Sub-fingerprint Masking for a Robust Audio Fingerprinting System in a Real-noise Environment for Portable Consumer Devices

Wooram Son, Hyun-Tae Cho, Kyoungro Yoon and Seok-Pil Lee

Abstract — *The robustness of audio fingerprinting system in a noisy environment is a principal challenge in the area of content-based music retrieval, especially for use in portable consumer devices. Our new audio fingerprint method using sub-fingerprint masking based on the predominant pitch extraction dramatically increases the accuracy of the audio fingerprinting system in a noisy environment, while requiring much less computing power for matching, compared to the expanded hash table lookup method, where the searching complexity increases by the factor of 33 times the degree of expansion.*[1].

Index Terms — Audio retrieval, Audio fingerprint, Noisy environment.

I. INTRODUCTION

Recently, content-based multimedia information retrieval gets more and more attention from wired/wireless service providers as well as application developers for portable consumer devices. For instance, it has been reported that there already are available services not only for providing information on the fragment of songs being played over public loudspeakers such as Shazam service for iPhone and other mobile devices, but also for monitoring broadcast for advertisement tracking and reporting. [1, 2] These music identification/retrieval systems are based on audio fingerprinting schemes and the Philips scheme proposed by Haitsma and Kalker [3] is proven to be the most robust and accurate audio fingerprint scheme.

In this scheme, the auditory range in frequency domain is logarithmically divided into 33 sub-bands and the hash values which is the sequence of signs of the differences in energy between each sub-band, both along the time and frequency axis, are obtained to be the sub-fingerprint. But when there is noise, the noise distorts the sub-fingerprints. To compensate the distorted sub-fingerprints, the query for the database lookup are expanded into hash values which are within the Hamming distance of a one-bit error from the original sub-fingerprint, causing additional 33 times of lookup time for audio identification. [4]

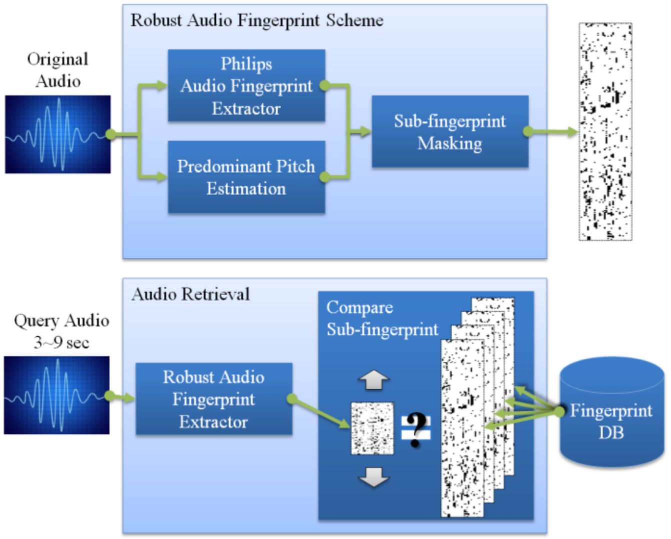
Robust Audio Fingerprint Scheme
Original Audio
Philips Audio Fingerprint Extractor
Predominant Pitch Estimation
Sub-fingerprint Masking
Audio Retrieval
Query Audio 3~9 sec
Robust Audio Fingerprint Extractor
Compare Sub-fingerprint
Fingerprint DB

Fig. 1. Overview of robust audio fingerprinting system.

Mansoo introduces frequency-temporal filtering for a robust fingerprinting scheme in real-noise environment based on Philips scheme and shows that frequency filtering improves noise robustness in audio identification. [5] In practice, however, it still needs further improvement to be used in a real environment. In this paper, we propose a new approach of sub-fingerprint masking using predominant pitch extraction, to improve robustness of the audio fingerprinting in a noisy environment for music identification/retrieval system.

