Comparison of Algorithms for Audio Fingerprinting

Heinrich A. van Nieuwenhuizen, Willie C. Venter and Leenta M.J. Grobler
School of Electrical, Electronic and Computer Engineering
North-West University, Potchefstroom Campus, South Africa
Email: {20252188; willie.venter; leenta.grobler}@nwu.ac.za
Telephone: (027) 18 299-4058 Fax: (027) 18 299-1977

***Abstract*-The practical implementation and application of audio fingerprinting to recognize specific audio segments are becoming more and more popular these days. Different research groups are working on implementations of audio fingerprinting, but they seldom, compare their algorithms to those of other people working in the field. A fair judgment can therefore not be made of which of the available algorithms is suitable for certain applications. In this paper, two audio fingerprinting algorithms are tested that of Avery Wang's and Haitsma and Kalker's in terms of accuracy, speed, versatility and scalability**

Keywords: Audio Fingerprinting; Automatic Music Recognition; Content-based Audio Identification; Perceptual Hashing; Robust Matching;

I. INTRODUCTION

Fingerprinting systems are not a new concept; it has been around for more than a hundred years. In 1893 Sir Francis Galton was the first to "prove" that no two fingerprints of human beings are alike [1]. This notion was taken further by using any unique feature to identify an object; this includes the iris and even ears. People also realized the potential of constructing fingerprints of audio signals to identify and compare them. This principle is called audio fingerprinting.

Audio fingerprints make use of short audio segments usually between 3-30 seconds in length (depending on the algorithm) to create an audio fingerprint. This audio fingerprint is compared to a database of known audio fingerprints to identify the original audio source. The audio fingerprints of the segments do not necessarily have to be of high quality to be a match. Distortions and interference of the original signal makes matching of the fingerprints less reliable, but (to a certain extent) it will still be recognizable. The distortions and interferences can be compared to a smudged or partial human fingerprint.

In this paper a brief description about audio fingerprinting, a generic code of Avery Wang's Shazam algorithm and that of Haitsma and Kalker's algorithm are discussed [2].

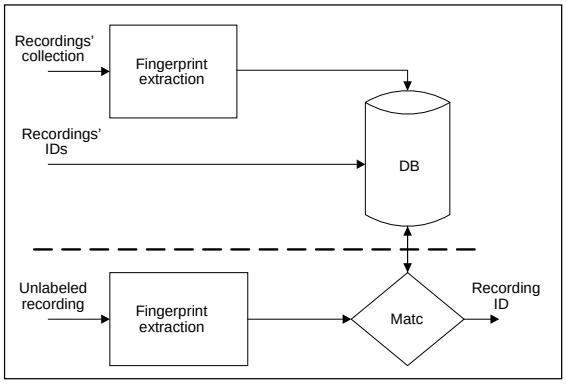

Figure 1 : Content-based audio identification framework [6]

In section II, the three main groups of audio fingerprinting techniques are briefly discussed, after which the operation of Avery Wang's Shazam [2] and Haitsma and Kalker [1] algorithm are briefly discussed. In section III the operation of the algorithms are discussed. In Section IV the validation and verification is presented. The paper is ended with a conclusion and future work in section V.

II. DIFFERENT AUDIO FINGERPRINTING TECHNIQUES

There are several applications for audio fingerprinting algorithms. According to Wes Hatch [3] the biggest benefactor would be the broadcast monitoring industry. Other applications would be playlist generation, royalty collection, program verification and advertisement verification.

According to P.J.O Doets, M. Menor Gisbert and R.L. Lagendijk there are three groups [4] which audio fingerprinting can be classified as:

Group 1: Systems that use features based on multiple subbands, namely Philips' Robust Hash algorithm, reported to be very robust [1] against distortions. Phillips uses Haitsma and Kalker's algorithm.

Group 2: Systems that use features based on a single band such as the spectral domain, namely Avery Wang's Shazam and Fraunhofer's AudioID algorithms.

Group 3: Systems using a combination of subbands or frames, which is optimized through training, namely Microsoft's Robust Audio Recognition Engine (RARE) that uses Hidden Markov Models (HMMs). [5]

This work was completed at the Telkom-Grintek Centre of Excellence
At the NWU, and is funded by the HTBO THRIP project

III. Operation

The Haitsma and Kalker's algorithm propose that a fingerprint extraction scheme should be based on a general streaming approach, taking an audio signal framing it into windows of 370ms length for every 11.6ms giving it an overlapping factor of 31/32 [1].

After which the FFT of every frame is computed filter though band division stored into 32-bit sub Fingerprints.

