http://dx.doi.org/10.5392/IJoC.2015.11.4.001

Conjoined Audio Fingerprint based on Interhash and Intra hash Algorithms

Dae-Jin Kim, Hong-Sub Choi
Electronic Engineering
Daejin University, Sundan-dong, Phochon-si, Kyunggi-do, 487-711, Korea

ABSTRACT

In practice, the most important performance parameters for music information retrieval (MIR) service are robustness of fingerprint in real noise environments and recognition accuracy when the obtained query clips are matched with the an entry in the database. To satisfy these conditions, we proposed a conjoined fingerprint algorithm for use in massive MIR service. The conjoined fingerprint scheme uses interhash and intrahash algorithms to produce a robust fingerprint scheme in real noise environments. Because the interhash and intrahash algorithms are masked in the predominant pitch estimation, a compact fingerprint can be produced through their relationship. Experimental performance comparison results showed that our algorithms were superior to existing algorithms, i.e., the sub-mask and Philips algorithms, in real noise environments.

***Key words**: Music Information Retrieval, Conjoined Fingerprint, Interhash, Intrahash.*

1. INTRODUCTION

With the recent rapid development of mobile devices, there has been a great deal of interest in content-based music information retrieval (MIR) services, which search for music that is playing over public loudspeakers in coffee shops, shopping malls, off-street venues, and so on. Music retrieval applications, such as Shazam, SoundHound, and GraceNote, are now offering services for iPhones, iPads, and other mobile devices. In real environments, however, such music is corrupted by background noise, i.e., people's voices or machine sounds. An MIR service must therefore have two key properties: accuracy and robustness [1]. Accuracy refers to the number of the correct, missed, and wrong identifications. Robustness is the ability to make accurate identifications even if there is distortion or compression between signal conversions. Most MIR services are based upon audio fingerprinting schemes, and the Philips scheme proposed by Haitsma has proven to be the most robust and accurate of these schemes. For a more robust fingerprinting system in real noise environments [2], Mansoo proposed a frequency-temporary filtering method [3]. In practice, however, this method still needs further improvement in order to be used in real noise environments. To this end, Wooram proposed a sub-fingerprint masking method for creating a robust fingerprint scheme in real noise environments [4]. This algorithm is very robust in real-noise environments, but uses only 5-bit hash values for its sub-fingerprint, using a mask generated by predominant pitch estimation. Thus, this scheme is more appropriate for a small-scale MIR service than for a massive MIR service. In this paper, we propose the new conjoined fingerprint scheme, which is based on an inter- and intrahashing algorithm that improves the robustness of the audio fingerprinting system in real noise environments, and can also be used in a massive MIR service.

* *Corresponding author, Email: hschoi@daejin.ac.kr*
Manuscript received Apr. 14, 2015; revised Sep. 18, 2015; accepted Sep. 25, 2015

2. RELATED WORKS

2.1 Philips Hashing Algorithm

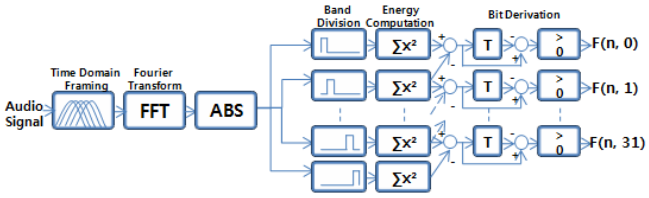


Fig. 1. Overview of Philips hash fingerprint scheme [2]

Fig. 1 shows an overview of the Philips scheme's hash extraction fingerprinting system. The audio signal framing is segmented into overlapping frames. Next, the audio signal is computed by applying Fourier transform to each frame. The results of the Fourier transform are divided into bands ranging from 300 Hz to 2000 Hz. The energy is calculated on the basis of each sub-band, and the energy of band m of frame n is denoted by $E(n,m)$. In order to create a fingerprint block, a 32-bit sub-fingerprint value is extracted for each frame.

