http://dx.doi.org/10.5392/IJoC.2012.8.2.001

Analysis of Pre-Processing Methods for Music Information Retrieval in Noisy Environments using Mobile Devices

Dae-Jin Kim, Ddeo-Ol-Ra Koo
Tech. Lab, DirectMedia, Seoul, 135-840, Republic of Korea

ABSTRACT

Recently, content-based music information retrieval (MIR) systems for mobile devices have attracted great interest. However, music retrieval systems are greatly affected by background noise when music is recorded in noisy environments. Therefore, we evaluated various pre-processing methods using the Philips method to determine the one that performs most robust music retrieval in such environments. We found that dynamic noise reduction (DNR) is the best pre-processing method for a music retrieval system in noisy environments.

***Keywords**: Music Information Retrieval, Audio Fingerprint, Noise Reduction, Mobile Devices, and Pre-Processing.*

1. INTRODUCTION

Recently, content-based music information retrieval (MIR) systems for mobile devices have attracted great interest. MIR systems perform various functionalities such as music recommendation and music recognition. MIR applications such as Shazam, SoundHound, and Gracenote have already been developed for the iPhone, iPad, and other such mobile devices.

Music retrieval involves searching for music that is played over loudspeakers in public places such as a coffee shop or shopping mall, or even on the street. However, in such environments, the music is accompanied by background noise such as people's voices, vehicular noises, or the sound of machinery.

To develop a music retrieval system, first, it is necessary to create an audio fingerprint that can be matched against those stored in a music database. An audio fingerprint contains short summary information of an audio or a perceptual piece of audio content.

Then, to improve the retrieval rate, when a query is input in a noisy environment, first, it is necessary to find candidate audio matching the query from a lookup table (LUT). This increases the probability of correct music identification.

In this study, we evaluate various pre-processing methods for a hash-based fingerprint system and determine the best one by determining the accuracy of searching for a query from an LUT.

Pre-processing can be carried out using various approaches such as normalization, noise reduction, and filtering. In this study, we evaluate the search accuracy of various such methods by calculating the number of exact matches when searching from robust fingerprints[1].

The hash-based fingerprinting system of the Philips method is used as the base system [2][3]. The experiments were performed for music from three genres—jazz, pop, and classical—recorded at distances of 1, 2, and 3 m from a stereo loudspeaker using a mobile device. The framework of the experiment is shown in Fig. 1. Our study focuses only on the pre-processing stage.

Chapter 2 describes the hash-based fingerprinting system of the Philips method. Chapter 3 describes the tested pre-processing methods. Chapter 4 presents the experimental results and analysis. Chapter 5 presents the conclusion of this study.

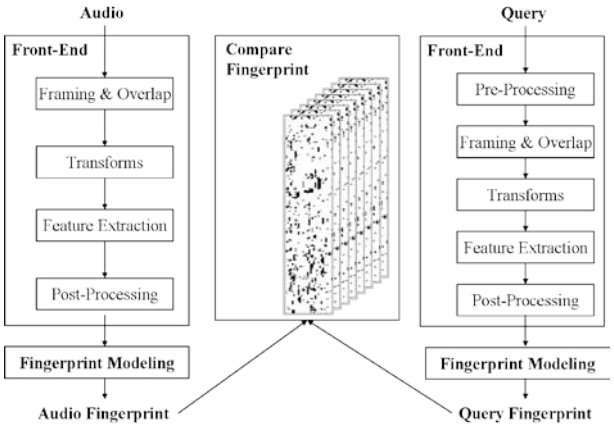


Fig. 1. Framework of experiment.

