A Robust Audio Fingerprinting System with Predominant Pitch Extraction in Real-Noise Environment

Wooram, Son
Seoul National University, Korea.
103 Yongon-dong, Jongno-Gu
Seoul, Korea
+82-2-2072-1762
sonwr@snu.ac.kr

Kyoungro, Yoon*
Konkuk University, Korea.
1 Hwayang-dong, Gwangjin-Gu
Seoul, Korea
+82-2-450-4129
yoonk@konkuk.ac.kr

ABSTRACT
The robustness of audio fingerprinting system in a noisy environment is a principal challenge in the area of content-based audio retrieval. The selected feature for the audio fingerprints must be robust in a noisy environment and the computational complexity of the searching algorithm must be low enough to be executed in real-time. 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 noise environment using predominant pitch which reduces the bit error of created hash values. The sub-fingerprint of our approach method is computed in each time frames of audio. The time frame is transformed into the frequency domain using FFT. The obtained audio spectrum is divided into 33 critical bands. Finally, the 32-bit hash value is computed by difference of each bands of energy. And only store bits near predominant pitch. Predominant pitches are extracted in each time frames of audio. The extraction process consists of harmonic enhancement, harmonic summation and selecting a band among critical bands.

Categories and Subject Descriptors
H.3.3 [INFORMATION STORAGE AND RETRIEVAL]: Information Search and Retrieval – *Retrieval model, Search process.*

General Terms
Algorithms, Performance.

Keywords
Audio classification, audio fingerprinting, audio searching.

1. INTRODUCTION
Recently, content-based music information retrieval becomes one of the most important services of wired/wireless communication service or consumer electronics 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]. A music information retrieval can also be used to prevent unauthorized peer-to-peer sharing of music files [3], [4]. Audio fingerprinting is a powerful tool for identifying audio, using a database of fingerprints. In developing commercial service using audio fingerprinting system recent reports focus on the several properties as follows [5, 6, 7, 8, 9].

- Robustness: the ability of the algorithm to accurately identify an item and still recognize correctly when it has been seriously degraded by distortion (pitching, equalization, audio coders(such as MP3 and GSM), background noise, A/D-D/A conversion, among others) or compression,.

- Accuracy: the ability of the algorithm to correctly identify an item and not to miss correct item (false negative rate, false positive rate),.

- Scalability: the ability to perform well or reasonably, when the number of items in the database grows. This property affects the accuracy and the complexity of the system,.

- Complexity: the computational costs of the fingerprint extraction and, search as well as the size of the fingerprint.

An audio fingerprinting scheme that has proved known to be the most robust is the so-called Philips scheme proposed by Haitsma and Kalker [5]. This scheme uses the energies of 33 logarithmically scaled bands to obtain their hash value which is the sign of the energy band differences (both in the time and the frequency axis). But in the noisy environment, the hash value or 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 [6] causing additional 33 times of lookup for audio identification. In this paper, we propose a novel approach to improve robustness of the audio fingerprinting in noisy environment, using predominant pitch extraction and enhancement as a high-level semantic attribute in music.

2. SYSTEM OVERVIEW

Our audio fingerprinting system is composed of three main modules, which are Philips' hashing algorithm module, predominant pitch extraction module, hashvalue masking module and search module as shown in Figure 1. 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 hash value masking module whose mask are dynamically created based on the result of the predominant pitch extraction module. Then an audio file or a segment of a audio file is represented by a fingerprint block which is a sequence of bit-masked 32-bit hash values. The search module compares the sequence of sub-fingerprints with the target fingerprint block, by simply applying bit-wise equality operation, as presented in [5].

The best-matched result is determined based on the number of matched hash value per fingerprint block. In the following sections, each module is described in detail.

