A METHOD FOR ASSESSING THE EFFECTIVENESS OF SIMILARITY MEASURE FOR AUDIO CLUSTERING USING AREA UNDER THE CURVE (AUC)

Sunun Tati[1*]

[1]Faculty of Industrial Technology, Pibulsongkram Rajabhat University, Muang, Phitsanulok, Thailand, 65000

*Corresponding author: sunun.t@psru.ac.th

วันที่เข้าระบบ 10 พฤษภาคม 2562
วันที่แก้ไขบทความ 7 มิถุนายน 2562
วันที่ตอบรับบทความ 7 สิงหาคม 2562

Abstract

Audio fingerprint analysis is a widely used method for identifying specific pieces of audios, and can be effective at analyzing specific audio tracks. However, this analysis is only effective in detecting exact duplicate content that matches another piece of audio. A method for finding similarity between audio pieces is reviewed, and an evaluation method to find how effectiveness of that algorithm is proposed. The evaluation method uses Area Under the Curve (AUC) which is calculated by using similarity ranking.

Keywords: Data clustering, Music recognition, Audio fingerprint

1. Introduction

In the information age, copyright infringement is easy to do and is resulting in significant financial damage to many original artists, filmmakers, singers and song writers, among others. For the music industry particularly, releasing new versions of songs without the permission of the owners is having a huge financial impact on the original copyright holders. The ability to detect copyright infringements and unauthorized reproductions is an important matter in these circumstances.

Methods to detect music copyright infringement are currently practiced, with the more usual method being random checking by a copyright officer. This approach has significant limitations due to the difficulty of analyzing large amounts of data, together with the need for an informed and experienced listener. Some computer-based, algorithmic approaches are also in use (Downie J.S., 2003).

Music Information Retrieval (often abbreviated to MIR) is an interdisciplinary research approach integrating different areas of research, such as digital signal processing, pattern recognition, music theory, and psychology. A subset of MIR, the Music Recognition Technique (Tao and Ogihara, 2006) is a technique to extract specific features of the music that can be used to detect duplicate content.

The audio content of any piece of audio has identifiable features, one of which is known as the audio fingerprint (Ouali *et al*., 2016). The audio fingerprint is a unique data feature of a piece of audio, similar to the uniqueness of human fingerprints used to identify a specific person. For any musical piece, the audio fingerprint of any input audio file can be generated and used for comparison against the audio fingerprints of other audio data. Copyright infringements can be detected by matching the audio fingerprints of an original piece, and the cover version (the term used for the suspected copy).

The problem in audio fingerprinting is that it can only be used to detect exact matches and is unable to be used for close copies or items that are only similar to each other. Copyright infringement can occur due to significant similarity, or different arrangements of the musical piece, which are not necessarily direct or exact copies of the original.

2. Research Objectives and Focus

We propose the evaluation method to find how effectiveness of an approach to find similarities between two songs by comparing their audio fingerprints, which can be used for copyright infringement detection or song identification.

This paper is structured as follows: Section 3 is literature review, Section 4 presents the proposed approach, Section 5 summarizes and discusses the results, and there is a final Conclusion section.

3. Literature Review

Many music-related processes such as copyright infringement detection, song identification or genre classification apply music recognition techniques. Previous research compared the effectiveness of human listening and music recognition techniques for musical genre classification. Perrot and Gjerdingen, 1999 showed that a non-expert person can identify the genre with 72% accuracy after listening to a 3-second segmentation of music. Tzanetakis and Cook (Tzanetakis and Cook, 2002) found that automatically classifying ten musical genres by computerized music recognition techniques achieved only 61% accuracy. They concluded that the main factor affecting the accuracy of automatic recognition accuracy is in the methods of data feature extraction.

There are various ways to extract data features from recorded music, which is available in many forms, both analog and digital, on a variety of media, CDs, tapes, vinyl records, or, more usually in this day and age, digital audio files. The main musical feature of interest here is the audio fingerprint of the musical piece or song, which is the most widely used feature to use for copyright infringement detection.