II. AUDIO FINGERPRINT SCHEME

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. Each audio frame produces 32-bit hash values called a sub-fingerprint using Philips' hashing algorithm and 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 also bit-masked using predominant pitch estimation. The search module compares the sequence of bit-masked sub-fingerprints with the

[1] This research was supported by Konkuk University.

W. Son is with Interdisciplinary Program in Radiation Applied Life Science, Seoul National University.

H.T. Cho and K. Yoon* are with Department of Computer Science and Engineering, Konkuk University. * Author for correspondence.

S.-P. Lee is with Korea Electronics Technology Institute.

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]. The audio signal is sampled at the rate of 44100 Hz and segmented into Hanning windowed overlapping frames, each of which contains 512 non-overlapped samples and 15872 overlapped samples. Each frame of 16384 samples is then Fast Fourier Transformed. By logarithmically dividing the obtained audio spectrum, 33 non-overlapping frequency bands from 300 Hz to 2000Hz are acquired. Then total of 32 hash bits are assigned for each frame to become a single sub-fingerprint. A single sub-fingerprint for frame *n*th frame is defined as a bit sequence of F(n,m) for $0 \le \mathrm{m} < 32$ where F(n,m) is defined as equation (1).

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

where F(n,m) is the hash bit for the *n*th frame of the *m*th band and E(n, m) is the energy of the *n*th frame and the *m*th band.

C. *Predominant Pitch Estimation*

The temporal sequence of the harmonic structures in frequency domain is the key to the human perception of music and the recognition of predominant pitch is the process of perceiving the harmonic characteristics of partials. Therefore, we use harmonic enhancement and harmonic summation in spectral domain to estimate predominant pitch in polyphonic sound source. The details of harmonic enhancement and summation algorithms are summarized in the equation (2) and described in previous reports. [6, 7]

$$\mathrm{E}_{\mathrm{t}}^{\mathrm{EP}}(\mathrm{k}) = \sum_{\mathrm{i=-W}}^{\mathrm{W}} \mathrm{A}(\mathrm{E}_{\mathrm{t}}(k) - E_t(k+i)), 0 \le k < N \tag{2}$$

where $A(x) = x, \forall x \ge 0$ and $A(x) = 0, \forall x < 0$.

In equation (2), *N* is the FFT index range, $E_t^{EP}(k)$ represents the degree of the predominance of the harmonic in the frequency index *k*, considering the spectral amplitudes $E_t(k)$ of surrounding signals within the frequency range *W* in time *t*. $E_t^{EP}(k)$ becomes large when the spectral amplitude of the frequency component with the frequency index *k* is larger than the frequency components of surrounding frequency index in the spectrum and it becomes much larger when the spectral amplitude of given index *k* is local maxima within the frequency index range given by the window size W.

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

In equation (3), $\lfloor x \rfloor$ is an integer which is not bigger than *x* and $F_t(p)$ is the average strength of harmonics having the fundamental frequency *p* in the audio frame at time *t*. Above equation shows that $F_t(p)$ is decided by averaging the degree of predominance value obtained by harmonic enhancement process in equidistanced frequency indexes. If there is predominant harmonics of fundamental frequency *p*, the value $F_t(p)$ becomes large and represents the possibility that the sound of fundamental frequency *p* would occur is large.

$$PredominantPitch = \arg\max F_t(p)$$

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 extract musically meaningful high-level attribute based on the predominant pitch estimation. After 32-bit hash values are extracted using ordinary Philips hashing algorithm, the bit-masking is applied to the hash values. The applied bit-mask is a bit pattern identifying the critical band and nearby bands, where the critical band is the band containing the frequency index of the predominant pitch. By doing so, only few bits, identifying the critical band and nearby bands, from the hash value mask are set to 1, e.g. the bit-mask used in our experiment III has only five bits set to 1. Then the sub-fingerprint of equation (1) is redefined as shown in equation (2).

$$F_R(n,m) = \begin{cases} F(n,m) \text{ if } m_p - k \le m < m_p + k \\ 0 \qquad \text{else} \end{cases} \tag{4}$$

$\mathrm{m_p}$ is frequency band which contains the predominant pitch.

