

# An infringement detection system for videos based on audio fingerprint technology

Yang Zheng

Institute of Automation Chinese Academy of Sciences  
Beijing, China  
yang.zheng@ia.ac.cn

Jie Liu, Shuwu Zhang

BJDCRC  
Beijing, China

**Abstract**—The system designs and implements a copyright detection system for videos based on audio fingerprint technology, which consists of two parts: the construction of the fingerprint dataset and the detection of infringing videos. In the audio fingerprint extraction algorithm, the audio fingerprint is constructed by using the peak point extraction method based on the auditory mechanism and the method of using the adjacent peak point pairs, which ensures that the fingerprint feature has a high differentiation and robustness. In the audio fingerprint retrieval algorithm, the stability and efficiency of fingerprint retrieval are ensured by using the combination of hash function and inverted indexing method and the method of flexible matching of fingerprints within a certain threshold range. The validity of the system method is proved by verification in two datasets.

**Keywords**- copyright detection system; fingerprint extraction; fingerprint retrieval

## I. INTRODUCTION

The rapid development of science and technology has made people come into contact with more and more video playback sites, on the one hand, promoting the dissemination of information and enriching people's lives. On the other hand, some websites illegally broadcast videos under unauthorized conditions, infringing the interests of others at the same time have a negative impact on the development of video industry. The illegal broadcasting of a large number of videos has seriously affected the mood and enthusiasm of the authors of videos, damaged the legitimate rights and interests of the owners of videos. It is necessary to protect the copyright of digital videos. Content-based audio retrieval technology can effectively address the infringement of videos, while audio fingerprint (AF) is the key to the content-based audio retrieval technology. Audio fingerprinting technology is mainly used in the recognition of audio content, especially for digital rights management (DRM) systems, with efficiency and accuracy, has gradually become a popular research topic.

In the process of industrial production, audio fingerprint technology can provide the advantages of small storage space and fast retrieval speed, so that it can be applied to music recognition, radio and television monitoring and tracking, copyright detection and personalized entertainment and interaction. Since it can be applied to a wide range of application scenarios, many application systems have been

developed based on audio fingerprint technology as demanded. Among them, Shazam<sup>[1]</sup> and SoundHound<sup>[2]</sup> are two very classic audio fingerprint-based retrieval systems.

In academia, Philips Research, Google, the University of Illinois, the Massachusetts Institute of Technology, the Institute of Acoustics of Chinese Academy of Sciences, Harbin Institute of Technology, Xi'an University of Electronic Science and Technology and other well-known research institutions have carried out research on audio fingerprint extraction and audio fingerprint retrieval and other related algorithms. They achieved fruitful results in improving the robustness of audio fingerprint, the differentiation of fingerprint, the speed of fingerprint retrieval and reducing the error rate. Among them, the Philips Robust Hash (PRH) algorithm<sup>[3]</sup> and Shazam algorithm<sup>[4]</sup> are two representatives of audio fingerprint technology, respectively.

The PRH algorithm is the fingerprint feature based on the multi-sub-belt, which plays as a connecting link between the preceding and the following, Qiang Wang<sup>[5]</sup> and Siyuan Wu<sup>[6]</sup> from Beijing University of Posts and Telecommunications, Qiming He<sup>[7]</sup> from Harbin Institute of Technology, Jie Guo<sup>[8]</sup> from the Institute of Acoustics of Chinese Academy of Sciences and Coover B<sup>[9]</sup>, Wu L M<sup>[10,11]</sup>, Ouali C<sup>[12]</sup>, Ning Sun and Weiping Zhao<sup>[13]</sup>, Shanshan Yao<sup>[14]</sup> and others improve their algorithms respectively based on the PRH algorithm. Their experiments prove the effectiveness of the improved algorithm in detection effect and realize the efficient retrieval of query audio. Shazam fingerprint algorithm<sup>[4]</sup> constructs fingerprint features by assuming that the local energy maximum point in the spectrogram is the spectral peak point and gets the final search result by matching the relationship between fingerprints, which plays a very important role in scientific research. Jie T<sup>[15]</sup>, Jiang T<sup>[16]</sup>, Duong<sup>[17]</sup>, Malekesmaeli M<sup>[18]</sup>, Jun Hu<sup>[19]</sup>, Yong Zhang<sup>[20]</sup> and others improve their algorithms respectively based on the Shazam algorithm, and the audio retrieval was effectively completed.

