Signal Processing 98 (2014) 308–321

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

A local fingerprinting approach for audio copy detection

Mani Malekesmaeili*, Rabab K. Ward

Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada

ARTICLE INFO

Article history:
Received 1 June 2013
Accepted 21 November 2013
Available online 1 December 2013

Keywords:
Audio copy detection
Music retrieval
Audio feature extraction
Pattern recognition
Dictionary learning
Linear regression
Multimedia forensics
Multimedia copy detection

ABSTRACT

This study proposes an audio copy detection system that is robust to various attacks. These include the severe pitch shift and tempo change attacks which existing systems fail to detect. First, we propose a novel two dimensional representation for audio signals called the time-chroma image. This image is based on a modification of the concept of chroma in the music literature and is shown to achieve better performance in song identification. Then, we propose a novel fingerprinting algorithm that extracts local fingerprints from the time-chroma image. The proposed local fingerprinting algorithm is invariant to time/frequency scale changes in audio signals. It also outperforms existing methods like SIFT to a great extent. Finally, we introduce a song identification algorithm that uses the proposed fingerprints. The resulting copy detection system is shown to significantly outperform existing methods. Besides being able to detect whether a song (or a part of it) has been copied, the proposed system can accurately estimate the amount of pitch shift and/or tempo change that were applied to a song.

1. Introduction

Audio copy detection has become a building block of many multimedia sharing websites. Whether users are searching for their favourite songs or whether content providers are looking for illegally distributed copies of their copyrighted songs, an audio copy detection system should be able to retrieve the sought-after songs. This is true even when the songs have been subjected to some content preserving modifications. Such modifications include compression, noise addition, frequency changes, and time scale changes. Compression effects arise from the use of different quantization levels and other processes involved in audio compression. Noise may be due to background voices recorded through a microphone, faulty song copying systems, etc. The most common frequency modification is pitch shift, where the pitches of a song are moved up or down without affecting the time characteristics of the signal. Pitch shift results from a change in the scale of the frequency axis (frequencies are shifted in the octave space). For example, with a 100% pitch shift, an audio signal is played at one octave higher, with the same speed. Time scale attacks include speed change and tempo change. To change the tempo of a song, its pace (time scale) is increased or decreased but its frequency content is not changed. In other words, the song is scaled along the time axis without any change in its pitches. When the speed of a song is modified both its tempo and its pitch are changed. Mash-up is another class of modifications where short snippets of different songs are put together as a single musical signal.

Most of the proposed solutions for audio copy detection are based on content-based fingerprinting [1,2]. Audio, image, and video fingerprinting have recently become a popular research topic due to the great interest in copy detection by industry. Audio fingerprints refer to compact signatures (extracted from an audio signal) that can distinguish between different songs based on their musical content. Traditionally the proposed audio fingerprinting methods have mainly focused on developing

* Corresponding author. Tel.: +1 6047108202.
E-mail addresses: manim@ece.ubc.ca (M. Malekesmaeili), rababw@ece.ubc.ca (R.K. Ward).

0165-1684/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2013.11.023

algorithms that are robust to attacks (effects) such as noise, compression, equalization, and echo [3]. However, due to the availability of powerful audio editing softwares, copy detection has become a much more sophisticated task extending far beyond these attacks [4]. Some of the emerging attacks are pitch shift and tempo change.

A practical audio fingerprinting algorithm should be robust to all content-preserving modifications including noise, compression effects, pitch shifts, and tempo changes. Moreover, to be able to deal with song mash-ups and other demands involving short snippets of audio, an audio fingerprinting algorithm should be based on *local* features (fingerprints) of the audio signal. Local features are signatures that represent a short segment of an audio signal (regardless of the rest of the signal). Many papers have proposed powerful and effective image and video fingerprinting algorithms based on local features. However, few papers have studied the use of local features for audio signals.