The hash fingerprint block consists of 256 sub-fingerprints. The sub-fingerprint denotes the m th bit of the fingerprint of frame n as $F(n,m)$, which is formally defined as

$$F(n,m)=\begin{cases}1 & \text{if } E(n,m)-E(n,m+1)-\left(E(n-1,m)-E(n-1,m+1)\right)>0 \\ 0 & \text{if } E(n,m)-E(n,m+1)-\left(E(n-1,m)-E(n-1,m+1)\right)\le 0\end{cases} \quad (1)$$

A lookup table (LUT) can be composed of the sub-fingerprint database that is created by (1). Using the LUT is approximately 800,000 times faster than using a music database (DB) for music retrieval [2]. The LUT is composed of entries for all possible 32-bit sub-fingerprints. These entries are listed with pointers to the positions in the real fingerprint block lists, where the respective 32-bit sub-fingerprints are located. In the LUT, compare with input query's and DB's sub-fingerprints of the fingerprint block. Then, the bit error rate (BER) is calculated. If the BER is below the threshold, the probability is high that the extracted fingerprint block originates from corresponding music stored in the DB.

2.2 Masking Algorithm

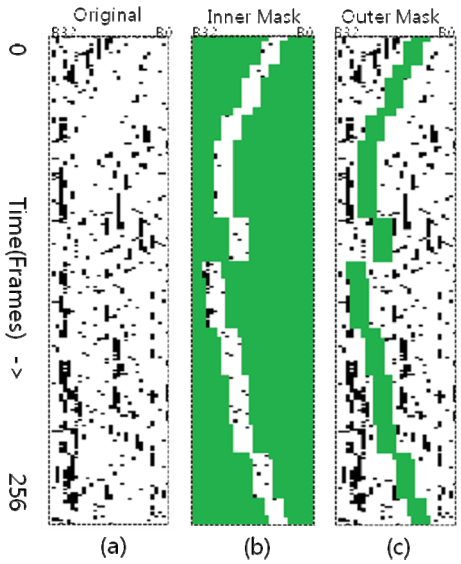


Fig. 2. (a) Fingerprint block of original music clip showing the bit errors in black, (b) inner-masked sub-fingerprint block, and (c) outer-masked sub-fingerprint block [4]

The audio fingerprint method using sub-fingerprint masking is based on predominant pitch estimation. The temporal sequence of the harmonic structures in the frequency domain is the key to the human perception of music, and the recognition of predominant pitch is a process of perceiving the harmonic characteristics of partials. This algorithm has a bit-masking step to extract musically meaningful high-level attributes based on the predominant pitch estimation.

However, this method uses only 5-bit hash values for the sub-fingerprint, using a mask generated by the predominant pitch estimation. If the DB in a music service is quite massive, the false accept rate (FAR) will be much higher. Therefore, this algorithm would exhibit low recognition accuracy for a massive music DB

3. PROPOSED AUDIO FINGERPRINT SYSTEM

3.1 System Overview

The proposed audio fingerprinting system is based on the Philips hashing algorithm. To obtain a robust fingerprint in real noise environments, we propose the new conjoined fingerprint scheme, which is based on an inter- and intrahashing algorithm for fingerprint extraction. The audio signal is segmented into overlapping frames, and the hash values are extracted for the fingerprint. The hash value is a 32-bit code derived from 33 perceptually divided frequency bands for each frame. We compose two kinds of hash values for the fingerprint. One is an interhash, which is obtained in a sequential frame, as in the Philips method. The other is an intrahash, which is obtained in the current frame. We then make a new, conjoined 32-bit sub-fingerprint that is compounded by sub-fingerprint masking. The search module compares the sequence of conjoined sub-fingerprints with the target fingerprint block by simply applying a bit-wise check. The search results can be obtained by the number of matches that are under the bit error threshold.

3.2 Inter- and Intrahashing Algorithm

We use an audio signal containing one channel sampled at a rate of 44100 Hz, with 16 bits per sample. The audio signal framing is segmented into Hanning-windowed overlapping frames. One frame is 371 ms long, and an overlapped length is taken every 11.6 ms. One important consequence of the frame's slice length/spacing combination (371 ms slices, each 11.6 ms long) is that the frequency varies slowly over time, affording sufficient robustness against position uncertainty in time. In addition, the signal is computed by applying Fourier transform to each frame. The frequency spectrum is divided logarithmically in a spectral range from 750 Hz to 2750 Hz. In the Philips algorithm, the spectral range is from 300 Hz to 2000 Hz. Although in this range, it can search for music well in the lyrics of a song, it has difficulty searching music in the melodic parts due to the high frequency generated by instruments. Therefore, we use a spectral range from 750 Hz to 2750 Hz as an experiment. The energy is calculated on the basis of each sub-band, and the energy of band m of frame n is denoted by $E(n,m)$. In order to generate a fingerprint block, a 32-bit hash value is extracted for each frame.