Audio fingerprint technology has been developed to a certain extent, but each algorithm has its advantages and disadvantages. Faced with the growth of data volume and the influence of external objective factors, it exposes the problems of incomplete fingerprint characteristics and inaccurate fingerprint retrieval. Therefore, the

corresponding methods of the system have been improved to adapt to the actual needs.

The remainder of the paper is organized as follows: The detection system is described in Section 2. Experiments are presented in Section 3. Finally, we give the conclusion and further work in Section 4.

## II. APPROACH

The infringement detection of videos consists of two parts: the construction of the fingerprint dataset and the detection of infringing videos, which are divided into five processes: audio signal extraction, peak point extraction, fingerprint construction, fingerprint index construction and fingerprint matching. As shown in Figure 2-1, these details will be described in this system.

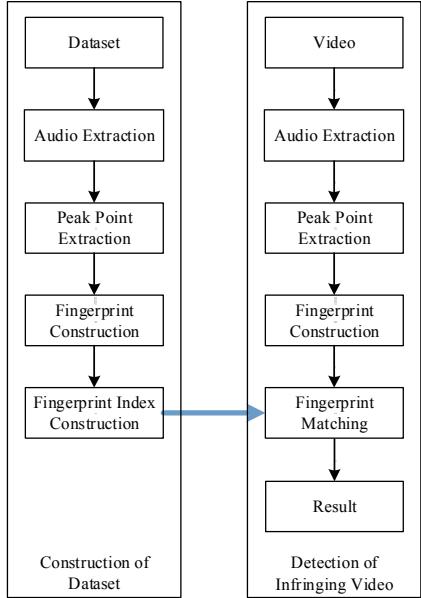


Figure 2-1 Chart of infringement detection of videos

### 2.1 Audio Signal Extraction

Audio signal in videos is an analog signal whose amplitude changes over time. It's a carrier of information transmission and a physical quantity with energy properties. Although the audio signals in different videos are expressed differently, the first step in processing is to convert audio analog signals into digital signals. Because the most sensitive frequency of human ears is below 4kHz, according to the Nyquist theorem, the system uses a sample rate of 8kHz to meet the minimum sampling rate requirements. At the same time, converting audio data to mono PCM encoded WAV formats does not largely lose the important characteristics of sound.

### 2.2 Audio Fingerprint Extraction

After extracted, the audio signal is converted to a fixed point in time and the corresponding amplitude, which is a time discrete signal. When analyzing signals by using the method of signal processing, the signal simply needs to be divided into stable regions. In the process of processing, the method of overlapping framing is adopted in order to ensure

the smooth transition of adjacent frames and the continuity of the signal. Generally, the framing is achieved by limited length of the window, that is, by multiplying the window function by the signal function to form the window-adding audio signal. The system uses the Hanming window as a window function, the frame length value is 512 points, the frame displacement is the half of the frame length.

