TNI Journal of Engineering and Technology
Vol.6 No.1 January - June 2018

การจัดกลุ่มเพลงโดยอาศัยความคล้ายคลึงของลายนิ้วมือทางเสียง

Song Clustering Using Similarity of Audio Fingerprint

สุนันท์ ธาติ[1] พงศ์พันธ์ กิจสนาโยธิน[2] วรลักษณ์ คงเด่นฟ้า[3]

ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยนเรศวร พิษณุโลก
[1]sununt56@email.nu.ac.th
[2]kphongph@nu.ac.th
[3]woralakk@gmail.com

บทคัดย่อ

วิธีการทั่วไปในการตรวจสอบการละเมิดลิขสิทธิ์ หรือการระบุข้อมูลเพลง คือการฟังโดยคน แต่การฟังมีข้อจำกัดในกรณีที่ต้องวิเคราะห์ข้อมูลเพลงจำนวนมากๆ อีกทั้งความถูกต้องแม่นยำยังขึ้นอยู่กับความเชี่ยวชาญของผู้ฟังแต่ละคน ด้วยข้อจำกัดดังกล่าว จึงมีการประยุกต์ใช้หลักการการรู้จำดนตรี (Music recognition) ในการแก้ไขปัญหานี้แทนและคุณลักษณะของข้อมูลที่ถูกนำมาใช้ในการวิเคราะห์อย่างกว้างขวาง คือลายนิ้วมือทางเสียง (Audio fingerprint) การวิเคราะห์ลายนิ้วมือทางเสียงมีประสิทธิภาพมากในการตรวจสอบว่าข้อมูลเสียงใดเป็นเพลงเดียวกัน ในกรณีที่มีเนื้อหาตรงกัน (Exactly match) แต่ยังไม่สามารถตรวจสอบได้ในกรณีที่เนื้อหาคล้ายคลึงกัน (Similar) งานวิจัยฉบับนี้นำเสนอวิธีการหาความคล้ายคลึงกันของเพลงโดยใช้ฟังก์ชันความสัมพันธ์ (Relation function) สำหรับการเปรียบเทียบลายนิ้วมือทางเสียงแทนการเปรียบเทียบข้อมูลเสียงขนาดใหญ่ทีละคู่ และทดลองใช้วิธีการที่นำเสนอเพื่อระบุเพลงต้นฉบับจากเพลงคัฟเวอร์ ผลการทดลองพบว่าวิธีการดังกล่าวสามารถระบุเพลงต้นฉบับได้อย่างมีประสิทธิภาพครอบคลุมกลุ่มเพลงทุกประเภท (Genre) โดยมีพื้นที่ใต้กราฟเฉลี่ย (Average AUC) เป็น 0.790

คำสำคัญ: การทำเหมืองข้อมูล, การจัดกลุ่มข้อมูล, การรู้จำข้อมูลดนตรี, สเปคโตแกรม, ลายนิ้วมือทางเสียง

ABSTRACT

Listening is the most common way to detect copyright infringement or identify unknown music data, but it is difficult to analyze a large amount of music data. The accuracy also depends on the listener's level of expertise. As mentioned earlier, music recognition is applied to solve this problem and the audio fingerprint is a widely used as data feature. Audio fingerprint analysis is effective at finding audio tracks which are duplicate content (exactly match) however it cannot detect in the case of similar content. This research proposes a method for finding the similarity between two songs using relation functions for comparing audio fingerprints instead of comparing bigger music content. For a case study, we try to find the original song from the cover song to assess the efficiency of our approach. The findings of this study indicate that proposed approach can be use effectively to identify the original song covered with many genres. Overall average area under curve (Average AUC) is 0.790.

Keywords: Data mining, Data clustering, Music recognition, Spectrogram, Audio fingerprint

1) บทนำ

ในยุคดิจิทัลการละเมิดลิขสิทธิ์ทำได้ง่ายและก่อให้เกิดความเสียหายแก่ผู้เป็นเจ้าของผลงานเป็นมูลค่ามหาศาล ซึ่งอุตสาหกรรมที่เกี่ยวข้องจำนวนมากได้รับผลกระทบนี้ อาทิ สิ่งประดิษฐ์ สื่อสิ่งพิมพ์ ภาพยนตร์ ดนตรี เป็นต้น สำหรับอุตสาหกรรมดนตรีนั้น การเผยแพร่ผลงานเพลงโดยผู้ที่ไม่ใช่เจ้าของผลงานถือเป็นปัญหาที่ส่งผลกระทบในวงกว้างอย่างมาก การตรวจสอบการละเมิดลิขสิทธิ์ทำได้หลายวิธี ซึ่งวิธีการทั่วไปใช้การสุ่มตรวจโดยการฟังของเจ้าหน้าที่ลิขสิทธิ์ แต่เนื่องด้วยข้อจำกัดหลายประการ อาทิ ความยากในการวิเคราะห์ข้อมูลจำนวนมากๆ โดยคนฟัง ประสบการณ์ของผู้ฟังที่ส่งผลต่อความถูกต้องของการวิเคราะห์ข้อมูลโดยตรง เป็นต้น ดังนั้นจึงมีการนำการสืบค้นข้อมูลดนตรี (Music information retrieval) [1] มาช่วยในการตรวจสอบการละเมิดลิขสิทธิ์แทนการฟัง

การสืบค้นข้อมูลดนตรีเป็นหนึ่งในกลุ่มการวิจัยที่บูรณาการความรู้จากหลายสาขาวิชา โดยอาศัยความรู้ในด้านต่างๆ อาทิ การประมวลผลสัญญาณดิจิทัล (Digital signal processing) การจดจำรูปแบบ (Pattern recognition) ทฤษฎีดนตรี (Music theory) เป็นต้น เทคนิคการรู้จำดนตรี (Music recognition technique) [2] ถูกนำมาประยุกต์ใช้เพื่อการดึงคุณลักษณะเฉพาะที่เป็นตัวแทนของเนื้อหาเสียงและการเปรียบเทียบเพื่อการตรวจสอบการทำซ้ำ คุณลักษณะเฉพาะที่ถูกเลือกเป็นตัวแทนของเนื้อหาเสียงมีหลากหลายรูปแบบทั้งนี้ขึ้นอยู่กับวัตถุประสงค์ของระบบ ซึ่งหนึ่งในคุณลักษณะเฉพาะที่ถูกนำมาใช้อย่างกว้างขวางในการตรวจสอบการละเมิดลิขสิทธิ์ คือ ลายนิ้วมือทางเสียง (Audio fingerprint) [3]

ลายนิ้วมือทางเสียงคือคุณลักษณะของข้อมูล (Data feature) ที่สามารถระบุถึงเนื้อหาของเสียงเช่นเดียวกับลายนิ้วมือที่แตกต่างกันของแต่ละบุคคลที่สามารถช่วยในการระบุตัวตนของมนุษย์ได้ สัญญาณเสียงที่เป็นข้อมูลเข้าของระบบจะถูกนำมาสร้างเป็นลายนิ้วมือทางเสียงเพื่อนำไปเปรียบเทียบกับลายนิ้วมือทางเสียงที่มีอยู่ในฐานข้อมูลอ้างอิง หากพบว่าเนื้อหาตรงกันถือว่าตรวจพบการละเมิดลิขสิทธิ์หรือสามารถระบุได้ว่าเป็นเพลงเดียวกันที่ถูกทำซ้ำนั้นเอง การประยุกต์ใช้ลายนิ้วมือทางเสียงที่มีอยู่ในปัจจุบันสามารถตรวจสอบข้อมูลเสียงว่าเป็นเพลง