This is an interhash, as in the Philips method. The interhash denotes the m th bit of the hash of frame n as $F_{inter}(n,m)$, which is formally defined as

$$F_{\text{inter}}(n,m)=\begin{cases}1 & \text{if } E(n,m)-E(n,m+1)-\left(E(n-1,m)-E(n-1,m+1)\right)>0 \\ 0 & \text{if } E(n,m)-E(n,m+1)-\left(E(n-1,m)-E(n-1,m+1)\right)\le 0\end{cases} \quad (2)$$

When a number of bands are corrupted by noise, the Hamming distance between an original and a distorted sub-fingerprint could be greater, owing to the correlation of the filter band energies (FBEs). Thus, frequency filtering is generally used to decorrelate the FBEs, an approach that has been verified to a certain degree in speech recognition systems (Chen *et al*, 2003), (H-Y. Jung, 2004). Therefore, we use a band-pass second-order

FIR filter(Mansoo *et al*, 2006) in an interhash. The filter is defined as

$$H_f(z) = z - z^{-1} \tag{3}$$

We also use an intrahash that imply information inside a frame. This functions as a low-pass filter, using the frequency average inside a frame. Therefore, the intrahash is a robust feature in terms of noise. It is particularly robust when the pitch changes suddenly between frames. For the intrahash, we first denote the frequency average inside the *n*th frame as $Freq_{ave}(n,m)$, which is formally defined as

$$Freq_{ave}(n,m) = \frac{1}{W}\sum_{i=0}^{W-1} x(n,m,i) \tag{4}$$

where $x(n,m,i)$ is the frequency value inside the n th frame, the m th sub-band, and the i th frequency bin index, and W is the sub-band length. Next, we calculate the summation of $Freq_{sub\_each}(n,m,i)$ for each frequency sub-band. The $Freq_{\text{sub\_each}}(n,m,i)$ is a unit separator by the average of frequency sub-band, which is formally defined as

$$Freq_{\text{sub\_each}}(n,m,i) = \begin{cases} 1 & \text{if } \text{Freq}(n,m,i) \geq \text{Freq}_{ave}(n,m) \\ -1 & \text{if } \text{Freq}(n,m,i) < \text{Freq}_{ave}(n,m) \end{cases} \tag{5}$$

$$F_{sub\_sum}(n,m) = \sum_{i=0}^{W-1} F_{sub\_each}(n,m,i) \tag{6}$$

Finally, the intrahash denotes the m th bit of the hash of frame n as $F_{intra}(n,m)$, which is formally defined as

$$F_{intra}(n,m) = \begin{cases} 1 & if \;\; F_{sub\_sum}(n,m) \geq 0 \\ 0 & if \;\; F_{sub\_sum}(n,m) < 0 \end{cases} \tag{7}$$

$F_{sub\_sum}(n,m)$ is the threshold for hash estimation inside a frame.

3.3 Conjoined fingerprint

We generate the new robust conjoined fingerprint by applying an inter- and intrahashing algorithm to each frame for the massive MIR service. For the new conjoined fingerprint, we must first estimate the predominant pitch(Song *et al*, 2002) in order to use harmonic enhancement and harmonic summation in the frequency domain. We also compose the critical band, which is the band that contains the frequency index of the predominant pitch. Because the sub-fingerprint bit count is 32 bits, we use a critical band size of 16 bits for each interhash and intrahash. The critical band filters are defined as

$$H_c(n,m) = \begin{cases} 1 & if \;\; m_p - k \leq m < m_p + k \\ 0 & else \end{cases} \tag{8}$$

where m_p is the frequency band that contains the predominant pitch, and k is 8. Fig. 3 shows a critical band filter based on predominant pitch estimation in the frequency domain.

