Review Of ReseaRch
ISSN: 2249-894X
Impact Factor : 5. 7631(UIF)
UGC Approved JoUrnAl no. 48514

AUDIO FINGERPRINTING: REVIEW AND COMPARISON

Kamana Tripathi[1], Mahendra Kanojia[2], Niketa Gandhi[3]
[1]Sheth L.U.J and Sir M.V. College, Mumbai, Maharashtra, India .
Email: kamanatripathi@gmail.com
[2]JJT University, Jhunjhunu, Rajasthan, India .
Email: kgkmahendra@gmail.com
[3]Machine Intelligence Research Labs (MIR Labs), Auburn, WA, USA.
Email: niketa@gmail.com

ABSTRACT

An audio frequency for the average human is characterized as a regular vibration with an audible frequency. Audio fingerprinting is recognizing a piece of audio through its audio sample. Audio fingerprinting includes identifying songs, melodies, tunes, advertisements, sound effects. Audio fingerprinting is latest inclusion in crime forensic and voice navigation. Emotions can also be detected with audio fingerprinting and extract emotional features such as babies crying to detect hunger and sleepiness. Audio fingerprinting for smart phone to provide security. It allows us to identify frequency over time and estimate each frequency and compare frequencies with the known audio fingerprint which is stored in database. This paper provides a conversant review on audio recognition research. It further reports the comprehensive and comparative study of the two most approved algorithms used for audio fingerprinting. The paper concludes with the scope of research in the audio fingerprinting and how it can be used as a robust biometric technique in the security arena.

KEYWORDS — *audio fingerprinting, DFT, Haitsma and Kalker's algorithm, Shazam's algorithm.*

I. INTRODUCTION

Audio signal is an electronic representation of longitudinal sound waves that travel through air and consist of compressions and rarefactions [1]. Audio signals usually comprise of two forms: Analog form and Digital form. The analog form is a smooth wave of energy and is continuous as shown in figure 1. Whereas digital is a wave form that contains binary numbers as shown in figure 2.

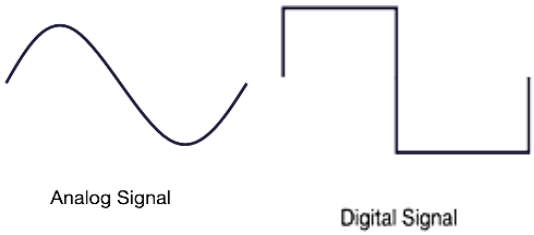

Fig. 1 Analog Signal Fig. 2 Digital Signal

Fingerprinting system is an old perception; they are now everywhere for not less than a hundred years [2]. Fingerprints of a two human being cannot be identical. This is further identified as a unique feature to identify a specific object. Audio Fingerprinting is basically constructing fingerprints from Audio signals samples to identify and even compare it according to one's convenience [2].

To identify the audio in presence of noise, the best solution is removing unique features using audio fingerprinting. The advertisement identification on this paper was inspired by Hatch's Research [2]. The two main algorithms for audio fingerprinting are discussed, which are algorithm of Avery Wang's Shazam and Haitsma and Kalker. Further, their comparative study is performed. Avery Wang's Shazam algorithm was more accurate as proposed by [5]. To find instances of repeated sounds events, Audio fingerprinting are being used. The results were however accurate, but the researcher's future study stated that storing data statically will increase the run time performance [8]. Copyright infringement can be tracked / traced using audio fingerprints. Songs and its cover song can be identified from the features extracted using MPEG-7 fingerprinting. Audio signature type is being represented for fingerprinting of the song. Audio spectrum projection and audio spectrum are being used to represent the fingerprint of a cover song. The result had an accuracy of 100% for a song and 85% for song cover recognition [10].

This research discusses and compares algorithms for audio fingerprinting in various domain and concluding the accurate algorithm for music, security and other domains.