As a result, the masked sub-fingerprint contains maximum of only 2k bits set to 1. For example, in our experiment, k is set to 2.

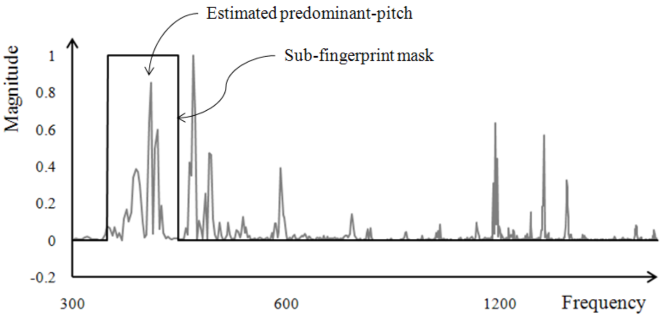

Fig. 2. Sub-fingerprint Masking in frequency domain.

TABLE I
BIT-ERROR-RATE OF SUB-FINGERPRINT BLOCK

Method	Bit-Error (rate)
Original	1026/8256 (12.4 %)
Inner Mask	81/1290 (6.3 %)
Outer Mask	945/6966 (13.6 %)

Figure 3a shows bit errors between original version and added noise version of same excerpt fingerprint block (256 subsequent sub-fingerprints), extracted with Philips scheme from a short excerpt of “Beethoven Sonata 21 (Waldstein) opus 53, in C Major”. Figure 3b and Figure 3c show a sub-fingerprint masked block and outer masked block of the Figure 3a, respectively. Inner mask area show smaller bit errors then outer mask area, as shown in Table I, and we believe that this smaller bit error of the inner mask area can enhance the matching accuracy in noisy environments.

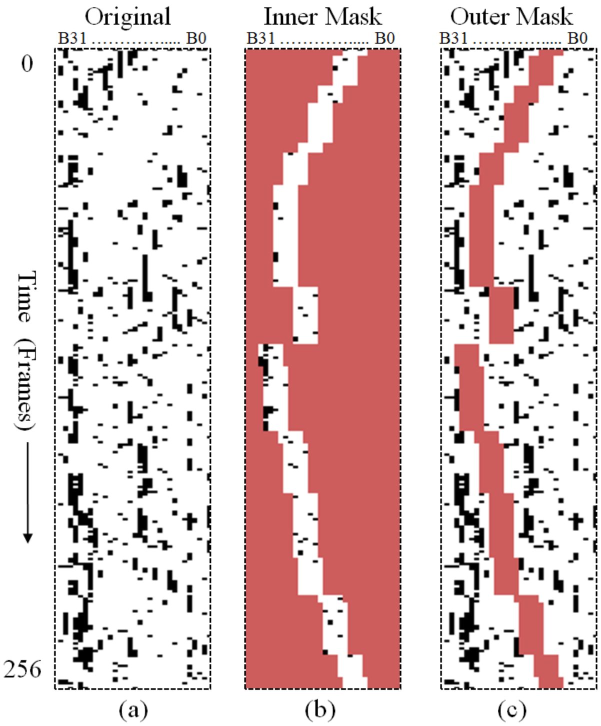

Fig. 3. (a) Fingerprint block of original music clip showing the bit errors in black (BER=0.12), (b) sub-fingerprint masked fingerprint block (BER=0.06), (c) outer sub-fingerprint masked fingerprint block (BER=0.14).