By doing fourier transformation of the window-adding signal, the spectrogram of the signal is obtained in the frequency domain. Based on the method of extracting peak points of Shazam algorithm, the energy-stable amplitude points are extracted from the spectrogram. In view of the masking effect of audio, the system uses the peak point extraction method based on auditory mechanism. Due to the instability of the beginning and end of the audio, the first and last ten frames of the audio are not processed. The detailed process of peak point extraction is as follows:

(1) Because the curve in the masking effect is similar to the Gaussian function curve, the masking effect is implemented by Gaussian function.

$$G(x) = \begin{cases} e^{-\frac{x^2}{2\sigma^2}}, & -4\sigma \leq x \leq 4\sigma \\ 0, & \text{other} \end{cases} \quad (2-1)$$

In Formula 2-1,  $G(x)$  stands for Gaussian function,  $\sigma$  stands for the standard deviation of Gauss curve. The larger the standard deviation  $\sigma$ , the wider the masking curve, the more hidden peak points, the fewer peak points we get, and the more we get conversely. Because the influence of each other is weak when the peak points are faraway, the range of the masking effect is fixed between  $-4\sigma$  and  $4\sigma$ .

(2) Suppose that the masking threshold of different points in the same frame is initialized to 0, with each frequency point as the center, and the amplitude of the other frequency points is calculated with the Gaussian function value centered on that point, updating the threshold for each point. Do this for all frames in the audio, in turn, to record the threshold for each frame after the update.

(3) If the energy value of a point in the same frame has the same data as the threshold of that point, the point is not affected by the other points, and the points are extracted as candidate points. These candidate points are compared to the thresholds of the previous frame and the latter frame respectively, and are peak points if they are greater than the threshold at the same time. During the extraction of peak points, due to the effect of attenuation, the attenuation factor  $\alpha$  can be used appropriately to reduce the value of each frame threshold curve.

In the process of selecting peak points, by setting the maximum number of PPNs (Peak Point Number, PPN) per frame peak point, the value of standard deviation  $\sigma$  of the Gaussian curve and the value of decay factors  $\alpha$ , the number of the final peak points is determined.

This system utilizes the triple group  $(f_1, f_2, \Delta t)$  in Shazam algorithm method to construct fingerprints. However, in practice, audio will be compressed and stretched on the timeline, resulting in a low correlation between the two peak points with a long time distance. Therefore, the system takes time and frequency as the sorting standard, takes a certain number of peak points after the peak point as the peak point of the target region, matches the peak point with the peak point of the target area, and constructs the fingerprint feature. Considering the accuracy of the algorithm and the computational complexity, the system sets the number of regional peak points to 15.

### 2.3 Audio Fingerprint Retrieval

The system uses HashMap to convert the audio fingerprint information to the key value of the hash by hash function. The calculation is as Formula 2-2 shown below.

$$key = f_1 * 2^{16} + \Delta f * 2^7 + \Delta t \quad (2-2)$$

In this formula, since the maximum value of the frequency is 255 and the maximum time difference is 100, the range of time difference, frequency difference, and frequency is set to 0 to 7 bits, 8 to 16 bits, and 9 to 24 bits in the binary, respectively, the uniqueness of the key value ensured. The value of keyword (*key*) obtained by hash function, and the value of *value* stored by the ID of audio, the frequency of the peak point, and time information. The corresponding relationship is shown in Formula 2-3.

$$key(f_1, \Delta f, \Delta t) \rightarrow value(num, songID, freq, offsetTime) \quad (2-3)$$

In order to retrieve the audio fingerprint more effectively, the system constructs the fingerprint dataset by means of inverted indexing, as shown in Figure 2-2.

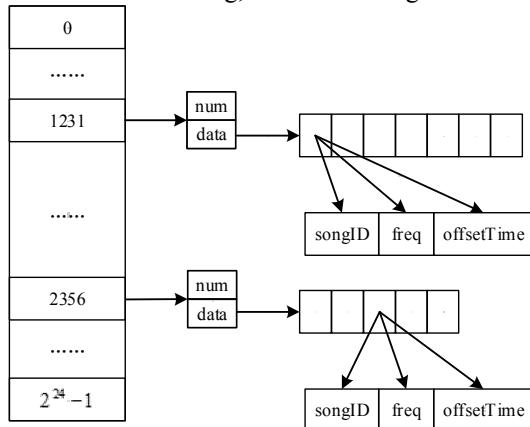


Figure 2-2 Structure of fingerprint hash table index

As shown in Figure 2-2, each value in the hash table index points to a structure that contains variables *num* and pointers *data*. Variable *num* represents the number of peak points to which the current index value points, pointer *data* points to the ID, frequency and time of all the peak points that make up the current index value. According to the definition of the hash formula, the maximum value set for the index value is  $2^{24} - 1$  to ensure that each fingerprint has a

unique corresponding *key* value.

