SPECTRO-TEMPORAL LANDMARKING WITH RANK-ORDERED LOCAL MAXIMA FOR AUDIO FINGERPRINTING

Toby Stokes
BAFTA Research
195 Piccadilly
London, UK
TobyS@bafta.org

ABSTRACT

Audio fingerprinting presents a way of searching a large body of media for a specific audio sequence in a more efficient manner than searching on a sample-by-sample basis. This paper presents an algorithm based on hashing landmarks in the time-frequency domain for use in the 2015 MIREX Audio Fingerprinting task. Previous work in the literature is augmented with the addition of a rank-ordering of maxima in a given time window to provide extra variability to the hashed data. An initial assesment of the accuracy of the system is made using the public part of the MIREX query set.

1. INTRODUCTION

Audio fingerprinting is an efficient method for searching a large audio database. Applications of audio fingerprinting include content identification, de-duplication and detection of copyright infringement. The most common technique is the hashing of spectro-temporal landmarks as originally documented by Avery Wang [2].

This paper presents a spectro-temporal landmarking approach to audio fingerprinting with the addition of a rank-ordering of local maxima that is being submitted to MIREX 2015 for the Audio Fingerprinting task [3]. This system is named STELLAR (Spectro-TEmporaL LAndmarking with Rank ordering).This task involves databasing ten thousand songs into a database of no more than 2GB taking no more than 24 hours. Having created a database, the system should identify approximately 6000 noisy queries within 24 hours.

The remainder of this paper is structured as follows: Section 2 details the system's database creation method; Section 3 describes the process of querying the database; Section 4 gives a summary of the expected performance of the system based on the MIREX pre-release dataset; and, Section 5 concludes the paper .

2. DATABASE CREATION

The database creation method presented here comprises four stages of processing to generate a fingerprint for each file:

1. spectrogram calculation via a short-time Fourier transform with logarithmically-spaced frequency values;
2. identification of the location and rank order of local maxima in each time window of the spectrogram;
3. landmark creation based on the relationship between principal peaks and subsequent peaks; and
4. hashing of the landmarks into 20-bit unsigned integers, which collectively make up the fingerprint of each file.

The following subsections describe each of these steps in turn.

2.1 Spectro-temporal transform

The spectrogram used is produced from audio resampled to 16 kHz to allow fingerprinting of audio below 8 kHz. The spectrogram is computed using a 256-point fast Fourier transform with logarithmic frequency spacing. A hamming window is applied with a 50% overlap.

2.2 Maxima identification and ranking

In each time window, local maxima are identified. At time m, the n^{th} frequency bin, $f[m,n]$, is considered a local maxima if it satisfies

$$f[m,n-1] < f[m,n] > f[m,n+1]. \qquad (1)$$

Note this definition does not use time data in identifying maxima.

2.3 Landmark creation

Landmarks are used to describe the relationship between two points. For each maxima, landmarks are created using future maxima that fall within a target zone. Future points are chosen from a target area. This area begins 16 windows beyond the anchor point and is 16 windows wide. The target area is 16 frequency bins wide. The features used in

describing the landmarks are the time and frequency location of the initial point and the respective differences in time, frequency and rank.

2.4 Landmark hashing

To allow an efficient search, the landmark features are hashed into a 20-bit integer according to the scheme in shown in Table 1. The time difference was allocated to the most significant bits as it has been observed to produce the highest entropy. The hashed landmarks are stored alongside a timestamp and the file that they are related to.

Bit Numbers	Landmark Features
20–17	Time difference
16–9	Start frequency
8–5	Frequency difference
4–1	Rank difference

Table 1. The bit allocations in the 20-bit hashed landmark.

2.5 Controlling the data rate

To control the data rate produced by each file and thus the size of the resultant database in memory two steps are used. Firstly, the number of maxima used for the process is limited to the eight most prominent in each time window. Secondly, the number of landmarks is limited to 10,000 per file. These are selected from uniform spacing throughout the list of landmarks generated.

3. DATABASE QUERYING

A query of the database should return the identified audio in as little time as possible. To query the database, the query audio is fingerprinted in the same manner as described for the database creation process.

A match is determined not by the number of matched landmarks but by finding a series of consecutive hits. The landmarks of the query are then compared with the landmarks stored in the database and when the landmarks are equal, a hit, the timestamps of both the database landmark and the query landmark are recorded.

The obtained fingerprint is then used to search as follows: firstly, landmark hashes are compared; secondly, the timestamps of the matching landmark hashes are compared and the modal time difference calculated for each file. The file with the highest frequency of its modal time offset is considered to be the match.

Currently, the system does not measure the significance of the most frequently occurring modal offset meaning the system is not yet able to reject out of vocabulary queries.

4. EXPECTED PERFORMANCE

Initial performance measurements have been created using the GTZAN genre dataset [1]. This corpus provides one thousand thirty-second music clips. This dataset has been used to test the system under both noisy and noiseless conditions.

Noiseless queries were generated by selecting random five-second excerpts from each track to query the database. Using one noiseless query per database entry, accuracy was measured to be around 95%.

Using the public MIREX query set, which comprises 1062 ten-second noisy recordings taken from the database and recorded in noisy environments using mobile phones, an accuracy of 65% was observed.

5. CONCLUSION

An audio fingerprinting system has been created using a rank-ordered spectro-temporal landmarking. The system allows a database to be both created and queried. Preliminary investigations suggest the system is 95% accurate under noiseless conditions and around 65% under noisy conditions. The system will be evaluated in the MIREX 2015 Audio Fingerprinting task.

6. REFERENCES

[1] George Tzanetakis. GTZAN genre collection. http://marsyasweb.appspot.com/download/data_sets/, 2002.

[2] Avery Wang. An industrial-strength audio search algorithm. In *Proceedings of the 4 th International Conference on Music Information Retrieval*, 2003.

[3] Chung-Che Wang. Mirex 2015 audio fingerprinting challenge. http://www.music-ir.org/mirex/wiki/2015:Audio_Fingerprinting, 2015.