III. EXPERIMENTAL RESULTS

Experiments were performed using a music database containing 2019 popular songs 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 2019 songs, 500 songs are selected. From the selected 500 songs, 1500 randomly created audio query clips of three, six and nine seconds each were captured using a microphone (Sennheiser ME66 model), which was placed 1.5m away from a stereo loudspeaker. With the randomly created 1500 queries, five 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 the proposed algorithm: 1) Our fingerprintng scheme (MBM); 2) Mansoo’s fingerprinting scheme [5]; 3) Philips fingerprinting scheme. The experiments are performed on a system with Windows Server 2003 O/S and Intel Xeon 2GHz CPU with 2GB memory and the followings are the description of the data used:

- *Clean* contains original 1500 query: music clips captured by Senheiser ME66 microphone in a quiet environment located 1.5m away from a stereo loudspeaker.
- *Others* are created by adding *Noise-data* to *Clean* with SNR of 19.1, 7.4, 0, -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 II
COMPARATIVE PERFORMANCE OF THREE SCHEMES

Method / Noise	MBM	Philips	Mansoo (F_2 + T_2)
Clean	97.4	90.5	92.9
19.1 dB	95.1	72.5	77.9
7.4 dB	82.3	32.4	41.7
0 dB	62.5	8.9	13.9
-3.5 dB	48.0	2.8	6.4

%: Recognition accuracy.
Database size = 500, Query length = 6 sec

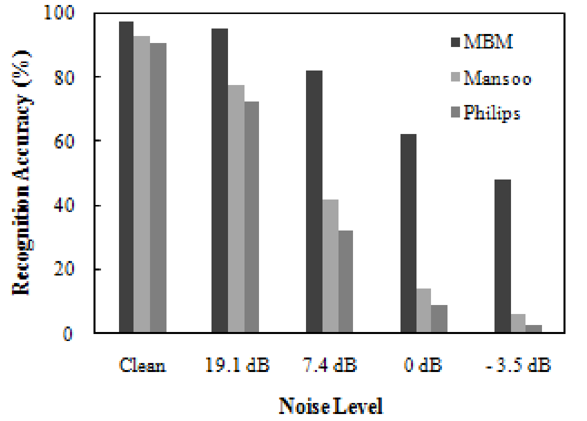

Fig 4. Recognition performance evaluation of MBM, Philips and Mansoo’s audio-fingerprint scheme.

Table II and Figure 4 show the results of the music retrieval experiments performed on a database with five hundred songs based on three different schemes, which are MBM, Mansoo and Philips schemes, using 1500 six seconds long queries.

These results clearly show that MBM scheme outperforms other two schemes in retrieval accuracy, especially in noisy environments.

Table III and Figure 5 show the recognition performance of three schemes, when the lookup candidates of the Philips scheme are expanded by one bit difference for 500 songs database as proposed by Mansoo [5]. As shown in Table III, whether the lookup candidates are expanded or not, MBM scheme show better recognition performance.

TABLE III

RECOGNITION PERFORMANCE WHEN THE LOOKUP CANDIDATES ARE EXPANDED.

Method Noise	MBM HD = 0	Mansoo ($F_2 + T_2$) HD = 0	Mansoo ($F_2 + T_2$) HD <= 1
Clean	97.4	92.9	96.3
19.1 dB	95.1	77.9	87.7
7.4 dB	82.3	41.7	55.5
0 dB	62.5	13.9	25.3
-3.5 dB	48.0	6.4	13.5

%: Recognition accuracy.
HD: Hamming distance.
Database size = 500, Query length = 6 sec

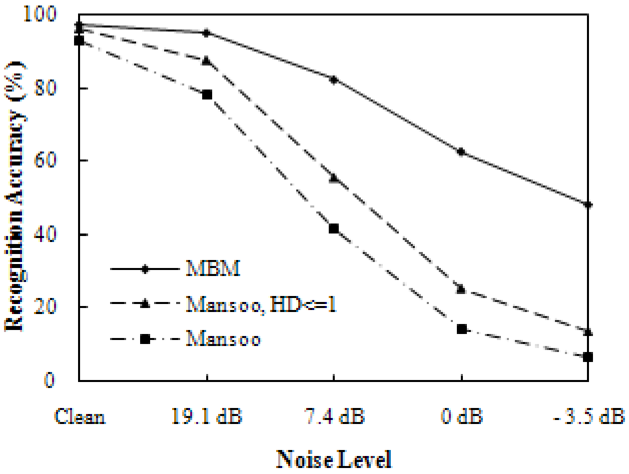

Fig 5. Comparison with MBM and Mansoo scheme when the lookup candidates are expanded.