In the process of fingerprint matching, the conventional method is to get the best audio ID and time information based on the number of fingerprint features, but because audio is often affected by external interference and other factors, resulting in some changes in both the peak point of the audio to be detected and the peak point in the corresponding original audio. Therefore, in view of the phenomenon of irregular stretching and compression of audio on the timeline and frequency axis, the system has redesigned the matching algorithm on the basis of a single peak point match, as follows:

(1) The values of frequency, frequency difference and time difference between two peak points in the audio fingerprint are set to a certain range. According to the experimental verification, the fingerprint features match best when the floating range of frequency and time difference is  $[-3, +3]$  and the floating range of frequency difference is  $[-15, +15]$  respectively. In the process of matching, when the number of peak point matches in the dataset is greater than a certain threshold, the peak point is judged to be the corresponding candidate peak point. The specific schematics are shown in Figures 2-3.

|                                       |
|---------------------------------------|
| pi: 2226, pj:205, pi-i: 55, pj-j: 14  |
| pi: 2229, pj:194, pi-i: 56, pj-j: 3   |
| pi: 2230, pj:212, pi-i: 57, pj-j: 21  |
| pi: 2234, pj:21, pi-i: 61, pj-j: 170  |
| counts: 15                            |
| frame: 2184, freq:21                  |
| pi: 2187, pj:123, pi-i: 3, pj-j: 102  |
| pi: 2197, pj:169, pi-i: 13, pj-j: 148 |
| pi: 2197, pj:219, pi-i: 13, pj-j: 198 |
| pi: 2199, pj:21, pi-i: 15, pj-j: 0    |
| pi: 2205, pj:144, pi-i: 21, pj-j: 123 |
| pi: 2205, pj:189, pi-i: 21, pj-j: 168 |
| pi: 2205, pj:224, pi-i: 21, pj-j: 203 |
| pi: 2226, pj:21, pi-i: 42, pj-j: 0    |
| pi: 2226, pj:224, pi-i: 42, pj-j: 203 |
| pi: 2228, pj:90, pi-i: 44, pj-j: 69   |
| pi: 2228, pj:205, pi-i: 44, pj-j: 184 |
| pi: 2229, pj:194, pi-i: 45, pj-j: 173 |
| pi: 2230, pj:212, pi-i: 46, pj-j: 191 |
| pi: 2234, pj:21, pi-i: 50, pj-j: 0    |
| pi: 2242, pj:142, pi-i: 58, pj-j: 121 |
| counts: 15                            |
| frame: 2187, freq:123                 |
| pi: 2197, pj:169, pi-i: 10, pj-j: 46  |
| pi: 2197, pj:219, pi-i: 10, pj-j: 96  |
| pi: 2199, pj:21, pi-i: 12, pj-j: 102  |
| counts: 15                            |
| frame: 2335, freq: 163                |
| pi: 2335, pj:180, pi-i: 0, pj-j: 17   |

Figure 2-3 Structure of fingerprint hash table index

(2) After peak point matching, if there is a certain number of peak points to be detected in the audio match the peak point of one audio in the audio database, the audio to be checked has some correlation with its corresponding audio. The calculation of peak point is shown in Formula 2-4.

$$Peak_{matched} = Peak_{total} \times Thre_{perNum} \quad (2-4)$$

In the formula 2-4, *Peak<sub>matched</sub>* stands for the number of peak points matched, *Peak<sub>total</sub>* stands for the total number of peak points in the audio to be detected, *Thre<sub>perNum</sub>* stands for a percentage threshold parameter, we can determine whether the work to be detected is infringing by setting the size of the parameter threshold.

