง

2.2.4.3 ฟังก์ชันหน้าต่างแฮนนิ่ง (Hanning Window) ดังสมการที่ (2-7)

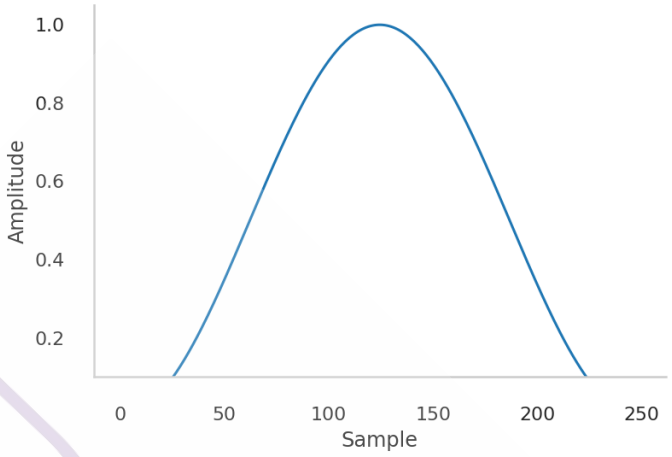
$$w(n) = \begin{cases} 0.5\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right], & 0 \le n \le N-1 \\ 0, & Otherwise \end{cases} \quad (2\text{-}7)$$

เมื่อ $w(n)$ เป็นผลลัพธ์ของฟังก์ชันกรอบตำแหน่งที่ n

N เป็นความกว้างของหน้าต่าง

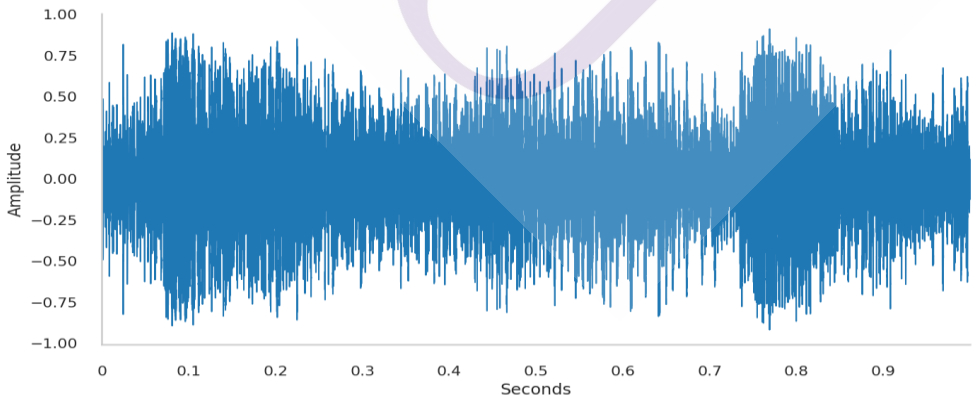
และ n เป็นข้อมูลในหน้าต่าง มีค่าตั้งแต่ 0 จนถึง $N-1$

ซึ่งมีลักษณะดังภาพที่ 2.5

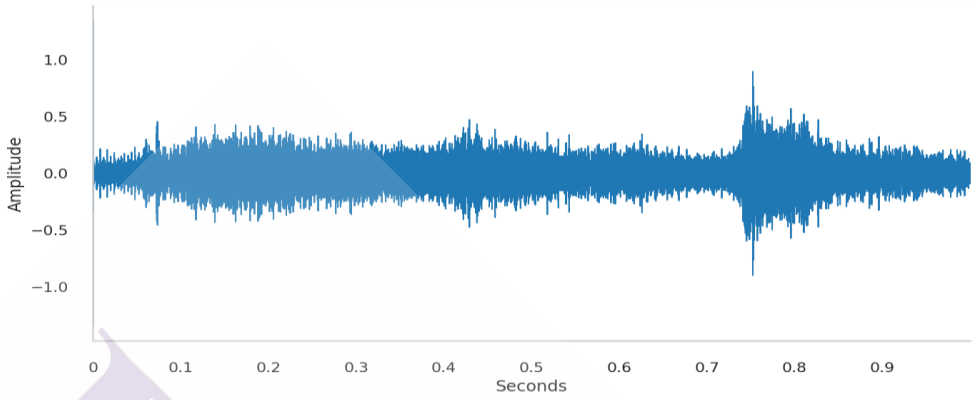


ภาพที่ 2.5 ฟังก์ชันกรอบสัญญาณแบบหน้าต่างแฮนนิ่ง

สำหรับในการประมวลผลสัญญาณเสียงฟังก์ชันหน้าต่างในแบบแฮมมิ่ง และแฮนนิ่ง เหมาะสมที่สุด เพราะสามารถเน้นสัญญาณเสียงในกรอบที่กำลังพิจารณาให้มีความสำคัญสูงสุดโดยจะลดความสำคัญของสัญญาณเสียงที่อยู่ในกรอบรอบข้าง ทำให้เสียงที่ผ่านการวางกรอบสัญญาณยังคงความครบถ้วนของข้อมูล



(ก)



(ข)

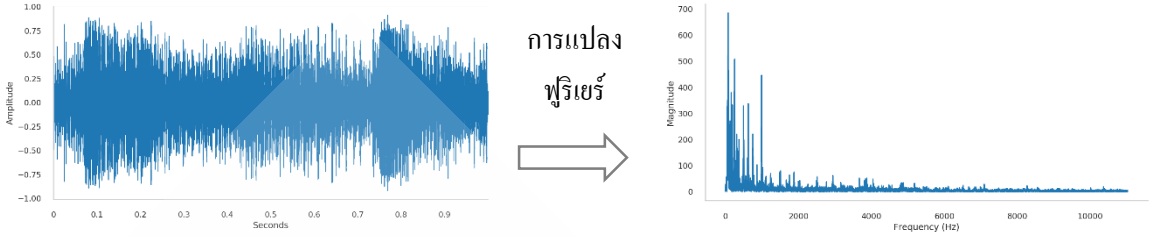
ภาพที่ 2.6 การประมวลผลสัญญาณเสียงเบื้องต้น

(ก) ก่อนประมวลผลสัญญาณเสียงเบื้องต้น

(ข) หลังประมวลผลสัญญาณเสียงเบื้องต้น

2.2.5 การวิเคราะห์ค่าเชิงสเปกตรัม (Spectrum Analysis)

ในการประมวลผลเสียงในโดเมนเวลา (Time Domain) นั้นค่อนข้างทำได้อยาก แต่หากแปลงจากสัญญาณเสียงให้มาอยู่ในโดเมนความถี่ (Frequency Domain) ก็จะสามารถประมวลผลได้ง่ายกว่า ดังนั้นจึงเกิดการดำเนินการทางคณิตศาสตร์เพื่อเปลี่ยนแปลงสัญญาณเสียงให้อยู่ในโดเมนความถี่ กระบวนการนี้เรียกว่า การแปลงฟูรีเยร์ (Fourier Transform) ซึ่งกำเนิดโดย ฟูริเยร์ (Fourier) นักคณิตศาสตร์ชาวฝรั่งเศสในปี 1820 แนวคิดพื้นฐานของการแปลงฟูรีเยร์มาจากสมมติฐานที่ว่าสัญญาณใดๆ ก็ตาม โดยปกติแล้วจะสามารถแยกองค์ประกอบออกเป็นกลุ่มของสัญญาณรูปคลื่นไซน์ (Sine Wave) หลายๆ ความถี่ ซึ่งเกิดจากกระบวนการโปรเจคชันสัญญาณบนกลุ่มของฟังก์ชันพื้นฐาน โดยแต่ละฟังก์ชันพื้นฐานจะสร้างมาจากสัญญาณรูปคลื่นไซน์ที่มีความถี่เดียว ค่าที่ได้จากการโปรเจคชันที่ความถี่หนึ่งๆ จะเป็นตัวบ่งชี้ถึงความใกล้เคียงของสัญญาณกับฟังก์ชันพื้นฐานรูปคลื่นไซน์ที่ความถี่นั้น จากนั้นนำมาจัดเรียงให้อยู่ในรูปสเปคตรัมความถี่ ผลจากการแปลงฟูรีเยร์ของสัญญาณใดๆ จะแสดงถึงองค์ประกอบความถี่ทั้งหมดของสัญญาณนั้นๆ



ภาพที่ 2.7 การแปลงฟูริเยร์เป็นการแสดงองค์ประกอบในโดเมนความถี่ของสัญญาณเสียง

การแปลงฟูรีเยร์เป็นการวิเคราะห์องค์ประกอบในโดเมนความถี่ของสัญญาณที่นำมาใช้ประโยชน์เป็นอย่างมากสำหรับสัญญาณคงที่ (Stationary Signal) ในขณะที่สัญญาณส่วนใหญ่ที่พบในโลกความจริง ค่อนข้างซับซ้อนและมีองค์ประกอบในโดเมนความถี่ที่มีการเปลี่ยนแปลงตลอดเวลา ในกรณีนี้การใช้กราฟรูปคลื่นไซน์อย่างง่ายมาแทนเป็นฟังก์ชันพื้นฐานของสัญญาณอาจไม่ง่ายนัก แต่ในขณะเดียวกันการอธิบายคุณลักษณะของสัญญาณด้วยสเปคตรัมความถี่เพียงอย่างเดียวอาจจะไม่เพียงพอ ดังนั้นการแปลงสัญญาณทั้งในเชิงเวลาและความถี่พร้อมกัน จึงถูกพัฒนาขึ้นมาใช้ในการอธิบายคุณลักษณะของสัญญาณที่มีองค์ประกอบในโดเมนความถี่ที่เปลี่ยนแปลงตามเวลา โดยจะเรียกการแสดงภาพองค์ประกอบของสัญญาณในลักษณะนี้ว่า สเปกโตรแกรม (Spectrogram) พื้นฐานแล้วการสร้างสเปกโตรแกรมจะอาศัยหลักการหาองค์ประกอบในโดเมนความถี่ของสัญญาณในแต่ละหน้าต่างของเวลา (Time Window) ที่มีการเคลื่อนที่ ดังนั้นสเปกโตรแกรมจะประกอบด้วยข้อมูลขององค์ประกอบในโดเมนความถี่ของสัญญาณ ที่เวลาในขณะใดขณะหนึ่งที่แตกต่างกัน วิธีการแปลงฟูรีเยร์ที่นิยมใช้รูปแบบหนึ่งคือ การแปลงฟูรีเยร์แบบไม่ต่อเนื่อง (Discrete Fourier Transform) ตามสมการที่ (2-8)

$$x_k = \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{2\pi i k n}{N}}, 0 \le n < N, 0 \le k < N \tag{2-8}$$

เมื่อ N เป็นจำนวนตัวอย่างในหนึ่งกรอบสัญญาณ

k เป็นลำดับของกรอบสัญญาณ $k = 1,2,\dots,K$

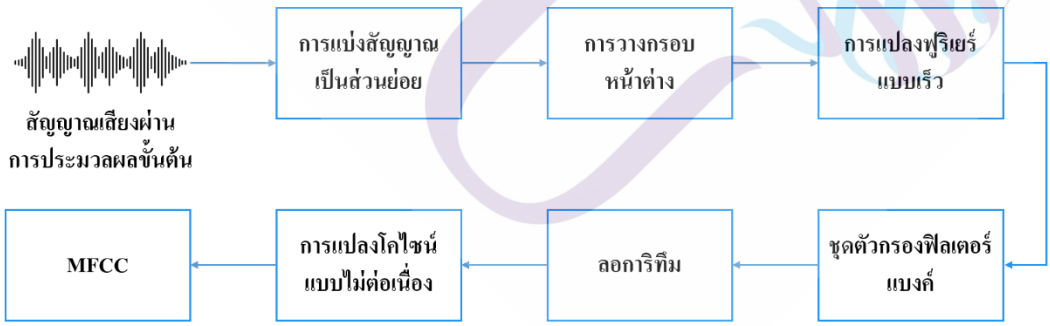
และ x_n เป็นสัญญาณอินพุตของสัญญาณเสียง

2.3 การแยกคุณลักษณะสำคัญของเสียง (Feature Extraction)

การแยกคุณลักษณะสำคัญเป็นการดึงเอาลักษณะเฉพาะของหน่วยเสียงแต่ละหน่วยเสียงที่แตกต่างกันออกมา แล้วให้ระบบทำการรู้จำลักษณะสำคัญของหน่วยเสียงแต่ละหน่วยเสียงไว้ เมื่อสัญญาณที่เข้ามาในภายหลัง มีลักษณะสำคัญที่เหมือนหรือใกล้เคียงกับลักษณะสำคัญของหน่วยเสียงใดๆ ระบบรู้จำจะสามารถบอกได้ว่าเป็นหน่วยเสียงในกลุ่มใด หรือใกล้เคียงกับหน่วยเสียงกลุ่มใดมากที่สุดและสามารถลดจำนวนข้อมูล โดยที่ข้อมูลเสียงจำนวนมากจะถูกแปลงเป็นชุดข้อมูลที่มีจำนวนน้อยลงและยังคงคุณสมบัติสำคัญของข้อมูลเดิมไว้ได้อย่างถูกต้อง

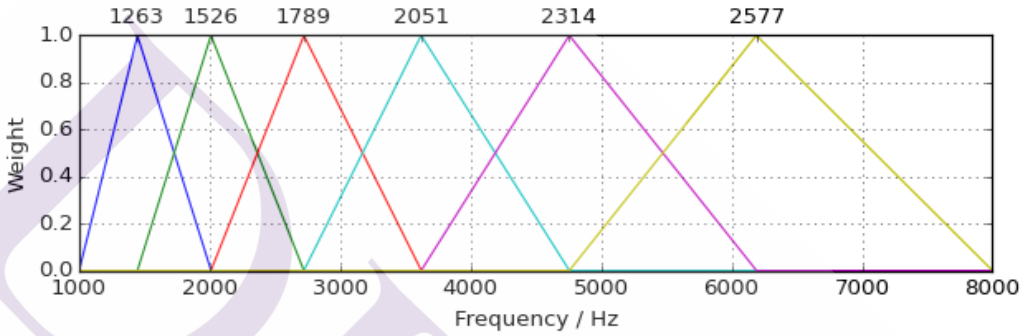
2.3.1 สัมประสิทธิ์เซปสตรัมบนสเกลเมล (Mel Frequency Cepstral Coefficient หรือ MFCC)

เซปสตรัม (Cepstral) เป็นการแปลงโคไซน์แบบไม่ต่อเนื่อง (Discrete Cosine Transform) จากค่าลอการิทึมจากสเปคตรัมสัญญาณในช่วงสั้นๆ สัมประสิทธิ์เซปสตรัมบนสเกลเมลเป็นเทคนิคที่พัฒนามาจากเทคนิคเซปตรัม ด้วยการปรับสเกลของสเปคตรัมให้อยู่บนสเกลที่เหมาะสมสำหรับการได้ยินของมนุษย์ซึ่งสามารถได้ยินเสียงเป็นเชิงเส้นตั้งแต่ 0 – 1,000 Hz โดยการดึงเอาสัญญาณเสียงในช่วงความถี่ต่ำให้มีความสำคัญกว่าช่วงความถี่สูง จึงเกิดการพัฒนาสเกลของสเปคตรัมที่สามารถเก็บรายละเอียดของสัญญาณเสียงในช่วงความถี่ต่ำได้มากกว่า ที่เรียกว่า สเกลเมล (Mel Scale) โดยมีขั้นตอนในการคำนวณเพื่อหาค่าสัมประสิทธิ์เซปตรัมบนสเกลเมลดังในภาพที่ 2.8



ภาพที่ 2.8 ขั้นตอนการคำนวณหาค่าสัมประสิทธิ์เซปตรัมบนสเกลเมล

2.3.1.1 Mel Frequency Filter Bank เป็นขั้นตอนการหาค่าสัมประสิทธิ์เซปสตรัมบนเกลเมล เริ่มต้นจากการนำสัญญาณเสียงมาผ่านการประมวลผลสัญญาณเสียงหลังจากนั้นส่งสัญญาณไปผ่านชุดตัวกรองฟิลเตอร์แบงค์ (Filter Bank) เพื่อเน้นความสำคัญของความถี่ที่อยู่ในช่วงกลางของชุดตัวกรองแต่ละตัวกรอง ชุดตัวกรองฟิลเตอร์แบงค์มีลักษณะตามรูปที่ 2.9



ภาพที่ 2.9 ชุดตัวกรองฟิลเตอร์แบงค์

โดยที่ความถี่กลางของตัวกรองแต่ละชุดนั้น เกิดจากการแปลงค่าความถี่ปรกติ (f) ให้อยู่บนสเกลเมล (f_{mel}) ดังสมการที่ (2-9)

$$f_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{2-9}$$

2.3.1.2 การคำนวณหาพลังงานสเปคตรัมที่ผ่านตัวกรอง ขั้นตอนนี้นำความถี่ที่ได้จากขั้นตอนคำนวณสเปคตรัมมาหาขนาดกำลังสองได้ $|\tilde{x}(k)|^2$ ส่งผ่านชุดตัวกรองแบบสามเหลี่ยมในสเกลเมล เพื่อเน้นความสำคัญของความถี่ที่อยู่ในช่วงกลางของชุดตัวกรองแต่ละตัวกรองตามสมการที่ (2-10)

$$E_j = \sum_{k=0}^{\frac{n}{2}-1} \Phi_j \cdot (k)|\tilde{x}(k)|^2; 0 \le j \le J \tag{2-10}$$

เมื่อ Φ_j เป็นค่าประจำตัวกรองที่ j

และ $\tilde{x}(k)$ เป็นสเปคตรัม

2.3.1.3 การคำนวณสัมประสิทธิ์เซปตรัมบนสเกลเมล (MFCC) ในขั้นตอนนี้จะเป็นการนำค่าลอการิทึมของผลลัพธ์ที่ได้จากชุดตัวกรองมาผ่านการแปลงโคไซน์แบบไม่ต่อเนื่อง (Discrete Cosine Transform หรือ DCT) ทำให้ได้ค่าสัมประสิทธิ์เซปสตรัมบนสเกลมเมล c ลำดับที่ m ตามสมการที่ (2-11)

$$c_m = w_t(m) \sum_{j=1}^{J} \log_{10}(E_j) \cos\left(\frac{\pi}{j}(j-0.5)m\right), \quad m = 0, 1, 2, \dots, J-1 \tag{2-10}$$

เมื่อ $$\mathrm{w}_t(m) = \begin{cases} \frac{1}{\sqrt{J}}, & m = 0 \\ \frac{\sqrt{2}}{J}, & 1 \le m < J \end{cases}$$

2.3.2 สัมประสิทธิ์เซปสตรัมบนสเกลเมลแบบความต่าง (MFCC Delta)

เป็นการหาค่าความแตกต่างระหว่างกรอบสัญญาณขนาด 5 เฟรมโดยนำค่าสัมประสิทธิ์ของ 2 กรอบสัญญาณทางด้านซ้ายลบด้วยค่าสัมประสิทธิ์ของ 2 กรอบสัญญาณทางด้านขวา ดังสมการที่ (2-11)

$$x_i(m) = \sum_{k=-2}^{2} kC_{i-k}(m) \tag{2-11}$$

เมื่อ $m = 1, \dots, M$ โดยที่ M เป็นค่าอันดับของสัมประสิทธิ์ MFCC

$i = 2, \dots, N-2$ โดยที่ N เป็นจำนวนเฟรมของข้อมูลเสียง

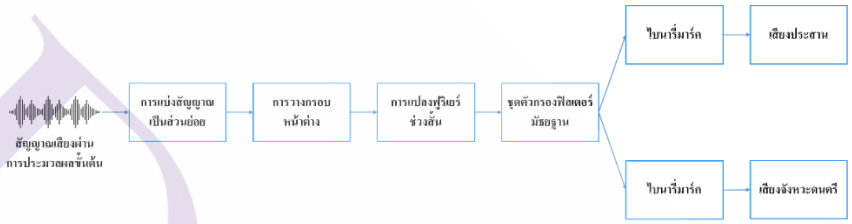
$C_i(m)$ เป็นสัมประสิทธิ์ MFCC

และ $x_i(m)$ เป็นสัมประสิทธิ์ MFCC Delta

2.3.3 ค่าเฉลี่ยของเสียงประสานและจังหวะดนตรี (Mean Harmonic and Percussive)

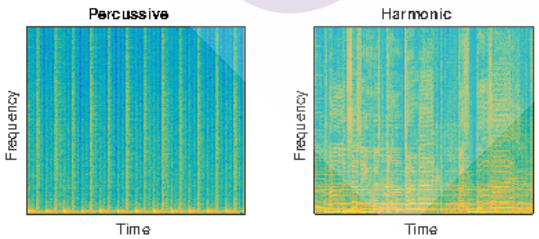
เสียงโดยทั่วไปมีแตกต่างและหลากหลายในเสียง อย่างไรก็ตามหากวิเคราะห์เสียงอย่างหยาบ จะพบว่าในเสียงส่วนใหญ่จะประกอบด้วยข้อมูลเสียงหนึ่งหรือสองอย่างนี้คือ เสียงประสาน (Harmonic) หรือจังหวะดนตรี (Percussive) ซึ่งเสียงประสานเป็นเสียงที่มีระดับเสียง (Pitch) ที่แน่นอนที่เราสามารถร้องตามได้ อย่างเช่นเสียงไวโอลิน ส่วนเสียงจังหวะดนตรีเป็นเสียงที่เกิดจากวัตถุสองอย่าง