II. LITERATURE REVIEW

Research was carried out by Nieuwenhuizen *et al.,* 2012 to identify advertisements by implementing Shazam's algorithm. The results were impressive as using the algorithm it was possible to identify radio signals that were 16 times faster than real time i.e allowing more data could be analysed with help of larger database. This is a profitable technique for the advertisement companies [2].

Work carried out by Parmar *et al*., 2012, studies voice printing as a tool for crime forensic. Researchers introduced that the spectrographic identification for criminal proceedings raised significant evidence problem. The reception of voiceprints has never been encouraged and that of fingerprints has been encouraged, therefore voiceprint analysis remains a controversial subject [3].

According to Nieuwenhuizen *et al*., 2010, audio fingerprinting has two algorithms Shazam's and Haitsma & Kalker's Algorithm. When compared Shazam proved to be better in accuracy [5].

Ogle and Ellis, 2007, had estimated a technique in private audio clips to notice reciting sound incidents. Experiments carried out by the researchers reported that the technique was practically well performed for sound incident variations, common in environmental recordings. The speed of search was also noteworthy [7].

A baby's cry was anticipated using voice frequency [9]. There are mainly two causes why baby would cry hunger and sleep. The emotion was detected using audio frequency recognition. Hunger has a high-

frequency region while sleep has low-frequency region. The 80% of baby's cry was positively recognized by computer simulations. The study claims that the baby's age in months defines the frequency of hunger and sleepiness. The tendency of the different characteristics like baby crying because of hunger or because of sleep in the frequency domain is clarified. It was seen that the hunger cry contains much more signal components than in higher frequency region of that of sleepy cry [9].

Sarno *et al.,* 2019, used audio fingerprinting to detect actual song and cover song for identifying copyright using Bhattacharya distance and k-nearest neighbour algorithm. The result turned out to be accurate for the songs [10].

III. RESEARCH METHODOLOGY

The two algorithms considered for this study are discussed below:

A. Haitsma and Kalker's Algorithm

Philip's uses features based on multiple sub-bands. They use Haitsma and Kalker's algorithm [6].

Haitsma and Kalker's algorithm suggests that the structure of fingerprint extracted should be based on a common streaming method in which after every 11.6 ms an overlap factor 31/32 is overlapped on 370ms long window frame [6].

FFT of every frame is calculated and filtered through band division deposited in a 32-bit sub-fingerprints.

For an algorithm to find a match in an unknown audio signal model minimum time frame required is 3 - 30 seconds.

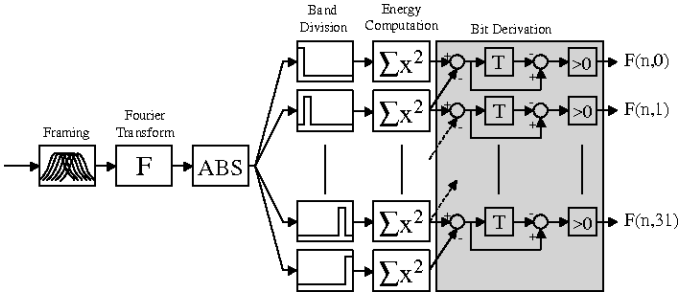

Fig.3 Haitsma and Kalker's algorithm [6]

Where,
F = Fourier Transform,
T = delay element,
ABS = absolute value,
$\sum x^2$: $Ene(i,j) = \sum |s(m,n)|^2$,
Where, Ene(i,j) = Sub-band energy, i and j is the n^{th} block
s= one-bit fingerprint,
n=sub-band,
m= fingerprint extracted from sub-band.
>0: 1 <=F <= 32.

A Hidden Markov model is a mixture model where the hidden variable controls the components to be selected for observation. Quantization is a conversion method from continuous set of values to a discrete set. Compressed image of a single frame is known as sub-fingerprint. Sub-fingerprint doesn't contain enough data to recognize audio clips. Fingerprint block contains satisfactory data to recognize an audio clip as they comprise of 256 sub-fingerprints and even a granularity of 3 seconds. For these frames, an audio signal is subdivided into frames with a distance of 0.37 seconds. For every 11.6 ms, one sub - fingerprint is retrieved.