### III. EXPERIMENTS

In order to make a fair and reasonable comparison of the algorithm, the system conducts experimental evaluation of the system method and Shazam method in accordance with the common evaluation criteria by using the datasets provided by two technology companies. The datasets include sample set and test set, the data in the sample set is copyrighted by law, the data in the test set contains infringing videos related to the sample set and videos not related to the sample set.

The sample of dataset No.1 consists of 50 videos in mp4 format, which account for 17.3G, as shown in Figure 3-1. The test set consists of 256 videos in the format of mp4, wmv and ts, which account for 7.53G, as shown in the Figure 3-2. Among them, there are 156 infringing samples and 100 unrelated videos. Among the 156 infringement samples, 128 were one-to-one infringements and 28 were one-to-many infringements.



Figure 3-1 The samples of dataset No.1



Figure 3-2 The test samples of dataset No.1

The detection results of the system method and Shazam method of dataset No.1 are shown in Table 3-1 and Table 3-2. From the analysis of data shown in Table 3-1, it can be seen that for 128 one-on-one infringing videos and 100 non-infringing videos, the system method and Shazam method have a good effect on the detection of infringing samples and non-infringement samples, and the difference between the recall rate and accuracy of the samples is small. From the analysis of data shown in Table 3-2, it can be seen that for the detection of 28 one-to-many infringement videos, the system method completed a one-to-many match for 26 samples. Although 2 infringement samples did not match the corresponding videos entirely, they matched part of the corresponding sample set. In the meanwhile, Shazam

method effectively completed a one-to-one match for all the 28 samples. From the analysis of the test results, it is shown that the system method has advantages to one-to-one and one-to-many types of infringing videos.

Table 3-1 Performance of different methods evaluated on the dataset NO.1(1)

| method     | Property         | Infringement | Non-infringement |
|------------|------------------|--------------|------------------|
| Our method | infringement     | 128          | 0                |
|            | Non-infringement | 0            | 100              |
| Shazam     | infringement     | 127          | 1                |
|            | Non-infringement | 0            | 100              |

Table 3-2 Performance of different methods evaluated on the dataset NO.1(2)

| method     | Property     | One-to-many | One-to-one |
|------------|--------------|-------------|------------|
| Our method | infringement | 26          | 2          |
| Shazam     | infringement | 0           | 28         |

The sample of dataset No.2 consists of 3052 videos in mp4 format, including 123 TV plays and movies, occupying size of 239G, as shown in Figure 3-3. A total of 7981 videos in mp4 formats are in the test set, which take up 76.5G, as shown in Figure 3-4. Among them, the test set has 4460 infringing videos and 3521 unrelated videos, all the infringing videos are one-to-one corresponding to a specific video in the sample set.



Figure 3-3 The samples of dataset No.2



Figure 3-4 The test samples of dataset No.2

The detection results of the system method and Shazam method of dataset No.2 are shown in Table 3-3. From the analysis of data, it can be seen that for 4460 infringing

videos and 3521 unrelated non-infringing videos detection, the system method can correctly detect 3403 infringing samples and 2805 non-infringement samples, Shazam method can correctly detect 2806 infringing samples and 2893 non-infringement samples. By analyzing the data of recall, although the detection of non-infringement samples of the system method is slightly worse than the Shazam method, but the advantages of the detection of infringing videos are obvious. From the analysis of the test results, the system method has a good detection effect for some audio which stretches or compresses in time axis and frequency axis.