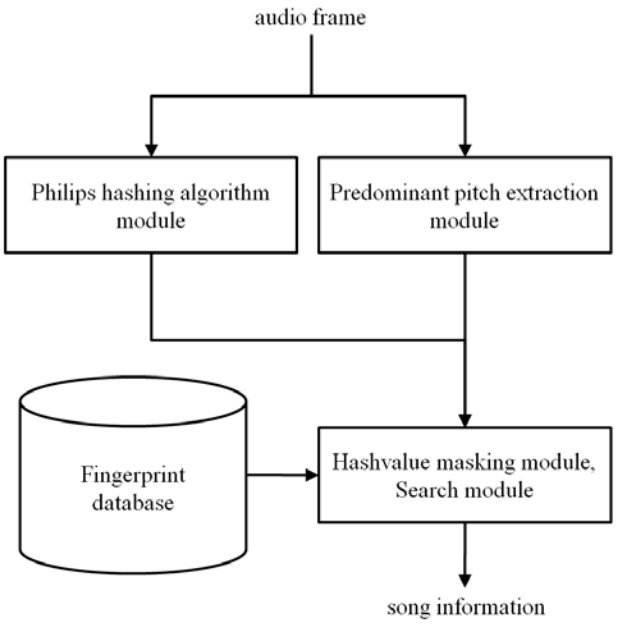

Figure 1. Overview of audio fingerprinting system.

3. PHILIPS' HASHING ALGORITHM

A detail of Philips' hashing algorithm is given in [5]. As In our implementation, the audio signal is first sampled at the rate of 44100 Hz and segmented into frames of 16384 samples, 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. By grouping 256 sub-fingerprints, a fingerprint block is generated.

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

4. PREDOMINANT PITCH EXTRACTION

The Philips' hashing algorithm may not be sufficiently rosbust enough in real-noise environment since the energy of 33 critical bands are still correlated. When some critical bands are affected by noise, such as hand clapping, car noise, subway noise or babble noise, extended comparing for the query sub-fingerprint to the sub-fingerprints within a certain hamming distance range could be required to enhance the robustness of the matching. But it needs 33 times more lookup for comparing each sub-fingerprint expanded by the hamming distance of one [5].

Therefore, a more robust audio fingerprint algorithm which does not require extended matching can be much beneficial. In an attempt to reduce the bit-error probability, we had pre-process of predominant pitch extraction and enhancement so that the audio fingerprint extraction can be based on pitch information.

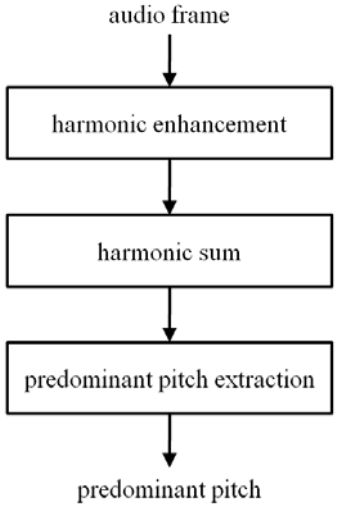

Figure 2. Predominant pitch extraction process.

In musical sound, fundamental frequency has harmonic structure. The harmonic structure is most important feature in recognizing musical sound. The polyphonic music show periodic peaks in the frequency domain according to the characteristics of each sound source. It is said that the recognition of the sound is the process of perceiving how much those partials have harmonic characteristics [8]. To estimate fundamental frequency, we propose to use predominant pitch extraction process as proposed in [10] and shown in Figure 2. The predominant pitch extraction module provides estimations of the predominant fundamental frequency in audio frame. In order to extract a predominant fundamental frequency for every audio frame, we extract harmonics (structure) in audio frame by harmonic enhancement process, and then estimate fundamental frequency in harmonic structure by harmonic summation process.

4.1 Harmonic Enhancement

Harmonic enhancement process, which is the first step of the predominant pitch extraction process, amplifies candidates of predominant pitches. In the harmonic enhancement step, audio frames are transformed into the frequency domain using Fast Fourier Transform and harmonic positions in the spectrum are estimated for enhancement of harmonics.

In musical sounds, audio frame have several harmonics. And most audible harmonics exhibits their harmonic structure and haves larger amplitudes in spectrum. Therefore the frequency components that exhibit harmonics of relatively larger amplitude have higher probability of being a predominant pitch. However, even though the amplitudes of harmonics are large, if their surrounding signals also have large amplitudes, those harmonics have less possibility of contributing to the predominant sound. In polyphonic music, such (most audible) harmonic structure and amplitude can be impaired by other harmonic structures, due to the frequency masking effect. For instance, if the fundamental frequency of most audible harmonic structure has large amplitude, but if it is impaired by noise or percussion sound, the predominant pitch of the sound cannot be easily detected.