2. RELATED WORK

2.1 Hash Fingerprint

* *Corresponding author, Email: sampoo00@hanmail.net*
Manuscript received Apr. 25, 2012; revised May 17, 2012; accepted Jun 19, 2012

In this section, we briefly describe hash fingerprints based on the work of Haitsma [2]. The hash fingerprint block consisted of 256 fingerprints. The audio signal framing is segmented into overlapping frames. Next, the audio signal is computed by Fourier transform on every frame. The Fourier transform results are divided into bands in the range of 300–2000 Hz. The energy is calculated on the basis of each sub-band; the energy of band m of frame n is denoted by $E(n,m)$. In order to create a fingerprint block, a 32-bit fingerprint value is extracted for every frame. The m th bit of the fingerprint of frame n is denoted by $F(n,m)$, and it is given as follows:

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

2.2 Music Matching

The LUT consists of the fingerprint DB. Using an LUT was found to be 800,000 times faster than using a music DB for music retrieval [2].

The LUT contains all possible 32-bit fingerprints as entries. These are listed using pointers to the positions in the real fingerprint block lists where the respective 32-bit fingerprints are located. Using the LUT, the input query and the DB's fingerprints from the fingerprint block are compared. Then, the bit error rate (BER) is calculated. If the BER is below a certain threshold, it was considered that there is a high probability that the extracted fingerprint block matches the music stored in the DB.

In our experiment, we find candidates that match the input query from the linked list of the LUT before the BER check in order to find an effective pre-processing method for robust retrieval in noisy environments.

3. PRE-PROCESSING METHODS

This study aims to evaluate various pre-processing methods from the viewpoint of application to an MIR system.

In this section, we briefly describe the pre-processing methods. We considered three main types of methods: normalization, finite impulse response (FIR) filtering, and others.

3.1 Normalization

In audio normalization, the peak amplitude is reduced to a target level or to the average of an audio signal in order to modify the amount of gain. Normalization methods include peak, dynamic range compression (DRC), and root mean square (RMS) normalization.

3.1.1 Peak Normalization

In peak normalization, the gain of an audio signal is modified such that the maximum peak level equals a desired level. In music, the maximum peak level must be lower than the allowed maximum level. Music occasionally has distortion called clipping, a strong type of waveform distortion. Peak normalization can be used to remove such distortion. In this study, we set the maximum peak level at 81 dB.

3.1.2 Dynamic Range Compression Normalization

In DRC, louder parts of the music are softened, effectively making the softer parts louder. The dynamic range is the difference between the loudest and the softest levels when the music is undistorted. These levels are measured in decibels (dB) [9]. The compression basically attenuates the dynamic range of a wave [9]. The threshold and ratio are two important parameters in DRC. The threshold dB is lower than the set threshold, as a result of which a larger portion of the signal is treated [10]. The ratio indicates the compression rate.

In this study, we tested DRC normalization for three sets of threshold and ratio values: 81 dB and 1.5, 89 dB and 1.5, and 89 dB and 4.0.

Fig. 2 shows the change in the output level (dB) for the abovementioned sets of values.

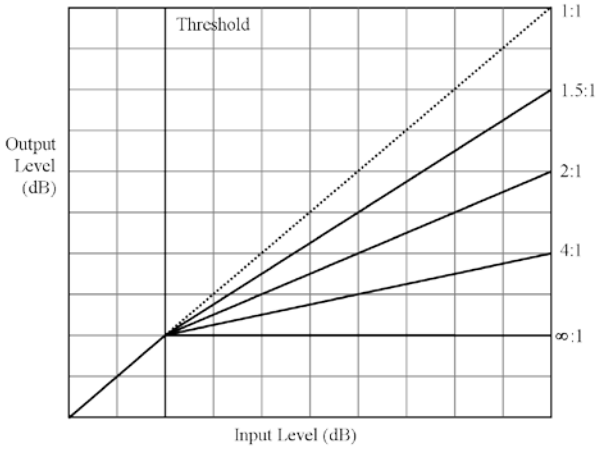


Fig. 2. Threshold and ration in DRC [7].