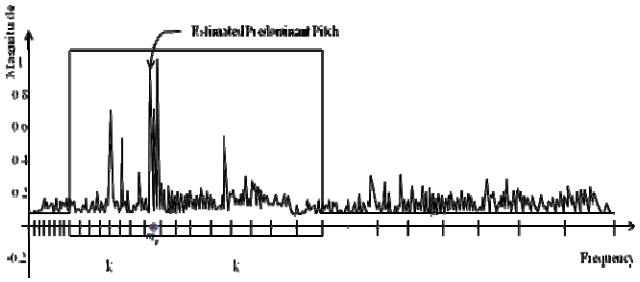


Fig. 3. Critical band filter based on predominant pitch estimation in frequency domain.

We also include a bit-masking step that extracts a musically meaningful fingerprint for each interhash and intrahash via critical band filtering, which is formally defined as

$$F_{pre\_\text{intra}}(n,m) = \{(F_{\text{intra}}(n,m)H_f(z)) << (32-(m_p+k))\} \& 0\times FFFF0000 \tag{9}$$

$$F_{pre\_\text{inter}}(n,m) = \{(F_{\text{inter}}(n,m)H_f(z)) >> (m_p-k)\} \& 0\times 0000FFFF \tag{10}$$

We then compose the new conjoined fingerprint for a sub-fingerprint that incorporates the relationship between the interhash and the intrahash. Because the conjoined fingerprint has information that is more meaningful in the predominant frequency area, we can search for music with higher precision in real noise environments. The conjoined fingerprint is formally defined as

$$F_{conjoined}(n,m) = F_{pre\_intra}(n,m) \mid F_{pre\_inter}(n,m) \tag{11}$$

where $F_{conjoined}(n,m)$ is composed of a filtered intrahash in the higher 16 bits and a filtered interhash in the lower 16 bits.

4. EXPERIMENTS AND RESULTS

In our experiment, we performed a simulation using a music database of 50,000 popular songs in the mp3 format (44100 Hz, 2 channels, 16 bits per sample) that were converted from an audio CD. They included various genres, such as rock, pop, hip-hop, dance, and classical. We randomly selected 500 songs for an audio query clip. The audio query clip was captured using an iPhone 5, which was placed 2.0 m away from a stereo speaker in real noise environments, such as a coffee shop, shopping mall, office, and street. Because the sound

recording format is m4a in the iPhone 5, all audio query clips were recorded in this format.

To evaluate our algorithm, we composed a query set classified according to signal-to-noise ratio (SNR) ranges

Table 1 and Fig. 4 show the results of the music retrieval experiments performed on a database with 25,000 songs based on three different algorithms—the conjoined, sub-mask and Philips algorithms—using 500 queries in one matching server. The query had 30 s offsets in m4a format audio and duration of 18 s. The results of a performance comparison show that our algorithm is superior to the existing algorithms in real noise environments.

Table 1. Comparative performance according to SNR

Noise \ Method	Conjoined	Sub-mask	Philips
Clean (Set 1)	100	100	100
15.0–20.0 dB (Set 2)	83.8	55.8	72.9
10.0–15.0dB (Set 3)	72.4	26.6	56.2
5.0–10.0 dB (Set 4)	52.6	12.2	22.8
0.0–5.0 dB (Set 5)	44.7	3.2	16.5

%: Recognition accuracy. DB size=25,000, Query length=18 s.

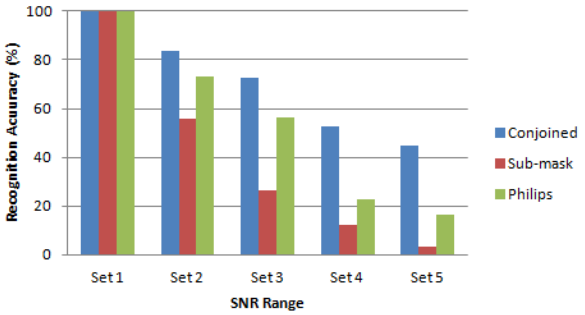


Fig. 4. Comparison between recognition performance of proposed and existing algorithms

Furthermore, we devise a service system that checks the degree of recognition accuracy in a massive MIR service. A single matching server has limits to its performance, so the matching process needs to be dispersed for a massive MIR service. We therefore use one clustering server and two matching servers in a server group. The matching server is connected to a MySQL DB containing 25,000 songs, and it evaluates the match rate for queries with the songs in this DB when it receives an input music query clip. These match results are then forwarded to a clustering server. The clustering server then checks the BER results obtained from each matching server and selects the song having the minimum BER. Fig. 5 presents a diagram of the service system for a massive MIR service.