ชนกันอย่างเช่น เสียงฉิ่ง ลักษณะเด่นของเสียงจังหวะดนตรีคือจะไม่มีระดับเสียงที่สูงแต่จะชัดเท่ากันเสมอ เสียงส่วนใหญ่ที่พบเจอโดยทั่วไปจะเกิดจากการผสมผลานระหว่างเสียงประสานและจังหวะดนตรี ตัวอย่างเช่น เสียงโน้ตตอนเล่นเปียโนที่ประกอบด้วยเสียงจังหวะดนตรีที่เกิดจากค้อนเคาะสาย (Hammer) กระทบกับสายเปียโน (Strings) และเสียงประสานที่เกิดจากการสั่นสะเทือนของสายเปียโน หากต้องการแยกคุณลักษณะของเสียงประสานและจังหวะดนตรีสามารถทำได้ดังภาพที่ 2.10



ภาพที่ 2.10 ขั้นตอนการคำนวณหาค่าเสียงประสานและจังหวะดนตรี

2.3.3.1 การแปลงฟูริเยร์ช่วงสั้น (Short-time Fourier Transform) ใช้เพื่อแปลงสัญญาณเสียงในโดเมนเวลาเป็นโดเมนความถี่ ตามสมการที่ (2-12) หลังจากแปลงฟูริเยร์ช่วงสั้นแล้วจะใช้ชุดกรองฟิวเตอร์มัธยฐานเพื่อคำนวณหาค่าเสียงประสานและจังหวะดนตรีตามภาพที่ 2.11 เป็นการแสดงถึงโครงสร้างของเสียงที่โดดเด่นโดยเสียงจังหวะดนตรีมักจะถูกแสดงในบางช่วงเวลาและกระจายข้ามความถี่ตามเส้นแนวตั้ง (Vertical Lines) ในขณะที่เสียงประสานจะแสดงทั่วไปในความถี่และจะแสดงเด่นชัดเมื่อเวลาผ่านไปตามเส้นแนวนอน (Horizontal Lines)



ภาพที่ 2.11 แสดงสเปกโตรแกรม เสียงจังหวะดนตรี (ซ้าย) และเสียงประสาน (ขวา)

$$x_{m,k} = \sum_{n=0}^{N-1} x_{n+mH} \cdot w_n \cdot e^{-\frac{2\pi ikn}{N}}, 0 \le n < N, 0 \le k < N \quad (2\text{-}12)$$

เมื่อ N เป็นจำนวนตัวอย่างในหนึ่งกรอบสัญญาณ

m เป็นจำนวนของเฟรม

k เป็นลำดับของกรอบสัญญาณ $k = 1,2, \dots, K$

H เป็นขนาดของการทับซ้อนกันของหน้าต่าง

และ x_{n+mH} เป็นสัญญาณอินพุตของสัญญาณเสียง

2.3.3.2 หลังจากได้ค่าสเปกโตรแกรมของเสียงประสานและจังหวะดนตรี จะใช้เทคนิคไบนารีมาร์คเพื่อเพิ่มประสิทธิภาพให้กับสเปกโตรแกรมของเสียงประสานและจังหวะดนตรี

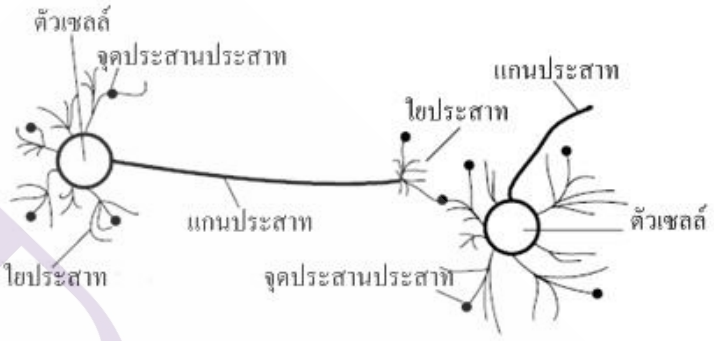
$$M_h(m,k) = \begin{cases} 1 & if\ \bar{y}_h(m,k) \ge \bar{y}_p(m,k) \\ 0 & else \end{cases} \quad (2\text{-}13)$$

$$M_p(m,k) = \begin{cases} 1 & if\ \bar{y}_p(m,k) \ge \bar{y}_h(m,k) \\ 0 & else \end{cases}$$

2.4 โครงข่ายประสาทเทียม (Artificial Neural Network)

เป็นศาสตร์แขนงหนึ่งของงานด้านปัญญาประดิษฐ์ (Artifitial Intelligence หรือ AI) มีโครงสร้างและการทำงานเหมือนกับสมองของสิ่งมีชีวิต ซึ่งมีการปรับเปลี่ยนตัวเองต่อการตอบสนองของอินพุตตามรูปแบบการเรียนรู้ (Learning rule) หลังจากที่โครงข่ายได้เรียนรู้สิ่งที่ต้องการแล้ว โครงข่ายนั้นๆ จะสามารถทำงานตามที่กำหนดไว้ได้ โครงข่ายประสาทเทียมนั้นได้ถูกพัฒนาจากการทำงานของสมองมนุษย์ซึ่งประกอบไปด้วยหน่วยประมวลผลที่เรียกว่า นิวรอน (เซลล์ประสาท หรือ Neuron) จำนวนนิวรอนในสมองมนุษย์มีอยู่ประมาณ 10^{11} และมีการเชื่อมต่อกันอย่างมากมาย จึงกล่าวได้ว่าสมองมนุษย์เปรียบเหมือนกับคอมพิวเตอร์ที่มีการปรับตัวเอง (Adaptive) ไม่เป็นเชิงเส้น (Nonlinear) และทำงานแบบขนาน (Parallel) ในการดูแลและจัดการการทำงานร่วมกันของนิวรอนในสมอง ดังนั้นการคำนวณเชิงนิวรอนจึงเป็นการคำนวณที่เลียนแบบมาจากการทำงานของสมองมนุษย์นั่นเอง โดยทั่วไปโครงข่ายประสาทเทียมจะมีทั้งแบบชั้นเดียว และแบบหลายชั้น สถาปัตยกรรม

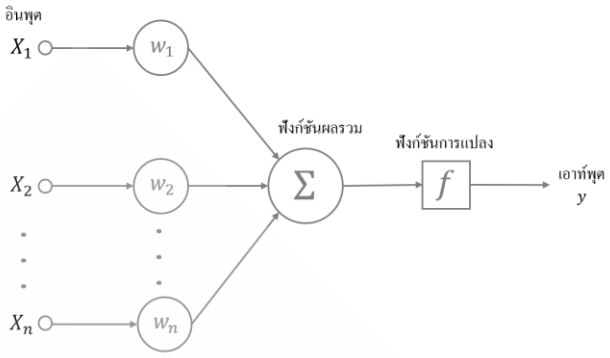
โครงข่ายประสาทเทียมมี 2 รูปแบบคือ แบบโครงข่ายไปข้างหน้า (Feed-forward Network) และแบบป้อนกลับ (Recurrent Network)



ภาพที่ 2.12 ตัวอย่างระบบโครงข่ายประสาทของมนุษย์

2.4.1 การทำงานของโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียมเป็นความพยายามในการเลียนแบบการทำงานของเซลล์ประสาทของมนุษย์ซึ่งประกอบด้วยนิวรอน หรือโหนด (Node) ต่างๆ โดยแต่ละโหนดจะเชื่อมต่อกัน (Interconnection) ที่มีลักษณะคล้ายร่างแห โหนดแต่ละโหลดจะประกอบด้วยอินพุตและเอาท์พุต โดยอินพุตในแต่ละหน่วยจะมีค่าน้ำหนัก (Weight) เป็นตัวกำหนดน้ำหนักของอินพุต ซึ่งโหลดแต่ละหน่วยจะมีฟังก์ชันกำหนดสัญญาณส่งออกที่เรียกว่า ฟังก์ชันการแปลง (Transfer Function) เป็นตัวกำหนดว่าน้ำหนักรวมของอินพุตต้องมากขนาดไหน จึงจะสามารถส่งเอาท์พุตไปยังโหนดอื่นได้ เมื่อนำโหนดแต่ละหน่วยมาต่อกันให้ทำงานร่วมกัน การทำงานนี้ในทางตรรกะแล้วจะเหมือนกับปฏิกิริยาเคมีที่เกิดในสมอง เพียงแต่ในทางคอมพิวเตอร์ทุกอย่างจะเป็นตัวเลข



ภาพที่ 2.13 กระบวนการประมวลผลของโครงข่ายประสาทเทียม

2.4.2 การเรียนรู้สำหรับโครงข่ายประสาทเทียม

2.4.2.1 การเรียนรู้แบบมีผู้สอน (Supervised Learning) เป็นการเรียนแบบที่มีการตรวจคำตอบเพื่อให้โครงข่ายปรับตัว ซึ่งชุดข้อมูลที่ใช้ฝึกสอนจะมีคำตอบไว้สำหรับใช้ตรวจสอบว่าโครงข่างให้คำตอบถูกต้องหรือไม่ ถ้าคำตอบที่ได้ไม่ถูกต้องโครงข่ายจะปรับตัวเองเพื่อให้ได้คำตอบที่ดีขึ้น

2.4.2.2 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นการเรียนแบบไม่มีการตรวจคำตอบว่าถูกหรือผิด แต่ตัวโครงข่ายจะจัดเรียงโครงสร้างด้วยตัวเองตามลักษณะของข้อมูล ซึ่งผลลัพธ์ที่ได้ตัวโครงข่ายจะสามารถจัดหมวดหมู่ของข้อมูลได้

2.5 การเรียนรู้เชิงอภิมาน (Meta Learning)

การเรียนรู้เชิงอภิมานเป็นขอบเขตงานวิจัยที่มีการศึกษามากที่สุดที่เกี่ยวข้องกับการเรียนรู้เชิงลึก (Deep Learning) โมเดลผลลัพธ์ที่ได้งานการเรียนรู้เชิงอภิมานเรียกได้ว่าเป็นโมเดลอเนกประสงค์เลยทีเดียว เพราะสามารถที่จะนำโมเดลไปประยุกต์ใช้ในงานด้านต่างๆ โดยไม่จำเป็นต้องมีการฝึกสอนอะไรตั้งแต่เริ่มต้น และในการฝึกสอนโมเดลการเรียนรู้เชิงอภิมานนั้นไม่จำเป็นต้องใช้ข้อมูลในปริมาณมากนักในแต่ละคลาส หลังจากฝึกสอนโมเดลไปแล้วหนึ่งครั้ง ก็สามารถที่จะนำโมเดลดังกล่าวไปใช้เพื่อแก้ปัญหาในงานที่มีลักษณะใกล้เคียงกับโมเดลนั้น โดยไม่จำเป็นต้องฝึกสอนอะไรเพิ่มเติมอีก ดังนั้นนักวิจัยและนักวิทยาศาสตร์หลายคนเชื่อว่า การเรียนรู้เชิงอภิมานเป็นเทคนิคที่มีความ

ใกล้เคียงกับการเรียนรู้ของมนุษย์มากที่สุด ซึ่งอาจจะเป็นแนวทางในการพัฒนาปัญญาประดิษฐ์ต่อไปในอนาคต

แนวความคิดของการเรียนรู้เชิงอภิมานเกิดขึ้นในปี ค.ศ. 1979 โดยนายโดนัล บี. มอส์สลีย์ (Donald B. Maudsley) ได้พูดถึงแนวความคิดใหม่เกี่ยวกับกระบวนการที่ผู้เรียนรู้ตระหนักถึงการควบคุมการรับรู้, การค้นหา, การเรียนรู้ และการเติมโตขึ้นจากความรู้ที่มีอยู่เดิม และต่อมาในปี ค.ศ. 1985 นายจอห์น บิ๊กส์ (John Biggs) ได้อธิบายคำกัดความของเทคนิคการเรียนรู้เชิงอภิมานในผลงานของเขาว่า การเรียนรู้และควบคุมการเรียนรู้ด้วยตัวเอง ซึ่งจากคำจำกัดความเหล่านั้นอาจจะถูกต้องตามหลักการในงานด้านวิทยาศาสตร์ทั่วไป แต่ก็อาจจะดูเหมือนยากที่จะนำมาปรับให้เข้ากับการทำงานของปัญญาประดิษฐ์ (AI)

ในบริบทของระบบปัญญาประดิษฐ์ การเรียนรู้เชิงอภิมานนั้นถูกให้คำนิยามว่าเปแ็นความสามารถในการับรู้ความรู้ที่หลายหลาย ถ้ามองกลับไปลักษณะการเรียนรู้ของมนุษย์นั้นสามารถที่จะเรียนรู้หลายๆ อย่างพร้อมกันและยังอาศัยข้อมูลในปริมาณน้อยก็สามารถที่จำแนกหรือรับรู้วัตถุชนิดใหม่ได้เพียงการเห็นเพียงภาพเดียวหรือสามารถเรียนรู้กิจกรรมที่ซับซ้อนหลายกิจกรรม เช่น การขับรถหรือการขับเครื่องบินได้ในครั้งเดียว สำหรับปัญญาประดิษฐ์นั้นสามารถทำงานที่ซับซ้อนได้ก็จริง แต่ก็ต้องอาศัยข้อมูลจำนวนมากในการฝึกสอน และผลลัพธ์ที่ได้จากการทำงานก็ไม่ได้ดีมากเท่าใดนัก ดังนั้นแนวทางการเพิ่มความสามารถให้กับปัญญาประดิษฐ์คือการเรียนรู้วิธีการเรียนรู้ หรือการเรียนรู้เชิงอภิมาน

สำหรับมนุษย์นั้นสามารถที่จะปรับวิธีการเรียนรู้ให้เหมาะกับสถานการณ์ ในทำนองเดียวกันเทคนิคเชิงอภิมานใช่ว่าจะเรียนรู้ด้วยวิธีเดียวกันในทุกโมเดล อาทิเช่น บางโมเดลจะเน้นไปที่การปรับปรุงประสิทธิภาพโครงสร้างของโครงข่ายประสาทเทียม และบางโมเดลก็เน้นไปที่การค้นหาชุดข้อมูลที่ถูกต้องสำหรับการฝึกฝน ในงานวิจัยล่าสุดจากห้องวิจัยของ UC Berkeley AI Lab ได้จำแนกประเภทของการเรียนรู้เชิงอภิมานออกเป็น 4 ประเภทหลักคือ

2.5.1 การเรียนรู้ด้วยตัวอย่างข้อมูลจำนวนน้อย (Few-shot Learning) แนวความคิดของการเรียนรู้ด้วยตัวอย่างข้อมูลจำนวนน้อยคือการสร้างโครงข่ายประสาทเทียมเชิงลึกที่สามารถเรียนรู้จากชุดข้อมูลขนาดเล็ก ตัวอย่างเช่นการระบุวัตถุโดยอาศัยการมองเห็นจากภาพเพียงหนึ่งหรือสองภาพ ซึ่งแนวความคิดนี้ทำให้เกิดการสร้างเทคนิคต่างๆ เช่น หน่วยความจำเสริมโครงข่ายประสาทเทียม

(Memory Augmented Neural Networks) [11] หรือการเรียนรู้ด้วยตัวอย่างข้อมูลเพียงหนึ่งตัวอย่าง (One-shot Learning) [8]

2.5.2 การเรียนรู้เพื่อปรับปรุงประสิทธิภาพ (Optimizer Learning) การเรียนรู้ในลักษณะนี้มุ่งเน้นที่การเรียนรู้วิธีการเพิ่มประสิทธิภาพโครงข่ายประสาทเทียมให้ดีขึ้น โดยทั่วไปแล้วโมเดลประเภทนี้มีเป้าหมายเพื่อปรับพารามิเตอร์ของโครงข่ายประสาทเทียมที่เหมาะสมสำหรับงานนั้นๆ ตัวอย่างงานวิจัยที่ใช้เทคนิคการเรียนรู้เชิงอภิมานเพื่อปรับปรุงประสิทธิภาพเช่น การใช้เทคนิคนี้ในการเรียนรู้วิธีการทำงานของ gradient descent [12]

2.5.3 การเรียนรู้โดยอาศัยตัวชี้วัด (Metric Learning) เป็นการเรียนรู้ที่มีพื้นฐานมาจากตัวชี้วัดตัวอย่างเช่น หากต้องการเรียนรู้เพื่อหาความคล้ายกันของภาพสองภาพ อาจจะใช้พื้นฐานตัวชี้วัดโดยใช้โครงข่ายประสาทเทียมง่ายๆ เพื่อใช้แยกคุณลักษณะสำคัญจากภาพแต่ละภาพ เพื่อนำมาหาความคล้ายคลึงกันของภาพ โมเดลที่นิยมใช้เช่น โครงข่ายสยาม (Siamese network) [13], โครงข่ายแม่แบบ (Prototypical network) [14] และโครงข่ายความสัมพันธ์ (Relation network) [15]

2.5.4 การเรียนรู้งานแบบซ้ำ (Recurrent Model Learning) การทำงานของการเรียนรู้ในลักษณะนี้เหมาะกับโครงข่ายประสาทเทียมงานแบบซ้ำ เช่น Long Shot Term Memory (LSTM) หรือ Gated Recurrent Units (GRU) ในโครงสร้างนี้การเรียงรู้เชิงอภิมานจะเรียนรู้ด้วยข้อมูลตามลำดับพร้อมกับประมวลผลกับข้อมูลที่เป็นข้อมูลอินพุต (Input) ใหม่ ตัวอย่างเช่น Meta Reinforcement Learning [16]

2.6 ฟังก์ชันสูญเสียแบบจัดอันดับ (Ranking Loss Function)

ฟังก์ชันนี้ต่างจากฟังก์ชันสูญเสียอื่นๆ อย่างเช่น cross-entropy loss หรือ mean square error loss เป็นต้น ที่มีวัตถุประสงค์เพื่อเรียนรู้การทำนายป้ายกำกับ, ค่าบางอย่าง จากข้อมูลขาเข้า ส่วนฟังก์ชันสูญเสียแบบจัดอันดับมีวัตถุประสงค์เพื่อหาความคล้ายคลึงกันของข้อมูลตัวอย่างเช่น ชุดข้อมูลสำหรับตรวจสอบข้อมูลใบหน้า ที่จะมีข้อมูลที่เป็นภาพใบหน้าที่เป็นคนคนเดียวกัน และไม่ใช่คนคนเดียวกัน ในการฝึกสอนด้วยโมเดลสามารถใช้ฟังก์ชันสูญเสียแบบจัดอันดับเพื่อบอกว่าว่าภาพสองภาพเป็นคนคนเดียวกันหรือไม่

2.6.1 การจัดอับดับสูญเสียแบบแฝด (Pairwise Ranking Loss)

การเตรียมข้อมูลเพื่อใช้งานจะต้องแบ่งข้อมูลเป็นชุด คือชุดข้อมูลที่เหมือนกันและชุดข้อมูลที่ต่างกัน แต่ละรายการจะมี 2 ตัว วัตถุประสงค์เพื่อลดระยะห่างในคู่ที่เหมือนกันและเพิ่มระยะห่างของคู่ที่ต่างกัน ยิ่งเหมือนกันระยะห่างก็ยิ่งเข้าใกล้ 0

$$L = \begin{cases} d(r_a, r_p) & if\ PositivePair \\ \max(0, m - d(r_a, r_n)) & if\ NegativePair \end{cases} \quad (2\text{-}14)$$

โดยค่า r_a, r_p และ r_nเป็นตัวแทนของตัวอย่างข้อมูล ส่วนค่า d เป็นฟังก์ชันวัดระยะห่าง สำหรับคู่ที่เหมือนกันค่าสูญเสียจะเป็น 0 ได้เมื่อทั้งสองตัวไม่มีระยะห่างต่อกัน และค่าสูญเสียจะเพิ่มขึ้นไปตามระยะห่างจริงที่เกิดขึ้น ส่วนสำหรับคู่ที่ต่างกันหากค่าสูญเสียน้อยกว่าค่าขอบ m ซึ่งจะนำค่านี้ไปเพื่อปรับปรุงระยะห่างของวัตถุทั้ง 2 ตัวนี้ และหากค่าสูญเสียมากกว่าค่าขอบ m จะใช้ค่า m เพื่อปรับปรุงระยะห่างแทน ซึ่งสามารถเขียนเป็นสมการที่ (2-15)

$$L(r_0, r_1, y) = y \parallel r_0 - r_1 \parallel + (1 - y)\max(0, m - \parallel r_0 - r_1 \parallel) \quad (2\text{-}15)$$

ให้ r_0 และ r_1 แทนตัวอย่างข้อมูลในคู่นั้นๆ และ y เป็นค่าไบนารี่หากมีค่าเป็น 0 จะหมายถึงคู่ที่ต่างกัน แต่หากเป็น 1 หมายถึงคู่ที่เหมือนกัน และระยะห่าง d ใช้วิธีการวัดแบบยุคลิด

2.6.2 การจัดอับดับสูญเสียแบบแฝดสาม (Triplet Ranking Loss)

การเตรียมข้อมูลเพื่อใช้งานจะต้องแบ่งข้อมูลเป็นชุด โดยแต่ละรายการจะมีข้อมูล 3 ตัว ประกอบไปด้วยตัวอย่างข้อมูลหลัก r_a , ตัวอย่างข้อมูลที่เหมือนกับข้อมูลหลัก r_p และตัวอย่างข้อมูลที่ต่างจากข้อมูลหลัก r_n วัตถุประสงค์เพื่อเพิ่มระยะห่างของตัวอย่างที่ต่างกันและลดระยะห่างของข้อมูลที่เหมือนกัน ซึ่งสามารถเขียนเป็นสมการที่ (2-16)

$$L(r_a, r_p, r_n) = max(0, m + d(r_a, r_p) - d(r_a, r_n) \quad (2\text{-}16)$$

2.7 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับการตรวจสอบความคล้ายคลึงของเสียงมีไม่มากนัก และยิ่งเป็นการตรวจสอบความคล้ายคลึงกันของเสียงเพลงแล้วมีไม่เยอะมากนัก ดังนั้นทางผู้วิจัยได้ศึกษางานวิจัยที่ได้รับการตีพิมพ์เหล่านั้นและสรุปได้ดังนี้