3.1.3 Root Mean Square Normalization

In RMS normalization, the positive and negative of a sinusoidal signal are measured. For a set of n numbers or values of a discrete distribution, we consider $x_1, x_2, \ldots, x_n$, and the RMS is given as follows:

$$x_{rms}=\sqrt{\frac{1}{n}\left(x_1^2+x_2^2+x_3^2+\cdots+x_{n-1}^2+x_n^2\right)} \quad (2)$$

3.2 FIR Filter

A FIR filter is a type of digital filter. An FIR filter allows only certain input signal levels to pass through. FIR Filters include moving average filters, and low-pass (Fig. 3a), high-pass (Fig. 3b), band-pass (Fig. 3c), and band-stop (Fig. 3d) filters.

Eq. (3) describe an FIR Filter; here, M_1 and M_2 are finite, $x[]$ is the input signal, $y[]$ is the output, and b_k is the impulse response.

$$y[n] = \sum_{k=-M_1}^{M_2} b_k x[n-k] \quad (3)$$

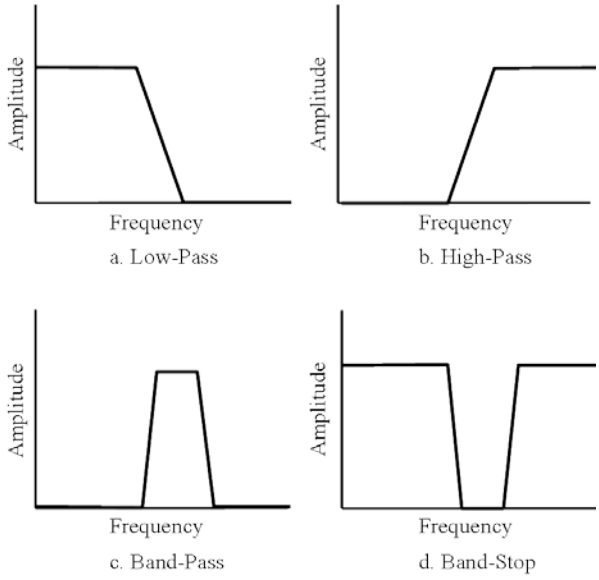


Fig. 3. FIR filters [4].

3.2.1 Moving Average Filter (MAF)

The MAF is a general filter used in digital signal processing (DSP). An MAF complements an average filter (AF). An AF can apply a static signal but not a dynamic one because it removes the element of a dynamic signal. The MAF has a cue and buffer of fixed size. The one applies average inner buffer size to do reflect lately data. The advantage of an MAF is that it is simple, and it reduces random noise while retaining a sharp step response. The MAF shows poor performance for frequency domain encoded signals, but it shows good performance for time domain encoded signals [4].

Eq. (4) describes the MAF; here, $y[]$ is the output signal, $x[]$ is the input signal, $b[]$ is the filter bank, M is the number of filter banks, and N is the sum of each filter bank value.

$$y[i] = \frac{1}{N} \sum_{j=0}^{M-1} x[i + (j-2)]b[j] \quad (4)$$

In this study, we used 5 filter banks having values of 1.0, 4.0, 6.0, 4.0, and 1.0.

3.2.2 Low-Pass FIR Filter: A low-pass filter is a type of FIR filter that passes signals having a frequency lower than a cutoff frequency (f_c) but attenuates signals having a higher frequency. A low-pass filter is also called a high-cut filter or treble cut filter [8]. It removes short-term fluctuations and smoothens the signal [8]. In this study, we set the cutoff frequency at 2kHz.

3.2.3 High-Pass FIR Filter: A high-pass filter is a type of FIR filter that passes signals having a frequency higher than a cutoff frequency (f_c) but attenuates signals having a lower frequency. In this study, we set the cutoff frequency at 2kHz.

3.2.4 Band-Pass FIR Filter

A band-pass filter is a type of FIR filter that passes signal frequencies of a certain range and attenuates other frequencies. Such a filter is obtained by combining a low-pass filter with a high-pass filter. In this study, we set the passband between 1 and 2kHz.