The process for using the audio fingerprint method of copyright infringement detection starts with the extraction of the data from an input audio file, during which action the waveform of the music is generated. Next, the audio fingerprint is generated from the waveform by using a mathematical function. This process is illustrated in Figure 1.

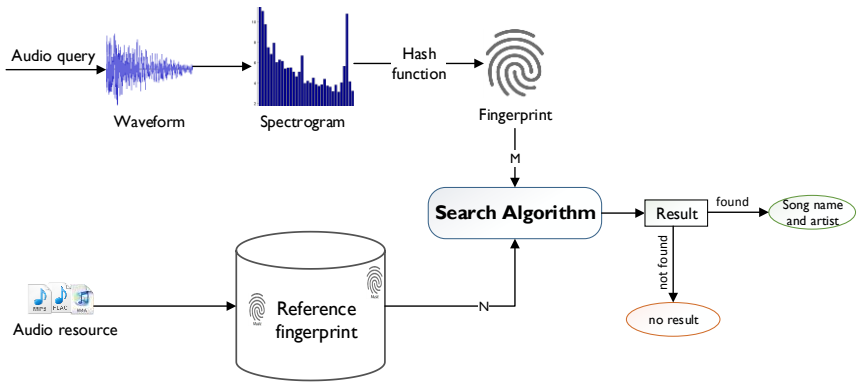

Figure 1 Music copyright infringement detection method using fingerprint schema

For song identification, the audio fingerprint schema can also be used to detect a song with the same audio fingerprint as that of the original or of another copy of the original. The effectiveness of audio fingerprinting was demonstrated as being superior to using binary images (Ouali *et al.*, 2015). Audio fingerprints can be generated from binary images. The significant problem with both binary images and audio fingerprints, however, is that they are only effective in detecting content that is an exact match to another piece of audio. A method for finding similarity between musical pieces can address this problem.

In data pre-processing method, Chahid *et al.*'s approach (Ouali *et al.*, 2016) generate the audio fingerprint. The pre-processing of the audio data has three main steps: generating the audio spectrogram, generating the binary image and generating the audio fingerprint. This is illustrated in Figure 2.

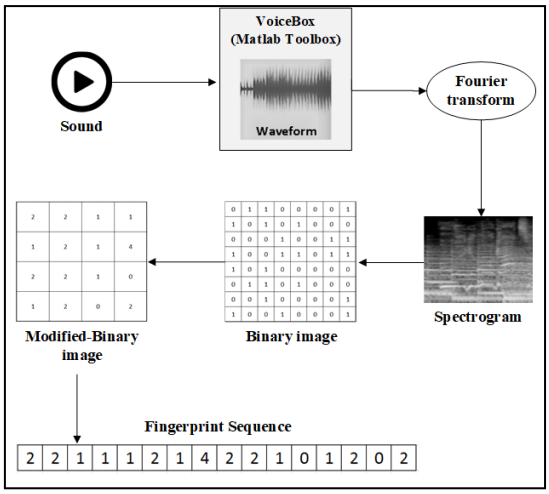

Figure 2 Data pre-processing

The first step is to generate the audio waveform from full-length song which is in a WAV format audio file. The output waveform shows how a signal changes with time (time-domain graph) and is split into 0.3-second frames, which are then individually transformed from the time domain waveform to the frequency domain audio waveform by Fast Fourier Transform (FFT). Then the frequency domain audio waveform by the FFT is used as input to an algorithm to generate a spectrogram of each audio frame.

The algorithm merges 0.3-second eight spectrogram frames, which is a spectrogram window (our terminology) of 0.24 seconds duration. The spectrogram image is an 8x8 matrix of pixel values. The spectrogram window is converted into a binary image by comparing each value in the spectrogram window matrix to the mean arithmetic value of the window matrix (Ouali *et al.*, 2016), replacing each value with 0, where the value is less than the mean, or 1 where the value is equal to or greater than the mean (Algorithm 2). Next, combine the values in each 2x2 sub-matrix tile into one value to calculate a 4x4 matrix from the 8X8 matrix, then convert the 4X4 binary image matrix in 16 values vectors as the fingerprint sequence.