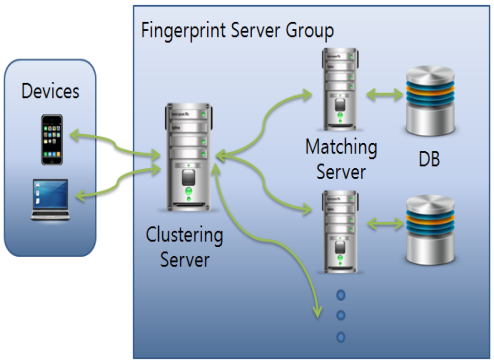


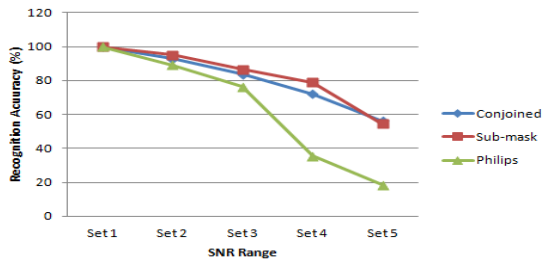
Fig. 5. Diagram for a massive MIR service

Table 2 show the results of the music retrieval experiments performed on a database with various numbers of songs based on three different algorithms—the conjoined, sub-mask and Philips algorithms—using 500 queries, when an MIR service is employed.

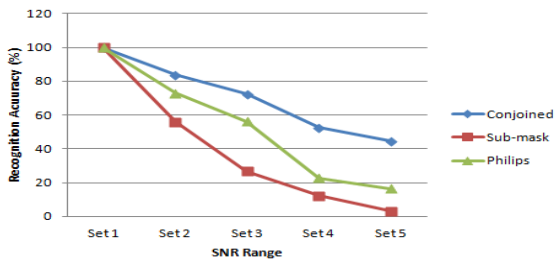
Table 2. Comparative performance according to SNR and database size

Noise \ Method (DB amount)	Conjoined			Sub-mask			Philips		
	2.5k	25k	50k	2.5k	25k	50k	2.5k	25k	50k
Clean (Set 1)	100	100	100	100	100	100	100	100	100
15.0–20.0 dB (Set 2)	92.8	83.8	77.2	95.1	55.8	37.4	89.2	72.9	53.3
10.0–15.0 dB (Set 3)	83.7	72.4	63.3	86.6	26.6	14.3	76.4	56.2	43.3
5.0–10.0 dB (Set 4)	72.1	52.6	31.4	78.9	12.2	6.8	35.6	22.8	24.4
0.0–5.0 dB (Set 5)	56.2	44.7	28.6	54.6	3.2	3.2	18.4	16.5	6.7

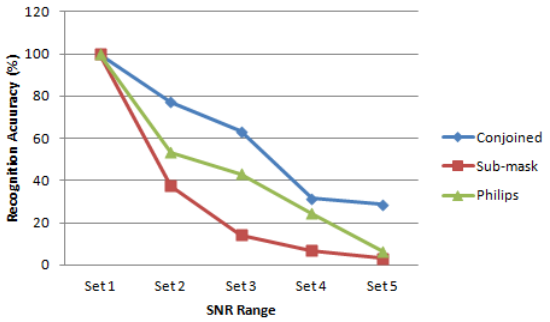
%: Recognition accuracy, Query length = 18 s



(a) 2,500 songs



(b) 25,000 songs



(c) 50,000 songs

Fig. 6. Comparative performance according to SNR for variously sized databases

Our experiments evaluate the recognition accuracy in DBs with 2,500, 25,000, and 50,000 songs. Figure 6 shows the results of their comparative performance in terms of the SNR for these databases. In the DB with 2,500 songs, the sub-mask method shows better recognition accuracy than the other methods in all SNR groups. However, recognition accuracy is more important when the DB size for an MIR service is fairly large. When the DB contains a larger amount of music, the conjoined and Philips methods afford a much higher recognition accuracy than the sub-mask method. Moreover, the conjoined method always produces a much higher degree of recognition accuracy than the Philips method, regardless of the DB size.