สุนันท์ ธาติ, พงศ์พันธ์ กิจสนาโยธิน, และ วรลักษณ์ คงเด่นฟ้า (2018) ในงานวิจัยนี้ได้นำเสนอการประยุกต์ใช้วิธีการหาความคล้ายกันของเพลงโดยใช้ลายนิ้วมือทางเสียง และฟังก์ชันความสัมพันธ์ (Relation function) ในข้อมูลเพลงจำนวน 120 ข้อมูลตัวอย่าง ประกอบด้วย 1) เพลงเวอร์ชั่นต้นฉบับ 20 ข้อมูลตัวอย่าง โดยเพลงที่ถูกเลือกเป็นเพลงที่มีจำนวนการถูกคัฟเวอร์จำนวนมากซึ่งประกอบด้วยเพลงสากลจำนวน 15 เพลงและเพลงเกาหลี 5 เพลง และ 2) เพลงเวอร์ชั่นคัฟเวอร์จำนวน 5 ข้อมูลตัวอย่างต่อหนึ่งเพลงต้นฉบับ รวมเป็น 100 ข้อมูลตัวอย่าง พบว่าสามารถระบุเพลงต้นฉบับจากเพลงคัพเวอร์ได้อย่างมีประสิทธิภาพครอบคลุมเพลงทุกประเภท โดยมีพื้นที่ใต้กราฟอยู่ระหว่าง 0.74 – 0.85 และการจัดกลุ่มเพลงมีประสิทธิภาพในกรณีที่มีการบรรเลงดนตรีใหม่ในรูปแบบคล้ายกับต้นฉบับ ในขณะที่การบรรเลงโดยเปลี่ยนรูปแบบดนตรีส่งผลให้ความถูกต้องของการจัดกลุ่มน้อยลง

Yichi Zhang, Bryan Pardo, และ Zhiyao Duan (2018) ในงานวิจัยนี้ได้นำเสนอการประยุกต์โมเดลโครงข่ายสยาม 2 รูปแบบคือ 1) IMINET โครงข่ายย่อยในโครงข่ายสยามทั้ง 2 ตัวมีการตั้งค่าและโครงสร้างเหมือนกัน 2) TL-IMINET เป็นโครงข่ายสยามที่มีโครงข่ายย่อยที่มีการตั้งค่าและโครงสร้างไม่เหมือนกันและประยุกต์ใช้ transfer learning จากงานวิจัยอื่นที่เกี่ยวข้องมาช่วยในการฝึกสอน จากการทดลองพบว่าโมเดลทั้ง 2 รูปแบบให้ประสิทธิภาพดีกว่าการใช้อัลกอริทึม IMISOUND ในการค้นหาเสียงเลียนแบบ และจากผลลัพธ์ยังแสดงให้เห็นว่าการประยุกต์ใช้ ltransfer learning ยังช่วยเพิ่มประสิทธิภาพในการดึงข้อมูลอย่างมีนัยสำคัญ

Pranay Manocha, Rohan Badlani, Anurag Kumar, Ankit Shah, Benjamin Elizalde และ Bhiksha Raj (2018) ในงานวิจัยนี้ได้นำเสนอการแยกคุณลักษณะเฉพาะเพื่อใช้เป็นตัวแทนเสียงเพื่อให้สามารถจับคู่เนื้อหาของเสียงที่มีความคล้ายคลึงกัน (เสียงเหตุการณ์เสียงเดียวกัน) โดยมุ่งเน้นไปที่เนื้อหาที่ไม่มีเสียงดนตรีและไม่มีการพูด ซึ่งเป้าหมายของงานวิจัยไม่ใช่การตรวจจับหรือจำแนกเสียงกิจกรรมต่างๆ แต่เป็นการพัฒนาวิธีการจับเนื้อหาของเสียงเพื่อค้นหาเหตุการณ์ต่างๆ ที่เกิดขึ้นในคลิปเสียง โดยใช้ชุดข้อมูลเสียง 3 ชุด คือ ESC-50, US8K, และ TUT2016 เป็นฐานข้อมูลเพื่อใช้อ้างอิงเสียง

โดยพบเสียงเหตุการณ์ทั้งหมด 76 คลาส ในการฝึกสอนโมเดลจะใช้ชุดข้อมูล Youtube หลังจากทดลองแล้วพบว่า การดึงข้อมูลเหตุการณ์จากเนื้อหาเสียงจากหลายๆ คลาส ผลลัพธ์ 25 ตัวแรกมีค่าความแม่นยำที่ค่อนข้างสูง สิ่งนี้แสดงให้เห็นความคุณลักษณะเวกเตอร์ที่ได้จากโครงข่ายสยามประสาทเทียมสามารถตัวจับความคลึงระหว่างคลิปที่เป็นเหตุการณ์ได้เป็นอย่างดี

Ivette Velez, Caleb Rascon และ Gibran Fuentes-Pineda (2018) ในงานวิจัยนี้เสนอวิธีการยืนยันตัวผู้พูดโดยประยุกต์ใช้โครงข่ายสยาม เพื่อระบุว่าเสียงพูดนั้นเป็นเสียงจากผู้พูดคนเดียวกันหรือไม่ โดยตรวจสอบข้อมูลจากฐานข้อมูลที่บันทึกเสียงของผู้พูดไว้ แต่หากเป็นผู้ใช้ใหม่ก็จะบันทึกเพิ่มเข้าไปในฐานข้อมูลโดยไม่ต้องฝึกอบรมใหม่ ในการทดลองใช้ชุดข้อมูล LibriSpeech โดยประกอบด้วยผู้พูดมากกว่า 6000 คน และเปรียบเทียบการทำงานของโมเดลโครงข่ายสยามที่มีโครงข่ายย่อยคนละรูปแบบกันประกอบด้วย โครงข่ายที่มีโครงสร้างแบบ VGG732, VGG7256, และ ResNet50 จากการทดลองพบว่าโมเดล VGG732 ให้ค่าความถูกต้องสูงที่สุดถึง 97.7% ตามด้วยโมเดล VGG7256 ให้ค่าความถูกต้อง 95% และ ResNet50 มีค่าความถูกต้องอยู่ที่ 94.5% ซึ่งโมเดลทุกตัวทนทานต่อเสียงรบกวนโดยให้ค่าความถูกต้องมากกว่า 80%

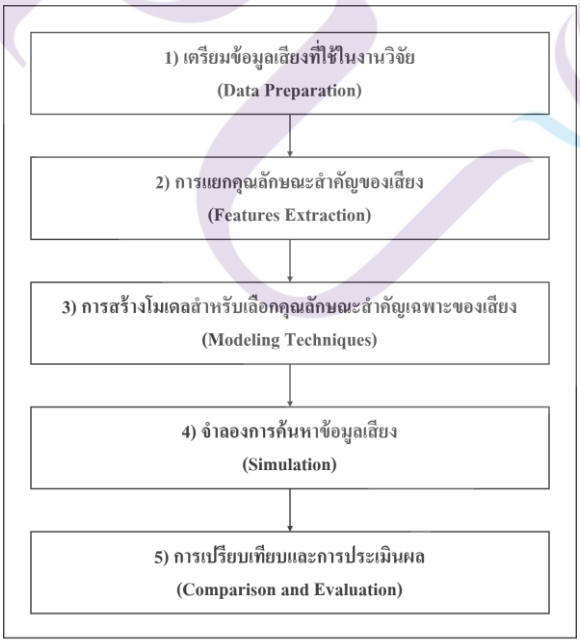
Kaavya Sriskandaraja, Vidhyasaharan Sethu และ Eliathamby Ambikairajah (2018) งานวิจัย นี้เสนอการแยกแยะการปลอมแปลงเสียงในระบบไบโอเมตริกซ์ โดยประยุกต์ใช้โครงข่ายสยาม โดยในการฝึกสอนโมเดลถ้าหากเป็นคู่เสียงแบบเดียวกันคือ เป็นเสียงแท้จริงทั้งคู่ หรือ เสียงเล่นซ้ำทั้งคู่ จะให้ความหมายว่าเหมือนกัน แต่หากเป็นคู่เสียงที่เสียงไม่เหมือนกันคือ เสียงมีความต่างกัน หลังจากทดสอบพอว่าโมเดลที่เสนอในงานวิจัยนี้ให้ผลลัพธ์ที่มีประสิทธิภาพสูงกว่าเทคนิคอื่นๆ ในปัจจุบัน โดยให้ค่าเปอร์เซ็นต์ความผิดพลาดอยู่ที่ 6.4 ซึ่งต่ำที่สุดในเทคนิคอื่นที่นำมาเปรียบเทียบ

บทที่ 3
ระเบียบวิธีวิจัย

งานวิจัยนี้เป็นการหาข้อมูลเพลงโดยอาศัยคุณลักษณะสำคัญเฉพาะของเสียง ในบทนี้จะอธิบายถึงวิธีการดำเนินงานวิจัย เพื่อให้บรรลุวัตถุประสงค์ตามที่กำหนดไว้ จึงแบ่งแนวทางการวิจัยดังภาพที่ 3.1 ได้แก่ การเตรียมข้อมูลเสียงที่ใช้ในงานวิจัย (Data Preparation) การแยกคุณลักษณะสำคัญของเสียง (Feature Extraction) การสร้างโมเดลสำหรับเลือกคุณลักษณะสำคัญเฉพาะของเสียงเพื่อใช้สำหรับการแสดงข้อมูลของเสียง (Modeling Technique) การจำลองการค้นหาข้อมูลเสียงด้วยเสียง (Simulation) และขั้นตอนสุดท้ายเป็นการเปรียบเทียบและการประเมินผล (Comparison and Evaluation)

3.1 แนวทางการวิจัย

แนวทางการวิจัยมีขั้นตอนดังภาพที่ 3.1



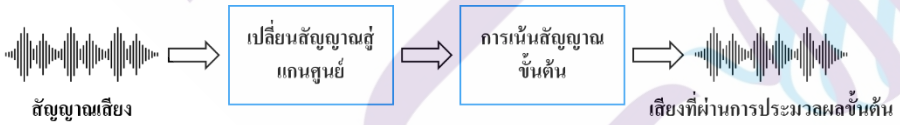
ภาพที่ 3.1 ภาพรวมวิธีการดำเนินงานวิจัย

3.1.1 การเตรียมข้อมูลเสียงที่ใช้ในงานวิจัย (Data Preparation)

ในงานวิจัยนี้จะใช้ข้อมูลที่เป็นข้อมูลเพลงจำนวน 200 เพลงเพื่อเป็นชุดข้อมูลสำหรับการฝึกสอนและจำนวน 100 เพลงสำหรับทดสอบประสิทธิภาพ เนื่องจากข้อมูลเพลงที่ได้มานั้นมีความหลากหลายทั้งชนิดและโปรไฟล์ของข้อมูลเพลง ดังนั้นก่อนอื่นจำเป็นต้องแปลงข้อมูลเพลงให้อยู่ในรูปแบบเดียวกันก่อน ในการทดลองนี้กำหนดข้อมูลเพลงให้อยู่ในรูปแบบไฟล์ WAV กำหนดอัตราสุ่มตัวอย่าง (Sampling Rate) 22.05 กิโลเฮิรตซ์ที่ระดับความลึกของเสียง (Bit Depth) 16 บิตต่อตัวอย่าง (Sampling Size) มีรูปแบบการบันทึกเสียงเป็นแบบโมโน (1 Channel)

การสร้างชุดข้อมูลเพลงนั้นกำหนดในเพลงหนึ่งเพลงคือหนึ่งคลาส (Classes) ซึ่งเมื่อตรวจสอบข้อมูลเพลงจะพบว่าความยาวของเพลงนั้นมีจำนวนไม่เท่ากัน และพบว่าความยาวของเพลงที่สั้นที่สุดจะอยู่ที่ประมาณ 2 นาที ดังนั้นจึงทำการสุ่มเพื่อตัดความยาวของเพลงทุกเพลงให้มีความยาว 2 นาทีต่อเพลง เพื่อเมื่อวางกรอบหน้าต่างแล้วจำนวนกรอบหน้าต่างจะได้เท่ากันในทุกเพลง

จากนั้นจะนำไฟล์เสียงที่ได้ นำเข้ากระบวนการปรับแต่งสัญญาณก่อนนำไปหาคุณลักษณะสำคัญ เนื่องจากข้อมูลเพลงที่มีอยู่นั้นมีความแตกต่างในระดับความดังของเสียงเนื่องมาจากแนวเพลงและลักษณะเครื่องดนตรีที่ประกอบในเพลง จึงต้องมีการปรับแต่งก่อนนำเข้าไปสู่กระบวนการต่อไป เพื่อลดทอนสัญญาณรบกวน โดยมีขั้นตอนและรายละเอียดดังนี้



ภาพที่ 3.2 กระบวนการประมวลผลสัญญาณเสียงเบื้องต้น

3.1.1.1 การเปลี่ยนสัญญาณสู่แกนศูนย์หรือการทำค่าเฉลี่ยเป็นศูนย์ (Zero Mean) เป็นการปรับสัญญาณที่นอกแกนศูนย์กลับเข้าสู่แกนศูนย์ตามสมการที่ (2-1)

3.1.1.2 การเน้นสัญญาณขั้นต้นเพื่อเป็นการกรองสัญญาณรบกวนที่เป็นความถี่ต่ำกว่าค่าที่กำหนดออกไป ตามสมการที่ (2-3) ในงานวิจัยนี้ได้เลือกค่าสัมประสิทธิ์วงจรกรอง $a = 0.97$ กรองความถี่สูงผ่านตามสมการดังนี้

$$\tilde{s}(n) = s(n) - 0.97s(n-1) \tag{3-1}$$

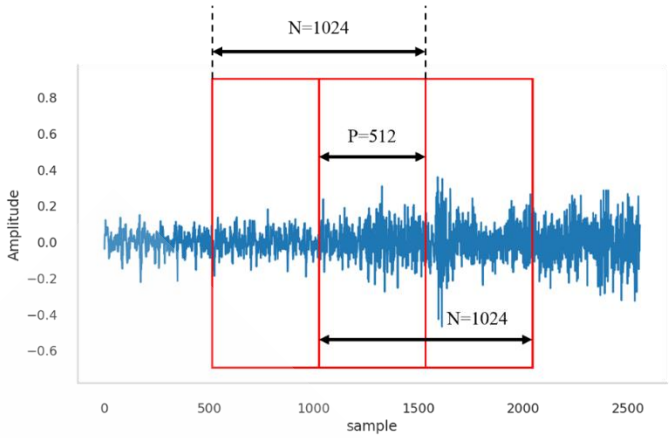
หลังจากนำสัญญาณเสียงผ่านกระบวนการประมวลผลสัญญาณเสียงเบื้องต้นแล้ว จะนำสัญญาณเสียงที่ได้มาแบ่งย่อยเป็นช่วงสั้นๆ เพื่อนำมาใช้เป็นตัวอย่างข้อมูลสำหรับเพลงในแต่ละเพลง (คลาส) โดยในแต่จะช่วงหรือแต่ละตัวอย่างข้อมูลจะมีความยาวประมาณ 1.46 วินาที (สัญญาณเสียงขนาด 32,256 ตัวอย่าง) และมีการทับซ้อนกัน 50% ของขนาดตัวอย่างหรือเท่ากับสัญญาณเสียงขนาด 16,128 ตัวอย่าง ดังนั้นในแต่ละเพลงจะมีจำนวนตัวอย่างข้อมูล 163 ตัวอย่าง

3.1.2 การแยกคุณลักษณะสำคัญ (Feature Extraction)

การแยกคุณลักษณะสำคัญของสัญญาณเสียงเพลงนั้น เป็นการหาค่าคุณลักษณะที่จะใช้แทนตัวอย่างของสัญญาณเสียงเพลง โดยที่คุณลักษณะดังกล่างจะบอกถึงลักษณะสำคัญของสัญญาณเสียงเพลงนั้นๆ ซึ่งคุณลักษณะสำคัญที่เหมาะสมนั้นก็จะทำให้ผลลัพธ์ที่ถูกต้องและแม่นยำ อีกทั้งในปัจจุบันมีวิธีการแยกคุณลักษณะสำคัญของสัญญาณเสียงหลากหลายรูปแบบ ซึ่งมีหลักการและวิธีการที่แตกต่างกันออกไป ในงานวิจัยนี้ได้เลือกใช้วิธีการแยกคุณลักษณะสำคัญของเสียงทั้งหมด 4 วิธีประกอบด้วย 1) ค่าลอการิทึมจากสเปคตรัมของเสียง, 2) ค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล, 3) ค่าเฉลี่ยของเสียงประสานและจังหวะดนตรี, 4) ค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลแบบความต่าง

3.1.2.1 ค่าลอการิทึมจากสเปคตรัมและค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล

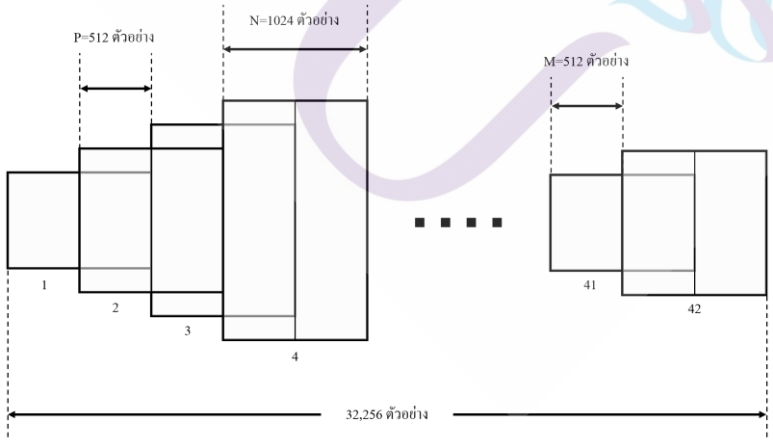
เริ่มต้นด้วยการแยกคุณลักษณะสำคัญด้วยวิธีหาค่าลอการิทึมจากสเปคตรัมของเสียงและตามด้วยวิธีการหาค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล เนื่องจากวิธีการหาค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลเป็นกระบวนการที่ทำเพิ่มเติมมาจากวิธีการหาค่าลอการิทึมจากสเปคตรัมของเสียง จากสัญญาณเสียงที่ผ่านกระบวนการประมวลผลสัญญาณเสียงเบื้องต้นในหัวข้อที่ 3.1.1 จะถูกแบ่งสัญญาณเป็นส่วนย่อยและวางกรอบสัญญาณหน้าต่าง โดยใช้ฟังก์ชันหน้าต่างแบบแฮนนิ่งขนาดของหน้าต่างกว้าง 46.43 มิลลิวินาทีหรือเท่ากับ $N = 1024$ ตัวอย่าง และมีการทับซ้อนกันของแต่ละกรอบหน้าต่างที่ 50% ของขนาดกรอบหน้าต่างหรือประมาณ 23.22 มิลลิวินาทีหรือเท่ากับ $P = 512$ ตัวอย่าง ตามที่แสดงในภาพที่ 3.3



ภาพที่ 3.3 การวางกรอบสัญญาณหน้าต่างเพื่อหาค่าค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล

ในกระบวนการนี้จะข้อมูลเสียงที่แบ่งย่อยเป็นช่วงสั้นๆ ช่วงละประมาณ 1.46 วินาที (สัญญาณเสียงขนาด 32,256 ตัวอย่าง) มาวางกรอบสัญญาณหน้าต่าง ซึ่งในแต่ละช่วงจะได้เป็นกรอบสัญญาณหน้าต่างจำนวน 42 กรอบ ที่จะใช้สำหรับหาค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลต่อไป ในการคำนวณหาจำนวนกรอบสัญญาณหน้าต่างสามารถคำนวณได้จากสมการที่ (3-2)

$$\text{จำนวนกรอบ } MFCC = trunc\left(\frac{Sample}{M} - 1\right) \tag{3-2}$$



ภาพที่ 3.4 การกำหนดส่วนย่อยและการทับซ้อนกันของส่วนย่อย

3.1.2.1.1 หลังจากสัญญาณเสียงได้ถูกแบ่งออกเป็นช่วงสั้นๆ ต่อไปจะเป็นการหาค่าสัมประสิทธิ์เซปตรัมบนสเกลเมล ตามขั้นตอนที่แสดงไปในหัวข้อที่ 2.3.1 เมื่อวางกรอบหน้าต่างสัญญาณเสียงแล้ว ข้อมูลของสัญญาณเสียงในแต่ละเฟรมจะถูกคูณด้วยฟังก์ชันกรอบหน้าต่างสัญญาณเสียง เพื่อลดทอนความไม่ต่อเนื่องของขอบสัญญาณ สำหรับเตรียมไปแปลงฟูริเยร์ต่อไป

3.1.2.1.2 การแปลงสัญญาณฟูริเยร์แบบเร็ว เป็นการเปลี่ยนสัญญาณเสียงจากโดเมนเวลาให้ไปอยู่ในโดเมนความถี่ สามารถแปลงโดยใช้สมการที่ (3-3)

$$\widetilde{x_k} = \sum_{n=0}^{N-1} x_n \cdot w_n e^{-\frac{j2mk}{N}}, 0 \le k < N \tag{3-3}$$