Harmonic enhancement process extracts and enhances harmonics that have outstanding peaks compared to the surroundings and the equation (2) represents enhanced harmonics.

$$E_t^{EP}(k) = \sum_{i=-W}^{W} A(E_t(k) - E_t(k+i)), 0 \le k < N \quad (2)$$
$$\text{where } A(x) = x, \forall x \ge 0, \quad A(x) = 0, \forall x < 0$$

In equation (2), N is the Fourier Transform index range, $E_t^{EP}(k)$ represents the magnitude of the predominance of the harmonic characteristic in the Fourier Transform Bin index k, considering the spectral amplitudes $E_t(k)$ of surrounding amplitudes within the bin range W in time t.

$E_t^{EP}(k)$ has large value when the spectral amplitude of the harmonic in the Fourier Transform Bin index k is a peak in the spectrum and it becomes much larger then adjacent amplitude when the spectral amplitude of given index k is large compared to its surroundings within the given window size W. Figure 3 and 4 shows spectrogram of a audio clip of musical sound and enhanced harmonics respectively. From these two figures, one can easily find that small peaks and wide band signals are eliminated and large and prominent peaks are emphasized.

4.2 Harmonic Sum

In musical sound, harmonics show periodic peaks in the frequency axis according to the characteristic of sound source. For that reason, our predominant pitch extraction algorithm has harmonic sum process based on above properties as shown in the equation (3).

$$F_t(p) = \frac{1}{N/p} \sum_{m=1}^{N/p} E_t^{EP}(mp) \quad (3)$$

In equation (3), p is the harmonic enhanced spectrum index. M is the multiplier of first harmonic (fundamental frequency) in harmonic structure. Consequently, $F_t(p)$ is the average strength of harmonics having the fundamental frequency p in an audio frame. If there is predominant pitch of frequency index p, the value $F_t(p)$, which represents the possibility that the sound of fundamental frequency p would occur, becomes large. Figure 5 illustrates this phenomenon by showing spectrogram and estimated pitch (white pixel in each time axis).

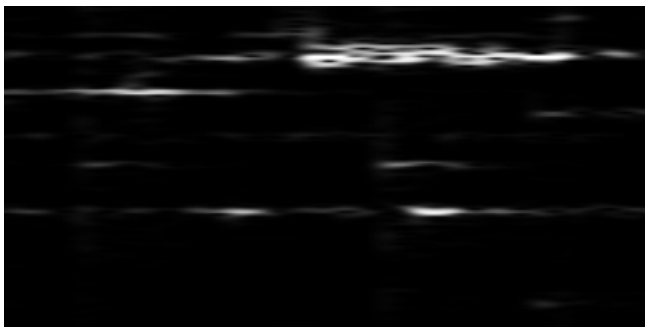

Figure 3. Spectrogram of audio spectrogram.

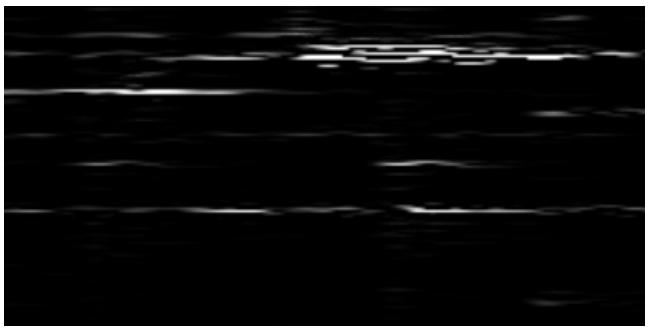

Figure 4. Enhanced harmonic of audio spectrogram.

Figure 5. Estimated pitch of audio spectrogram.