Table 3-3 Performance of different methods evaluated on the dataset NO.2

| method     | Property         | infringement | Non-infringement |
|------------|------------------|--------------|------------------|
| Our method | infringement     | 3403         | 1057             |
|            | Non-infringement | 716          | 2805             |
| Shazam     | infringement     | 2806         | 1654             |
|            | Non-infringement | 628          | 2893             |

#### IV. CONCLUSION

As per the analysis of the results of the two different types of dataset based on system method and Shazam method, the system method has better results for the detection of videos with one video corresponding to multiple videos at the same time and videos with irregular stretching, compression and other changes caused by external factors. The system method could match the sample data successfully and improve the efficiency of infringement detection of videos effectively, however, it still needs to be further improved on the detection of the audios in videos with large changes.

#### ACKNOWLEDGMENT

This work is supported by the China Postdoctoral Science Foundation (2018M641524) and the Science and Technology Program of Beijing (Z201100001820002). It was also the research achievement of the Key Laboratory of Digital Rights Services, which is one of the National Science and Standardization Key Labs for Press and Publication Industry.

#### REFERENCES

- [1] Shazam Music Recognition Service [EB/OL].
- [2] SoundHound [EB/OL]. <http://www.soundhound.com/>.
- [3] Haitsma J, Kalker T. A Highly Robust Audio Fingerprinting System[C]. Ismir. 2002, 2002: 107-115.
- [4] Wang A. The Shazam music recognition service[J]. Communications of the ACM, 2006, 49(8): 44-48.
- [5] Qiang Wang. A Study of Content Based Massive Music Retrieval Technology[D]. Beijing University of Posts and Telecommunications, 2013.
- [6] Siyuan Wu. A Multi Modeal Contend-based Copy Detection Approach[D]. Beijing University of Posts and Telecommunications, 2013.
- [7] Qiming He. Research on Indexing Method of Audio Sample Retrieval[D]. Harbin Institute of Technology, 2013.
- [8] Jie Guo, Zhiyu Wang. An improved algorithm of audio fingerprinting used in a fast music information retrieval system information retrieval system[C]. Technical Acoustics, 2007, 129-130.
- [9] Coover B, Han J. A power mask based audio fingerprint[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014: 1394-1398.
- [10] Wu L M, Han W, Zhou S, et al. A compressed-domain audio fingerprint algorithm for resisting linear speed change[J]. Computer Modeling & New Technologies, 2014, 18(10):192-196.
- [11] Wu L M, Han W, Li Y F, et al. A Resilient Novel Compressed-domain Audio Recognition Method for Anti-Linear Speed Change[C]. Key Engineering Materials. Trans Tech Publications Ltd, 2014, 620: 613-618.
- [12] Ouali C, Durmouchel P, Gupta V. A robust audio fingerprinting method for content-based copy detection[C]. 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI). IEEE, 2014: 1-6.
- [13] Ning Sun, Weiping Zhao, et al. An Improved Algorithm of Philips Audio Fingerprint Retrieval[J]. Computer Engineering, 044(001): 280-284.
- [14] Shanshan Yao. Research on the Key Technology of Big Audio Retrieval[D]. Taiyuan University of Technology, 2018.
- [15] Jie T, Gang L, Jun G. Improved algorithms of music information retrieval based on audio fingerprint[C]. 2009 Third International Symposium on Intelligent Information Technology Application Workshops. IEEE, 2009: 367-371.
- [16] Jiang T, Xiang K, Lu J, et al. A Large Scale Audio Fingerprinting System[C]. Pacific-rim Conference on Multimedia. Springer International Publishing, 2013.
- [17] Duong N Q K, Thudor F. Movie synchronization by audio landmark matching[C]. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 3632-3636.
- [18] Malekesmaeli M, Ward R K. A local fingerprinting approach for audio copy detection[J]. Signal Processing, 2014, 98: 308-321.
- [19] Jun Hu. Research on audio technology of query by example based on shazam algorithm[D]. Chongqing University of Posts and Telecommunications, 2018.
- [20] Yong Zhang. The research of segmented audio retrieval algorithm base on audio fingerprint[D]. Hunan University, 2017.