เมื่อ N เป็นจำนวนตัวอย่างในหนึ่งกรอบสัญญาณ

k เป็นลำดับของกรอบสัญญาณ $k = 1,2, \dots, K$

x_n เป็นสัญญาณอินพุตของสัญญาณเสียง

และ w_n เป็นฟังก์ชันกรอบสัญญาณ

3.1.2.1.3 นำสัญญาณสเปคตรัมที่ผ่านการแปลงสัญญาณฟูริเยร์แบบเร็วส่งผ่านชุดกรองฟิลเตอร์แบงค์ เพื่อเน้นความสำคัญของความถี่ที่อยู่ในช่วงกล่างของชุดตัวกรองแต่ละตัวกรอง โดยใช้ชุดตัวกรองฟิลเตอร์แบงค์ 64 ตัวหรือ 64 สัมประสิทธิ์ คำนวณหาค่าพลังงานในแต่ละชุดตัวกรองได้โดยใช้สมการที่ (2-10) ในการออกแบบชุดตัวกรองฟิลเตอร์แบงค์ นำความถี่ที่สุ่มตัวอย่าง (fs/2) มาคำนวณหาความถี่บนสเกลเมลตามสมการที่ (2-9) เมื่อได้ความถี่ที่อยู่บนสเกลเมลแล้ว ก็นำมาแบ่งเท่าๆ กันบนสเกลเมลขนาดชุดตัวกรองฟิลเตอร์แบงค์ แล้วแปลงความถี่บนสเกลเมล (Mel) มาเป็นความถี่ปกติ (f) ตามสมการที่ (3-4)

$$f_{Hz} = 700\left(10^{\frac{m}{2595}} - 1\right) \tag{3-4}$$

3.1.2.1.4 คำนวณค่าสัมประสิทธิ์เซปสตรัมโดยการแปลงแบบโคไซน์ไม่ต่อเนื่อง (DCT) สัญญาณเสียงในกรอบสัญญาณหน้าต่าง N=1024 หลังจากแปลงให้อยู่ในโดเมนความถี่ โดยหาขนาดกำลังสอง และผ่านการกรองด้วยตัวชุดกรองฟิลเตอร์แบงค์ นำมาหาค่าลอการิทึม (หลังจากหาค่าลอการิทึมแล้วจะได้คุณลักษณะในแบบแรก) และเข้าสู่วิธีการแปลงโคไซน์แบบไม่ต่อเนื่องตามสมการที่ (2-11) ทำการวางกรอบสัญญาณหน้าต่างทับซ้อนกันไป 512 ตัวอย่างไปเรื่อยๆ ไปจนถึงจุดสิ้นสุดของสัญญาณเสียง

3.1.2.2 ค่าเฉลี่ยของเสียงประสานและจังหวะดนตรี

การแยกคุณลักษณะสำคัญของเสียงด้วยวิธีการนี้จะใช้อินพุตที่เป็นสัญญาณเสียงที่แบ่งเป็นช่วงสั้นๆ ที่เตรียมไว้ในหัวข้อที่ 3.1.1 1 หลังจากที่ได้ข้อมูลเสียงมาก็จะนำมาผ่านกระบวนการแบ่งย่อยส่วนและวางกรอบสัญญาณหน้าต่างเช่นเดียวกับการแยกคุณลักษณะสำคัญด้วยค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล

3.1.2.2.1 หลังจากสัญญาณเสียงได้ถูกแบ่งออกเป็นช่วงสั้นๆ ต่อไปจะเป็นการหาค่าเฉลี่ยของเสียงประสานและจังหวะดนตรี ตามขั้นตอนที่แสดงไปในหัวข้อที่ 2.3.3 เมื่อวางกรอบหน้าต่างสัญญาณเสียงแล้ว ข้อมูลของสัญญาณเสียงในแต่ละเฟรมจะถูกคูณด้วยฟังก์ชันกรอบหน้าต่างสัญญาณเสียง เพื่อลดทอนความไม่ต่อเนื่องของขอบสัญญาณ สำหรับเตรียมไปแปลงฟูริเยร์ต่อไป

3.1.2.2.2 การแปลงสัญญาณฟูริเยร์ช่วงสั้น เป็นการเปลี่ยนสัญญาณเสียงจากโดเมนเวลาให้ไปอยู่ในโดเมนความถี่ สามารถแปลงโดยใช้สมการที่ (3-5)

$$x_{m,k} = \sum_{n=0}^{N-1} x_{n-m} \cdot w_n e^{-\frac{2\pi i k n}{N}}, 0 \le k < N \tag{3-5}$$

เมื่อ N เป็นจำนวนตัวอย่างในหนึ่งกรอบสัญญาณ

k เป็นลำดับของกรอบสัญญาณ $k = 1,2, \dots, K$

m เป็นจำนวนของกรอบสัญญาณ

x_n เป็นสัญญาณอินพุตของสัญญาณเสียง

และ w_n เป็นฟังก์ชันกรอบสัญญาณ

3.1.2.2.3 นำสัญญาณสเปคตรัมที่ผ่านการแปลงสัญญาณฟูริเยร์ช่วงสั้นส่งผ่านชุดกรองฟิลเตอร์มัธยฐาน เพื่อเน้นความสำคัญของความถี่ที่อยู่ในช่วงกล่างของชุดตัวกรองแต่ละตัวกรอง โดยใช้ชุดกรองฟิลเตอร์มัธยฐาน คำนวณหาค่าพลังงานในแต่ละชุดตัวกรอง หลังจากผ่านชุดกรองนนี้จะได้สเปกโตรแกรมฟิลเตอร์ของเสียงประสานและจังหวะดนตรี จากนั้นใช้ไบนารีมาสก์สร้าง STFTs ของเสียงประสานและจังหวะดนตรี

3.1.2.2.4 ในขั้นตอนสุดท้ายนำข้อมูลของเสียงประสานและจังหวะดนตรีมาหาค่าลอการิทึมจากสเปคตรัมของเสียง หลังจากได้ค่าลอการิทึมของสเปคตรัมของเสียงประสานและจังหวะดนตรีแล้ว จึงนำค่าทั้งสองมาหาค่าเฉลี่ยอีกครั้ง

3.1.2.3 ค่าลอการิทึมจากสเปคตรัมของเสียงแบบความต่างและค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลแบบความต่าง

การคำนวณเพื่อหาคุณลักษณะสำคัญทั้ง 2 แบบมีกระบวนการคำนวณแบบเดียวกัน ต่างกันแค่ข้อมูลอินพุตที่เข้ามา ซึ่งในการหาค่าลอการิทึมจากสเปคตรัมของเสียงแบบความต่างจะใช้อินพุตที่เป็นค่าลอการิทึมจากสเปคตรัมของเสียง ส่วนการหาค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลก็จะใช้ค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลมาใช้เป็นอินพุต

การสร้างคุณลักษณะสำคัญของเสียงที่จะป้อนเข้าสู่โครงข่ายประสาทเทียมตามขั้นตอนที่ได้แสดงในภาพที่ 3.5 กำหนดให้มีขนาด {64, 96, 128, 192} แถว x 64 คอลัมน์ x จำนวนชั้นของข้อมูล แถวจะเป็นค่าจำนวนชุดตัวกรองฟิวเตอร์แบงค์ (จำนวนสัมประสิทธิ์) และคอลัมน์เป็นจำนวนเฟรมของสัญญาณเสียง ซึ่งจะประกอบด้วยชุดของคุณลักษณะสำคัญเสียงดังต่อไปนี้

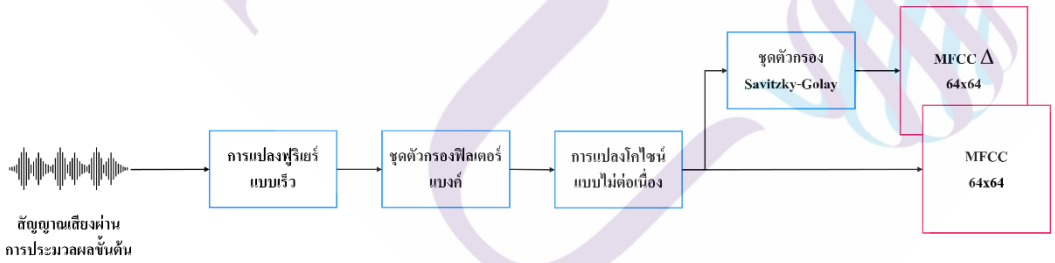
(ก) ค่าสัมประสิทธิ์สเซปสตรัมบนสเกลเมล และค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลแบบความต่าง

(ข) ค่าลอการิทึมจากสเปคตรัมของเสียง และค่าลอการิทึมจากสเปคตรัมของเสียงแบบความต่าง

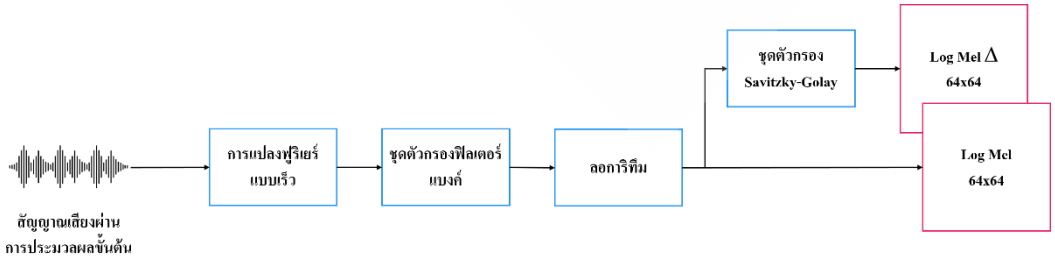
(ค) ค่าสัมประสิทธิ์สเซปสตรัมบนสเกลเมล และค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลแบบความต่าง และค่าเฉลี่ยเสียงประสานและจังหวะดนตรี

(ง) ค่าลอการิทึมจากสเปคตรัมของเสียง และค่าลอการิทึมจากสเปคตรัมของเสียงแบบความต่าง และค่าเฉลี่ยเสียงประสานและจังหวะดนตรี

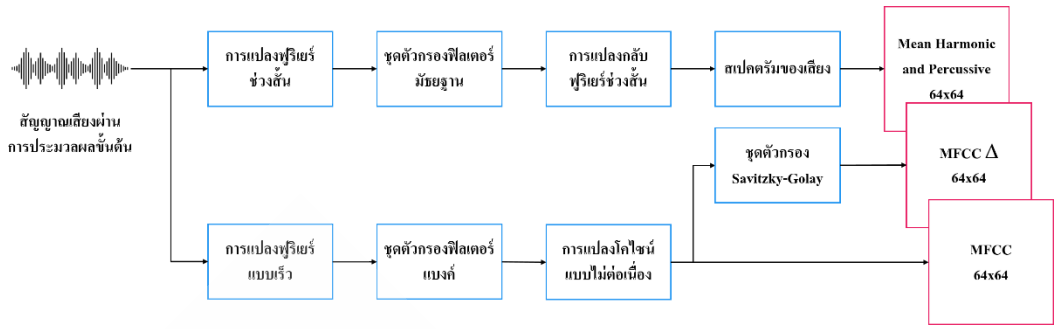
(จ) ใช้ทุกคุณลักษณะสำคัญของเสียงที่ใช้ในงานวิจัยนี้



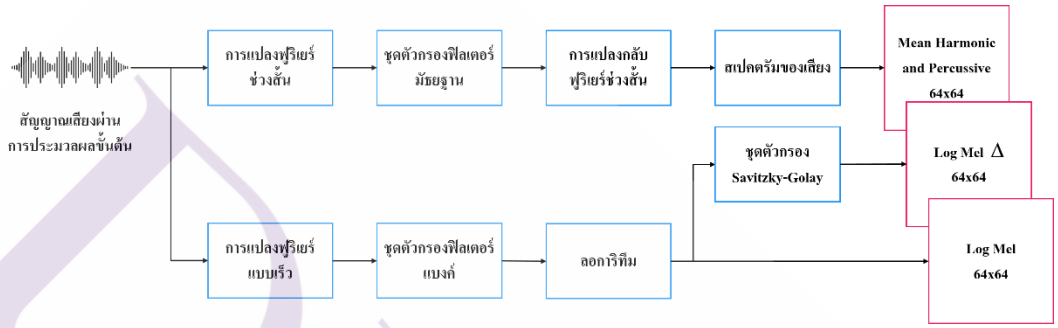
(ก)



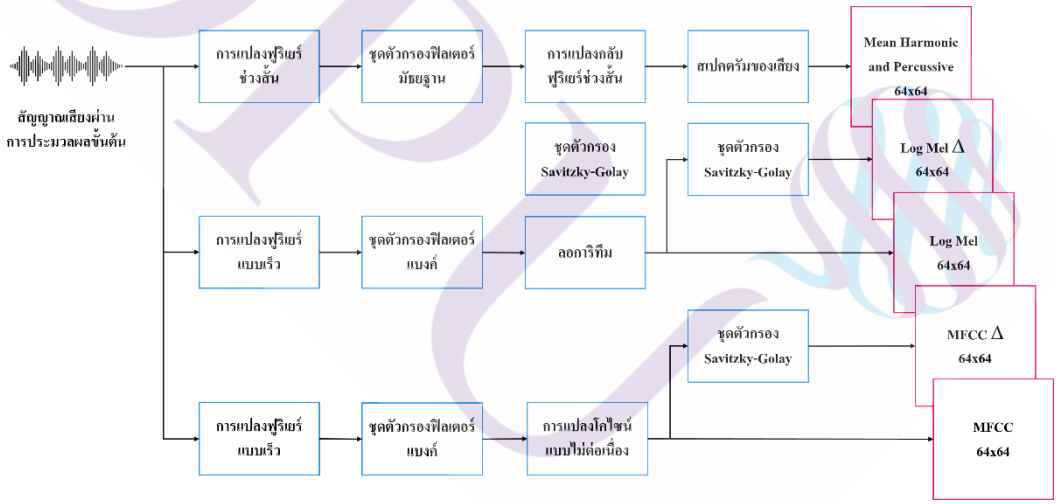
(ข)



(ค)



(ง)



(จ)

ภาพที่ 3.5 คุณลักษณะสำคัญของเสียงที่ใช้ในงานวิจัย

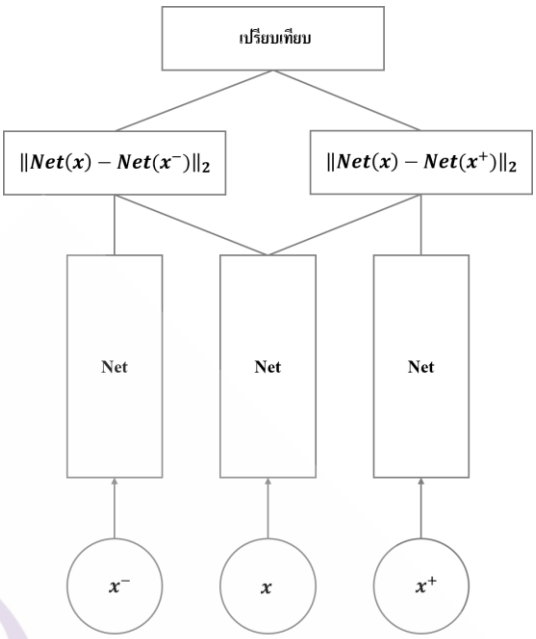
3.1.3 ขั้นตอนการเลือกคุณลักษณะสำคัญเฉพาะ

การทดสอบความคล้ายคลึงกันของเสียง และกฎเกณฑ์การตัดสินใจเข้าสู่กระบวนการเรียนรู้ โดยใช้โครงข่ายประสาทเทียม ซึ่งเป็นการจดจำรูปแบบที่ไม่แน่นอนเนื่องจากความสามารถในการจำลองพฤติกรรมทางกายภาพของระบบ ที่มีความซับซ้อนจากข้อมูลที่ป้อนให้เรียนรู้ การประยุกต์ใช้โครงข่ายประสาทเทียมจึงเป็นทางเลือกที่เหมาะสมกับงานวิจัยนี้

แต่เนื่องด้วยในงานวิจัยนี้ในแต่ละคลาส (เพลง) มีจำนวนตัวอย่างข้อมูลที่ค่อนข้างน้อย หากใช้โครงข่ายประสาทเทียมแบบคอนโวลูชันโดยทั่วไปเพื่อใช้ในการจำแนกคลาสมีโอกาสเสี่ยงที่อาจจะเกิดปัญหาโอเวอร์ฟิตติ้ง (Overfitting) ที่สูง ดังนั้นในงานวิจัยนี้จึงเลือกใช้เทคนิคการเรียนรู้เชิงอภิมาน โดยประยุกต์ใช้งานผ่านโมเดลโครงข่ายทริปเล็ต (Triplet Network)

โครงข่ายทริปเล็ต (ได้รับแรงบันดาลใจจาก โครงข่ายสยาม) ประกอบด้วยโครงข่ายประสาทเทียมแบบป้อนไปหน้า (Feed-forward Network) ย่อย 3 โครงข่ายดังภาพที่ 3.6 โดยเมื่อป้อนอินพุต 3 ตัวอย่างประกอบด้วย คุณลักษณะสำคัญของเสียงอ้างอิง x, คุณลักษณะสำคัญของเสียงที่เหมือนกับคุณลักษณะสำคัญของเสียงอ้างอิง x^+, และคุณลักษณะสำคัญของเสียงที่ต่างกับคุณลักษณะสำคัญของเสียงอ้างอิง x^- แต่ละตัวอย่างผ่านโครงข่าย $Net(x)$ เพื่อสร้างคุณลักษณะเวกเตอร์เพื่อแสดงถึงเนื้อหาของอินพุตนั้นๆ และสุดท้ายจะได้เอาต์พุตเป็นค่ากลาง 2 ค่า คือ 1) ระยะห่างระหว่างคุณลักษณะเวกเตอร์อ้างอิง (Anchor) และคุณลักษณะเวกเตอร์ที่เหมือนกับคุณลักษณะเวกเตอร์อ้างอิง (Positive) 2) ระยะห่างระหว่างคุณลักษณะเวกเตอร์อ้างอิง และคุณลักษณะเวกเตอร์ที่ต่างกับคุณลักษณะเวกเตอร์อ้างอิง (Negative) การหาระยะห่างระหว่างคุณลักษณะเวกเตอร์จะใช้การคำนวณระยะทางแบบยุคลิด (Euclidean Distance) ตามสมการที่ (3-6)

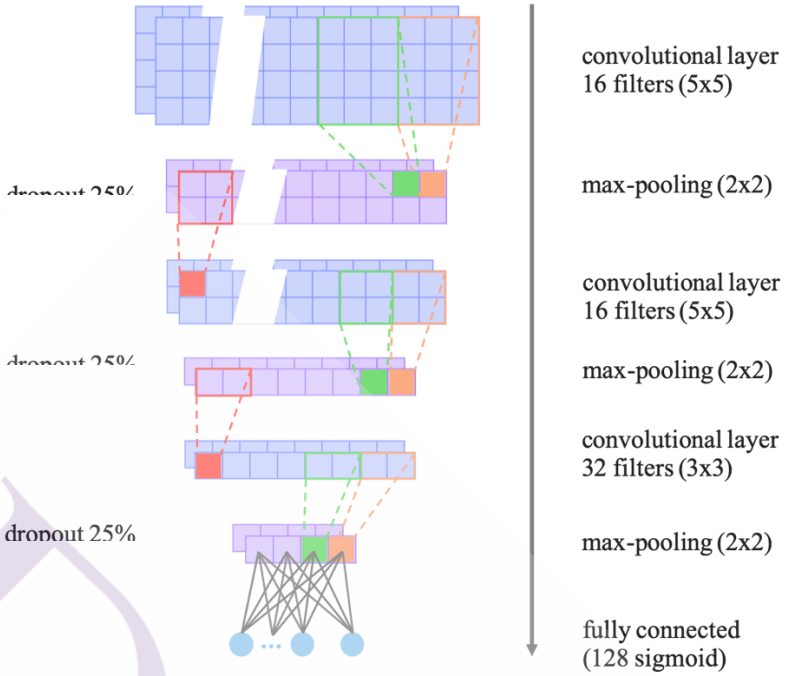
$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3-6}$$



ภาพที่ 3.6 สถาปัตยกรรมของโครงข่ายทริปเล็ต

3.1.3.1 สถาปัตยกรรมของโครงข่ายทริปเล็ต

สถาปัตยกรรมของโครงข่ายทริปเล็ตที่นำเสนอในงานวิจัยนี้จะประยุกต์ใช้โครงข่ายประสาทเทียมแบบคอนโวลูชันมาเป็นโครงข่ายย่อยในโครงข่ายทริปเล็ตตามภาพที่ 3.7 ในแต่ละโครงข่ายย่อยจะประกอบไปด้วยชั้นคอนโวลูชัน (Convolutional) ทั้งหมดจำนวน 3 ชั้นการประมวลผล โดยแต่ละคอนโวลูชันจะถูกคำนวณค่าบนข้อมูลคุณลักษณะสำคัญของเสียงที่ได้ในแต่ละตัวกรอง (Filter) กับค่าน้ำหนัก โดยผ่านฟังก์ชันกระตุ้น (Activation Function) เพื่อส่งต่อไปยังกระบวนการนำออกกลางคัน (Dropout) เพื่อตัดคุณลักษณะบางอย่างออกไป โดยจะอยู่ต่อจากชั้นคอนโวลูชันในแต่ละชั้น จากนั้นจึงจะเป็นชั้นบ่อรวมสูงสุด (Max Pooling) ทั้งหมด 3 ชั้นการคำนวณ จะอยู่ระหว่างชั้นคอนโวลูชัน ซึ่งจะเป็นกระบวนการในการลดรายละเอียดของคุณลักษณะลง โดยการเลือกค่าที่มาที่สุดในแต่ละตัวกรอง และในชั้นสุดท้ายเป็นชั้นการเชื่อมต่ออย่างสมบูรณ์ (Fully Connected) และใช้ฟังก์ชันกระตุ้นเป็นซิกมอยด์ (Sigmoid) ในชั้นนี้เป็นการสร้างคุณลักษณะเวกเตอร์ที่จะใช้เป็นตัวแทนของเนื้อหาของเสียง ในงานวิจัยนี้จะทดสอบผลกับคุณลักษณะเวกเตอร์ขนาด {32, 64, 128, 256, 512}