เดียวกันได้ในกรณีที่เนื้อหาตรงกัน (Exactly match) แต่ยังไม่สามารถตรวจสอบได้ในกรณีที่เนื้อหาทางเสียงคล้ายคลึงกัน (Similar) ทั้งที่การละเมิดลิขสิทธิ์ไม่ใช่เพียงการทำซ้ำโดยอ้างอิงตรงตามเนื้อหาเดิมเท่านั้น แต่การดัดแปลงรูปแบบของเพลงและนำมาเผยแพร่เพื่อแสวงหาผลกำไรโดยไม่จ่ายค่าลิขสิทธิ์แก่ผู้เป็นเจ้าของผลงานก็ถือเป็นการละเมิดลิขสิทธิ์รูปแบบหนึ่งเช่นกัน เนื่องด้วยกลุ่มเพลงดังกล่าวมีเนื้อหาคล้ายกันเท่านั้นจึงทำให้ไม่สามารถตรวจสอบได้เมื่อใช้วิธีการค้นหาเนื้อหาที่ซ้ำกัน เพลงคัฟเวอร์ (Cover song) ซึ่งเป็นเพลงที่ถูกนำมาร้องใหม่และบันทึกเสียงอีกครั้งโดยผู้อื่นที่ไม่ใช่ศิลปินหรือนักแต่งที่เป็นเจ้าของถือเป็นตัวอย่างหนึ่งของการดัดแปลงรูปแบบของเพลงที่ไม่สามารถระบุข้อมูลเพลงที่เป็นต้นฉบับได้

2) วัตถุประสงค์ของการวิจัย

งานวิจัยนี้เสนอวิธีการในการหาความคล้ายกันของเพลงโดยหาระยะห่างระหว่างเพลงสองเพลงโดยการเปรียบเทียบลายนิ้วมือทางเสียง และนำวิธีการที่นำเสนอมาทดลองเพื่อค้นหาเพลงต้นฉบับจากเพลงคัฟเวอร์ โดยมีสมมติฐานตั้งต้นว่าระยะห่างระหว่างเพลงต้นฉบับและเพลงคัฟเวอร์ของเพลงเดียวกันจะมีค่าน้อยกว่าระยะห่างระหว่างเพลงคัฟเวอร์นั้นและเพลงต้นฉบับอื่นที่ไม่ใช่เพลงๆ นั้น นอกจากนี้ ยังออกแบบการทดลองโดยคำนึงถึงปัจจัยที่เกี่ยวข้องต่างๆ อาทิ ช่วงความถี่ เวลาในการประมวลผล และความถูกต้องแม่นยำในการระบุเพลงต้นฉบับ เป็นต้น

งานวิจัยนี้อธิบายตามลำดับดังต่อไปนี้ ส่วนที่ 3 อธิบายวิธีดำเนินการวิจัย ส่วนที่ 4 แสดงผลการวิจัย ส่วนที่ 5 สรุปและอภิปรายผล และท้ายสุด ส่วนที่ 6 เป็นบทสรุปและข้อเสนอแนะทั้งหมดของงานวิจัย ฉบับนี้

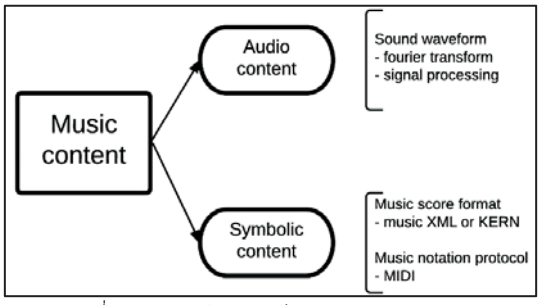
3) วิธีดำเนินการวิจัย

3.1) การทบทวนวรรณกรรม (Literature review)

การฟังโดยมนุษย์เป็นวิธีการดั้งเดิมในการวิเคราะห์ข้อมูลดนตรี ซึ่งวิธีการดังกล่าวไม่สามารถทำได้ในกรณีที่มีข้อมูลจำนวนมาก อีกทั้งประสบการณ์ของผู้ฟังส่งผลต่อความถูกต้องของการวิเคราะห์ข้อมูลโดยตรง เนื่องด้วยข้อจำกัดที่กล่าวมาแล้ว วิธีการค้นหาข้อมูลเพลงด้วยเทคนิคทางคอมพิวเตอร์จึงถูกนำมาประยุกต์ใช้ในการวิเคราะห์ข้อมูลดนตรีแทนการฟัง เทคนิคทางการรู้จำดนตรีสามารถช่วยจัดกลุ่ม แยกกลุ่ม และระบุลักษณะเฉพาะของงานดนตรีได้ โดยที่กระบวนการรู้จำดนตรีจะดึงข้อมูลจากเพลงเพื่อเป็นข้อมูลเข้าและหารูปแบบ (Pattern) ที่เหมือนหรือคล้ายกันของกลุ่มข้อมูลนั้นๆ เพื่อนำมาใช้ในการวิเคราะห์ ตัวอย่างของการนำความรู้เรื่องการรู้จำดนตรีมาประยุกต์ใช้ เช่น การตรวจหาการละเมิดลิขสิทธิ์ การระบุข้อมูลเพลงในกรณีที่ไม่ทราบแหล่งที่มา การระบุประเภทเครื่องดนตรีจากเสียง และการแยกประเภท เป็นต้น

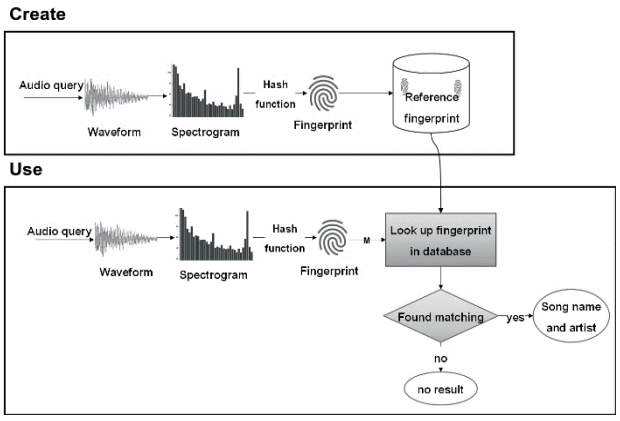
งานวิจัยก่อนหน้าที่ได้ศึกษาผลการวัดประสิทธิภาพวิธีการฟังโดยคนเทียบกับการประยุกต์ใช้เทคนิคการรู้จำดนตรี จากการทดลองของ Perrot และ Gjerdingen [4] พบว่ากรณีการจำแนกประเภทดนตรีด้วยวิธีการฟังโดยคน หลังจากฟังสามวินาที ผู้ที่ไม่เชี่ยวชาญด้านดนตรีสามารถระบุประเภทดนตรีได้ถูกต้อง 72% และการฟังมากกว่า สามวินาทีไม่ส่งผลให้ความถูกต้องเพิ่มขึ้น ข้อจำกัดในการฟังโดยคนประการแรกในกรณีที่ต้องแยกประเภทเพลงจำนวนมาก เวลาในการฟังจะเพิ่มขึ้นตามไปด้วย ประการที่สองความถูกต้องในการแยกประเภทขึ้นอยู่กับความเชี่ยวชาญและประสบการณ์ในการฟังเป็นหลัก ในขณะที่ Tzanetakis และ Cook [5] ได้ทดลองจำแนกประเภทดนตรีสิบประเภทโดยประยุกต์ใช้เทคนิคการรู้จำดนตรีเพื่อแยกประเภทเพลงจำนวนมาก ความถูกต้องที่ได้เป็น 61% และบ่งชี้ว่าปัจจัยสำคัญที่ส่งผลต่อความถูกต้องมากที่สุดคือการเลือกข้อมูลที่เหมาะสมเพื่อเป็นข้อมูลเข้าของระบบ