3.2.5 Band-Stop FIR Filter

A band-stop filter, also called a band-rejection filter, is a type of FIR filter that is the opposite of a band-pass filter. It passes most signal frequencies but attenuates those within a certain range. For example, if the stopband is from 70 to 100 Hz, a band-stop filter passes all signal frequencies except those in the range of 70–100 Hz. In this study, we set the stopband between 1 and 2kHz.

3.3 Dynamic Noise Reduction (DNR)

Dynamic noise reduction (DNR) is used to remove background noise, and it can be used in conjunction with other noise reduction techniques. DNR is typically used to block high-frequency hissing noises. DNR can be described as an automatically variable band-pass filter. The bandwidth of a music signal changes rapidly depending on each note, and DNR attenuates frequencies that lie beyond the low- and high-frequency cutoffs that are independently controlled by the input signal [5].

A DNR system operates based on three principles of psychoacoustics. First, white noise can mask pure tones. In order to mask a pure tone, the total noise energy required is equal to the energy of the tone itself. If the masking noise has a wide band, it is necessary to have a low noise amplitude within a certain limit. Unfortunately, the energy of the tone is lower than or equal to the total energy of the noise, which is inaudible. Second, the ear cannot detect a distortion lasting less than 1 ms. Third, reducing the audio bandwidth reduces the audibility of noise. Audibility depends on the noise spectrum. The sensitivity of the human ear increases between 2 and 10kHz. Therefore, noise is audible to humans in this range [6].

Fig. 4 shows the DNR system algorithm. Condition 1 indicates whether a high-pass filtered value lies in the range -*threshold[step]* ≤ value < *threshold[step]*. In this study, the DNR system has a setting gain value of 18 dB. In order to increase the DNR's performance, the output is obtained after the audio signal is passed through eight low- and high-pass filters.

In this system, we set eight thresholds using the setting gain value. The threshold values are set at fractions of the exponent value. The first threshold value is 0. The second threshold value is 1/3 of the exponent of the gain value. The third threshold value is 1/2 of the exponent of the gain value. The others values have existing exponents of the gain value. At each threshold, it is checked whether Condition 1 is satisfied; if it is not, low and high-pass filtering is performed again, and if it is, the DNR system performs noise reduction. The first filter uses the input audio signal; the others use the output of the previous step. The final output is obtained by the summation of each filtered value.

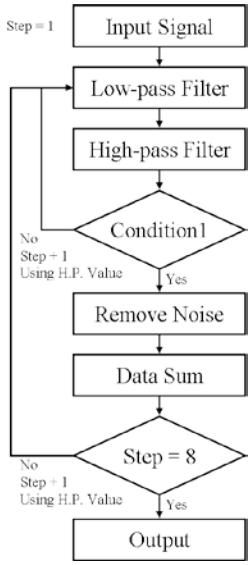


Fig. 4. DNR system algorithm.

4. EXPERIMENT

Experiments were performed based on the Philips method for jazz, pop, and classical music. We tested 3000 songs for each genre. A query is recorded using an iPhone at distances of 1, 2, and 3 m from a stereo loudspeaker.

In order to increase the search accuracy, a query fingerprint should be exactly matched to target music from the LUT. Thus, in this study, we investigate how many exact matches are found when music searching is performed using various pre-processing methods.

In the search, 16 iterations are performed using the input query fingerprint block. In the experiment, searches were performed using queries without filtering and with the various pre-processing methods. In addition, we investigated a combination of pre-processing methods. First, we tested a combination of peak normalization (normalization) and MAF (FIR filter). Second, we tested a combination of peak normalization (normalization) and DNR (noise reduction). This combination removes background noise and modifies the gain of the music.

The pre-processing methods used are described as follows: M1 is no filtering. M2 is DNR with a gain value of 18 dB. M3 is MAF with a filter bank with values of 1.0, 4.0, 6.0, 4.0, and 1.0. M4, M5, M6, and M7 are low-pass, high-pass, band-pass, and band-stop FIR filters, respectively, and these are applied between 30 Hz and 2kHz. M8 is peak normalization with a maximum peak level of 81 dB. M9, M10, and M11 are DRC normalization with a threshold and ratio of 81 dB and 1.5, 89 dB and 1.5, and 89 dB and 4.0, respectively. M12 is RMS normalization. M13 is a combination of peak normalization and MAF, and M14, a combination of peak normalization and DNR; in both these cases, the conditions used for each individual method are applied.