ภาพที่ 3.7 สถาปัตยกรรมโครงข่ายย่อยในโครงข่ายทริปเล็ต

3.1.3.2 การฝึกสอนโมเดล

การฝึกสอนจะดำเนินการโดยการป้อนข้อมูลคุณลักษณะสำคัญของเสียงเข้าสู่โครงข่ายตามที่ได้อธบายไว้ข้างต้นคือ คุณลักษณะสำคัญของเสียงอ้างอิง x, คุณลักษณะสำคัญของเสียงที่เหมือนกับคุณลักษณะสำคัญของเสียงอ้างอิง x^+, และคุณลักษณะสำคัญของเสียงที่ต่างกับคุณลักษณะสำคัญของเสียงอ้างอิง x^- สถาปัตยกรรมโครงข่ายช่วยแก้ปัญหา 2 คลาส ซึ่งมีวัตถุประสงค์เพื่อจำแนกประเภทของ x^+หรือ x^- ว่าสิ่งใดมีคลาสเดียวกับ x หรือเพื่อวัตถุประสงค์ในการเรียนรู้เพื่อหาคุณลักษณะเวกเตอร์ตัวชี้วัดที่มีความใกล้เคียงกับ x เพื่อที่สุดท้ายแล้วจะนำเอาท์พุตทั้ง 2 ไปเปรียบเทียบเพื่อสร้างอัตราส่วนชี้วัด เช่นเดียวกับคอนโวลูชันแบบดั้งเดิมการฝึกสอนยังทำโดยใช้การเลื่อนลงตามความชัน (Gradient Descent) ในการสูญเสีย Negative Log Likelihood จากปัญหา 2 คลาส แต่เพื่อให้ผลลัพธ์ที่ดีจะใช้ฟังก์ชันสูญเสียตามสมการที่ (3-7) แทนที่ฟังก์ชันสูญเสียแบบเดิม

$$Loss(d_+, d_-) = \|(d_+, d_- - 1)\|_2^2 = const \cdot d_+^2 \tag{3-7}$$

โดย

$$d_+ = \frac{e^{\|Net(x)-Net(x^+)\|_2}}{e^{\|Net(x)-Net(x^+)\|_2} + e^{\|Net(x)-Net(x^-)\|_2}} \tag{3-8}$$

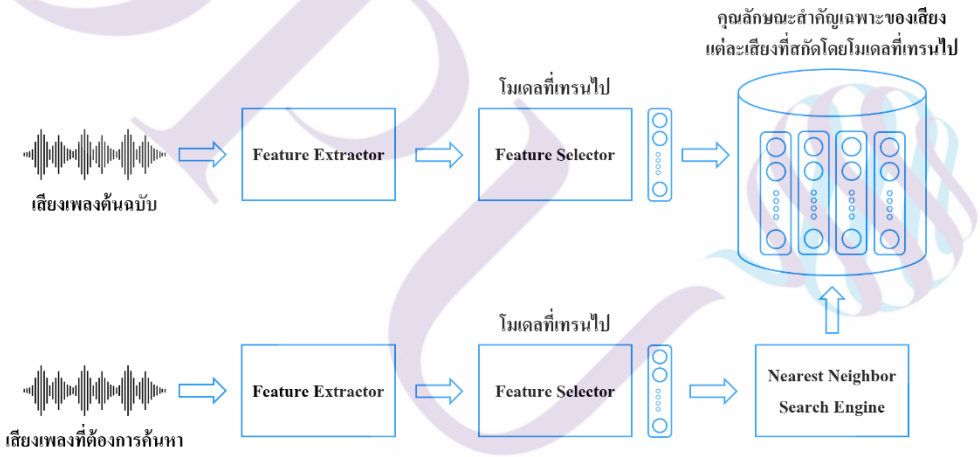
และ

$$d_- = \frac{e^{\|Net(x)-Net(x^-)\|_2}}{e^{\|Net(x)-Net(x^+)\|_2} + e^{\|Net(x)-Net(x^-)\|_2}} \tag{3-9}$$

เป้าหมายคือพยายามให้ $Loss(d_+, d_-) \rightarrow 0$ หรือ $\frac{\|Net(x)-Net(x^+)\|}{\|Net(x)-Net(x^-)\|} \rightarrow 0$ ผลลัพธ์เข้าใกล้ 0 การฝึกสอนโครงข่ายนั้นในทุกโครงข่ายย่อยนั้นจะแชร์การตั้งค่าและพารามิเตอร์

3.1.4 การจำลองการค้นหาเสียงเพลง

ในงานวิจัยนี้จะจำลองการค้นหาเพลงจากคุณลักษณะสำคัญเฉพาะของเสียงที่สกัดจากโมเดลที่ได้จากการวิเคราะห์ข้อมูลคุณลักษณะสำคัญของเสียง โดยทดลองกับข้อมูลเสียงที่ความยาวเสียง 3 วินาที จากตัวอย่างเสียงจำนวน {100, 200, 300, 500} ตามเฟรมเวิร์ค (Framework) ดังภาพที่ 3.8



ภาพที่ 3.8 เฟรมเวิร์คที่ใช้ในการจำลองการค้นหาเสียงเพลง

การค้นหาข้อมูลคุณลักษณะสำคัญเฉพาะจากฐานข้อมูล ในงานวิจัยนี้ประยุกต์ใช้วิธีการค้นหาเพื่อนบ้านใกล้สุด (Nearest Neighbor Search) การค้นหาข้อมูลด้วยวิธีการค้นหาเพื่อนบ้านใกล้สุดจะเป็นการเปรียบเทียบกันระหว่างข้อมูลที่ใช้ค้นหากับข้อมูลในฐานข้อมูลทั้งหมดว่ามีลักษณะเหมือนกันหรือใกล้เคียงกันด้วยการพิจารณาข้อมูลแอทริบิวต่างๆ โดยข้อมูลเรคคอร์ดหนึ่งๆ จะสามารถมองว่าเป็นจุดหนึ่งในระนาบ n มิติ (เมื่อ n คือจำนวนแอทริบิวทั้งหมด) ถ้านำ

ข้อมูลในฐานข้อมูลในทุกๆ เรคคอร์ดมาวางในระนาบ n มิติ จากนั้นนำข้อมูลที่ใช้ค้นหามาวางในระนานด้วยเช่นกัน จากนั้นพิจารณาหาว่ามีข้อมูลจุดใดบ้าง (เรคคอร์ดใดบ้าง) มีคุณลักษณะใกล้เคียงกันข้อมูลที่ใช้ค้นหาบ้างเป็นจำนวน k เรคคอร์ด

การที่จะหาความเหมือนกันต่างกันหรือค่าความคล้ายคลึงกันระหว่างข้อมูลใดๆ เราสามารถประยุกต์ใช้มาตรวัดความแตกต่าง อาทิเช่น การคำนวณระยะทางแบบยุคลิด ซึ่งจะทำการพิจารณาความแตกต่างระหว่างค่าที่ปรากฏขึ้นในแต่ละแอทริบิวของข้อมูลทั้งสองแล้วนำมารวมเป็นค่าความแตกต่างรวม ซึ่งสามารถจะคำนวณได้ตามสมการที่ (3-6)

ในการค้นหาหนึ่งคลิปเสียงจะถูกแบ่งเป็นส่วนย่อย 3 ส่วนเพื่อใช้ในการค้นหาเสียงเพลงจากฐานข้อมูล ในการเลือกที่จะตอบว่าคลิปเสียงนี้เป็นเพลงอะไรพิจารณาจากข้อมูลเพลงที่ใช้ค้นหาทั้ง 3 ส่วนให้คำตอบมา หากผลลัพธ์ที่ออกมาเหมือนกัน 2 ใน 3 จะถือว่าเพลงนี้เป็นคำตอบของคลิปเสียงนี้

3.1.5 การเปรียบเทียบและการประเมินผล

การประเมินผลจะพิจารณาจากผลการค้นหาวิธีการค้นหาเพื่อนบ้านใกล้สุดว่าถูกต้องมากน้อยเพียงใด เพื่อนำมาคำนวณค่าความแม่นยำตามสมการที่ (3-10)

$$\text{accuracy} = \frac{\text{จำนวนเพลงที่ตอบถูกต้อง}}{\text{จำนวนเพลงทั้งหมด}} \tag{3-10}$$

3.2 เครื่องมือที่ใช้ในการวิจัย

3.2.1 ฮาร์ดแวร์ที่ใช้ในงานวิจัย

1. หน่วยประมวลผล: Intel® Core i5 9400F
2. หน่วยความจำสำรอง: 500 GB
3. หน่วยความจำหลัก: 32 GB
4. การ์ดจอ NVIDIA GeForce RTX 2060 SUPER 8GB

3.2.2 ภาษา Python

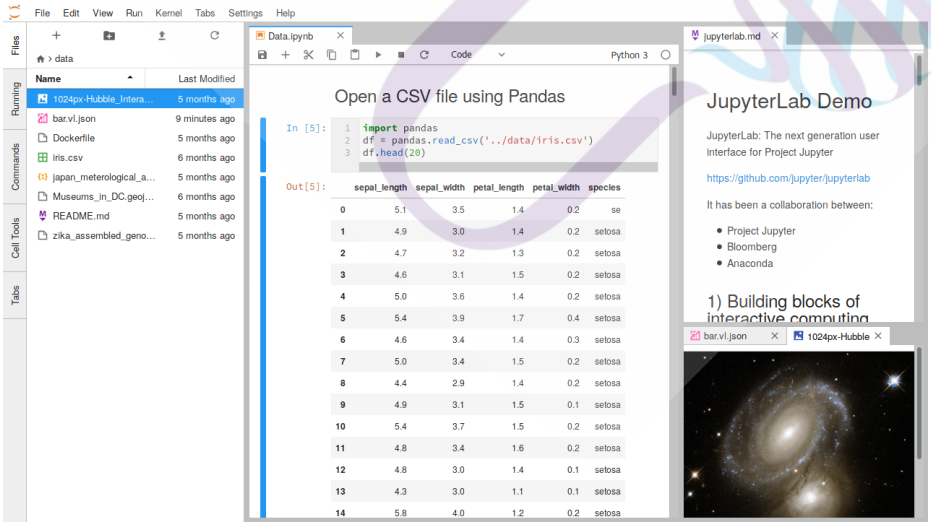
Python เป็นภาษาเขียนโปรแกรมระดับสูงที่ใช้กันอย่างกว้างขวางในการเขียนโปรแกรมสำหรับวัตถุประสงค์ทั่วไป ภาษา Python นั้นสร้างโดย Guido van Rossum และถูกเผยแพร่ครั้งแรกในปี 1991 Python นั้นเป็นภาษาแบบ interprete ที่ถูกออกแบบโดยมีปรัญชาที่จะทำให้โค้ดอ่านได้ง่ายขึ้น และโครงสร้างของภาษานั้นจะทำให้โปรแกรมเมอร์สามารถเข้าใจแนวคิดการเขียนโค้ดโดยใช้บรรทัดที่น้อยลงกว่าภาษาอย่าง C++ และ Java ซึ่งภาษานั้นถูกกำหนดให้มีโครงสร้างที่ตั้งใจให้การเขียนโค้ดเข้าใจง่ายทั้งในโปรแกรมเล็กไปจนถึงโปรแกรมขนาดใหญ่

Python นั้นมีคุณสมบัติเป็นภาษาเขียนโปรแกรมแบบไดนามิกส์และมีระบบการจัดการหน่วยความจำอัตโนมัติและสนับสนุนการเขียนโปรแกรมหลายรูปแบบ ที่ประกอบไปด้วย การเขียนโปรแกรมเชิงวัตถุ imperative การเขียนโปรแกรมแบบฟังก์ชัน และการเขียนโปรแกรมแบบขั้นตอน มีไลบรารี่ที่ครอบคลุมการทำงานอย่างหลากหลาย ซึ่งไลบรารี่ที่สำคัญในงานวิจัยนี้ได้แก้

1. Librosa (0.7.2) เป็นไลบรารี่ที่ใช้ในงานด้านการวิเคราะห์ดนตรีและเสียง
2. Pandas (1.0.1) เป็นไลบรารี่ที่ใช้การจัดกับข้อมูล
3. Numpy (1.17) เป็นไลบรารี่ที่ใช้จัดการเกี่ยวกับการดำเนินการข้อมูลที่เป็น อาเรย์ (Array) หรือเมทริกซ์ (Matrix)
4. Tensorflow (2.0) เป็นไลบรารี่ที่ใช้สำหรับสร้างโมเดลโครงข่ายทริปเล็ต

3.2.3 JupyterLab

JupyterLab เป็นเครื่องมือที่พัฒนาต่อยอดมาจาก Jupyter Notebook ซึ่งเป็นเครื่องมือหนึ่งที่ได้รับความนิยมใช้ในสายงาน Data Science เปิดให้ใช้งานตั้งแต่เดือนกุมภาพันธ์ 2018 ในเวอร์ชั่น 0.32 JupyterLab ออกแบบให้สามารถทำงานได้ทั้งงานเอกสารและกิจกรรม เช่น Jupyter Notebook, โปรแกรมแก้ไขข้อความ (Text Editors), เทอร์มินอล (Terminal) และองค์ประกอบ (Component) เสริมอื่นๆ



ภาพที่ 3.9 ตัวอย่างหน้าจอของ JupyterLab

บทที่ 4
ผลการศึกษา

งานวิจัยนี้แบ่งออกเป็น 2 ส่วนคือ การวิเคราะห์คุณลักษณะสำคัญของเสียงเพื่อเลือกคุณลักษณะสำคัญเฉพาะของเสียงและการค้นหาเพลงจากคุณลักษณะสำคัญเฉพาะโดยใช้โมเดลที่ได้จากการวิเคราะห์ ซึ่งในการทดลองวิเคราะห์คุณลักษณะสำคัญของเสียงจะประยุกต์ใช้โมเดลโครงข่ายทริปเล็ต เพื่อใช้เลือกคุณลักษณะสำคัญเฉพาะของเสียง จากข้อมูลเสียงเพลงจำนวน 200 เพลง หลังจากวิเคราะห์และได้โมเดลสำหรับเลือกคุณลักษณะสำคัญเฉพาะของเสียงแล้ว จะนำโมเดลดังกล่าวมาสกัดคุณลักษณะสำคัญเฉพาะจากเสียงที่จะสำหรับจำลองการค้นหาเพลงด้วยวิธีการค้นหาเพื่อนบ้านใกล้สุด เพื่อคำนวณหาค่าความแม่นยำในการค้นหาเสียงเพลง

4.1 การประยุกต์ใช้โมเดลโครงข่ายทริปเล็ต

การประยุกต์ใช้โมเดลโครงข่ายทริปเล็ตเพื่อใช้เลือกคุณลักษณะสำคัญเฉพาะของเสียงเริ่มต้นด้วยข้อมูลของสัญญาณเสียงที่ผ่านกระบวนการประมวณผลสัญญาณเสียงเบื้องต้น กระบวนการแยกคุณลักษณะสำคัญของเสียง ซึ่งได้กล่าวไว้เบื้องต้น ในขั้นตอนถัดไปเป็นการนำข้อมูลเสียงเพลงที่ผ่านจากกระบวนการข้างต้น มาแบ่งออกเป็น 2 กลุ่ม คือข้อมูลที่ใช้สำหรับฝึกสอน (Training) โครงข่ายทริปเล็ต และข้อมูลสำหรับตรวจสอบ (Validation)

4.1.1 ทดสอบเพื่อเลือกคุณลักษณะสำคัญ

การเลือกคุณลักษณะสำคัญนั้นส่งผลต่อความถูกต้องในการเลือกคุณลักษณะสำคัญเฉพาะของโมเดลที่จะสร้าง ดังนั้นในงานวิจัยนี้จึงเลือกจัดเตรียมข้อมูลเพื่อใช้ในการทดสอบดังต่อไปนี้ ซึ่งคุณลักษณะสำคัญจะถูกจัดเรียงข้อมูลในรูปแบบเมตริกซ์ขนาดต่างๆ ดังต่อไปนี้

- เมตริกซ์ขนาด 64 x 64 x 2
- เมตริกซ์ขนาด 64 x 64 x 3
- เมตริกซ์ขนาด 64 x 64 x 5

ตารางที่ 4.1 การจัดเรียงข้อมูลก่อนนำเข้ากระบวนการเรียนรู้และตรวจสอบ

คุณลักษณะสำคัญ	ขนาด	จำนวนตัวอย่างต่อเพลง
MFCC + MFCC-Δ	64 x 64 x 2	163
Log Mel + Log Mel-Δ	64 x 64 x 2	163
MFCC + MFCC-Δ + HPSS	64 x 64 x 3	163
Log Mel + Log Mel-Δ + HPSS	64 x 64 x 3	163
MFCC + Log Mel + MFCC-Δ + Log Mel-Δ + HPSS	64 x 64 x 5	163

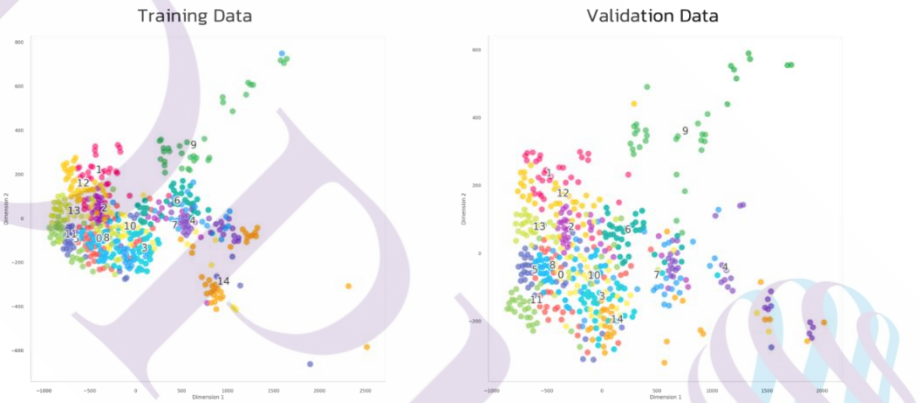
ในการทดสอบเพื่อเลือกคุณลักษณะสำคัญได้เตรียมโครงข่ายทริปเล็ต โดยกำหนดขนาดของคุณลักษณะเวกเตอร์ที่เป็นเอาท์พุตไว้ที่ 128 x 1, ไม่เพิ่มชั้นการนำออกกลางคันในโครงข่ายย่อย และกำหนดการเรียนรู้จำนวน 50 Epoch โดยแต่ละรอบการเรียนรู้จำกำหนดสุ่มข้อมูลชุดฝึกสอนจำนวน 64 ตัวอย่าง มีการฝึกสอนซ้ำในแต่ละรอบการเรียนรู้จำนวน 1000 รอบ หลังจากทดสอบพบว่าคุณลักษณะสำคัญโดยใช้ค่าสัมประสิทธิ์สเซปสตรัมบนสเกลเมล และค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลแบบความต่าง ให้ค่าเปอร์เซ็นต์ความคลาดเคลื่อนต่ำที่สุดอยู่ที่ 6.7% ดังแสดงในตารางที่ 4.2

ตารางที่ 4.2 ผลการทดสอบและเปรียบเทียบคุณลักษณะสำคัญที่ใช้เป็นอินพุตให้กับโมเดลโครงข่ายทริปเล็ต

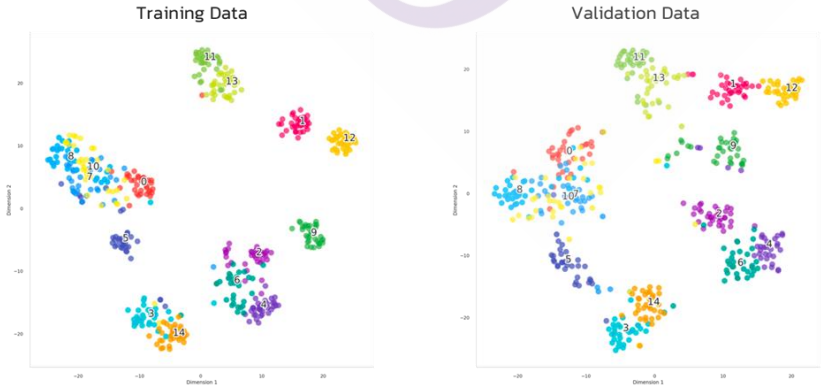
คุณลักษณะสำคัญ	รอบการเรียนรู้	เปอร์เซ็นต์ (%) ความคลาดเคลื่อน
MFCC + MFCC-Δ	47	6.7%
Log Mel + Log Mel-Δ	46	8.9%
MFCC + MFCC-Δ + HPSS	47	7.3%
Log Mel + Log Mel-Δ + HPSS	45	8.4%
MFCC + Log Mel + MFCC-Δ + Log Mel-Δ + HPSS	49	7.4%

จากตารางที่ 4.2 จะเห็นได้ว่าการจัดเรียงข้อมูลที่ประกอบด้วยคุณลักษณะจากค่าสัมประสิทธิ์สเซปสตรัมบนสเกลเมล และค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลแบบความต่างให้ค่าเปอร์เซ็นต์ความคลาดเคลื่อนที่ต่ำอย่างมีนัยสำคัญ อนึ่งในหลายงานวิจัยที่เกี่ยวกับการรู้จำเสียงและใช้การแยกคุณลักษณะโดยใช้ค่าสัมประสิทธิ์สเซปสตรัมบนสเกลเมลให้ผลลัพธ์ที่ดี