เนื่องด้วยข้อมูลดนตรีถูกแสดงในรูปแบบที่แตกต่างกัน ดังนั้นการเลือกข้อมูล (Data feature extraction) จึงขึ้นอยู่กับวัตถุประสงค์ของการนำข้อมูลนั้นไปใช้เป็นหลัก โดยการแสดงข้อมูลดนตรีมีสองประเภท ประการแรกการแสดงข้อมูลดนตรีด้วยสัญลักษณ์ (Symbolic representation) [1,6] ใช้สำหรับการอธิบายเนื้อหาทางดนตรีเพื่อให้นักดนตรีสามารถสื่อสารกันเพื่อบรรเลงเพลงได้ ตัวอย่างเช่น โน้ตสกอร์เพลง และการเข้ารหัสคอมพิวเตอร์ (Music notation protocol) เป็นต้น ประการที่สองการแสดงข้อมูลดนตรีด้วยเสียง (Audio representation) [1] ใช้สำหรับการบันทึกเสียง ซึ่งก็คือข้อมูลสัญญาณเสียงที่แสดงในรูปแบบอนาล็อกหรือดิจิทัล ตัวอย่างเช่น แผ่นซีดี (CD) เทป แผ่นเสียงในรูปแบบอนาล็อก และไฟล์เสียงรูปแบบดิจิทัล เป็นต้น ประเภทของการแสดงข้อมูลดนตรีพร้อมด้วยตัวอย่าง ดังแสดงในรูปที่ 1



รูปที่ 1: การแสดงข้อมูลดนตรี (Music representation)

งานวิจัยนี้เน้นการศึกษาการใช้ข้อมูลดนตรีที่แสดงด้วยเสียงเพื่อการวิเคราะห์ข้อมูล ซึ่งกระบวนการดึงข้อมูล เริ่มต้นด้วยการนำเข้าข้อมูลตั้งต้นซึ่งก็คือไฟล์เสียง (Audio file) จากนั้นคุณลักษณะของข้อมูลจะถูกสร้างขึ้นหลังจากไฟล์เสียงถูกนำไปผ่านกระบวนการแปลงข้อมูลให้อยู่ในรูปแบบที่สามารถนำไปวิเคราะห์ได้โดยพิจารณาลำดับเวลา ตัวอย่างของคุณลักษณะของข้อมูล อาทิ รูปแบบของคลื่น (Waveform) สเปคโตแกรม (Spectrogram) ภาพไบนารี (Binary Image) ลายนิ้วมือทางเสียง เป็นต้น ลายนิ้วมือทางเสียงเกิดจากการแปลงรูปแบบของคลื่น โดยใช้หลักการทางคณิตศาสตร์ในการคิดคำนวณเพื่อใช้เป็นตัวแทนข้อมูลเสียงและเป็นหนึ่งในคุณลักษณะของข้อมูลที่ถูกนำไปใช้อย่างกว้างขวางในการตรวจสอบการละเมิดลิขสิทธิ์และการระบุข้อมูลเพลง (Song Identification) สำหรับขั้นตอนการใช้ลายนิ้วมือทางเสียงที่กล่าวมาแล้วข้างต้นแสดงดังในรูปที่ 2

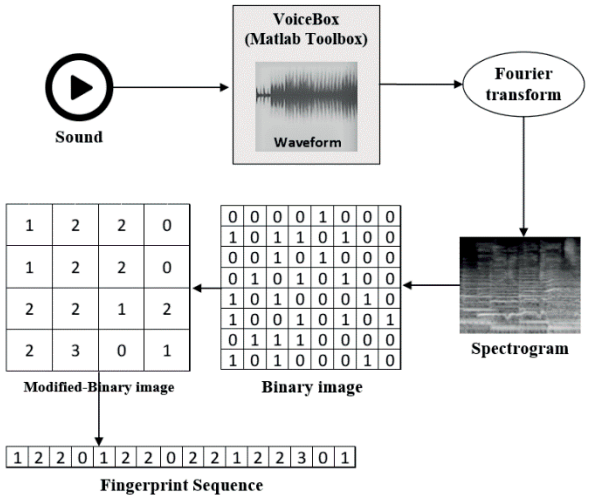


รูปที่ 2: ขั้นตอนการใช้ลายนิ้วมือทางเสียงในการตรวจสอบการละเมิดลิขสิทธิ์และการระบุข้อมูลเพลง

ทั้งนี้ การใช้ลายนิ้วมือทางเสียงในการระบุข้อมูลเพลงที่ไม่ทราบแหล่งที่มา โดยใช้การค้นหาข้อมูลจากเนื้อหาบางส่วน (Query-by-Example) [7] และการตรวจสอบการทำซ้ำของเพลง [3] มีประสิทธิภาพสูงในการตรวจสอบแท็กเสียงที่ผลิตจากแหล่งเดียวหรือเนื้อหาซ้ำกัน (Duplicate content) การใช้ลายนิ้วมือทางเสียงที่สร้างจากภาพไบนารีมีประสิทธิภาพดีกว่าการใช้ภาพไบนารีที่สร้างจากสเปคโตแกรมของข้อมูลเพลงโดยตรง [8] แต่กลับไม่สามารถตรวจสอบกรณีที่ข้อมูลเสียงมีเนื้อหาคล้ายคลึงกันแต่ไม่ตรงกันได้

3.2) กระบวนการเตรียมข้อมูล (Data pre-processing)

สำหรับงานวิจัยนี้การเตรียมข้อมูล (Data pre-processing) ประยุกต์แนวคิดการสร้างลายนิ้วมือทางเสียง [3] ซึ่งมีสามขั้นตอนประกอบด้วย (1) การสร้างสเปคโตแกรม (2) การสร้างภาพไบนารี และ (3) การสร้างลายนิ้วมือทางเสียงจากภาพไบนารี ภาพรวมของกระบวนการเตรียมข้อมูลดังแสดงในรูปที่ 3



รูปที่ 3: ภาพรวมของกระบวนการเตรียมข้อมูล (Data pre-processing)