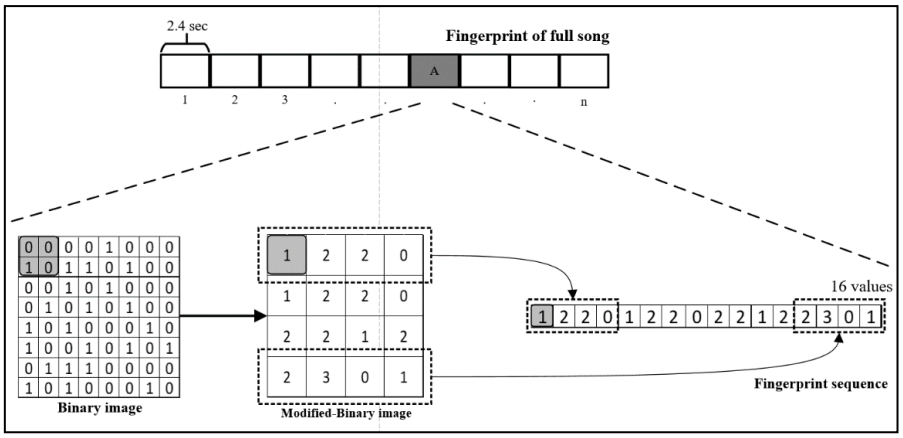

Figure 3 Audio fingerprint window of full-length song and sequence value in each window.

Each audio has a set of n fingerprint sequence windows that were constructed in the pre- processing process, previously described. The length of the song will determine n. For brevity, we refer to each of these n fingerprint sequence windows as a window vector, each of which is comprised of 16 bits, and, if we refer to Song A and Song B, each window vector is designated A_n or B_n in the following calculations, where n is determined by the length of each song. These are the data input used in our approach (Figure 3). Each A is compared to the associated window vector B, as illustrated in Figure 4.

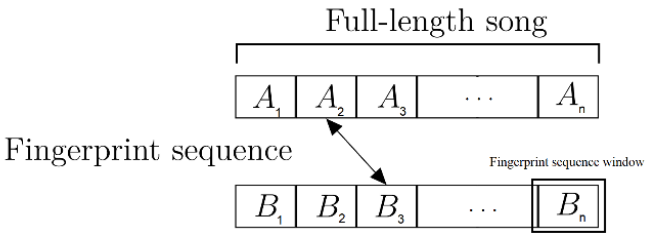

Figure 4 fingerprint sequence window vectors of two different songs

Because a song consists of many fingerprint windows, similarity measurement between two songs is evaluating the similarity scores of all fingerprint window. Relationship Functions [7] for comparison of fingerprints, instead of comparing large data pairs are shown in Figure 5.

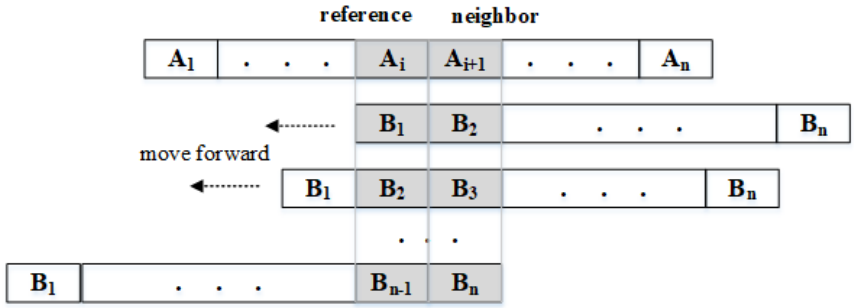

Figure 5 Measuring song similarities

After that calculate two statistical significance values: sum similarity (σ) and degree of difference (δ). Sum similarity is the sum of matching score of reference window $(A_i - B_i)$ and neighbor window $(A_{i+1} - B_{i+1})$. The degree of difference is absolute of difference between matching score of reference window $(A_i - B_i)$ and neighbor window $(A_{i+1} - B_{i+1})$.