Even in the worst-case scenario the large intersection ensures that the sub - fingerprints of the audio clip given as input that are to be known are very similar to the sub - fingerprints of the same clip with respect to database. Audio features in frequency domain. On every frame the Fourier transformation takes place because of the sensitivity of the higher frequencies in phase of fourier transformation, different frame boundaries are generated that are audible to the human auditory system.

Algorithm as shown in fig 3.

Step 1: Audio signals are divided into frames; features set is calculated for each frame.

Step 2: Fourier coefficients features are termed as audio features. Extracted Features are compact representation by using HMM (Hidden Markov Model) and Quantization.

Step 3: The ABS range value is kept. Human auditory system works only on frequencies less than 2k Hz.

Step 4: For each frame, 33 non-overlapping frequency band is selected and 32-bit sub-fingerprint value is extracted. Bands are located at 300 hz-2000 Hz (HAS range).

Step 5:

$Sen(i,j) = \sum |s(m,n)|^2$

Step 6: 32-bit sub-fingerprints extracted. A '1' and a '0' bit till 31bits.

B. Shazam's Algorithm

Audio fingerprints usually uses audio sections within a span of 3 - 30 seconds to generate an audio fingerprint. Compared to the audio fingerprints in database, the original audio section is recognized.
The algorithm creates a fingerprint of the song using 20 seconds audio clip, the fingerprint generated is matched with the database to infer the result.

Avery Wang states that the time taken to search in the database with 20000 audio samples is between 5 to 500 milliseconds [5].

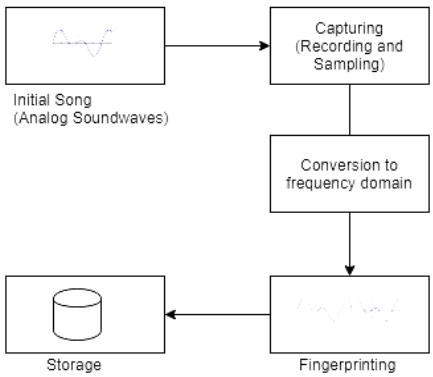

Fig 4. Processing pipeline

As shown in fig 4 analog signal is captured converted to frequency domain. They extract features known as fingerprinting and stores in storage (database).

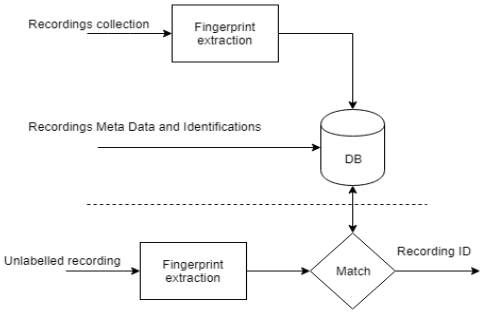

Fig 5. Content-based audio identification framework [5]

Audio is captured so that the features are extracted and are kept in database. The unlabelled recording of the audio's features is extracted and then matched with the database. If the fingerprint matches with audio samples in database, the details of the identified song is reported as shown in fig 5. This algorithm uses STFT (Short - time Fourier Transform) spectrogram squared magnitude.

$$Spectrogram\ (t, w) = |STFT(t, w)|^2$$

Where,
t = time,
w =frequency.