เพลงที่มีความยาวเต็มทั้งเพลงในรูปแบบไฟล์เสียงเป็นข้อมูลดั้งต้นที่ถูกนำไปสร้างรูปคลื่นของสัญญาณเสียง ซึ่งแสดงลักษณะสัญญาณที่เปลี่ยนไปในช่วงเวลาต่างๆ ด้วยอัตราการสุ่มตัวอย่าง 44,100 ครั้งต่อวินาที โดยใช้ VOICEBOX [9] ที่เป็นกล่องเครื่องมือการประมวลผลเสียง (Speech processing toolbox) ในโปรแกรม MATLAB [10] หลังจากนั้นแบ่งรูปคลื่นของสัญญาณเสียงที่มีความยาวทั้งเพลงเป็นเฟรมเสียง (Spectrogram frame) และแปลงรูปคลื่นสัญญาณของเฟรมเสียงบนโดเมนเวลา (Time domain) ให้อยู่บนโดเมนความถี่ (Frequency domain) โดยการแปลงฟูเรียร์แบบเร็ว (FFT) รูปคลื่นสัญญาณเสียงที่ได้ของแต่ละเฟรมจะถูกนำไปสร้างสเปคโตแกรมดังขั้นตอนที่อธิบายในอัลกอริทึมที่ 1

Algorithm 1: Generate spectrogram

```
Data: x : matrix represent audio file
Result: y : matrix of spectrogram
Set initial value ;
Set frequency band i=1,2,3,...8;
Slice full length waveform to frame;
fftdata=fourier transform of each frame;
for j=1 to length(fftdata) do
    currentSample=fftdata(j);
    if currentSample is in the i band then
        y(j,i)=y(j,i)+currentSample;
    else
        do nothing;
    end
end
```

ในขั้นตอนต่อมาการสร้างภาพไบนารี เริ่มจากการสร้างหน้าต่างของสเปคโตแกรม (Spectrogram window) โดยการนำแปดเฟรมของสเปคโตแกรมที่ได้จากขั้นตอนก่อนหน้ามารวมเข้าด้วยกัน จากนั้นนำมาสร้างเป็นภาพไบนารีโดยการเปรียบเทียบกับค่าเฉลี่ยเลขคณิต (Arithmetic mean) ซึ่งถูกคำนวณมาจากค่าของสมาชิกทุกตัวในหน้าต่างนั้นๆ [3] ภาพไบนารีจะถูกสร้างโดยกำหนดค่าศูนย์หรือหนึ่งแทนที่ค่าเดิมของสมาชิกแต่ละตัวในหน้าต่างตามเงื่อนไขดังอธิบายในอัลกอริทึมที่ 2

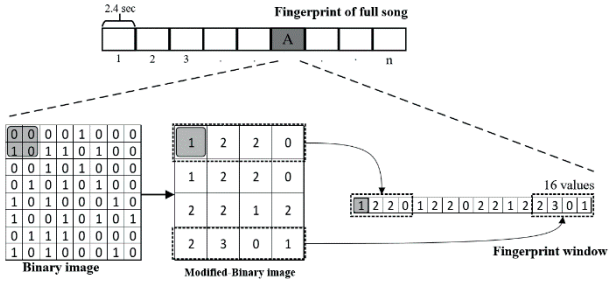
Algorithm 2: Generate binary image

```
Data: x : matrix of spectrogram
Result: y : matrix of binary image
split spectrogram (x) to eight frames window (y);
for i=1 to length(y) do
    if y(i) <mean of y window then
        replace y(i) by 0;
    else
        replace y(i) by 1;
    end
end
```

ในขั้นตอนสุดท้ายเป็นการสร้างลายนิ้วมือทางเสียงจากภาพไบนารี ซึ่งทำโดยการบวกรวมค่าของภาพไบนารีขนาดสองคูณสองเป็นค่าเดียว ดังนั้น หนึ่งหน้าต่างจะได้ค่าทั้งหมด 16 ค่าตามลำดับ เรียกลำดับดังกล่าวว่า ลายนิ้วมือทางเสียง

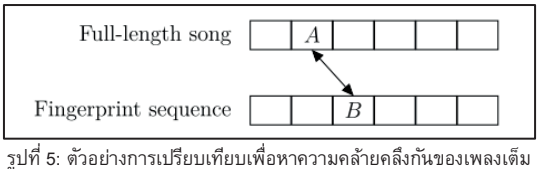
3.3) วิธีการที่นำเสนอ (Proposed Approach)

หลังจากที่ข้อมูลเพลงเต็มผ่านกระบวนการเตรียมข้อมูลแล้วจะถูกแบ่งเป็นหน้าต่างลายนิ้วมือทางเสียงขนาดเล็กที่ประกอบไปด้วย 16 ค่าข้อมูลต่อหนึ่งหน้าต่าง ดังแสดงรายละเอียดในรูปที่ 4



รูปที่ 4: หน้าต่างลายนิ้วมือทางเสียงของเพลงเต็มและลำดับค่าในแต่ละหน้าต่าง

การเปรียบเทียบเพื่อหาความคล้ายคลึงกันของแต่ละหน้าต่างเมื่อกำหนดให้ A และ B คือหน้าต่างลายนิ้วมือทางเสียงของเพลงสองเพลงที่แตกต่างกัน ดังแสดงในรูปที่ 5



รูปที่ 5: ตัวอย่างการเปรียบเทียบเพื่อหาความคล้ายคลึงกันของเพลงเต็ม

ความคล้ายคลึงกันของสองหน้าต่างลายนิ้วมือทางเสียงจะถูกคำนวณโดยการวัดคะแนนที่ตรงกัน (Matching score) เมื่อหน้าต่างลายนิ้วมือทางเสียง A และ B มีค่าดังสมการที่ (1) และ (2) จะสามารถนิยามค่า (A-B) ได้ดังสมการที่ (3) โดยที่ w_i เป็นค่าที่สอดคล้องกับเงื่อนไขในสมการที่ (4)

$$A = < a_1, a_2, ..., a_n > \quad (1)$$

$$B = < b_1, b_2, ..., b_n > \quad (2)$$

$$A - B = < w_1, w_2, ..., w_n > \quad (3)$$

$$w_i = \begin{cases} 1, & if\ (a_i - b_i) = 0 \\ 0, & otherwise \end{cases} \quad (4)$$

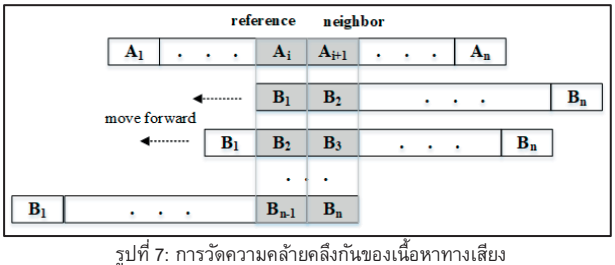
สำหรับคะแนนที่ตรงกันของหน้าต่างลายนิ้วมือทางเสียง A และ B เป็นผลรวมค่าของสมาชิกทุกตัวใน (A-B) ซึ่งคะแนนที่ตรงกันจะเพิ่มขึ้นทีละหนึ่งเมื่อค่าในลำดับเดียวกันของทั้งสองหน้าต่างเท่ากัน ดังสอดคล้องกับสมการที่ (5)

$$\mu(A - B) = \sum_{i=1}^{n} w_i \mid w_i = (A - B)_{i^{th}} \quad (5)$$

ตัวอย่างในการคำนวณคะแนนที่ตรงกันของหน้าต่างลายนิ้วมือทางเสียง A และ B แสดงไว้ดังรูปที่ 6