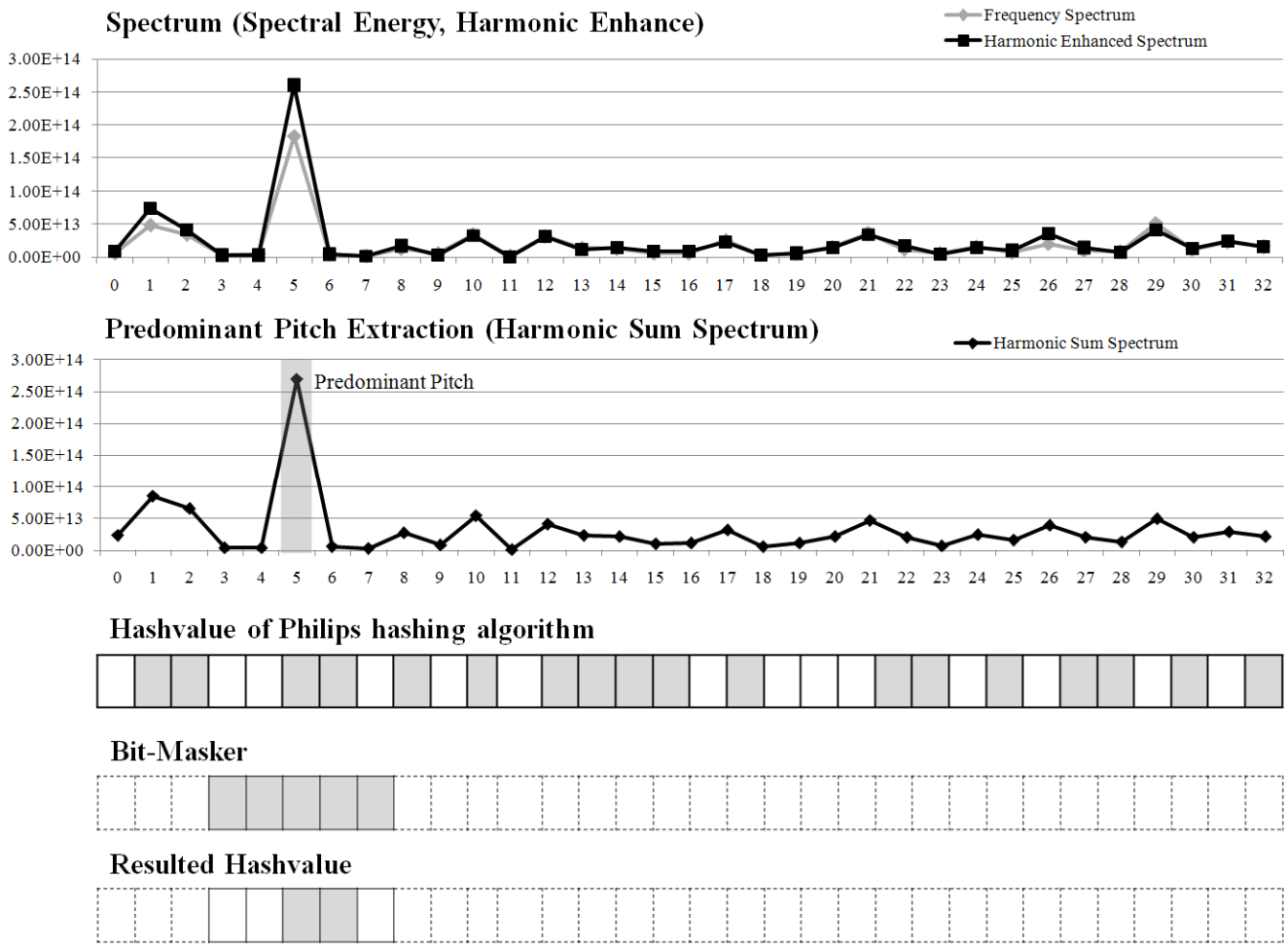

Figure 6. Process of generating masked hash value.

5. HASH VALUE MASKING

Hash values only contain spectral energy band differences in the time and the frequency axis as a cryptographical approach. Therefore, our approach has bit-masking process to contain musically meaningful high-level attribute. Our Fingerprint Extractor's final process is hash value masking operations as shown in Figure 6. 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. For example, in Figure 6, only five bits of the hash value mask are set to 1. And the resulted hash value contains differences of each bands of energy which also have predominant pitch information.

6. EXPERIMENTATION

6.1 Experiment Configuration

Experiment was executed at the music database that is composed of 2,019 popular songs. These songs are selected from Korean and Western popular songs and classics include various genres such as pop, hip-hop, jazz and classic. All the audio data is stored in PCM format with mono, 16 kHz bit rate and 44.1 kHz sampling rate converted from audio CDs. From these 2,019 songs, 1500 randomly created audio query clip of nine seconds each were captured using a microphone (Sennheiser ME66 model), which was placed 1.5m from an stereo loudspeaker. With the randomly created 1500 queries, four sets of distinguishable query sets are created by adding different level of noise, which is acquired on-line [11].