4. Proposed approach

We compared the audio input file of a song against all of the audio files in our dataset to calculate the similarity score between the example input file and each file in the database, to create a similarity ranking, which we ordered in descending order of similarity score. To measure the effectiveness, calculate the area under the curve (AUC), shown in Figures 6, 7, 8 and 9 is calculated from the similarity ranking. The area under the graph (AUC) is calculated from the similarity ranking to measure the effectiveness of the method we use. AUC is an area under the curve which plot between true positive rate (TPR) and false positive rate (FPR). Of concern in these comparisons is the possibility of True Positives where the process either correctly identifies a song as being a copy of another song, or where a song is wrongly identified as not being the same. In our experiments, we tested an audio input file that was known to have a copy cover song in the database. We could then test for the proportion of true and false positives to assess the overall accuracy of our method, which we calculated as the Area Under the Curve (AUC)

The AUC space has 2 axes, each axis having a maximum value of one. The x-axis value increases for each false positive identified, and the y-axis value increases for

each true positive identified. When the input song is cover or original of a song in similar ranking, it means a true positive occur, the y-axis line will step up, and if it is a false positive the x-axis line steps to the right (only positives are being tested). If all correct songs are in the top of ranking, the graph will sit along the 0 value of the y-axis, at the maximum y-axis value, and the area AUC will be 1.0 (Figure 6).

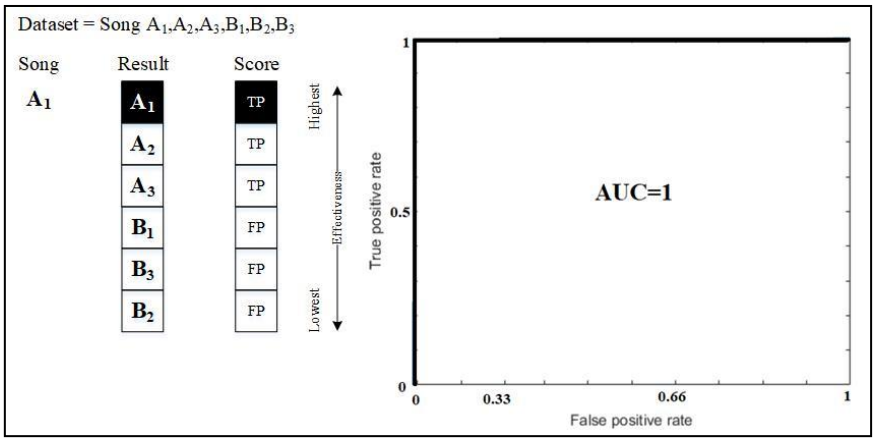

Figure 6 Example results: AUC=1

The other extreme case is where every comparison is shown to be a false positive, in which case the graph line will be 0 on the x-axis, and the AUC is equal to 0.0 (Figure 7).

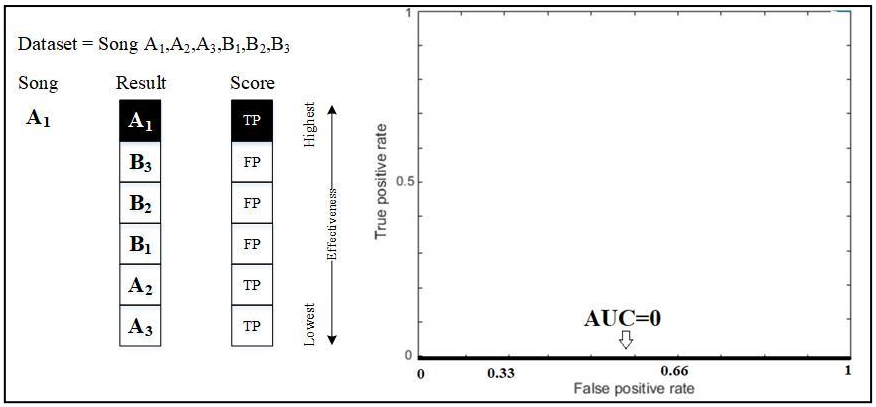

Figure 7 Example results: AUC=0

In Figure 8, each part of the graph line indicates either a True Positive (move along the x-axis) or a False Positive (move up the y-axis), resulting in an AUC of 0.5. Figure 9, using the same description, shows an AUC of 0.66.