รูปที่ 6: ตัวอย่างการเปรียบเทียบหน้าต่างลายนิ้วมือทางเสียง A และ B

เนื่องด้วยเพลงหนึ่งประกอบไปด้วยหลายหน้าต่างลายนิ้วมือทางเสียง การวัดความคล้ายคลึงกันระหว่างเพลงสองเพลงนั้นจึงทำได้โดยการประเมินค่าคะแนนที่ตรงกันของทุกๆหน้าต่างลายนิ้วมือทางเสียง แต่งานวิจัยฉบับนี้ได้นำเสนอฟังก์ชันความสัมพันธ์ (Relation function) สำหรับการเปรียบเทียบลายนิ้วมือทางเสียงแทนการเปรียบเทียบข้อมูลขนาดใหญ่ทีละคู่ ดังแสดงในรูปที่ 7



รูปที่ 7: การวัดความคล้ายคลึงกันของเนื้อหาทางเสียง

ฟังก์ชันความสัมพันธ์ (Relation function) ที่นำเสนอประกอบด้วยค่านัยสำคัญทางสถิติที่ถูกคำนวณคือค่าความคล้ายคลึงกัน (Sum of similarity: σ) และค่าความแตกต่าง (Degree of difference: δ) สำหรับค่าความคล้ายคลึงกันคือผลรวมของคะแนนที่ตรงกันของหน้าต่างอ้างอิง $(A_i - B_i)$ และหน้าต่างเพื่อนบ้าน $(A_{i+1} - B_{i+1})$ ในขณะที่ค่าความแตกต่างคือสัมบูรณ์ของความแตกต่างระหว่างคะแนนที่ตรงกันของหน้าต่างอ้างอิงและหน้าต่างเพื่อนบ้านดังสมการที่ (6) และ (7)

$$\sigma_i = \mu(A_i - B_i) + \mu(A_{i+1} - B_{i+1}) \quad (6)$$

$$\delta_i = \left| \mu(A_i - B_i) - \mu(A_{i+1} - B_{i+1}) \right| \quad (7)$$

จากสมการข้างต้น σ_i และ δ_i หมายถึงค่าความคล้ายคลึงกันและความแตกต่างของหน้าต่างลายนิ้วมือทางเสียงที่ i เมื่อคำนวณค่าความคล้ายคลึงกันและค่าความแตกต่างแล้ว ทั้งนี้เพื่อวิเคราะห์ความคล้ายกันของทั้งเพลงจึงนิยามค่าสหสัมพันธ์ระหว่างสองเพลง (Correlation: C) ขึ้น โดยค่าสหสัมพันธ์ระหว่างเพลงสองเพลงถูกกำหนดให้เป็นเซตของข้อมูลที่มีสมาชิกเป็นคู่ลำดับของค่าความคล้ายคลึงกันและค่าความแตกต่าง ดังสมการที่ (8)

$$C_{A,B} = \left\{ (\sigma_{11}, \delta_{11}), (\sigma_{12}, \delta_{12}), ..., (\sigma_{(n-1)(n-1)}, \delta_{(n-1)(n-1)}) \right\} \quad (8)$$

โดยที่ $C_{A,B}$ มีจำนวนสมาชิกเป็น $(n-1)^2$ เมื่อเปรียบเทียบเพลงที่ประกอบด้วยหน้าต่างลายนิ้วมือทางเสียงจำนวน n หน้าต่าง สมาชิกหนึ่งตัวที่มีค่าความคล้ายคลึงกันมากที่สุดโดยที่มีค่าความแตกต่างน้อยที่สุดด้วยจะถูกเลือกให้เป็นตัวแทนของคะแนนความคล้ายกันของเพลง A และ B รายละเอียดดังแสดงในสมการที่ (9) (10) และ (11)

$$max(C_{A,B}) = \left\{(\sigma,\delta) \mid (\sigma,\delta),(\sigma',\delta') \in C_{A,B} \text{ and } \sigma \geq \sigma'\right\} \quad (9)$$

$$min(C_{A,B}) = \left\{(\sigma,\delta) \mid (\sigma,\delta),(\sigma',\delta') \in C_{A,B} \text{ and } \delta \leq \delta'\right\} \quad (10)$$

$$similarity(A-B) \in min(max(C_{A,B})) \quad (11)$$

ตัวอย่างในการคำนวณค่านัยสำคัญทางสถิติที่เกี่ยวข้องกับฟังก์ชันความสัมพันธ์ (Relation function) เพื่อหาค่าความคล้ายกันของเพลงดังแสดงในรูปที่ 8

$$C_{A,B} = \left\{(20,5),(28,1),(28,5),(28,1)\right\}$$
$$max(C_{A,B}) = \left\{(28,1),(28,5),(28,1)\right\}$$
$$min(max(C_{A,B})) = \left\{(28,1),(28,1)\right\}$$
$$similarity(A-B) = (28,1)$$

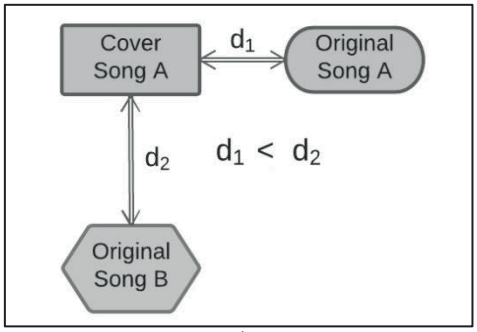
รูปที่ 8: ตัวอย่างในการคำนวณค่านัยสำคัญทางสถิติที่เกี่ยวข้อง

4) ผลการวิจัย

การทดลองประยุกต์ใช้วิธีการหาความคล้ายกันของเพลงเพื่อระบุเพลงต้นฉบับจากเพลงคัฟเวอร์ ข้อมูลที่ใช้สำหรับการทดลองเป็นไฟล์เสียงจำนวน 120 ข้อมูลตัวอย่าง ซึ่งประกอบด้วย 1) เพลงเวอร์ชั่นต้นฉบับ 20 ข้อมูลตัวอย่าง โดยเพลงที่ถูกเลือกเป็นเพลงที่มีจำนวนการถูกคัฟเวอร์จำนวนมากซึ่งประกอบด้วยเพลงสากลจำนวน 15 เพลงและเพลงเกาหลีจำนวน 5 เพลง และ 2) เพลงเวอร์ชั่นคัฟเวอร์จำนวนห้าข้อมูลตัวอย่างต่อหนึ่งเพลงต้นฉบับ รวมเป็น 100 ข้อมูลตัวอย่าง