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 9.54dB.

- Set III is created by adding noise data to Set I with SNR of 3.68dB.

- Set IV is created by adding noise data to Set I with SNR of 3.01dB.

- Noise Data is acquired from [11], 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).

In the hash extraction step, the sub-fingerprints for every audio frame of interval 11.6 milliseconds are extracted. The overlapping frames have a length of 0.37 seconds and are weighted by a Hanning window, the selected frequency bands lie in a range from 300Hz to 2000Hz (the most relevant spectral range for the HAS) and have a logarithmic spacing.

6.2 Experiment Results

Experiments were executed with various conditions of database size and noise level. Experiments in Tables 1 and, 2 were performed with database of 2,019 music items stored, and query sets of 1,500 audio clips each. Table 3 and 4 were shows the experiment results with database of 500 music items stored, and query sets of 1,500 audio clips each. These experimental results clearly show that our fingerprinting algorithm generally performs better when the query set is corrupted by real environmental noise. These results also show that the recognition accuracy partially depend on the size of the music database. Table 5 shows comparison of the proposed algorithm and the Philips algorithm with expanded lookup of hamming distance of one-bit errors [6].

Table 1. Philips Hashing Algorithm

	Correct	Incorrect	Percent
	Database Size: 2019, Hamming Distance = 0		
Set I	1224	276	81.60
Set II	1094	406	72.93
Set III	753	747	50.20
Set IV	413	1087	27.53

Table 2. Our Hashing Algorithm

	Correct	Incorrect	Percent
	Database Size: 2019, Hamming Distance = 0		
Set I	1370	130	91.33
Set II	1352	148	90.13
Set III	1202	298	80.13
Set IV	975	525	65.00

Table 3. Philips Hashing Algorithm

	Correct	Incorrect	Percent
	Database Size: 500, Hamming Distance = 0		
Set I	1272	228	84.80
Set II	1162	338	77.47
Set III	854	646	56.93
Set IV	499	1001	33.27

Table 4. Our Hashing Algorithm

	Correct	Incorrect	Percent
	Database Size: 500, Hamming Distance = 0		
Set I	1408	92	93.87
Set II	1389	111	92.60
Set III	1250	250	83.33
Set IV	1028	472	68.53

Table 5. Philips Hashing Algorithm

	Correct	Incorrect	Percent
	Database Size: 500, Hamming Distance <= 1		
Set I	1391	109	92.73
Set II	1342	158	89.47
Set III	1148	352	76.53
Set IV	788	712	52.53

Table 6. Our and Philips Hashing Algorithm

	Our	Philips	Philips
	Database Size: 500		
	HD = 0		HD <= 1
Set I	93.87 %	84.80 %	92.73 %
Set II	92.60 %	77.47 %	89.47 %
Set III	83.33 %	56.93 %	76.53 %
Set IV	68.53 %	33.27 %	52.53 %

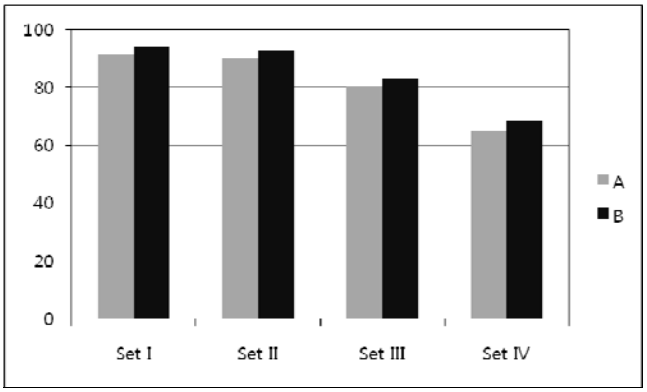

Figure 7. Scalability of the database's music items size.