ในภาพที่ 4.1 และ 4.2 ใช้การจัดเรียงข้อมูลที่เป็นค่าสัมประสิทธิ์สเซปสตรัมบนสเกลเมล และค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลแบบความต่าง โดยสุ่มหยิบตัวอย่างข้อมูลเพลงจำนวน 15 เพลง เพื่อดูการกระจายตัวของข้อมูลทั้งก่อนฝึกสอนและหลังฝึกสอน โดยใช้เทคนิค PCA เพื่อลดมิติของข้อมูล จะเห็นได้ว่าหลังฝึกสอนข้อมูลมีการเกาะกลุ่มกันแน่นขึ้น และมีการทับซ้อนที่น้อยลง



ภาพที่ 4.1 แสดงการกระจายตัวของข้อมูลก่อนการฝึกสอน



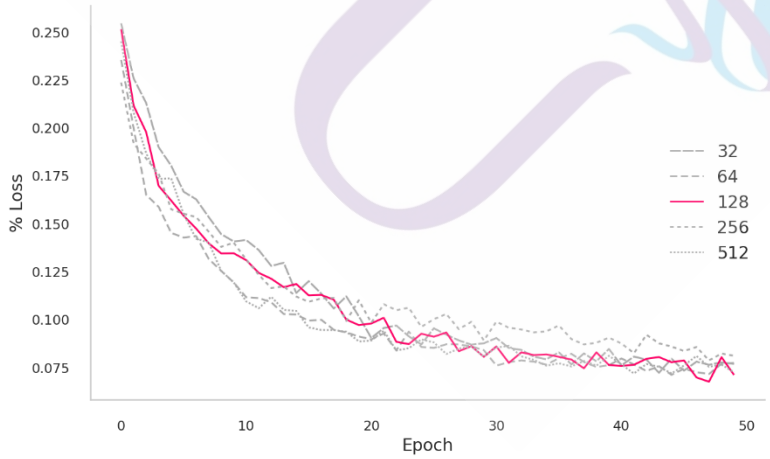
ภาพที่ 4.2 แสดงการกระจายตัวของข้อมูลหลังการฝึกสอน

4.1.2 ทดสอบเพื่อเลือกขนาดของคุณลักษณะเวกเตอร์

ในการวิเคราะห์เพื่อเลือกคุณลักษณะสำคัญเฉพาะของเสียงขนาดของคุณลักษณะเวกเตอร์มีผลต่อความถูกต้องของข้อมูล ซึ่งหากแคบไปก็จะเกิดการทับซ้อนกันของข้อมูลที่ค่อนข้างสูง แต่หากกว้างไปก็จะเกิดการกระจายตัวของข้อมูลที่สูง ในงานวิจัยนี้พิจารณาขนาดของคุณลักษณะเวกเตอร์ที่ 32, 64, 128, 256 และ 512 ซึ่งพบว่าคุณลักษณะเวกเตอร์ขนาด 128 มีเปอร์เซ็นต์ความคลาดเคลื่อนต่ำสุดที่ 6.7% ดังแสดงในตารางที่ 4.3

ตารางที่ 4.3 ผลการทดสอบและเปรียบเทียบขนาดคุณลักษณะเวกเตอร์ที่ใช้เป็นเอาท์พุตของโครงข่ายย่อยในโครงข่ายทริปเล็ต

ขนาดคุณลักษณะเวกเตอร์	รอบการเรียนรู้	เปอร์เซ็นต์ (%) ความคลาดเคลื่อน
32	43	7.2%
64	44	7.1%
128	47	6.7%
256	47	7.8%
512	41	7.1%



ภาพที่ 4.3 กราฟแสดงการเปรียบเทียบความคลาดเคลื่อนของโมเดลที่มีขนาดของคุณลักษณะเวกเตอร์ที่ต่างกัน

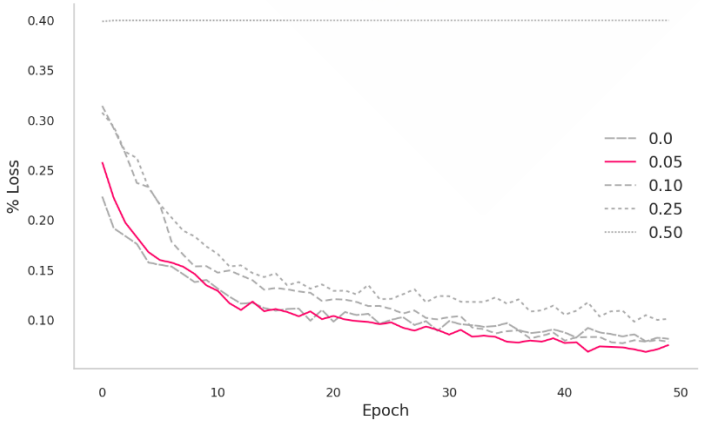
4.1.3 ทดสอบเพิ่มชั้นการนำออกกลางคันในโครงข่ายย่อย

จากการทดสอบเพิ่มชั้นการนำออกกลางคันในโครงข่ายย่อย โดยพิจารณาที่เปอร์เซ็นต์การนำออกกลางคันที่ 5%, 10%, 25%, 50% และไม่มีการเพิ่มชั้นนี้ พบว่าการเพิ่มชั้นการนำออกกลางคันไม่มีผลกับการเรียนรู้ของโมเดลกับโมเดลที่มีคุณลักษณะเวกเตอร์ขนาดน้อยกว่า 256 อาจจะเนื่องจากข้อมูลมีจำนวนที่น้อย เลยทำให้เมื่อเพิ่มชั้นการนำออกกลางคันเข้าไปทำให้ข้อมูลสำคัญบางอย่างอาจถูกตัดทิ้งไป ตามที่แสดงในตารางที่ 4.4

ตารางที่ 4.4 ผลการทดสอบและเปรียบเทียบขนาดของชั้นการนำออกกลางคันของโครงข่ายย่อยในโครงข่ายทริปเล็ตที่ใช้คุณลักษณะเวกเตอร์ขนาด 128

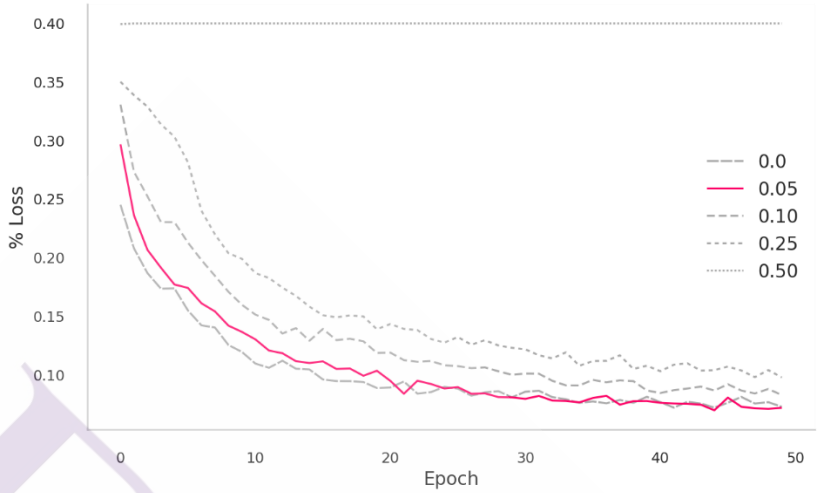
ขนาดคุณลักษณะเวกเตอร์	รอบการเรียนรู้	เปอร์เซ็นต์ (%) ความคลาดเคลื่อน
None	47	6.7%
5%	45	7.3%
10%	49	7.7%
25%	48	10.4%
50%	1	39.9%

จากภาพที่ 4.4 และ 4.5 จากจะเห็นว่าการเพิ่มชั้นการนำออกกลางคันส่งผลต่อโมเดลที่มีคุณลักษณะเวกเตอร์ขนาดใหญ่



ภาพที่ 4.4 กราฟแสดงการเปรียบเทียบความคลาดเคลื่อนของโมเดลที่มีขนาดชั้นการนำออก

กลางคันที่ต่างกันกับคุณลักษณะเวกเตอร์ที่เป็นเอาท์พุตขนาด 256



ภาพที่ 4.5 กราฟแสดงการเปรียบเทียบความคลาดเคลื่อนของโมเดลที่มีขนาดชั้นการนำออก กลางคันที่ต่างกันกับคุณลักษณะเวกเตอร์ที่เป็นเอาท์พุตขนาด 512

4.2 การจำลองการค้นหาเพลง

เมื่อทดสอบและเปรียบเทียบผลการประยุกต์ใช้โมเดลโครงข่ายทริปเล็ตเพื่อหาพารามิเตอร์และการจัดเรียงคุณลักษณะสำคัญของเสียงเพื่อจะใช้เป็นอินพุตของโมเดล จะใช้การจัดเรียงข้อมูลโดยใช้ค่าสัมประสิทธิ์สเซปสตรัมบนสเกลเมล และค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลแบบความต่าง ในส่วนของโมเดลจะไม่มีการเพิ่มชั้นการนำออกกลางคันเข้าไปและกำหนดเอาท์พุตให้มีขนาดของคุณลักษณะเวกเตอร์ที่ 128 x 1 กำหนดการเรียนรู้จำนวน 100 Epoch โดยแต่ละรอบการเรียนรู้จำกำหนดสุ่มข้อมูลชุดฝึกสอนจำนวน 64 ตัวอย่าง มีการฝึกสอนซ้ำในแต่ละรอบการเรียนรู้จำนวน 1000 รอบ หลังจากฝึกสอนเป็นที่เรียบร้อยแล้ว สิ่งที่จะนำมาใช้เพื่อสกัดคุณลักษณะสำคัญเฉพาะคือโมเดลย่อยภายในโมเดลโครงข่ายทริปเล็ต ตามที่ได้แสดงในภาพที่ 3.8

4.2.1 เปรียบเทียบผลการค้นหาเพลงที่ความยาวเสียงที่ต่างกัน

ทดสอบและเปรียบเทียบผลการค้นหาเพลงด้วยโมเดลโครงข่ายทริปเล็ตที่ผ่านการฝึกสอนแล้ว ด้วยข้อมูลเพลงจำนวน 100 เพลง ด้วยความยาวเพลงที่ใช้ค้นหาที่ต่างกัน จากผลการทดสอบในตารางที่ 4.4 จะเห็นว่ายิ่งข้อมูลเสียงมีความยาวมากขึ้นความถูกต้องในการค้นหาเพลงก็ยิ่งเพิ่มสูงตามไปด้วย

ตารางที่ 4.5 ผลการทดสอบและเปรียบเทียบการค้นหาเพลงที่ความยาวเพลงต่างกัน

ความยาวเพลง	จำนวนเพลงที่ทายถูก	เปอร์เซ็นต์ (%) ความแม่นยำ
1 วินาที	82	82%
3 วินาที	86	86%
5 วินาที	97	97%
10 วินาที	100	100%
15 วินาที	100	100%

4.2.2 เปรียบเทียบผลการค้นหาเพลงที่มีสัญญาณรบกวน

ทดสอบและเปรียบเทียบผลการค้นหาเพลงด้วยโมเดลโครงข่ายทริปเล็ตที่ผ่านการฝึกสอนแล้ว ด้วยข้อมูลเพลงจำนวน 100 เพลง ด้วยความยาวเพลงที่ใช้ค้นหาที่ต่างกัน พร้อมด้วยเพิ่มสัญญาณรบกวนควอน (Quantization Noise) ที่มีความถี่สม่ำเสมอมีขนาดสูงสุดที่ 512

ตารางที่ 4.6 ผลการทดสอบและเปรียบเทียบการค้นหาเพลงที่มีสัญญาณรบกวน

ความยาวเพลง	จำนวนเพลงที่ทายถูก	เปอร์เซ็นต์ (%) ความแม่นยำ
1 วินาที	76	76%
3 วินาที	83	83%
5 วินาที	91	91%
10 วินาที	99	99%
15 วินาที	99	99%

4.2.3 เปรียบเทียบผลการค้นหาเพลงด้วยอัลกอริทึมที่เกี่ยวข้อง

ทดสอบและเปรียบเทียบผลการค้นหาเพลงระหว่างโมเดลโครงข่ายทริปเล็ตและอัลกอริทึมลายนิ้วมือเสียง (Audio Fingerprint) ซึ่งเป็นอัลกอริทึมหนึ่งที่นิยมใช้เพื่อค้นหาและเปรียบเทียบความคล้ายคลึงกันของข้อมูลเสียง การทดสอบนี้จะเลือกทดสอบที่จำนวน {100, 200, 300, 400, 500} เพลงตามลำดับ ด้วยความยาวเสียง (ความยาวเพลงสุ่ม) 3 วินาที

ตารางที่ 4.7 ผลการทดสอบและเปรียบเทียบการค้นหาเพลง

จำนวนเพลงในฐานข้อมูล	Audio Fingerprint	Triplet Network
100 เพลง	97	86
200 เพลง	197	174
300 เพลง	291	223
400 เพลง	386	312
500 เพลง	481	359

จากในตารางที่ 4.7 จะเห็นว่าอัลกอริทึมลายนิ้วมือเสียงยังให้ผลดีกว่าโมเดลโครงข่ายสยามในงานวิจัยนี้ จากผลการทดสอบจะเห็นว่ายิ่งจำนวนเพลงเพิ่มมากขึ้นความถูกต้องในการทำนายผลยิ่งน้อยลงตามไปด้วย จะเห็นว่าอัลกอริทึมลายนิ้วมือเสียงให้เปอร์เซ็นต์ความแม่นยำประมาณ 97% ส่วนโมเดลโครงข่ายทริปเล็ตให้เปอร์เซ็นต์ความแม่นยำประมาณ 79% ซึ่งหากตรวจสอบข้อมูลเพลงที่ใช้ในการทดสอบพบว่าแนวเพลงหรือเสียงดนตรีในบางเพลงมีความแตกต่างจากข้อมูลเพลงที่ใช้สำหรับฝึกสอนนั้นอาจเป็นหนึ่งสาเหตุที่ทำให้โมเดลมีความถูกต้องน้อยลง

บทที่ 5
บทสรุปและข้อเสนอแนะ

งานวิจัยในวิทยานิพนธ์เล่มนี้ได้กล่าวถึงการวิจัยทางด้านการประมวลผลสัญญาณดิจิตอล และโครงข่ายประสาทเทียม เพื่อนำมาพัฒนาและประยุกต์ใช้ในการเลือกคุณลักษณะสำคัญของเสียงเพื่อใช้เป็นตัวแทนของเสียงด้วยโมเดลโครงข่ายทริปเล็ต โดยใช้ตัวอย่างจากเพลงจำนวน 200 เพลงเพื่อสร้างเป็นชุดข้อมูลสำหรับฝึกสอน ในการพิจารณาเพื่อเลือกคุณลักษณะสำคัญมีปัจจัยอย่างเช่น วิธีการแยกคุณลักษณะสำคัญของเสียง, ขนาดของคุณลักษณะเวกเตอร์ที่เป็นเอาท์พุต, ขนาดของตัวอย่างเสียงในแต่ละเพลงที่ใช้เป็นอินพุต

งานวิจัยนี้ผู้วิจัยได้นำเอาขั้นตอนและวิธีการต่างๆ มาใช้เพื่อพัฒนาให้โมเดลโครงข่ายทริปเล็ตสามารถแยกเอาคุณลักษณะสำคัญเฉพาะของเสียงนั้นๆ โดยเริ่มต้นตั้งแต่การนำสัญญาณเสียงที่ผ่านขั้นตอนและกระบวนการต่างๆ เช่น การประมวลผลสัญญาณเสียง เพื่อลดทอนสัญญาณรบกวนของสัญญาณเสียง การแยกคุณลักษณะสำคัญของเสียงด้วยวิธีการต่างๆ อาทิ ค่าลอการิทึมของเสียงสเปคตรัมบนสเกลเมล (Log scaled Mel-spectrogram), ค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล (Mel Frequency Cepstral Coefficient), ค่าเสียงประสานและจังหวะดนตรี (Harmonic and Percussive) เพื่อให้ได้ชุดข้อมูลของสัญญาณ และเพื่อทำให้ได้คุณลักษณะสำคัญเฉพาะของเสียงจึงใช้โมเดลโครงข่ายทริปเล็ตเพื่อใช้เลือกคุณลักษณะนั้นจากคุณลักษณะสำคัญของเสียง ซึ่งหลังจากทดสอบแล้วสามารถนำมาใช้งานได้จริง และสามารถนำความรู้นี้ไปประยุกต์ใช้และพัฒนากับงานที่เกี่ยวข้องกับการค้นหาข้อมูลด้วยใช้เสียงได้ เช่น การค้นหาเพลง, การค้นหาข้อมูลจากคลังข้อมูลเสียง เป็นต้น

5.1 สรุปผลการศึกษา

วิธีการเลือกคุณลักษณะสำคัญเฉพาะในงานวิจัยนี้ ทางผู้วิจัยได้ศึกษาและพัฒนาระบบโครงข่ายทริปเล็ต เพื่อนำมาประยุกต์ใช้เพื่อค้นหาข้อมูลเพลงจากข้อมูลเสียงสั้นๆ (~ 3 วินาที) ซึ่งพบว่าสามารถนำมาใช้งานได้จริง แนวความคิดนี้ยังเป็นแนวทางการพัฒนาระบบที่ใช้ตรวจสอบจำนวนเพลงที่เปิดผ่านสถานีวิทยุ หรือใช้เพื่อตรวจสอบลิขสิทธิ์เพลง

การทดสอบกระบวนการเพื่อใช้เลือกคุณลักษณะสำคัญเพื่อใช้เป็นตัวแทนของเนื้อหาเพลงด้วยโมเดลโครงข่ายทริปเล็ต โดยนำข้อมูลเพลงจำนวน 200 เพลง มาแบ่งเป็นตัวอย่างข้อมูลจำนวน 163 ตัวอย่างในแต่ละเพลง จากนั้นนำมาผ่านกระบวนการประมวลผลสัญญาเสียงเบื้องต้น กระบวนการหาคุณลักษณะสำคัญด้วยวิธีการต่างๆ ซึ่งจะได้เป็นข้อมูลในรูปแบบเมตริกซ์ขนาด 64 x 64 ที่มีจำนวน {2, 3} ชั้น และแบ่งข้อมูลเป็นสองกลุ่มคือ กลุ่มข้อมูลสำหรับฝึกสอน 60 เปอร์เซ็นต์ และกลุ่มข้อมูลสำหรับตรวจสอบ 40 เปอร์เซ็นต์ โดยใช้โมเดลโครงข่ายทริปเล็ตที่มีโครงข่ายย่อยเป็นโครงข่ายประสาทเทียมแบบคอนโวลูชัน ซึ่งโครงข่ายย่อยแต่ละตัวจะให้เอาท์พุตเป็นคุณลักษณะเวกเตอร์ขนาด 128 x 1 จากการทดลองพบว่า ข้อมูลที่มีรูปแบบ 64 x 64 x 2 ที่ใช้วิธีการแยกคุณลักษณะสำคัญด้วยค่าสัมประสิทธิ์สเซปสตรัมบนสเกลเมล และค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลแบบความต่าง ให้เปอร์เซ็นต์ความคลาดเคลื่อนน้อยที่สุดที่ 6.7 เปอร์เซ็นต์

หลังจากฝึกสอนโมเดลโครงข่ายทริปเล็ตเรียบร้อยแล้ว จึงนำโมเดลที่ได้มาใช้เพื่อแยกคุณลักษณะสำคัญเฉพาะจากเสียง ซึ่งส่วนที่จะใช้เป็นเพียงโมเดลย่อยเท่านั้น จากนั้นทำการทดสอบด้วยเฟรมเวิร์คตามภาพที่ 3.8 พบว่าโมเดลสามารถทายข้อมูลเพลงได้ถูกต้องร้อยละ 79.43 ซึ่งหากตรวจสอบข้อมูลเพลงที่ใช้ในการทดสอบพบว่าแนวเพลงหรือเสียงดนตรีในบางเพลงมีความแตกต่างจากข้อมูลเพลงที่ใช้สำหรับฝึกสอนนั้นอาจเป็นหนึ่งสาเหตุที่ทำให้โมเดลมีความถูกต้องน้อยลง

5.2 ข้อเสนอแนะ

ในการวิจัยพบว่าผลของงานวิจัยที่ได้เป็นที่ยอมรับได้ในระดับหนึ่ง แต่ก็ยังไม่ครอบคลุมเพลงได้ในทุกๆ แนวเพลง อาจต้องปรับปรุงปัจจัยอื่นเพื่อให้ได้ผลลัพธ์ที่ดีขึ้นดังนี้

5.2.1 ชุดข้อมูลสำหรับฝึกสอน

จากผลการทดสอบและเปรียบเทียบผลการค้นหาเพลงพบว่าข้อมูลเพลงในบางเพลงยังไม่ครอบคลุม ดังนั้นอาจจะต้องเพิ่มจำนวนเพลงหรือเลือกประเภทเพลงหรือแนวเพลงเพื่อนำมาใช้ในการฝึกสอนข้อมูล เพื่อเพิ่มประสิทธิภาพของโมเดลให้สูงขึ้น

5.2.2 วิธีการแยกคุณลักษณะสำคัญ

ในงานวิจัยนี้ได้ทดลองวิธีการแยกคุณลักษณะสำคัญของเสียงหลายรูปแบบเพื่อสร้างเป็นชุดข้อมูล จากการทดลองพบว่าคุณลักษณะที่ให้ผลดีที่สุดคือการใช้ค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลและค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมลแบบความต่าง แต่เมื่อทดสอบค้นหาเพลงกลับพบว่ายังได้ผลลัพธ์ที่ไม่ดีนัก อาจจะเป็นเพราะวิธีการแยกคุณลักษณะยังไม่ครอบคลุมแนวเพลงหรือลักษณะของเสียงดนตรี ดังนั้นถ้าสามารถเพิ่มหรือเปลี่ยนวิธีการอื่นๆ ที่สามารถทำให้การ

แยกคุณลักษณะสำคัญที่เป็นเอกลักษณ์ของเสียงได้มากขึ้น ก็อาจจะทำให้ผลลัพธ์ที่ได้มีประสิทธิภาพเพิ่มขึ้น

5.2.3 การปรับปรุงโครงข่ายย่อยในโครงข่ายทริปเล็ต

ในงานวิจัยนี้สร้างโครงข่ายย่อยในโครงข่ายทริปเล็ตด้วยโครงข่ายประสาทเทียมแบบคอนโวลูชัน ซึ่งอาจจะให้ผลลัพธ์ที่ไม่ดีมากนักสำหรับข้อมูลที่ความต่อเนื่องอย่างข้อมูลที่เป็นเพลง เมื่อเทียบกับโครงข่ายประสาทเทียมแบบรีเคอร์เรนท์ (Recurrent Neural Network) ดังนั้นหากสามารถเปลี่ยนโครงข่ายย่อยเป็นโครงข่ายประสาทเทียมแบบรีเคอร์เรนท์ ก็อาจจะเพิ่มประสิทธิภาพในการเลือกคุณลักษณะสำคัญเฉพาะของเสียงได้

5.2.4 การใช้งานโครงข่ายทริปเล็ต

ในการใช้งานโครงข่ายทริปเล็ตเพื่อใช้แยกคุณลักษณะสำคัญเฉพาะของเสียงนั้น จำเป็นต้องใช้ข้อมูลที่มีความยาวของเสียงที่เท่ากันทั้งหมดเพื่อใช้เป็นอินพุตให้กับโครงข่ายย่อย ดังนั้นหากสามารถเปลี่ยนวิธีการเป็นวิธีการอื่น เช่น การใช้โครงข่ายย่อยที่มีโครงสร้างที่ต่างกัน ก็อาจจะเพิ่มประสิทธิภาพในการเลือกคุณลักษณะสำคัญเฉพาะของเสียงได้

5.2.5 ค่าพารามิเตอร์

ในงานวิจัยนี้มีค่าพารามิเตอร์หลายค่าที่ต้องมีการปรับเปลี่ยนให้เหมาะสม ซึ่งในหลายค่าพารามิเตอร์เป็นการอ้างอิงจากงานวิจัยที่เกี่ยวข้องก่อนหน้า และมีอีกหลายค่าพารามิเตอร์เป็นค่าเริ่มต้น (Default) ของฟังก์ชัน ซึ่งอาจจะไม่ใช่ค่าพารามิเตอร์ที่เหมาะสมที่สุดในงานวิจัยนี้ ดังนั้นหาสามารถปรับเปลี่ยนค่าพารามิเตอร์ดังกล่าวได้ ก็อาจจะเพิ่มประสิทธิภาพในงานนี้ได้

บรรณานุกรม

บรรณานุกรม

สุนันท์ ธาติ, พงศ์พันธ์ กิจสนาโยธิน และ วรลักษณ์ คงเด่นฟ้า (2018). “Song Clustering Using Similarity of Audio Fingerprint,” วารสารสถาบันเทคโนโลยีไทย-ญี่ปุ่น, ฉบับ 6, ครั้ง 1, หน้า 49-55.