See **figure 2** below

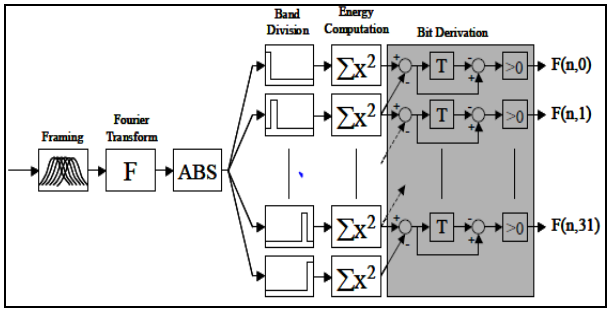

Figure 2 : Haitsma and Kalker's algorithm [1]

An unidentified sample of an audio signal should be 3-30 seconds in length for their algorithm to identify a match.

Avery Wang claims for a database of 20 thousand tracks implemented on a PC, the search time is 5 to 500 milliseconds [2]. As the code is not available, generic code for MATLAB was produced by Dan Ellis [5]. Robert Macrae of C4DM Queen Mary University London altered the code for use in the windows environment.

The latter algorithm proposed to make use of a spectrogram. The spectrogram is the squared magnitude of the STFT (Short-time Fourier Transform).

$$spectrogram(t,\omega) = \left|STFT(t,\omega)\right|^2 \quad (1)$$

Usually the spectrogram is divided into small sizes (typically 512 points) which are called windows or frames. This is the shared basis of group 2. The differences between the fingerprint algorithms in the group typically involve how much the frames overlap, and how the fingerprint is defined in the frame and the storing and searching of the fingerprints. Avery Wang Shazam algorithm uses the energy peaks in the frame and form spectral pair landmarks. The local maxima within a defined section are grouped into pairs [4].

The hash values are computed and compared the entry with the most hits is returned as the match. (Typically more than 9 spectral peaks are considered a match [7].)

IV. Validation and Verification

The two algorithms are implemented. Haitsma and Kalker's results are compared to that of their own "A Highly Robust Audio Fingerprinting System" [1] and Avery Wang's compared to his own "An Industrial-Strength Audio Search Algorithm" [2] and that of Dan Ellis "Robust Landmark-Based Audio Fingerprinting" [5]. Verifying that the algorithms are correctly implemented.

The two algorithms are then compared to each other in terms of accuracy, speed, versatility and scalability. The algorithms are further tested by subjecting the data to noise and compression and the algorithms parameters are tweaked for better results. Comparing the algorithms in the latter terms but with different data ranging from classical music to pop and advertisements.

V. Conclusion and Future Work

The following were observed in the study.

Increasing and decreasing both algorithms frame size will decrease speed but increase the accuracy respectfully.

Defining more peaks in the Shazam's algorithm's frame (normally 5) would also result in better accuracy but decrease speed. Increasing the overlapping regions in both algorithms will increase robustness but decrease speed.

The nearest false positive is on average 5 or 6, which strengthens James .P Ogle and Daniel P.W. Ellis theory that 9 peaks[7] are needed to be a match.

In the future the following should be tried

Translating the algorithms to a faster programing environment. Further research should be done on different techniques to access the database quicker.

More application uses should be investigated e.g. gunshots, engine noise etc.

VI. Bibliography

[1] J. Haitsma and A. Kalker, "A Highly Robust Audio Fingerprinting System," *International Symposium on Music Information Retrieval (ISMIR)*, pp. 107-115, 2002.

[2] A. L.-C. Wang, "An Industrial-Strength Audio Search Algorithm," *ISMIR*, 2003.

[3] W. Hatch, "A Quick Review of Audio Fingerprinting," Mar. 2003.

[4] P. J. O. Doets, M. M. Gisbert, and R. L. Lagendijk, "On the comparison of audio fingerprints for extracting quality parameters of compressed audio," vol. 6072, 2006.

[5] D. P. W. Ellis. (2009) Robust Landmark-Based Audio Fingerprinting. http://labrosa.ee.columbia.edu/matlab/fingerprint/

[6] P. Cano, E. Batlle, E. Gómez, L. de C.T.Gomes, and M. Bonnet, "Audio Fingerprinting: Concepts And Applications," *Studies in Computational Intelligence (SCI)*, no. 2, pp. 233-245, 2005.

[7] J. P. Ogle and D. P. W. Ellis, "Fingerprinting to identify repeated sound events in long-duration personal audio recordings," 2007.

Heinrich van Nieuwenhuizen received his B.Eng degree in 2009 and is currently pursuing his M.Eng at the North West University, Potchefstroom campus. His research interests include software design, audio fingerprinting and implementation and comparison of audio fingerprinting algorithms for industrial use.

This work was completed at the Telkom-Grintek Centre of Excellence
At the NWU, and is funded by the HTBO THRIP project