To check the scalability of the proposed MBM scheme, exactly same experiments were also performed on a database with 2019 songs. The Table IV and Figure 6 show the performance comparisons of the three schemes, when the database size changes from 500 songs to 2019 songs. This result clearly shows that MBM scheme still outperforms the other two schemes with the increased database size. Table V shows recognition performance of MBM scheme when query length are changed. This result shows that the performance increases as the length of the query increases. Also, the proposed scheme shows satisfactory performance with just three seconds long query.

Table VI shows effectiveness of the size of sub-fingerprint mask on MBM scheme. The selection of the sub-fingerprint size can be critical to the performance of the system. The experimental result shows an improvement of the accuracy when the mask size increases in clean environment. In noisy environment, however, the accuracy of the retrieval decreases as the mask size increases. We believe that this is due to the fact that the proposed scheme converges to the original Philips scheme as the mask size increases.

TABLE IV

PERFORMANCE EVALUATION WITH DATABASE SCALABILITY

DB size Noise	MBM		Philips		Mansoo ($F_2 + T_2$)	
	500	2,019	500	2,019	500	2,019
Clean	97.4	96.3	90.5	89.5	92.9	91.7
19.1 dB	95.1	94.2	72.5	69.1	77.9	75.1
7.4 dB	82.3	80.5	32.4	28.1	41.7	37.5
0 dB	62.5	59.8	8.9	6.3	13.9	11.1
-3.5 dB	48.0	45.5	2.8	1.3	6.4	4.1

%: Recognition accuracy.
Query length = 6 sec

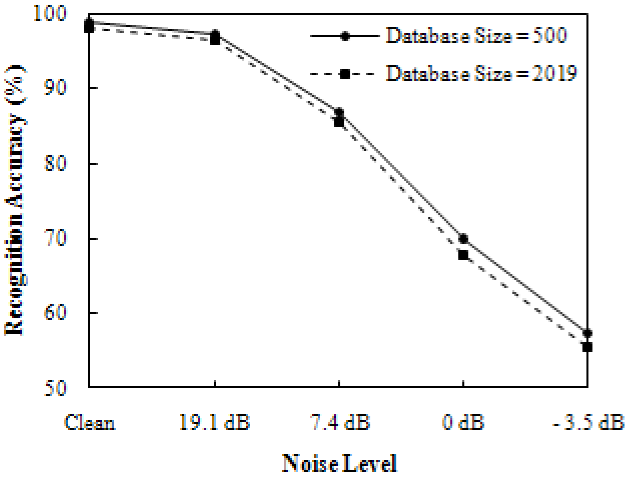

Fig 6. Recognition performance evaluation according to database size.

TABLE V

PERFORMANCE EVALUATION ACCORDING TO QUERY LENGTH

Query Noise	3 sec	6 sec	9 sec
Clean	91.4	96.3	98.2
19.1 dB	83.5	94.2	96.7
7.4 dB	63.2	80.5	85.5
0 dB	39.8	59.8	67.8
-3.5 dB	26.1	45.5	55.5

%: Recognition Accuracy
Database size = 2,019

TABLE VI
PERFORMANCE EVALUATION ACCORDING TO SUB-FINGERPRINT MASK SIZE

Mask	MBM	MBM	MBM
Noise	3-Bit	5-Bit	7-Bit
Clean	95.8	96.3	96.4
19.1 dB	93.9	94.2	92.3
7.4 dB	83.9	80.5	73.4
0 dB	65.9	59.8	47.2
-3.5 dB	55.1	45.5	33.1