Because the sub-mask method exhibits a high false accepted rate (FAR), it affords low recognition accuracy. Table 3 shows the average FAR results of music retrieval experiments performed on databases of various sizes using three different algorithms—the conjoined, sub-mask and Philips algorithms—with 500 queries.

All the experiments are performed on a system composed of a matching server running the Windows 7 operating system (OS) and MySQL DB with an Intel Xeon 2.4 GHz 2EA processor and 128 GB of memory as well as a clustering server running the Windows 7 OS with an Intel i5 2.66 GHz processor and 4 GB of memory.

Table 3. Comparative average FAR according to database size

Method (DB amount) / Noise	Conjoined			Sub-mask			Philips		
	2.5k	25k	50k	2.5k	25k	50k	2.5k	25k	50k
FAR (%)	2.5	12.8	18.4	2.8	63.7	82.1	2.4	14.2	18.8

5. CONCLUSIONS

We performed experiments using our newly proposed conjoined fingerprint algorithm, which uses interhashes and intrahashes to produce a robust fingerprint scheme in real noise environments. The experimental results show that the recognition accuracy of the proposed algorithm is much higher than that of the original Philips and sub-mask fingerprint algorithms in a massive MIR service. Because the conjoined fingerprint is based on the use of interhashes and intrahashes for each frame in predominant pitch estimation in the frequency domain, the relationship between frames is much stronger, which makes for a more robust fingerprint in real noise environments. In the future, we will seek to create methods that can improve both the retrieval time as well as the robustness in such environments.

REFERENCES

[1] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, “A Review of Audio Fingerprinting,” J. VLSI Signal Processing Systems for Signal Image Video Technology, vol. 41, no. 3, 2005, pp. 271-284.

[2] J. Haitsma and T. Kalker, “A Highly Robust Audio Fingerprinting System,” Proc. Of the 3rd Int. Symposium on Music Information Retrieval, 2002, pp. 144-148.

[3] Mansoo Park, Hoi-Rin Kim, and Seung Hyun Yang, “Frequency Temporal Filtering for a Robust Audio Fingerprinting Scheme in Real-Noise Environments,” ETRI Journal, vol. 28, no. 4, 2006, pp. 509-512.

[4] Wooram Son, Hyun-Tae Cho, Kyoungro Yoon, and Seok-Pil Lee, “Sub-fingerprint Masking for a Robust Audio Fingerprinting System in Real-noise Environment for Portable Consumer Devices,” IEEE Transactions on Consumer Electronics, vol. 56, no. 1, 2010, pp. 156-160.

[5] 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.

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

[7] J. Chen, K. Paliwal, and S. Nakamura, “Cepstrum derived from Differentiated Power Spectrum for Robust Speech Recognition,” Speech Communication, vol. 41, 2003, pp. 469-484.

[8] H.-Y. Jung, “Filtering of Filter-Bank Energies for Robust Speech Recognition,” ETRI Journal, vol. 26, no. 3, 2004, pp. 273-276.

Dae-Jin Kim

He received the B.S., M.S in electronics from DaeJin university, Korea in 1998, M.S in electronics from Dongkuk university, Korea in 2000, and Ph.D. in electronics from DaeJin university, Korea in 2010. Since then, he has been with Media Solution Business Dev., Markany. His main research interests include multimedia information retrieval, codec, fingerprint, watermark, video cartooning, and multimedia system.

Hong-Sub Choi

He received the B.S., M.S. and Ph.D. degrees from Seoul National University, Seoul, Korea in 1985, 1987 and 1994, respectively all in Electronic Engineering. In March 1995, he joined the faculty of the Daejin University, Pocheon, where he is currently a Professor at the Department of Electronic Engineering. He was a visiting Scholar in the Department of Electrical and Computer Engineering at CMU, Pittsburgh, PA in 1999. and another visiting scholar in CSLR at University of Colorado, Boulder, CO, 2006. His current research interests include multimedia, speech signal processing and communications.