MIREX 2017 Audio Fingerprinting System

Peng Li
NetEase
lip0620@gmail.com

Guanglong Hu
NetEase
hgllgh007@163.com

Junyan Jiang
NetEase
at2jjy@gmail.com

ABSTRACT

This document describes our submission to audio fingerprinting task of MIREX 2017. Our approach is based on robust landmark detection in spectrogram and efficient matching strategy. It follows the basic idea introduced by Wang [1] and some modifications have been made to improve system robustness, especially in noisy environments. To speed up the matching process, we adopt two methods. One is to simply prune those songs which are unlikely to match the input query. The other one is to do matching simultaneously in index replicas, with each replica covering a part of the entire music database.

1. INTRODUCTION

Audio fingerprinting has been widely used in various applications such as music retrieval for end users and detection of copyright infringement for music companies, etc. The main idea behind it is to extract the so called "fingerprints" from an audio segment. These fingerprints can be considered as a set of identifiers of this segment which are highly distinguishable, even when corrupted by reverberation and noise, either stationary or transient (e.g., recording the sound played by a speaker 5m away in a noisy supermarket). Thus matching becomes possible if some fingerprints of the query segment overlaps with those of a song in the music database.

The main challenge in developing a successful audio fingerprint system is how to extract robust fingerprints and how to perform real time matching in a music database with millions of songs. In [1], Wang proposed a method to detect the landmarks in spectrogram in which a landmark was defined as a local maxima. Then these landmarks are made as pairs according to a predefined scheme and hashed afterwards. As a result, each pair corresponds to a hash value, along with a time stamp of the first landmark of this pair. With this idea, it is possible to perform fast matching since searching could be very efficient with hashing. It has also been shown that this kind of fingerprinting is robust to noise. Therefore, we followed this approach and have made some modifications in order to get better performance. In the rest of this document, we will describe our attempts.

2. FINGERPRINTS EXTRACTION

Audio segments (songs in db or query recordings) are firstly re-sampled to make sure the sampling rate are the same for all the segments. Usually 8khz is enough for audio fingerprinting but one can use higher sampling rate at the expense of higher computational cost.

To extract fingerprints, we convert the re-sampled audio data to time-frequency domain with STFT. Then, following Wang's idea, we detect the local maxima in spectrogram. a local maxima is detected if its value is higher than any other values in a small region. When doing this, we divide the entire frequency bands into several groups. Each group has its own region shape. Thus the density of landmarks can be different for different frequencies by simply adjusting region shapes. Besides, we removed some low frequencies from spectrogram since we have found that the microphones of some types of smart phones have high response at low frequencies in its recordings, which has adverse impact on landmark detection.

Region shapes are also adjusted by the local energy of the segments. For example, when detecting landmarks at the *j*th frame, we calculate the energy of the time range [*j*-*k*, *j*+*k*], where *k* is the context window size. If this energy value is small, we use a smaller region shape in order to extract more maxima. According to our preliminary experiments, this is helpful for the matching of the beginnings and endings of songs since both parts usually have low energies.

3. MATCHING STRATEGY

When dealing with a music database of millions of songs, one has to take computation capacity into account. It is rather wasting if each song in db is matched carefully with input query since most songs have nothing in common with input query even though they do have a few fingerprints that are also included in the query. Therefore, we prune the songs which share few fingerprints with input query before checking the time difference (see [1][2] for more details on the time difference used in matching), which greatly reduces computational cost. This idea is also adopted by Ellis in [2].

Secondly, to speed up matching process, we divide the entire music database into *K* disjoint sets and build one matching index for each set. When a query is received, matching is done simultaneously in each index replica. Then the match results of all replicas are collected and sorted accordingly.

4. DISCUSSION

In this document we present our submission to audio fingerprinting task of MIREX 2017. We have built a system that is capable of extracting robust fingerprints from audio segments. The main idea is to adaptively adjust the region shapes when detecting local maxima in spectrogram. Preliminary experimental results on the publicly accessible recordings of George's music genre dataset (from MIREX website) have shown promising performance.

5. REFERENCES

[1] A. Wang: “An industrial strength audio search algorithm,” *Proceedings of the International Symposium on Music Information Retrieval*, 2003.

[2] D. Ellis (2009), "Robust Landmark-Based Audio Fingerprinting", web resource, available: http://labrosa.ee.columbia.edu/matlab/fingerprint/