Our Hashing Algorithm
A: Database Size: 2019
B: Database Size: 500

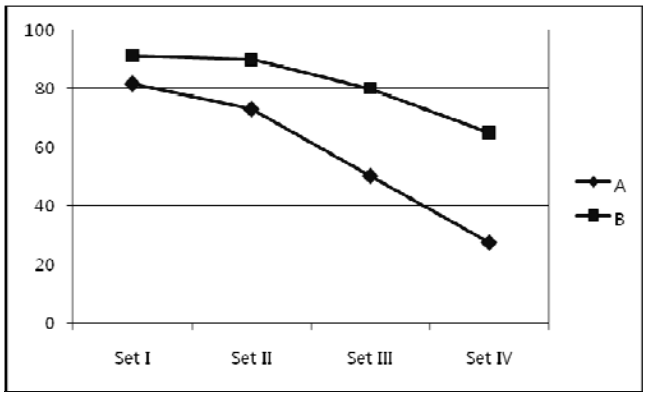

Figure 8. Comparison of performance.

A: Philips Hashing Algorithm
B: Our Hashing Algorithm

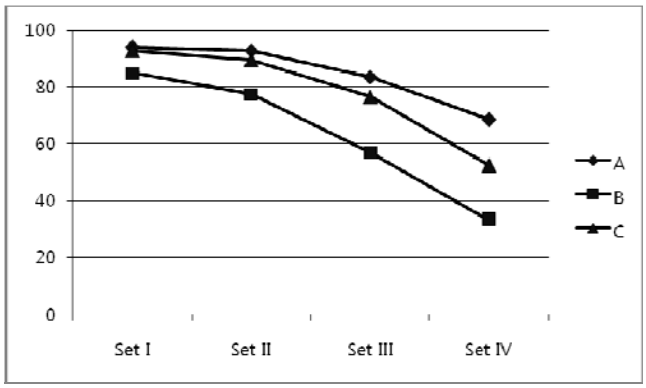

Figure 7. The case of expanded lookup candidates.

A: Our Hashing Algorithm (Hamming Distance = 0)

B: Philips Hashing Algorithm (Hamming Distance = 0)

C: Philips Hashing Algorithm (Hamming Distance <= 1)

7. CONCLUSION

This paper presented a novel audio fingerprinting algorithm to recognize songs in real noisy environment. The proposed algorithm enhances the Philips fingerprint algorithm by creating hash 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 III or IV are much higher than the original Philips fingerprint.. In addition, our system has the scalability of the database size. 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 research.

8. ACKNOWLEDGMENTS

We thank Seoul R&BD Program (10581) for supporting this work.

9. REFERENCES

[1] Alexander Sinitsyn. 2006. Duplicate Song Detection using Audio Fingerprinting for Consumer Electronics Devices. In Proceedings of the 10th IEEE Symposium on Consumer Electronics, pp.1-6.

[2] Jose, Ramon CERQUIDES. 2007. A Real Time Audio Fingerprinting System for Advertisement Tracking and Reporting in FM radio. In Proceedings of the 17th International Conference Radioelektronika 2007. Radioelektronika (17). Num. 17. Brno, Czech Republic. Fryza. 2007. Pag. 455-459

[3] Shazam. 2008. http://www.shazam.com/. Shazam Entertainment Ltd.

[4] Gracenote. 2008. http://www.gracenote.com/. Gracenote Ltd.

[5] J. Haitsma and T. Kalker. 2002. A Highly Robust Audio Fingerprinting System. In Proceedings of the ISMIR 2002, pp. 144-148.

[6] C. Burges, J. Platt, and S. Jana. 2003. Distortion Discriminant Analysis for Audio Fingerprinting. IEEE Trans. Speech and Audio Processing, vol. 11, Mar. 2003, pp. 165-174.

[7] M. L. Miller, M. A. Rodriguez, and I. J. Cox. 2002. Audio Fingerprinting: Nearest Neighbor Search in High-dimensional Binary Space. IEEE Multimedia Signal Processing Workshop, Dec. 2002, pp. 182-185.

[8] D. Kirovski and H. Attias. 2002. Beat-ID: Identifying Music via Beat Analysis. IEEE Multimedia Signal Processing Workshop, Dec. 2002, pp. 190-193.

[9] M. K. Mihcak and R. Venkatesan. 2001. A Perceptual Audio Hashing Algorithm: A Tool for Robust Audio Identification and Information Hiding. LNCS, vol. 2137, 2001, pp. 51-65.

[10] Jungmin Song, So Young Bae, Kyoungro Yoon. 2002. Mid-Level Music Melody Representation of Polyphonic Audio for Query-by-Humming System. In Proceedings of the ISMIR 2002. pp. 133-139

[11] SPIB. 2007. http://spib.rice.edu/. Rice University.