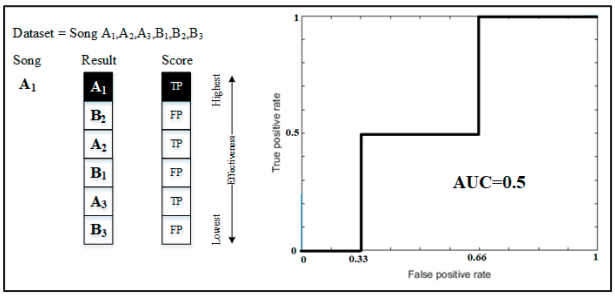
สำหรับกระบวนการเตรียมข้อมูลนั้น ข้อมูลตัวอย่างที่เป็นไฟล์เสียงถูกนำมาสร้างเป็นเฟรมของรูปคลื่นสัญญาณความยาว 0.3 วินาที จากนั้นนำไปสร้างสเปคโตแกรมที่ช่วงความถี่ 40 ถึง 1200 Hz ซึ่งเป็นช่วงความถี่ของเสียงร้องต่ำที่สุดของผู้ชายถึงเสียงร้องสูงสุดของผู้หญิงในเพลง (Vocal frequency range) [11] ขั้นตอนถัดมานำเฟรมของสเปคโตแกรมทีละแปดเฟรมมาสร้างเป็นภาพไบนารี และขั้นตอนสุดท้ายสร้างหน้าต่างของลายนิ้วมือทางเสียงจากภาพไบนารีดังกล่าว

คะแนนความคล้ายกันระหว่างเพลงต้นฉบับและเพลงคัฟเวอร์ของเพลงเดียวกันจะมีค่ามากกว่าคะแนนความคล้ายกันระหว่างเพลงคัฟเวอร์นั้นและเพลงต้นฉบับอื่นที่ไม่ใช่เพลงๆ นั้น ซึ่งนั้นก็หมายความว่าในกรณีเพลงเดียวกันแต่คนละเวอร์ชั่นจะมีระยะห่างระหว่างกันน้อยกว่ากรณีที่เป็นคนละเพลง ซึ่งสอดคล้องกับตัวอย่างในรูปที่ 9

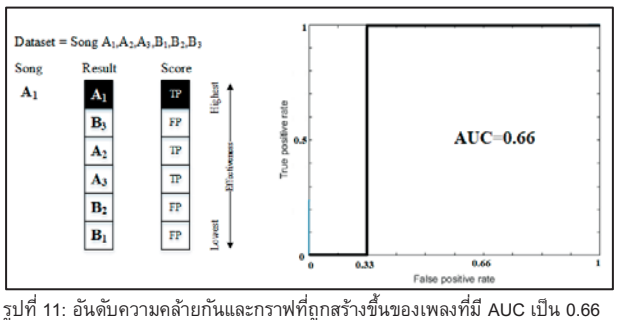


รูปที่ 9: การเปรียบเทียบข้อมูลเสียงเพื่อระบุเพลงต้นฉบับจากเพลงคัฟเวอร์

การประเมินประสิทธิภาพของวิธีการที่นำเสนอทำได้โดยการเปรียบเทียบข้อมูลขาเข้า (Input) กับข้อมูลทุกตัวในกลุ่มตัวอย่าง ซึ่งการเปรียบเทียบแต่ละครั้งจะมีการคำนวณค่าคะแนนความคล้ายกันของเพลงเพื่อจัดอันดับเพลงที่มีคะแนนจากมากไปหาน้อย ดังนั้นเพลงที่มีความคล้ายกับเพลงที่เป็นข้อมูลขาเข้ามากกว่าจะถูกจัดให้อยู่ในลำดับที่ดีกว่า อันดับความคล้ายของเพลง (Similarity rank) ที่ได้มาจะถูกนำไปสร้างกราฟเปรียบเทียบระหว่างอัตราผลบวกจริง (True positive rate) และอัตราผลบวกเท็จ (False positive rate) โดยที่เส้นกราฟจะขยับไปตามแนวแกน X เมื่อเพลงที่อยู่ในลำดับดังกล่าวเป็นเพลงต้นฉบับเพลงอื่นที่ไม่ถูกต้อง (False positive) และกราฟจะขยับไปตามแนวแกน Y เมื่อเพลงที่อยู่ในลำดับดังกล่าวเป็นเพลงต้นฉบับที่ถูกต้อง (True positive) ค่าของทั้งสองแกนอยู่ระหว่างศูนย์ถึงหนึ่งและประสิทธิภาพของการการจัดกลุ่มแปรผันตรงกับพื้นที่ใต้กราฟที่ได้ การจัดกลุ่มที่ถูกต้องสมบูรณ์จะมีพื้นที่ใต้กราฟเท่ากับหนึ่ง ตัวอย่างอันดับความคล้ายกันของเพลงและกราฟที่ถูกสร้างขึ้น รวมทั้งพื้นที่ใต้กราฟที่คำนวณได้ดังแสดงในรูปที่ 10 และ 11 ตามลำดับ



รูปที่ 10: อันดับความคล้ายกันและกราฟที่ถูกสร้างขึ้นของเพลงที่มี AUC เป็น 0.5



รูปที่ 11: อันดับความคล้ายกันและกราฟที่ถูกสร้างขึ้นของเพลงที่มี AUC เป็น 0.66

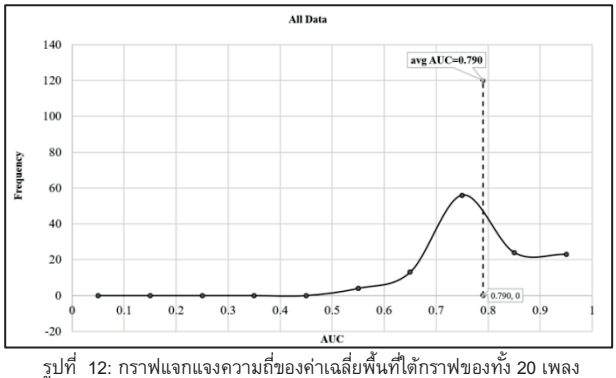
5) อภิปรายผล

พื้นที่ใต้กราฟเป็นตัวบ่งชี้สำคัญต่อประสิทธิภาพของวิธีการที่นำเสนอ ค่าเฉลี่ยของพื้นที่ใต้กราฟ (Average AUC) โดยรวมของทั้ง 20 เพลงในกลุ่มตัวอย่างอยู่ที่ 0.790 และค่าเฉลี่ยของพื้นที่ใต้กราฟของแต่ละเพลงดังแสดงไว้ในตารางที่ 1

ตารางที่ 1: แสดงค่าเฉลี่ยของพื้นที่ใต้กราฟของ 20 เพลงในกลุ่มตัวอย่าง

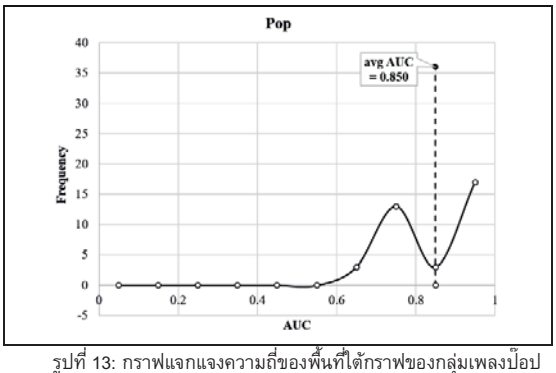
#Song	V1	V2	V3	V4	V5	V6	Avg AUC
1	0.779	0.737	0.818	0.626	0.804	0.812	0.763
2	0.756	0.725	0.796	0.651	0.753	0.743	0.737
3	0.746	0.723	0.675	0.751	0.791	0.732	0.736
4	0.975	0.963	0.904	0.967	0.940	0.988	0.956
5	0.967	0.981	0.784	0.923	0.916	0.930	0.917
6	0.991	0.995	0.993	0.974	0.995	0.993	0.990
7	0.789	0.807	0.654	0.819	0.737	0.547	0.726
8	0.707	0.739	0.747	0.701	0.793	0.725	0.735
9	0.893	0.695	0.896	0.781	0.696	0.935	0.816
10	0.746	0.746	0.807	0.795	0.754	0.888	0.789
11	0.911	0.826	0.904	0.814	0.791	0.704	0.825
12	0.791	0.626	0.715	0.701	0.774	0.702	0.718
13	0.717	0.678	0.690	0.718	0.802	0.774	0.730
14	0.870	0.830	0.754	0.721	0.589	0.812	0.763
15	0.826	0.893	0.805	0.828	0.821	0.637	0.802
16	0.733	0.720	0.589	0.696	0.703	0.547	0.665
17	0.781	0.833	0.774	0.746	0.720	0.699	0.759
18	0.739	0.781	0.809	0.788	0.795	0.763	0.779
19	0.754	0.746	0.772	0.668	0.712	0.782	0.739
20	0.925	0.904	0.809	0.916	0.868	0.707	0.855

