>Paper ID <

Sub-fingerprint masking for a robust audio fingerprinting system in a real-noise environment for portable consumer devices

Wooram Son[1)], Hyun-Tae Cho[2)] and Kyoungro Yoon[2)*]
[1)] Program in Radiation Applied Life Science, Seoul National University, Korea
[2)] Department of Computer Science and Engineering, Konkuk University, Korea

Abstract

The robustness of audio fingerprinting system in a noisy environment is a principal challenge in the area of content-based audio retrieval, especially for use in portable consumer devices. The audio fingerprint proposed by Philips uses expanded hash table lookup to compensate errors introduced by noise. The expanded hash table lookup increases the searching complexity by a factor of 33 times the degree of expansion defined by the hamming distance. We propose a new method to improve noise robustness of audio fingerprinting in a noisy environment using predominant pitch which reduces the bit error of created sub-fingerprint.

I. INTRODUCTION

Recently, content-based music information retrieval attracts interests as a new addition to the wired/wireless communication service or consumer electronics devices such as the Shazam service for iPhone and other mobile devices. For example, a user may be interested in finding information of a music based on only a small fragment of the overall tune. For instance, it has been reported that there already are available services not only providing information on songs being played over public loudspeakers, but also monitoring broadcast for advertisement tracking and reporting. [1, 2] An audio fingerprinting scheme that has proved known to be the most robust is the so-called Philips scheme proposed by Haitsma and Kalker. [3] This scheme uses the energies of 33 logarithmically scaled subbands to obtain their sub-fingerprint which is the sign of the energy band differences (both in the time and the frequency axis). But in the noisy environment, the sub-fingerprint tends to be distorted. To compensate the error caused by the distortion, the candidate positions for the database lookup are expanded into hash values with a Hamming distance of a one-bit error causing additional 33 times of lookup time for audio identification. [4] Mansoo introduces frequency-temporal filtering for a robust fingerprinting scheme in real-noise environment and shows that frequency filtering improves noise robustness in audio identification. [5] In practice, however, the robustness still needs further improvement to be used in a real environment. In this paper, we propose a novel approach to improve robustness of the audio fingerprinting in noisy environment, using predominant pitch as a high-level semantic attribute in music.

II. THE ALGORITHM

A. System Overview

The proposed audio fingerprinting system is based on the Philips' hashing algorithm. For the robust fingerprint extraction in real-noise environment, we propose to use a mask generated by predominant pitch estimation on each sub-fingerprint of Philips' hashing algorithm. The Philips' hashing algorithm module extracts 32-bit hash values (also called a sub-fingerprint) from each audio frame. Every extracted sub-fingerprint is masked by masking module whose masks are dynamically created based on the result of the predominant pitch estimation algorithm. Every audio file or every segment of an audio file in target database is also represented by a fingerprint block which is a sequence of sub-fingerprints, which is bit-masked using predominant pitch estimation. The search module compares the sequence of bit-masked sub-fingerprints with the target fingerprint block by simply applying bit-wise equality operation. The best-matched result is determined based on the number of matched sub-fingerprints per fingerprint block. In the following sections, each module is described in detail.

B. Philips Hashing Algorithm

A detail of Philips' hashing algorithm is given in [3]. As In our implementation, the audio signal is first sampled at the rate of 44100 Hz and segmented into overlapping frames of 16384 samples, only 512 of which are not overlapped and Hanning windowed. Each frame is then transformed using FFT. The obtained audio spectrum is divided into 33 non-overlapping frequency bands of logarithmic spacing from 300 Hz to 2000Hz. For each frequency band, the energy difference in time is compared with the adjacent band. If the energy difference in one frequency band is greater than the one in next higher frequency band, the corresponding hash bit is set to 1. The hash bit is generated for each frequency band, resulting in 32 bits of sub-fingerprint.

C. Predominant Pitch Estimation

In musical sound, fundamental frequency has harmonic structure in frequency domain according to the characteristics of each sound source. It is said that the recognition of the predominant pitch is the process of perceiving how much those partials have harmonic characteristics. Therefore, we use harmonic enhancement and harmonic summation in spectral domain to estimate predominant pitch. The details of harmonic enhancement and summation algorithms are described in previous reports. [6, 7]

* Author for correspondence. Supported by Seoul R&BD Program # 10581.

D.Sub-fingerprint Masking

One sub-fingerprint only contains spectral energy band differences in time and frequency axis as a cryptographical approach. We have bit-masking step to contain musically meaningful high-level attribute based on the predominant pitch estimation. Philips hashing algorithm extracts 32-bit hash value, then that is bit masked, where a mask is a bit pattern indicating each critical bands with predominant energy of summed harmonics and nearby critical bands. Other bits are set to 0. Only few bits of the hash value mask are set to 1 (we have used five bits for the experiments III).