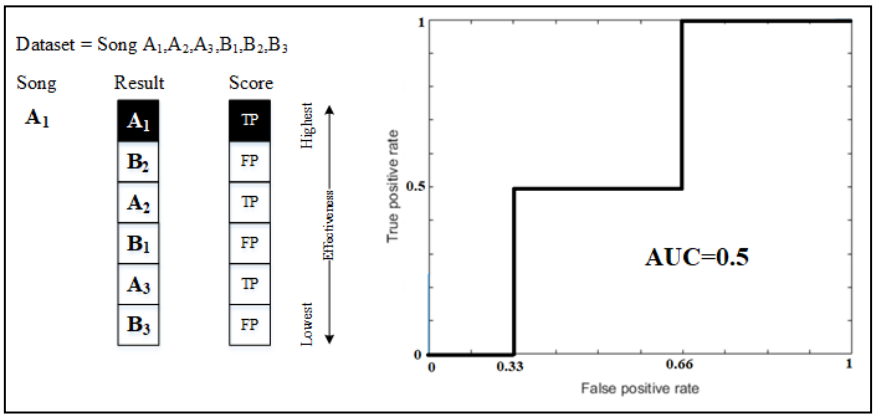

Figure 8 Example results: AUC=0.5

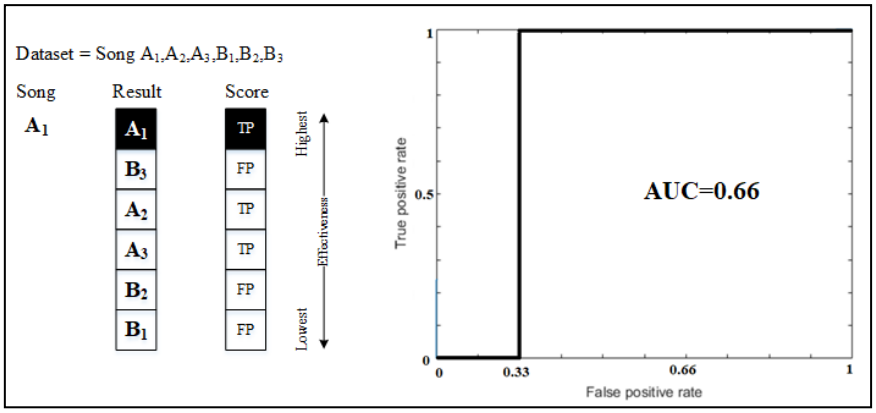

Figure 9 Example results: AUC=0.66

5. Research Conclusions

The digital distance between a same and similar audio can be calculated, and the shortest digital distance calculated will identify the closest copy, therefore the most likely copyright infringement. Effective of clustering algorithm can measure by calculating the area under the curve (AUC). AUC is an area under the curve which plot between true positive rate (TPR) and false positive rate (FPR). The algorithm which have more average AUC, that indicates the algorithm is more effective than other.

6. Reference

Downie, J. S. (2003). Music information retrieval. **Annual Review of Information Science and Technology.** 37, 295--340

Ouali, C., Dumouchel, P., & Gupta, V. (2015). Efficient spectrogram-based binary image feature for audio copy detection. **IEEE International Conference on Acoustics, Speech and Signal Processing *(ICASSP).*** 1792-1796.

Ouali, C., Dumouchel, P., & Gupta, V. (2016) A spectrogram-based audio fingerprinting system for content-based copy detection. **Multimedia Tools Appl.** 75, 9145-9165

Perrot, D., & Gjierdingen, R. O. (1999). **Scanning the dial: An exploration of factors in the identification of musical style.** Proceeding of International Conference on Music Perception and Cognition.

Tao, L., & Ogihara, M. (2006). Toward intelligent music information retrieval. **Multimedia, IEEE Transactions on.** 8, 564-574

Tati, S., Kijsanayothin, P., & Kongdenfha, W. (2018, January - June). Song clustering using similarity of audio fingerprint. **TNI Journal of Engineering and Technology.** 6(1), 49-55

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. **Speech and Audio Processing, IEEE Transactions on.** 10, 293-302