%: Recognition Accuracy
Database size = 2,019, Query length = 6 sec

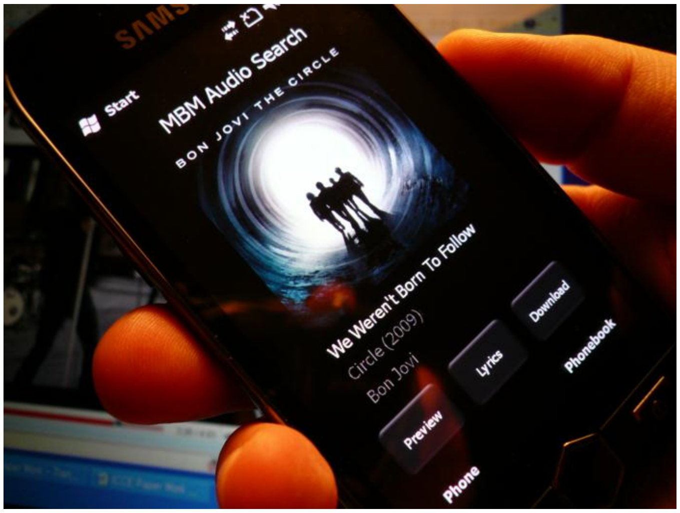

Fig 7. Mobile Application Demonstration

IV. CONCLUSIONS

This paper presented a new modified 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, which, ideally, conveys melody information of a music piece. The proposed algorithm clearly outperforms original Philips algorithm in recognizing polyphonic music in real noisy environment. 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. We believe that this dramatic improvement in the high noise environments is mainly due to the fact that the hash bit masking based on the predominant pitch has the effect of selecting only the sub-bands which have more energy in terms of harmonic structure, emphasizing the most audible sound of music, at the same time decreasing the effect of added noise. Enhancement and optimization of the predominant pitch extraction and hash mask creation can be considered for future work. The analysis and improvement in retrieval time are also considered for future work.

REFERENCES

[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.

BIOGRAPHIES

Wooram Son received a B.S. degree in Computer Science and Engineering from Konkuk University, Seoul, Korea in 2008. He is currently a graduate student working towards his M.S. at the Seoul National University, Seoul, Korea. His research interests include audio identification and content-based multimedia information retrieval.

Hyun-Tae Cho He is currently attending graduate school for M.S degree in Konkuk University, Seoul, Korea. His research interests are in multimedia information retrieval and GPU computing for high-performance multimedia processing and system implementation. He is planning to graduate in February 2011.

Kyoungro Yoon received a B.S. degree in Computer and Electronic Engineering from Yonsei University, Seoul, Korea in 1987, a M.S.E. degree in Electrical Engineering/Systems from University of Michigan, Ann Arbor in 1989, and a Ph.D. degree in Computer and Information Science from Syracuse University in 1999. He was a principal researcher and a group leader in Mobile Multimedia Research Lab, LG Electronics Institute of Technology from 1999 to 2003. He joined the school of Computer Science and Engineering in 2003 as an assistant professor and is an associate professor now. He served as a chair of Ad Hoc Groups on User Preferences and MPEG Query Format of ISO/IEC JTC1 SC29 WG11 (a.k.a. MPEG) and is currently serving as a chair of JPSearch Ad Hoc Group of ISO/IEC JTC1 SC29 WG1 (a.k.a. JPEG). He is also an editor of various international standards such as ISO IS 15938-12, 23005-2, 23005-5 and 24800-3.

Seok-Pil Lee received the B.S., M.S., and Ph.D. degrees in electrical engineering from Yonsei University, Seoul, Korea, in 1990, 1992, and 1997, respectively. From 1997 to 2002, he has been working as a senior research engineer in Digital TV Research Center of DAEWOO Electronics Co., Ltd., Seoul, Korea. Since 2002, he has been working at Korea Electronics Technology Institute (KETI), where he is presently the Director of Digital Media Research Center. His research interests are in the areas of multimedia communication and data broadcasting protocol, personalized broadcasting service. He is a chairman of TVA-Korea and a member of committee of MPEG-Korea.