III. EXPERIMENTS

Experiments were performed using a music database containing 500 popular songs. These songs are selected from Korean and Western popular songs and Classics of various genres such as pop, hip-hop, jazz and classic. All the audio data are stored in PCM format with mono, 16 bit depth and 44.1 kHz sampling rate converted from audio CDs. From these 500 songs, 1500 randomly created audio query clips of three seconds each were captured using a microphone (Sennheiser ME66 model), which was placed 1.5m away from an stereo loudspeaker. With the randomly created 1500 queries, five sets of distinguishable query sets are created by adding noise of different levels, which is acquired from the signal processing society's online database[8]. To compare the performance between algorithms, we implement the following three schemes including proposed algorithm. 1) Our fingerprintng scheme, 2) Mansoo's fingerprinting scheme [5] and 3) Philips fingerprinting scheme. The experiments are performed on a system with windows XP O/S and the followings are the description of the data used:

- Set I contains 1500 query: clips captured by microphone in a quiet environment.
- Set II is created by adding noise data to Set I with SNR of 19.1dB.
- Set III is created by adding noise data to Set I with SNR of 7.4dB.
- Set IV is created by adding noise data to Set I with SNR of 0dB.
- Set V is created by adding noise data to Set I with SNR of -3.5dB.
- Noise Data is acquired from [8], which contains: voice babble of 100 people speaking in a canteen, acquired by recording from 1/2" B&K condenser microphone onto digital audio tape (DAT).

TABLE I
COMPARATIVE PERFORMANCE ANALYSIS OF THREE ALGORITHMS

Data Set	**Our scheme**	Mansoo(H_{F2}+H_{T2})	Philips
Set I	**93.13%**	89.87%	85.33%
Set II	**86.13%**	68.27%	59.13%
Set III	**67.07%**	29.87%	19.67%
Set IV	**44.67%**	8.00%	2.73%
Set V	**31.13%**	3.27%	0.53%

%: RECOGNITION ACCURACY

IV. CONCLUSION

This paper presented a novel audio fingerprinting algorithm to recognize songs in real noisy environments. The proposed algorithm enhances the Philips fingerprint algorithm by creating sub-fingerprint-mask based on the extracted predominant pitch of the audio segment. The proposed algorithm clearly outperforms original Philips algorithm in recognizing polyphonic music in real noisy environment. The hash bit masking based on the predominant pitch has the effect of emphasizing the most audible sound of music as a high-level musically meaningful attribute. Experimental results show that the recognition qualities of the proposed algorithm in very high noise environment, such as Set IV or V are much higher than the original Philips and Mansoo fingerprinting scheme. The enhanced performance due to mask of the extracted predominant pitch suggests that introduction of other musically meaningful high level attribute to the fingerprint may also enhance the performance of the fingerprint system in noisy environment. Enhancement and optimization of the predominant pitch extraction and creating hash mask can also be considered for future work.

V. REFERENCE

[1] A. Sinitsyn, "Duplicate Song Detection using Audio Fingerprinting for Consumer Electronics Devices", *IEEE International Symposium on Consumer Electronics*, 2006,1-6

[2] J. Cerquides, "A Real Time Audio Fingerprinting System for Advertisement Tracking and Reporting in FM Radio", *Radioelektronika, 2007. 17th International Conference*, 2007,1-4

[3] J. Haitsma, and T. Kalker, “A highly robust audio fingerprinting system with an efficient search strategy,” *Journal of New Music Research,* vol. 32, no. 2, pp. 211-221, 2003.

[4] C. Burges, J. Platt, and S. Jana, “Distortion discriminant analysis for audio fingerprinting,” *IEEE Transactions on Speech and Audio Processing,* vol. 11, no. 3, pp. 165-174, 2003.

[5] M. Park, H. Kim, and S. Yang, “Frequency-temporal filtering for a robust audio fingerprinting scheme in real-noise environments,” *ETRI journal,* vol. 28, no. 4, pp. 509-512, 2006.

[6] J. Song, S. Bae, and K. Yoon, "Mid-level music melody representation of polyphonic audio for query-by-humming system", *International Symposium on Music Information Retrieval*, 2002,

[7] J. Song, S. Bae, K. Yoon, "Query by humming: matching humming query to polyphonic audio", *IEEE International Conference on Multimedia and Expo*, 2002,

[8] D. Johnson, “Signal processing and the World Wide Web,” *IEEE Signal Processing Magazine,* vol. 12, no. 5, pp. 53-57, 1995.