This paper proposes a local feature extraction algorithm for audio signals and shows how local features can be helpful in the task of audio copy detection. The proposed algorithm is based on a new time–frequency representation of audio signals which we call the time-chroma representation. Chroma, also known as the pitch class profile (PCP), was originally proposed in [5]. It is the set of all pitches that are perceived by the human ear to have similar musical notes. In terms of frequency, a chroma is a set of frequencies that are apart by one or more octaves. A chroma set is usually represented by a value such as its main pitch or the sum of the energy of the pitches it includes. It follows from the definition of the chroma that shifting the pitches of an audio signal by multiples of an octave does not change the chroma values of the signal. In other words, chroma is a single octave representation of the frequency (pitch) content of an audio signal. The time-chroma image is a two dimensional representation of an audio signal that shows its chroma values at different time instances. Shifting the pitch of the audio by one or more octaves has no effect on the time-chroma image. Pitch shift by a fraction of an octave circularly shifts the time-chroma image along the chroma axis. If the tempo of the audio signal is changed, its time-chroma image scales accordingly along the time axis.

Other commonly studied audio representations such as the Bark scale or the Mel scale are non-linearly distorted by a pitch shift attack. In other words, the original representation cannot be retrieved from the distorted one, while for the time-chroma image a simple shift can generate the original representation. This makes time-chroma a suitable platform for designing pitch invariant audio detection algorithms.

In this paper, we introduce an algorithm that extracts time-scale invariant local fingerprints from the time-chroma image. The invariance to time-scale characteristic and the fact that they are extracted locally make such fingerprints robust against tempo change and pitch shift attacks. The locality characteristic of the fingerprints also enables us to detect short snippets of a song mashed up with other songs. This paper is organized as follows. In Section 2, some of the related works are reviewed. In Section 3, a local audio fingerprinting method is proposed. The robustness of the proposed method is evaluated and compared to the state-of-the-art. In Section 4, a novel audio copy detection system that is based on the proposed local audio fingerprints is presented. This system can precisely locate the query audio in a given database of songs. The system can further estimate the amount of pitch shift as well as the amount of tempo change that might have affected a copied song. The proposed copy detection system is also evaluated through a series of severe audio attacks. We conclude the paper in Section 5.

2. Related work

Probably the most well-known publicly available audio fingerprinting algorithm is Shazam [6]. Shazam is based on local audio fingerprints. With Shazam, people can find the song they are looking for, using their smart phones. Shazam uses the peaks (maxima) observed in the spectrogram of an audio signal as the local feature points of a song. Feature descriptors (fingerprints) are then generated from the attributes of pairs of these points. The frequency of every point in each pair as well as their time difference form a compact fingerprint for each pair. The extracted fingerprints are shown to be highly robust to audio compression, foreground noises, and other types of noise. However, they are not robust to tempo changes or pitch shifts.

An audio copy detection algorithm based on global fingerprints (extracted from the spectrogram of the audio signal) has been proposed by Haitsma and Kalker [7]. Because the human auditory system (HAS) approximately operates on logarithmic bands, in [7], the frequency axis of the spectrogram is transformed to a subjectively developed scale called the Bark scale. The signs of the energy differences between two adjacent bands (in time and frequency) generate the fingerprints. The proposed fingerprints are proven to be very robust to compression, but as in [6] they are not robust to large tempo changes and pitch shifts and can only tolerate very small amounts of such modifications (up to around 4%). The generated fingerprints are also very long (8 kbits).

Two time-based fingerprinting algorithms are proposed by Özer et al. [8]. Both algorithms derive a periodicity score from a given audio stream to generate fingerprints. They also propose a spectrogram-based fingerprinting algorithm, using the Mel-frequency cepstral coefficients (MFCCs). The time-based periodicity score is shown to have promising results for speech signals. However, as shown in their paper, for music signals the MFCC based approach outperforms the periodicity based approaches. Their proposed algorithms were tested in the presence of noise as well as very small pitch (up to 2%) and time scale (up to 6%) changes.

A wavelet based audio fingerprinting algorithm called Waveprint is proposed by Baluja et al. [9]. Waveprint divides a logarithmically scaled spectrogram (based on the Bark scale) into smaller spectral images along time. The top *t*-wavelet coefficients of such images are then binary embedded to generate intermediate fingerprints. The final fingerprints are the Min-Hash [10] values of these