อภิวัฒน์ จันโท และ อาทิตย์ ศรีแก้ว (2556). “ระบบตรวจจับการจราจรบนถนนเชิงเสียงด้วยวิธีทางปัญญาประดิษฐ์,” บัณฑิตวิทยาลัยมหาวิทยาลัยเทคโนโลยีสุรนารี.

J. Stephen Doownie (2003). “Music information retrieval”, Annual Review of Information Science and Technology, vol. 37, pp. 295-340.

Avery Li-Chun Wang (2003). “An industrial-strength audio search algorithm”, Proc. Int. Conf. Music Information Retrieval, pp. 713-718.

D. Ellis (2009). “Robust landmark-based audio fingerprinting,” 09 2009.

B. G Sherlock, DM Monro, and K Millard (1994). “Fingerprint enhancement by directional fourier filtering,” IEE ProceedingsVision, Image and Signal Processing, vol. 141, no. 2, pp. 87–94

P. Cano, M. Koppenberger, and N. Wack (2005). “Content-based music audio recommendation,” in Proceedings of the 13th annual ACM international conference on Multimedia. ACM, pp. 211–212.

G. Koch, R. Zemel and R. Salakhutdinov (2015). “Siameseneural networks for one-shot image recognition,” ICML Deep Learning Workshop, vol. 2

I. Ve´lez, C. Rascon and G. Fuentes-Pineda (2018). “One-Shot Speaker Identification for a Service Robot using a CNN-based Generic Verifier,” arXiv preprint arXiv:1809.04115.

Y. Zhang, B. Pardo and Z. Duan (2019). “Siamese Style Convolutional Neural Networks for Sound Search by Vocal Imitation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 2, pp. 429-441.

A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra amd T. Lillicrap (2016). “Meta-learning with memory-augmented neural networks,” International conference on machine learning 2016, pp. 1842-1850.

Y. Chen, M. W. Hoffman, S. Gomez Colmenarejo, M. Denil, T. P. Lillicrap, M. Botvinick and N. de Freitas (2017). “Learning to learn without gradient descent by gradient descent,” Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 748-756.

P. Manocha, R. Badlani, A. Kumar, A. Shah, B. Elizalde and B. Raj (2017). “Content-based Representations of audio using Siamese neural networks,” arXiv preprint arXiv:1710.10974.

J. Snell, K. Swersky and R. Zemel (2017). “Prototypical Networks for Few-shot Learning,” Advances in neural information processing systems, pp. 4077-4087.

F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H.S. Torr and T. M. Hospedales (2018). “Learning to compare: Relation network for few-shot learning,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 1199-1208.

A. Gupta, B. Eysenbach, C. Finn and S. Levine (2018). “Unsupervised Meta-Learning for Reinforcement Learning,” arXiv preprint arXiv:1806.04640.

S. Florian, D. Kalenichenko, and J. Philbin (2015). “Facenet: A unified embedding for face recognition and clustering,” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp. 815-823.

H. Song, M. Willi, Jayaraman J. Thiagarajan, V. Berisha1, A. Spanias1 (2018). “Triplet network with attention for speaker diarization,” arXiv preprint arXiv:1808.01535.

Kumar, B. G., G. Carneiro, and I. Reid (2016). “Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 5385-5394.

ภาคผนวก

ภาคผนวก ก

ผลงานตีพิมพ์

การแยกแยะเสียงโฆษณาบนวิทยุโดยการประยุกต์ใช้โครงข่ายสยาม
Commercial Spot Detection Using Siamese Networks

ศักรินทร์ นุ้ยพิน (Sakkarin Nupin) และดวงใจ จิตคงชื่น (Duangjai Jitkongchuen)
หลักสูตรวิศวกรรมข้อมูลขนาดใหญ่ วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
605162020005@dpu.ac.th, duangjai.jit@dpu.ac.th

บทคัดย่อ

เนื่องจากข้อจำกัดทางด้านกฎหมายที่กำหนดจำนวนครั้งในการเปิดโฆษณาบนวิทยุ และความถี่ในการเปลี่ยนเสียงโฆษณานั้นค่อนข้างต่ำ จึงทำให้จำนวนโฆษณาที่เกิดขึ้นนั้นมีค่อนข้างน้อย งานวิจัยนี้จึงนำเสนอการแยกแยะเสียงโฆษณาออกจากเสียงอื่นบนวิทยุ โดยนำทฤษฎีการเรียนรู้ด้วยตัวอย่างจำนวนน้อย (Few-shot learning) มาประยุกต์ใช้งานด้วยโมเดลโครงข่ายสยาม (Siamese network) ภายในประกอบด้วยโครงข่ายย่อยสองโครงข่ายที่เหมือนกันทั้งค่าน้ำหนักและโครงสร้าง ซึ่งทั้งสองโครงข่ายมีหน้าที่สร้างคุณลักษณะเวกเตอร์ (feature vectors) ขนาด 128 มิติ จากนั้นนำผลลัพธ์ที่ได้มาหาความคลายคลึงกัน (Similar) โดยอาศัยการคำนวณยุคลิด (Euclidean distance) จากการทดลองพบว่าวิธีการดังกล่าวสามารถแยกแยะเสียงโฆษณาได้อย่างมีประสิทธิภาพและครอบคลุมเสียงเพลงและเสียงพูดของนักจัดรายการวิทยุด้วย โดยให้ค่าความแม่นยำ (Accuracy) 0.907 และค่าประสิทธิภาพโดยรวม (F-measure) 0.907

คำสำคัญ: การทำเหมืองข้อมูล, การจำแนกเสียง, การเรียนรู้ด้วยตัวอย่างจำนวนน้อย, โครงข่ายสยาม

Abstract

There are a few commercial spots due to legal restrictions that limit the number of times to open it. In addition, the frequency of commercial spot changes is relatively low. This paper presents a commercial spot detection from another sound on the radio by applying few-shot learning with the Siamese network. The Siamese network consists of two symmetrical neural networks that are identical in both weight and structure. Each network generates a 128-dimensional feature vector then finds a similar result using Euclidean distance. The experiment results showed that the proposed algorithm could distinguish effectively commercial spots and covers both the song and the voice of the radio presenter (speech). The accuracy rate is 0.907 and f-measure is 0.907.

Keyword: Data mining, Audio classification, Few-shot learning, Siamese network.

1. บทนำ

ในการบันทึกและจัดเก็บเสียงการออกอากาศของสถานีวิทยุหากมีการติดป้ายกำกับเพื่อบอกว่าไฟล์เสียงนั้นเป็นเสียงแบบใดบ้าง อาทิ เสียงเพลง (Song), เสียงโฆษณา (Commercial spot) และเสียงจากนักจัดรายการวิทยุ (Speech) พร้อมระบุได้ว่าเสียงนั้นเกิดในช่วงเวลาใดของการบันทึก ก็จะทำให้ง่ายต่อการค้นหาและนำไปใช้งานในภายหลัง ซึ่งวิธีการทั่วไปที่ใช้แยกแยะเสียงนั้นจะการฟังของพนักงาน แต่เนื่องจากข้อจำกัดอาทิ ความยากในการแยกแยะเมื่อเสียงมีความยาวหลายชั่วโมง, ความสามารถของผู้ฟังที่มีผลต่อความถูกต้อง เป็นต้น ซึ่งการเรียนรู้เชิงลึกนั้นเป็นวิธีที่น่าสนใจเพื่อใช้แยกแยะเสียง

การเรียนรู้เชิงลึก (Deep learning) ได้ถูกพัฒนาเพื่อแก้ไข้อจำกัดของการเรียนรู้ด้วยวิธี Artificial neural network คือหากโครงข่ายเส้นใยประสาทเทียมมีความลึกหลายชั้น จะส่งผลให้ใช้เวลาในการประมวลผลนานและอาจพบคำตอบที่เป็นค่าที่ดีที่สุดเฉพาะที่ (Local minimum) ซึ่งไม่ใช่คำตอบที่ดีที่สุด โดยมีสาเหตุจากโครงข่ายในระดับลึกมีอัตราการเรียนรู้ต่ำ จึงส่งผล

ให้การปรับค่าถ่วงน้ำหนักในระดับลึกมีการเปลี่ยนแปลงทีละน้อยทำให้ต้องใช้เวลานานในการเรียนรู้

การเรียนรู้เชิงลึกถูกนำมาประยุกต์ใช้ในงานหลายด้าน อาทิ การแก้ปัญหาทางธุรกิจ การวิเคราะห์ภาพ การแยกสัญญาณเสียงผสม เป็นต้น แต่การเรียนรู้เชิงลึกต้องอาศัยข้อมูลจำนวนมากฝึกสอนโมเดล เพื่อให้ได้โมเดลที่ทำงานได้ใกล้เคียงกับมนุษย์ แต่ไม่ใช่ทุกปัญหาจะมีข้อมูลจำนวนมากพอที่จะทำให้ได้โมเดลที่แก้ปัญหานั้นได้อย่างถูกต้อง ดังนั้นในช่วงไม่กี่ปีที่ผ่านมาเริ่มมีงานวิจัยเพื่อแก้ไขข้อจำกัดดังกล่าว โดยอาศัยข้อมูลตัวอย่างเพียงไม่กี่ตัวอย่างในแต่ละคลาสเพื่อทำนายข้อมูลที่ไม่มีป้ายกำกับ เทคนิคนั้นคือ การเรียนรู้ด้วยตัวอย่างจำนวนน้อย (Few-shot learning) [1] [2]

งานวิจัยที่นำเทคนิคการเรียนรู้ด้วยตัวอย่างจำนวนน้อยไปประยุกต์ใช้ส่วนใหญ่จะเป็นงานด้านคอมพิวเตอร์วิทัศน์ (Computer vision) เช่น งานวิจัยเพื่อการจำแนกภาพตัวอักษรและตัวเลขที่เขียนด้วยลายมือ [3], งานวิจัยเพื่อยืนยันลายเซ็นต์ของบุคคล [4], งานวิจัยเพื่อการจำแนกโครโมโซม [5] เป็นต้น แต่ระยะหลังมีหลายงานวิจัยที่นำเทคนิคการเรียนรู้ด้วยตัวอย่างจำนวนน้อยไปใช้ในงานด้านเสียงเพื่อตอบว่าเสียงนั้นคือเสียงอะไร [6] หรือใช้เพื่อระบุเสียงพูดว่าเป็นเสียงใคร [7][8] ในงานวิจัยเหล่านั้นใช้โมเดลที่มีชื่อว่าโครงข่ายสยาม (Siamese network) เพื่อประยุกต์ใช้การเรียนรู้ด้วยตัวอย่างจำนวนน้อย

สืบเนื่องจากจำนวนของโฆษณาที่เกิดขึ้นบนวิทยุนั้นมีจำนวนน้อย เป็นเหตุมาจากข้อกำหนดทางกฎหมายที่กำหนดให้สถานีวิทยุสามารถเปิดโฆษณาได้ชั่วโมงละไม่เกิน 10 นาที และการเปลี่ยนตัวโฆษณานั้นไม่ได้ถูกเปลี่ยนกันบ่อย โฆษณาหนึ่งตัวอาจจะถูกใช้งานซ้ำไปมากกว่าหนึ่งเดือน งานวิจัยนี้จึงเสนอวิธีการแยกแยะเสียงโฆษณาออกจากเสียงอื่นบนวิทยุด้วยโมเดลโครงข่ายสยาม โมเดลนี้อาศัยการเปรียบเทียบความคล้ายคลึงกันของข้อมูล

2. งานวิจัยที่เกี่ยวข้อง

2.1 การเรียนรู้อภิมาน

การเรียนรู้อภิมาน (Meta learning) เป็นงานวิจัยในโดเมนหนึ่งของสาขาของปัญญาประดิษฐ์ (AI) ที่น่าสนใจในขณะนี้ จนมีหลายคนกล่าวว่าการเรียนรู้อภิมานจะเป็นกุญแจสำคัญที่ทำให้ปัญญาประดิษฐ์มีระดับปัญญาใกล้เคียงกับมนุษย์ได้ในอนาคต [9] ด้วยความสามารถในการเรียนรู้สองชั้น (การเรียนรู้เพื่อที่จะแสดงอะไรบางอย่างในแต่ละงานและในขณะเดียวกันนั้นก็พยายามที่จะเก็บเกี่ยวความรู้เพื่อให้สามารถบอกความเหมือนหรือความต่างในระหว่างงานนั้นได้)

จากการงานวิจัยของ Santoro [2] ใช้เทคนิคการเรียนรู้อภิมานมาฝึกฝนเพื่อเพิ่มหน่วยความจำโครงข่ายประสาทเทียมโดยการให้เรียนรู้วิธีการจัดเก็บและดึงความทรงจำที่ใช้สำหรับในแต่ละงานจำแนก และจากงานวิจัยของ Andrychowicz [10] ใช้ LSTM เพื่อฝึกสอนโครงข่ายประสาทเทียม ซึ่งพวกเขามีความสนใจในการปรับปรุงอัลกอริทึมในการฝึกสอนโครงข่ายประสาทเทียมเพื่องานจำแนกขนาดใหญ่โดยอาศัยเทคนิคการเรียนรู้ด้วยตัวอย่างจำนวนน้อย

2.2 การเรียนรู้ด้วยตัวอย่างจำนวนน้อย

การเรียนรู้จากข้อมูลจำนวนน้อยถูกเรียกว่า การเรียนรู้ด้วยตัวอย่างจำนวนน้อย (Few-shot learning) หรือ การเรียนรู้ด้วยตัวอย่างจำนวน k ตัว (k-shot learning) โดยที่ k จะหมายถึงจำนวนข้อมูลในแต่ละคลาสในชุดข้อมูล ตัวอย่างเช่น หากต้องการจำแนกภาพว่าเป็นแมวหรือสุนัข โดยที่มีภาพแมวและสุนัขแค่อย่างละภาพ วิธีการแบบนี้จะเรียกว่า การเรียนรู้ด้วยตัวอย่างหนึ่งตัวอย่าง (One-shot learning) แต่หากมีภาพแมวหรือสุนัขอย่างละ 10 ภาพ แบบนี้จะเรียกว่า การเรียนรู้ด้วยตัวอย่าง 10 ตัวอย่าง (10-shot learning) ดังนั้นก็จะเห็นแล้วว่าค่า k ในการเรียนรู้ด้วยตัวอย่างจำนวน k ตัว เป็นตัวเลขที่แสดงถึงจำนวนข้อมูลในแต่ละคลาส

จากงานวิจัยของ Koch [3] ใช้โมเดลโครงข่ายสยามเชิงลึก (Deep siamese networks) เพื่อฝึกสอนด้วยโครงข่ายแบบสังวัตนาเพื่อสร้างตัวอย่างแบบฝังตัว (Embed examples) พบวัตถุในคลาสเดียวกันจะอยู่ใกล้กัน แต่ในทางกลับกันวัตถุที่ต่างกันจะอยู่ไกลกัน และงานวิจัยของ Vinyals [11] เป็นโครงข่ายการจับคู่วัตถุซึ่งมาจากแนวคิดที่ว่าชุดฝึกสอนและชุดทดสอบที่เหมือนกันย่อมอยู่ใกล้กันและมีความคล้ายคลึงกันโดยพิจารณาจากการคำนวณโคไซน์ (Cosine similarities)

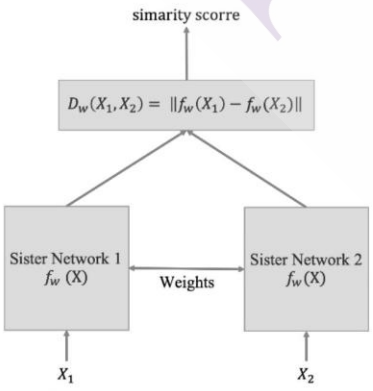
2.3 โครงข่ายสยาม

โครงข่ายสยาม (Siamese network) ประกอบด้วยโครงข่ายย่อยสองโครงข่าย ที่มีค่าน้ำหนักและโครงสร้างเหมือนกัน

งานวิจัยก่อนหน้าโครงข่ายสยามถูกนำมาใช้หาความคล้ายคลึงของวัตถุสองชนิดหรือมากกว่านั้น ในงานวิจัยของ Dey [4] ใช้โครงข่ายสยามเพื่อใช้ยืนยันลายเซ็นต์ของบุคคล งานวิจัยของ Muller [12] ก็ใช้โครงข่ายสยามเพื่อให้คะแนนการแปลโดยการพิจารณาจากคู่ของประโยคที่ใช้เป็นอินพุต ในงานวิจัยด้านเสียงประยุกต์ใช้โครงข่ายสยามเพื่อหาความคล้ายในเนื้อหาของเสียง เช่น งานวิจัยของ Raffel และ Ellis [13] ใช้เพื่อระบุเสียงพูดว่าเป็นเสียงของใคร [7][8]