Fast Fourier Transform (FFT) calculates the strength of signal i.e. amplitude at a specific frequency. The red colour indicates high amplitude whereas green indicates low amplitude value. The amplitude can be plot on time and frequency coordinates. The variations between fingerprint algorithms in the cluster includes the overlap of frames, defining the frame, searching and storing of the fingerprint.
Step 1: The Spectrogram is basically separated into smaller fragment that are called windows or frames using FFT.
Step 2: Using energy units in the frame, this algorithm creates a spectral pair landmark.
Step 3: It selects landmark as an anchor point and matches it within the target zone with nearby landmarks.
Step 4: Using combination of landmarks different hashes are formed.
Step 5: In database each hash value is deposited with an audio file with the time of occurrence.
Step 6: The background noise is filtered out by selecting the maxima values in the spectrum as landmarks.
Step 7: Stored hashes are compared with the hashes generated from fingerprinted query signal.
Step 8: In archive the hashes are stored as a pair, if the time of offset is similar of resulted and saved hashes.
Step 9: From the histogram of common hash count the peak value is identified statistically and selected as match.

RESULTS

Audio fingerprinting used to recognize frequent sound events in long duration results as follows [7]

TABLE I

IDENTIFYING LONG DURATION RESULTS

Audio length (min)	60	120	180	390
Time (ms)	21	31	37	131

Comparison of two algorithms [4]:

TABLE 2
COMPARISON OF TWO ALGORITHMS

Algorithm	Result
Avery Wang (Shazam)	5 min unknown audios recognized in 19.8s.
Haitsma and Kalker's	5 min unknown audios recognized in 20.696 ms.

The database is populated with the audio length of 300 minutes (approx. 60 songs). Accuracy of Shazam's algorithm to search in an unknown 5 minute audio is 19.8 seconds. Whereas, accuracy for Haitsma and Kalker's is 20.696 seconds [1].

Long - term sound event identification:

30 audio samples were introduced in the recording with 45 telephone rings each, door closures and beeps were introduced after every 10 occurrences in data set.

IV. CONCLUSIONS

It is observed that Avery Wang Shazam's algorithm has better accuracy than Haitsma & Kalker's algorithm. This algorithm can also be used in other field such as emotion detector, criminal identification, etc. Nowadays these algorithms are mostly used in personal assistant to recognize the voice. The future scope of the audio recognition is in music industry with its various possible applications.

REFERENCES

1. *https://www.its.bldrdoc.gov/publications/2641.aspx accessed on March 01, 2019.*
2. *Van Nieuwenhuizen, Heinrich A., Willie C. Venter, and Leenta MJ Grobler. "The study and implementation of Shazam's audio fingerprinting algorithm for advertisement identification." Proceedings of SATNAC 2011. 2012.*
3. *Parmar, P. (2012). Voice Fingerprinting: A Very Important Tool against Crime. Journal of Indian Academy of Forensic Medicine, 34(1), 70-73.*
4. *Van Nieuwenhuizen, H. A., Venter, W. C., & Grobler, L. M. Comparison of Algorithms for Audio Fingerprinting. 2010.*
5. *Cano, P., Batlle, E., Kalker, T., & Haitsma, J. (2005). A review of audio fingerprinting. Journal of VLSI signal processing systems for signal, image and video technology, 41(3), 271-284.*
6. Haitsma, J., & Kalker, T. (2002, October). A highly robust audio fingerprinting system. In Ismir (Vol. 2002, pp. 107-115.
7. *Ogle, J. P., & Ellis, D. P. (2007, April). Fingerprinting to identify repeated sound events in long-duration personal audio recordings. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 (Vol. 1, pp. I-233). IEEE.*
8. *Cano, Pedro, et al. "Audio fingerprinting: Concepts and applications." Computational intelligence for modelling and prediction. Springer, Berlin, Heidelberg, 2005. 233-245(2002) The IEEE website. [Online]. Available: http://www.ieee.org/*
9. *Arakawa, Kaoru. "Recognition of the cause of babies' cries from frequency analyses of their voice classification between hunger and sleepiness." Proc. International Congress on Acoustics. 2004.*
10. *Sarno, R., Wijaya, D. R., & Mahardika, M. N. (2019). Music fingerprinting based on bhattacharya distance for song and cover song recognition. International Journal of Electrical and Computer Engineering (IJECE), 9(2), 1036-1044.*