จากตารางที่ 1 แต่ละแถวของตารางแสดงพื้นที่ใต้กราฟของเพลงหนึ่งเพลง ซึ่งประกอบไปด้วยหกเวอร์ชั่น โดยที่ V1 เป็นเวอร์ชั่นต้นฉบับ(Original version) ของเพลงดังกล่าว V2-V6 เป็นเวอร์ชั่นคัฟเวอร์ (Cover version) ของเพลงดังกล่าวจำนวนห้าเวอร์ชั่นตามลำดับ และคอลัมน์สุดท้ายของตารางเป็นค่าเฉลี่ยของพื้นที่ใต้กราฟของเพลงนั้นๆ เมื่อนำค่าเฉลี่ยของพื้นที่ใต้กราฟของแต่ละเพลงมานับความถี่ที่เกิดขึ้นโดยการแบ่งช่วงกว้าง (Range) ของค่าเฉลี่ยของพื้นที่ใต้กราฟออกเป็นช่วงๆ ดังนี้ 0-0.1, 0.1-0.2, 0.3-0.4 ,..., 0.9-1.0 ตามลำดับจะได้กราฟแสดงการแจกแจงความถี่ของค่าเฉลี่ยของพื้นที่ใต้กราฟของทุกเพลงในกลุ่มตัวอย่างดังแสดงในรูปที่ 12

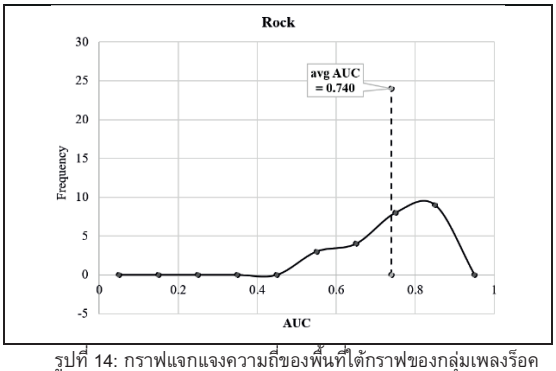


รูปที่ 12: กราฟแจกแจงความถี่ของค่าเฉลี่ยพื้นที่ใต้กราฟของทั้ง 20 เพลง

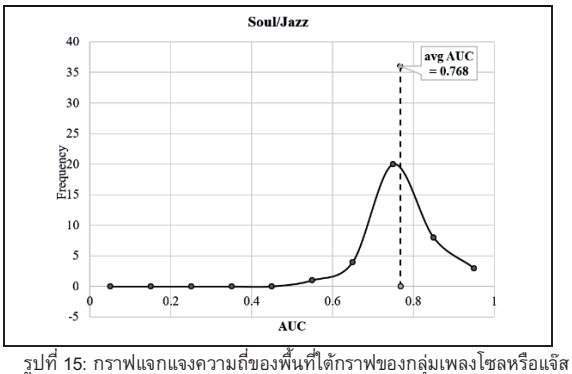
กลุ่มข้อมูลเพลงตัวอย่างถูกแบ่งเป็นสี่กลุ่มเมื่อตามประเภทของเพลง (Genre) ดังนี้ กลุ่มที่หนึ่งเพลงป๊อป (Pop) จำนวน 36 ข้อมูลตัวอย่าง กลุ่มที่สองเพลงร็อค (Rock) จำนวน 36 ข้อมูลตัวอย่าง กลุ่มที่สามเพลงโซลหรือแจ๊ส (Soul/Jazz) จำนวน 24 ข้อมูลตัวอย่าง และสุดท้ายกลุ่มที่สี่เพลงริทึมแอนด์บลูส์ (R&B) ฮิปฮอป (Hip-Hop) หรืออีดีเอ็ม (EDM) จำนวน 24 ข้อมูลตัวอย่าง เมื่อพิจารณาพื้นที่ใต้กราฟแยกตามกลุ่มจะได้กราฟแจกแจงความถี่ ดังแสดงในรูปที่ 13 ถึง 16



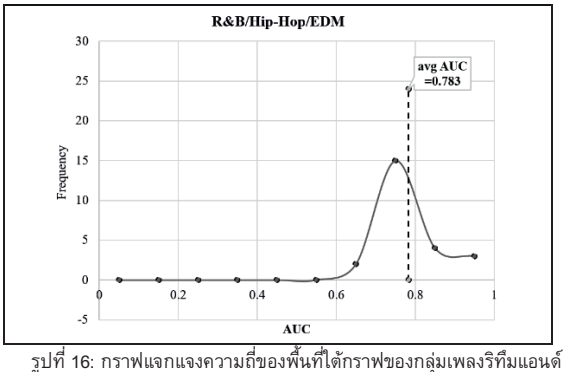
รูปที่ 13: กราฟแจกแจงความถี่ของพื้นที่ใต้กราฟของกลุ่มเพลงป๊อป



รูปที่ 14: กราฟแจกแจงความถี่ของพื้นที่ใต้กราฟของกลุ่มเพลงร็อค



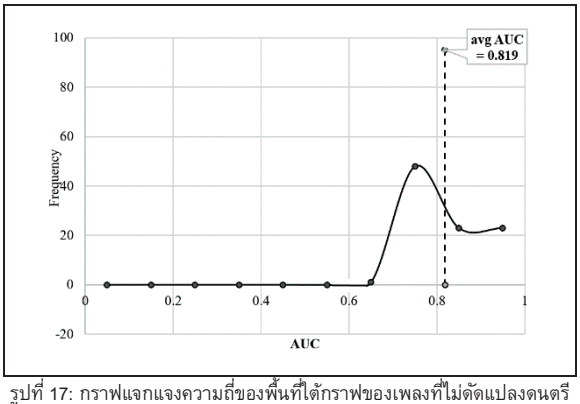
รูปที่ 15: กราฟแจกแจงความถี่ของพื้นที่ใต้กราฟของกลุ่มเพลงโซลหรือแจ๊ส