เนื่องจากโครงข่ายย่อยในโครงข่ายสยามนั้นใช้ค่าน้ำหนักที่เหมือนกันจึงทำให้ต้องการปริมาณข้อมูลที่ใช้ฝึกสอนน้อยลง และมีแนวโน้มการเกิด overfitting ที่น้อย ในการฝึกสอนโครง-ข่ายสยามเป้าหมายคือการพยายามเรียนรู้พารามิเตอร์ W ของฟังก์ชัน f_w เพื่อให้รู้ว่าใครมีความคล้ายคลึงกับเราบ้าง เพื่อที่จะบรรลุเป้าหมาย จำเป็นจะต้องกำหนดฟังก์ชันวัตถุประสงค์ (Objective function) D_w ที่เหมาะสม โดยให้ $X_1, X_2 \in P$ แทนคู่ตัวอย่างและ Y แทนป้ายกำกับ (Label) ของคู่ตัวอย่างนี้ โดย Y=1 เมื่อ X_1 และ X_2 มีความคล้ายคลึงกันถ้าไม่ใช่ Y=0 ในการหาความคล้ายคลึงระหว่าง X_1 และ X_2 จะอาศัยการคำนวณยุคลิด (Euclidean distance) เพื่อหาระยะห่างของวัตถุทั้งสองที่ผ่านฟังก์ชัน f_w

$$D_w(X_1, X_2) = \|f_w(X_1) - f_w(X_2)\| \quad (1)$$



ภาพที่ 1: โครงสร้างพื้นฐานของโครงข่ายสยาม

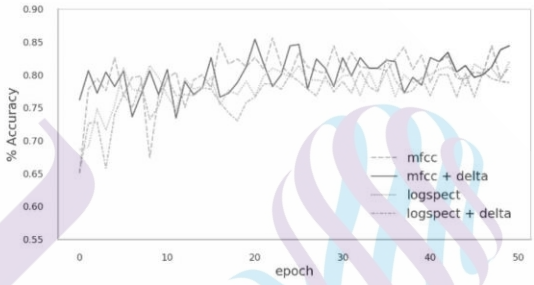
3. วิธีการดำเนินการวิจัย

3.1 ชุดข้อมูล

ชุดข้อมูลที่ใช้เป็นเสียงที่บันทึกจากสถานีวิทยุคลื่น Mono Fresh 91.5 จากนั้นทำการตัดเสียงด้วยโปรแกรม Audacity โดยเสียงที่จะใช้งานในงานวิจัยนี้มี 3 รูปแบบคือ เสียงเพลง, เสียงพูดของนักจัดรายการวิทยุและเสียงโฆษณา โดยตัดเป็นคลิปเสียงความยาว 5 วินาทีต่อไฟล์ โดยจะได้ไฟล์เสียงรูปแบบละ 637 ไฟล์ เหตุที่ใช้จำนวนไฟล์เสียงจำนวน 637 ไฟล์เพราะเมื่อรวมระยะเวลาของคลิปเสียงแล้วจะได้ 53 นาที เท่ากับเวลาของโฆษณาทั้งหมดที่พบในสถานีวิทยุ (กันยายน 2561 - มกราคม 2562) แปลงไฟล์เสียงให้อยู่ในรูปแบบ WAV กำหนดอัตราสุ่มสัญญาณ (Sampling rate) 22.05 กิโลเฮิรตซ์ที่ 16 บิตต่อตัวอย่าง (Sampling size) และมีคุณภาพเสียงการบันทึกเสียงเป็นแบบโมโน (1 Channel) จากนั้นแบ่งกลุ่มไฟล์เสียงที่ได้เป็น 3 ประเภทคือชุดฝึกสอน (Train), ชุดตรวจสอบ (Validation), ชุดทดสอบ (Test) แบ่งด้วยอัตรา 60%, 20% และ 20% ตามลำดับ โดยเลือกสุ่มไม่ให้เสียงโฆษณาที่เป็นโฆษณาตัวเดียวกันอยู่ในกลุ่มเดียวกัน

3.2 กระบวนการเตรียมข้อมูล

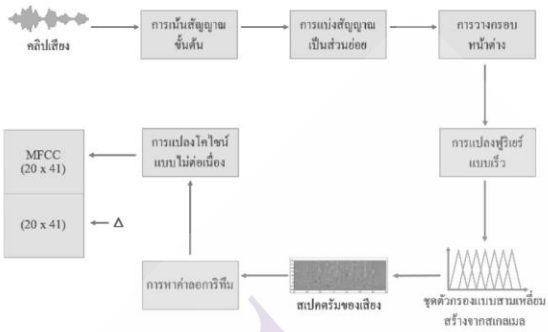
การเลือกวิธีการสกัดคุณลักษณะที่สำคัญของเสียงเพื่อใช้ในการแยกแยะเสียงโฆษณา ซึ่งหากเลือกวิธีการสกัดคุณลักษณะที่เหมาะสมก็จะทำให้ความแม่นยำของโมเดลเพิ่มขึ้น



ภาพที่ 2: ความแม่นยำของวิธีการสกัดคุณลักษณะที่สำคัญของเสียง

จากภาพที่ 2 เป็นการเปรียบเทียบความแม่นยำของวิธีการสกัดคุณลักษณะที่สำคัญของเสียงประกอบด้วย ค่าสัมประสิทธิ์เซปสตรัมบนแกนความถี่เมล (MFCC), ค่าสัมประสิทธิ์เซป-สตรัมบนแกนความถี่เมล + อนุพันธ์อันดับหนึ่งของค่าสัมประ-สิทธิ์เซปสตรัมบนแกนความถี่เมล (MFCC + delta), ค่าลอการิ-ทึมของสเปคตรัมบนสเกลเมล (Log scaled mel-spectrogram) และค่าลอการิทึมของสเปคตรัมบนสเกลเมล + อนุพันธ์อันดับหนึ่งของค่าลอการิทึมของสเปคตรัมบนสเกลเมล (Log-scaled mel-spectrogram + delta) จะเห็นว่าค่าสัมประสิทธิ์เซปสตรัมบนแกนความถี่เมล + อนุพันธ์อันดับหนึ่งของค่าสัมประสิทธิ์เซปสตรัมบนแกนความถี่เมลมีความแม่นยำมากกว่าการสกัดคุณลักษณะที่สำคัญของเสียงวิธีการอื่น ดังนั้นในงานวิจัยนี้จึง

เลือกใช้วิธีการสกัดคุณลักษณะที่สำคัญของเสียงเป็นค่าสัมประ-สิทธิ์เซปสตรัมบนแกนความถี่เมล (MFCC) + อนุพันธ์อันดับหนึ่งของค่าสัมประ-สิทธิ์เซปสตรัมบนแกนความถี่เมล โดยวิธีการสกัดคุณลักษณะสำคัญของเสียงได้แสดงดังในภาพที่ 3

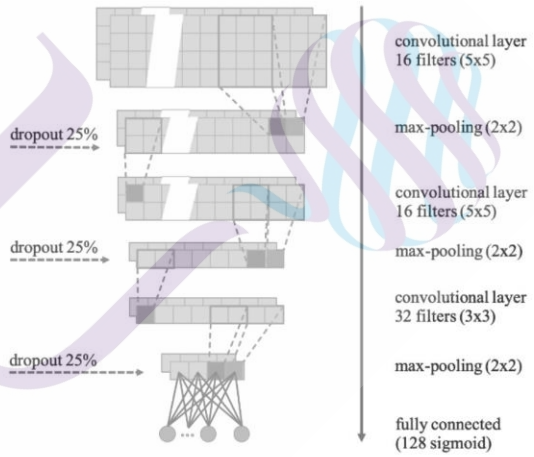


ภาพที่ 3: วิธีการสกัดคุณลักษณะสำคัญของเสียง

เพื่อให้ได้คุณลักษณะที่สำคัญของเสียง โดยเสียงจะถูกนำมาผ่านการประมวลผลสัญญาณเบื้องต้นโดยการเน้นสัญญาณขั้น-ต้น (Pre-emphasis) ด้วยการนำสัญญาณเสียงผ่านวงจรกรองอันดับที่ 1 เพื่อให้สัญญาณเสียงชัดเจนขึ้นและสัญญาณรบกวนน้อยลง จากนั้นแบ่งสัญญาณเป็นส่วนย่อย (Block to Frame) ขนาด 41 เฟรมหรือ 930 วินาที โดยแต่ละส่วนมีส่วนซ้อนทับกัน 50% หลังจากแบ่งสัญญาณเป็นส่วนย่อยแล้ว ในแต่ละส่วนย่อยจะวางกรอบแฮมมิ่ง (Hamming Window) เพื่อทำให้สัญญาณต่อเนื่องและชัดเจนมากขึ้น นำสัญญาณที่ได้มาผ่านการแปลงฟูริเยร์แบบเร็ว (FFT) เพื่อให้ได้ข้อมูลของเสียงในโดเมนความถี่ และนำข้อมูลความถี่ของเสียงมาผ่านการกรองด้วยชุดตัวกรองแบบสามเหลี่ยมที่สร้างจากสเกลเมล (Mel scale) ซึ่งการใช้ตัวกรองแบบสามเหลี่ยมที่สร้างจากสเกลเมลเป็นเทคนิคที่ใช้ปรับสเกลของสเปคตรัมให้อยู่บนสเกลที่เหมาะสมสำหรับการฟังของมนุษย์ โดยสัญญาณเสียงในช่วงความถี่ต่ำจะมีความสำคัญมากกว่าช่วงความถี่สูงจึงมีการออกแบบสเกลของสเปคตรัมให้สามารถเก็บรายละเอียดของสัญญาณเสียงช่วงความถี่ต่ำได้มากกว่า จากนั้นนำค่าที่คำนวณได้จากแต่ละช่วงของความถี่ของตัวกรองแบบสามเหลี่ยมมาหาค่าลอการิทึม (Logarithm) และนำเข้าสู่การแปลงโคไซน์แบบไม่ต่อเนื่อง (DCT) ซึ่งจะได้ค่าสัมประสิทธิ์เซปสตรัมบนแกนความถี่เมล อย่างไรก็ตามค่าสัมประสิทธิ์เซปสตรัมบนแกนความถี่เมลยังขาดลักษณะที่สำคัญคือการเปลี่ยนแปลงของค่าสัมประสิทธิ์เซปสตรัมบนแกนความถี่เมลกับเวลา ซึ่งจะได้จากการหาอนุพันธ์อันดับที่หนึ่ง (Delta) ซึ่งสุดท้ายแล้วจะได้คุณลักษณะสำคัญของเสียงในแต่ละส่วน (แต่ละเวลา) 20 แถว/ช่วงคลื่น x 41 คอลัมน์/เฟรม เพื่อใช้ในการแยกแยะเสียงต่อไป

3.3 สถาปัตยกรรมของโครงข่าย

สถาปัตยกรรมของโครงข่ายย่อยในโครงข่ายสยายที่นำเสนอในงานวิจัยนี้ประยุกต์ใช้โครงข่ายประสาทเทียมแบบสังวัตนาการดังภาพที่ 4 โดยจะประกอบไปด้วยชั้น Convolutional ทั้งหมดจำนวน 3 ชั้นการประมวลผล โดยแต่ละ Convolutional จะคำนวณค่าบนข้อมูลเสียงที่ได้ในแต่ละตัวกรอง (filter) กับค่าน้ำหนัก โดยผ่าน Activation function เพื่อส่งต่อไปยังกระบวนการทำ Drop out เพื่อตัดค่าบางส่วนของคุณลักษณะออก โดยจะอยู่ต่อจากชั้น Convolutional ในแต่ละชั้น จากนั้นจึงนำส่งไปยังกระบวนการ Pooling โดยในสถาปัตยกรรมนี้นั้นจะมีชั้น Max pooling ทั้งหมด 3 ชั้นการคำนวณ จะอยู่ต่อจากชั้น Convolutional ในแต่ละชั้น ซึ่งเป็นกระบวนการในการลดรายละเอียดของคุณลักษณะลง ดังนั้นจะทำการคัดเลือกค่ามากที่สุดในแต่ละตัวกรอง ส่วนในชั้นสุดท้ายใช้เป็น Fully Connected และชั้น Sigmoid ซึ่งเป็นการสร้างตัวแทนเพื่ออธิบายถึงข้อมูลเสียงนั้น



ภาพที่ 4: สถาปัตยกรรมของโครงข่ายย่อยในโครงข่ายสยาม

3.4 ขั้นตอนการวิจัย

ในการทดสอบแยกแยะเสียงด้วยโมเดลโครงข่ายสยาม ในแต่ละชุดฝึกสอน (Batches) จะใช้เสียงที่มีความเหมือนกันจำนวน 32 คู่เสียงและเสียงที่ต่างกันจำนวน 32 คู่เสียง โดยฝึกสอนจำนวน 50 รอบ (Epochs) ในแต่ละรอบใช้ชุดฝึกสอน (Number batches per epoch) จำนวน 1,000 ชุด ซึ่งในแต่ละชุด

เสียงกำหนดป้ายกำกับ (Label) เป็น 1 สำหรับคู่เสียงที่เหมือนกันและเป็น 0 สำหรับคู่เสียงที่ต่างกัน เช่น เสียงเพลงกับเสียงโฆษณา, เสียงนักจัดรายการวิทยุกับเสียงโฆษณา เป็นต้น ในการทดสอบเลือกใช้ออฟติไมเซอร์อัลกอริทึมเป็น Adam ด้วยอัตราการเรียนรู้ 0.001 หากค่าความแม่นยำไม่เปลี่ยนแปลงค่าจากก่อนหน้าในทุก 10 รอบ อัตราการเรียนรู้จะลดลง 10 เท่า

และใช้เทคนิค Grid search เพื่อค้นหาพารามิเตอร์ของขนาดคุณลักษณะเวกเตอร์ (32, 64, 128, 256, 512) และ Drop out (0.0, 0.25, 0.5) ซึ่งผลลัพธ์ที่ดีที่สุดของขนาดคุณลักษณะเวกเตอร์เป็น 128 และมี Drop out เป็น 0.25

3.5 การวัดค่าประสิทธิภาพ

ในงานวิจัยนี้ผู้วิจัยทำการวัดประสิทธิภาพของโมเดลแยกแยะเสียงโดยใช้เกณฑ์การวัดประสิทธิภาพของตัวโมเดลด้วย 4 ตัวชี้วัดประกอบด้วยค่าความแม่นตรง (Precision) ค่าความระลึก (Recall) ค่าประสิทธิภาพโดยรวม (F-measure) และค่าความแม่นยำ (Accuracy) ดังสมการ (2) (3) (4) และ (5) ตามลำดับ

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

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

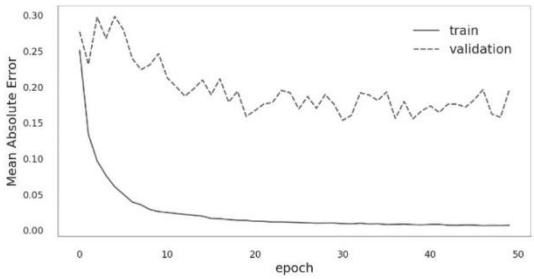
$$\text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

กำหนดให้ค่า True Positive (TP) เป็นค่าที่ทำนายถูกว่าเหมือน, True Negative (TN) เป็นค่าที่ทำนายถูกว่าต่างกัน, False Positive (FP) เป็นค่าที่ทำนายผิดว่าเหมือนและ False Negative (FN) เป็นค่าที่ทำนายผิดว่าต่างกัน

4. ผลการวิจัย

ผลทดสอบการแยกแยะรูปแบบเสียงด้วยโครงข่ายสยามฝึกสอนจำนวน 50 รอบ ในแต่ละหาค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ (Mean absolute error) เห็นได้ว่าเมื่อมีจำนวนรอบการฝึกสอนเพิ่มขึ้นค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ก็จะลดลง

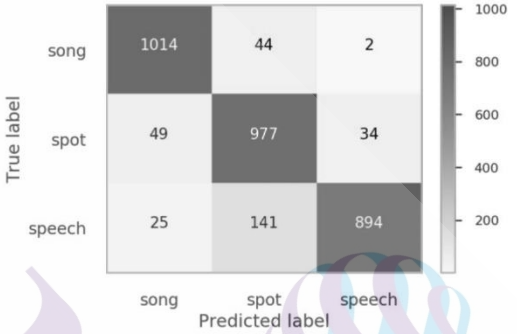


ภาพที่ 5: เปรียบเทียบค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ระหว่างชุดฝึกสอนและชุดตรวจสอบ

หลังจากฝึกสอนเรียบร้อยแล้วนำโมเดลที่ได้ไปทดสอบกับข้อมูลในชุดทดสอบ เพื่อวัดประสิทธิภาพของโมเดล ซึ่งจะได้ผลลัพธ์ดังแสดงไว้ในตารางที่ 1

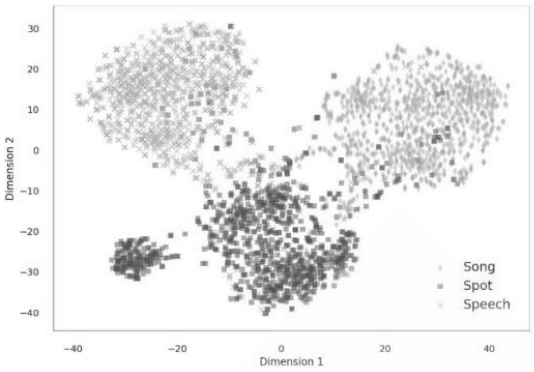
ตารางที่ 1: แสดงผลการวัดค่าประสิทธิภาพของโมเดล

	Accuracy	Precision	Recall	F-Measure
test	0.9072	0.9114	0.9072	0.9073



ภาพที่ 6: Confusion Matrix แสดงผลการแยกแยะรูปแบบของเสียง

ในภาพที่ 7 เป็นการแสดงความสัมพันธ์ของคุณลักษณะเวกเตอร์ที่เกิดจากการเรียนรู้ของโครงข่ายย่อยในโมเดลโครงข่ายสยาม ในข้อมูลชุดทดสอบจำนวน 3,180 ตัวอย่าง โดยใช้เทคนิค t-SNE แสดงคุณลักษณะเวกเตอร์ขนาด 128 มิติ ให้แสดงเป็นแผนภาพ 2 มิติ กำหนดให้สัญลักษณ์ข้าวหลามตัดแทนตัวอย่างข้อมูลเพลง, สัญลักษณ์สี่เหลี่ยมแทนตัวอย่างข้อมูลโฆษณา และสัญลักษณ์รูปตัวเอ็กซ์แทนตัวอย่างข้อมูลเสียงพูดของนักจัดรายการวิทยุ



ภาพที่ 7: t-SNE แสดงความสัมพันธ์ที่เกิดขึ้นในกลุ่มของเพลง, โฆษณา และเสียงพูดของนักจัดรายการวิทยุ

5. อภิปรายผล

จากการทดสอบพบว่าในในขั้นตอนการฝึกสอนโมเดลนั้น ชุดข้อมูลฝึกสอน (Batches) จำนวน 64 คู่เสียงที่เกิดจากการสุ่มประเภทและจับคู่เสียงซึ่งจะเป็นคู่เสียงที่เหมือนกันจำนวน 32 คู่เสียงและคู่เสียงที่ต่างกัน 32 คู่เสียง ซึ่งมีโอกาสที่จะสุ่มได้ประเภทของเสียงและคู่เสียงที่ซ้ำกัน จึงทำให้การเรียนรู้ของโมเดลทำได้ช้า

6. บทสรุป และข้อเสนอแนะ

ข้อสรุปที่ได้จากการวิจัยบ่งชี้ว่าการแยกแยะเสียงโฆษณาโดยใช้โมเดลโครงข่ายสยามสามารถแยกแยะเสียงโฆษณาออกจากเสียงอื่นได้อย่างมีประสิทธิภาพและยังครอบคลุมถึงการแยกแยะเสียงเพลงและเสียงพูดของนักจัดรายการวิทยุ โดยให้ค่าความแม่นยำ 0.907 และค่าประสิทธิภาพโดยรวม 0.907 แต่หากเสียงโฆษณานั้นมีรูปแบบที่เป็นการสนทนาแบบกลุ่มหรือเสียงดนตรีเพียงอย่างเดียวจะส่งผลให้ค่าความแม่นยำน้อยลงจะเห็นได้จากในภาพที่ 7 ที่ยังมีการทับซ้อนกันของเสียง ซึ่งข้อจำกัดดังกล่าวผู้วิจัยจะนำมาศึกษาและปรับปรุงต่อไป

เอกสารอ้างอิง

[1] L. Fei-Fei, R. Fergus and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594-611, 2006.

[2] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra and T. Lillicrap, "Meta-Learning with Memory-Augmented Neural Networks," *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, pp. 1842-1850, 2016.

[3] G. Koch, R. Zemel and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," *ICML Deep Learning Workshop*, vol. 2, 2015.

[4] S. Dey, A. Dutta, J. I. Toledo, S. K. Ghosh, J. Lladós and U. Pal, "SigNet: Convolutional Siamese Network for Writer Independent Offline Signature Verification," *arXiv preprint arXiv:1707.02131*, 2017.

[5] S. Jindal, G. Gupta, M. Yadav, M. Sharma and L. Vig, "Siamese Networks for Chromosome Classification," *Proceedings of ICCV Workshops*, pp. 72-81, 2017.

[6] P. Manocha, R. Badlani, A. Kumar, A. Shah, B. Elizalde and B. Raj, "Content-based Representations of audio using Siamese neural networks," *arXiv preprint arXiv:1710.10974*, 2017.

[7] I. Ve´lez, C. Rascon and G. Fuentes-Pineda, "One-Shot Speaker Identification for a Service Robot using a CNN-based Generic Verifier," *arXiv preprint arXiv:1809.04115*, 2018.

[8] Y. Zhang, B. Pardo and Z. Duan, "Siamese Style Convolutional Neural Networks for Sound Search by Vocal Imitation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 429-441, 2019.

[9] B. M. Lake, R. Salakhutdinov and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science Magazine, vol. 350, no. 6266*, pp. 1332-1338, 2015.

[10] M. Andrychowicz, M. Denil, S. Gomez, M. W Hoffman, D. Pfau, T. Schaul, B. Shillingford and N. De Freitas, "Learning to learn by gradient descent by gradient descent," *Advances in Neural Information Processing Systems*, pp. 3981-3989, 2016.

[11] O. Vinyals, C. Blundell, T. Lillicrap and D. Wierstra, "Matching networks for one shot learning," *Advances in neural information processing systems*, pp. 3630-3638, 2016.

[12] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," *Thirtieth AAAI Conference on Artificial Intelligence*, pp. 2786-2792, 2016.

[13] C. Raffel and D. PW Ellis, "Large-Scale Content-Based Matching of MIDI and Audio Files," *ISMIR*, pp. 234-240, 2015.

ประวัติผู้เขียน

ชื่อ-นามสกุล	นาย ศักรินทร์ นุ้ยพิน
ประวัติการศึกษา	วิศวกรรมศาสตรบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยธุรกิจบัณฑิตย์ ปีการศึกษา 2553
ตำแหน่งและสถานที่ทำงานปัจจุบัน	Programmer Supervisor , บริษัท โมโน เจนเนอเรชั่น จำกัด

ผลงานทางวิชาการ

- **ศักรินทร์ นุ้ยพิน** และ ดวงใจ จิตคงชื่น (2019). การแยกแยะเสียงโฆษณาบนวิทยุโดยการประยุกต์ใช้โครงข่ายสยาม. การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 15 (NCCIT 2019), หน้า 309-314