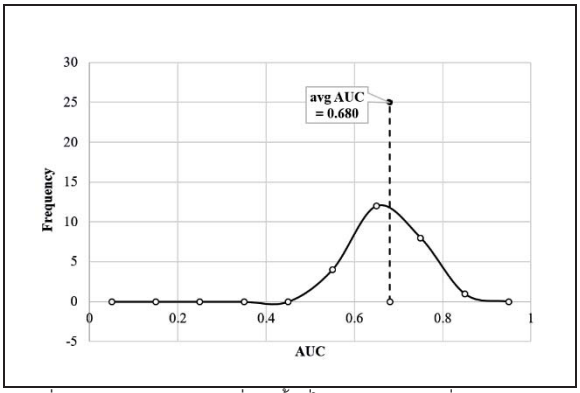
รูปที่ 16: กราฟแจกแจงความถี่ของพื้นที่ใต้กราฟของกลุ่มเพลงริทึมแอนด์บลูส์ ฮิปฮอป หรืออีดีเอ็ม

จากกราฟแจกแจงความถี่ของพื้นที่ใต้กราฟ กลุ่มเพลงป๊อปมีค่าเฉลี่ยพื้นที่ใต้กราฟมากที่สุดเมื่อเทียบกับเพลงในกลุ่มอื่นเป็น 0.85 และเพลงที่มีพื้นที่ใต้กราฟสูงที่สุดมีพื้นที่ใต้กราฟถึง 0.995 ซึ่งเกือบเป็นการจัดกลุ่มแบบสมบูรณ์แบบ ดังกราฟที่แสดงในรูปที่ 13 ในขณะที่อีกสามกลุ่มที่เหลือมีค่าเฉลี่ยพื้นที่ใต้กราฟใกล้เคียงกันอยู่ระหว่าง 0.74-0.78 ดังแสดงในรูปที่ 14 ถึง 16 จากผลการทดลองดังกล่าวบ่งชี้ว่า วิธีการที่นำเสนอสามารถจัดกลุ่มเพลงเดียวกันแต่คนละเวอร์ชั่นให้อยู่ในกลุ่มเดียวกันได้โดยไม่ขึ้นอยู่กับประเภทของเพลง

นอกจากการพิจารณาพื้นที่ใต้กราฟตามประเภทของเพลงแล้ว ยังมีการวิเคราะห์ผลการจัดกลุ่ม โดยพิจารณาลักษณะของการบันทึกเสียงใหม่จากต้นฉบับเดิม ซึ่งแบ่งออกเป็นสองกลุ่มตามลักษณะการบรรเลงดนตรี ลักษณะแรกเป็นเวอร์ชั่นคัฟเวอร์ที่เพียงแค่บรรเลงซ้ำเดิมและร้องใหม่เท่านั้น ไม่ดัดแปลงการบรรเลงดนตรี ลักษณะที่สองมีการดัดแปลงการบรรเลงดนตรีเมื่อบันทึกเสียงอีกครั้งเพื่อสร้างเวอร์ชั่นคัฟเวอร์ ตัวอย่างเช่น การนำเพลงต้นฉบับที่เป็นเพลงป๊อปมาบรรเลงใหม่เป็นเพลงร็อคแทน ผลการทดลองพบว่า ลักษณะแรกที่ไม่ดัดแปลงการบรรเลงดนตรีมีพื้นที่ใต้กราฟมากกว่าลักษณะที่สองอย่างเห็นได้ชัด ดังแสดงในรูปที่ 17 และ 18 ตามลำดับ



รูปที่ 17: กราฟแจกแจงความถี่ของพื้นที่ใต้กราฟของเพลงที่ไม่ดัดแปลงดนตรี



รูปที่ 18: กราฟแจกแจงความถี่ของพื้นที่ใต้กราฟของเพลงที่ดัดแปลงดนตรี

6) บทสรุป และข้อเสนอแนะ

ข้อสรุปที่ได้จากการวิจัยบ่งชี้ว่า การหาความคล้ายกันของเพลงโดยใช้ฟังก์ชันความสัมพันธ์ (Relation function) สามารถระบุเพลงต้นฉบับจากเพลงคัฟเวอร์ได้อย่างมีประสิทธิภาพครอบคลุมเพลงทุกประเภท โดยมีพื้นที่ใต้กราฟอยู่ระหว่าง 0.74-0.85 และการจัดกลุ่มเพลงมีประสิทธิภาพในกรณีที่มีการบรรเลงดนตรีใหม่ในรูปแบบคล้ายกับต้นฉบับเดิม ในขณะที่การบรรเลงโดยเปลี่ยนรูปแบบดนตรีส่งผลให้ความถูกต้องของการจัดกลุ่มน้อยลง ซึ่งข้อจำกัดดังกล่าวผู้วิจัยจะนำมาศึกษาและปรับปรุงวิธีการต่อไป

กิตติกรรมประกาศ

งานวิจัยนี้สำเร็จลุล่วงไปได้ด้วยดีด้วยความอนุเคราะห์ทุนสนับสนุนการศึกษา สำหรับนิสิตระดับปริญญาเอกจากคณะวิศวกรรมศาสตร์ มหาวิทยาลัยนเรศวร ผู้เขียนจึงขอกราบขอบพระคุณ ณ โอกาสนี้ ขอขอบคุณคณาจารย์ทุกท่านที่คอยให้คำปรึกษาและข้อเสนอแนะ สุดท้ายนี้ขอขอบคุณครอบครัวและคนรอบข้างที่เป็นแรงสนับสนุนสำคัญ

เอกสารอ้างอิง

[1] J. S. Downie, "Music information retrieval," *Annual Review of Information Science and Technology*, vol. 37, pp. 295-340, 2003.

[2] L. Tao and M. Ogihara, "Toward intelligent music information retrieval," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 564-574, 2006.

[3] C. Ouali, P. Dumouchel, and V. Gupta, " A spectrogram-based audio fingerprinting system for content-based copy detection," *Multimedia Tools and Application*, vol. 75, no. 15, pp. 9145-9165, 2016.

[4] D. Perrott and R. Gjerdingen, “Scanning the dial: An exploration of factors in the identification of musical style.,” in *The 8th international conference on music perception & cognition*, Evanston, Illinois, USA, 2004.

[5] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 15, pp. 293-302, 2002.

[6] D. C. Correa and F. A. Rodrigues, "A survey on symbolic data-based music genre classification," *Expert Systems with Applications*, vol. 60, pp. 190-210, 2016.

[7] V. Chandrasekhar, M. Sharifi, and D. A. Ross, “Survey and Evaluation of Audio Fingerprinting Schemes for Mobile Query-by-Example Applications.,” in *the 12th International Society for Music Information Retrieval Conference*, Miami, Florida, USA, 2011, pp. 801–806.

[8] C. Ouali, P. Dumouchel, and V. Gupta, "Efficient spectrogram-based binary image feature for audio copy detection," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane Convention & Exhibition Centre Brisbane, Queensland, Australia, 2015, pp. 1792-1796.

[9] “VOICEBOX: Speech Processing Toolbox for MATLAB.” [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. [Accessed: 24-Feb-2017].

[10] “MATLAB,” *MathWorks*. [Online]. Available: https://www.mathworks.com/products/matlab.html. [Accessed: 24-Jan-2017].

[11] W. B. Snow, "Audible Frequency Ranges of Music, Speech and Noise," *The Journal of the Acoustical Society of America*, vol. 3, no. 10, pp. 10-10, 1931.