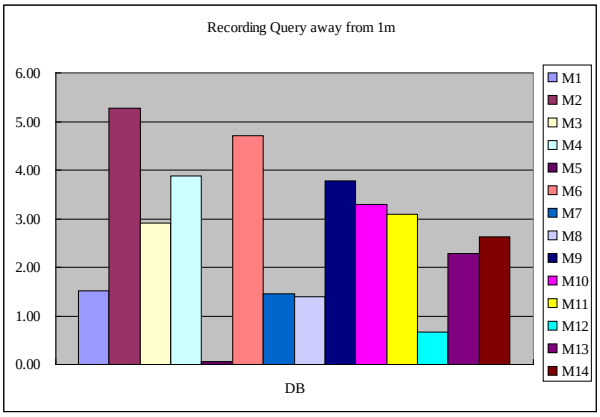
Table. 1. Results without filtering and with the various pre-processing methods.

	DB								
Methods	Pop			Jazz			Classic		
(Query)	1m	2m	3m	1m	2m	3m	1m	2m	3m
M1	0.2	0.1	0.5	1.5	1.4	1.5	2.7	2.5	1.6
M2	6.0	3.6	1.4	3.1	1.9	2.1	6.6	4.0	2.3
M3	0.2	0.1	2.3	4.4	3.4	1.6	4.0	3.1	1.2
M4	2.0	1.3	1.0	3.3	2.1	1.6	6.2	3.9	2.0
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.0
M6	4.5	2.9	2.3	3.6	2.3	2.0	5.9	3.8	1.5
M7	0.5	0.4	0.2	1.7	1.3	0.6	2.1	1.6	1.8
M8	0.0	0.0	1.1	2.0	1.9	0.8	2.2	2.1	2.0
M9	0.8	0.2	0.1	3.9	1.1	0.2	6.5	1.8	0.5
M10	0.9	0.3	0.1	2.7	1.1	0.8	6.1	2.5	0.6
M11	1.2	0.8	0.3	5.2	3.5	1.7	2.8	1.8	2.0
M12	0.0	0.0	0.1	1.2	1.1	0.9	0.7	0.6	0.4
M13	0.4	0.3	3.1	3.4	3.1	1.5	2.9	2.6	1.6
M14	2.4	2.0	1.7	2.1	1.7	1.8	3.3	2.7	1.8

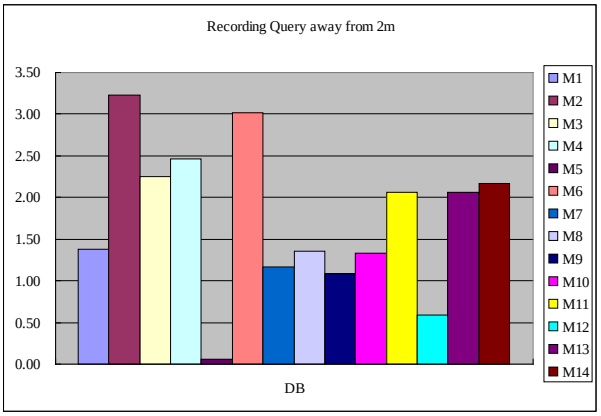
(Unit: Average count of exact matches during music retrieval)

Table 1 shows the average of the number of exact matches found when performing a search using queries without filtering and with the various pre-processing methods. M2 shows the best performance for pop 1 m, pop 2 m, jazz 3 m, classic 1 m, classic 2 m, and classic 3 m. M11 shows the best performance for jazz 1 m and jazz 2 m. M13 shows the best performance for pop 3 m. Overall, M2 shows the best performance, and therefore, its use would be preferable for searching a DB using an input query.

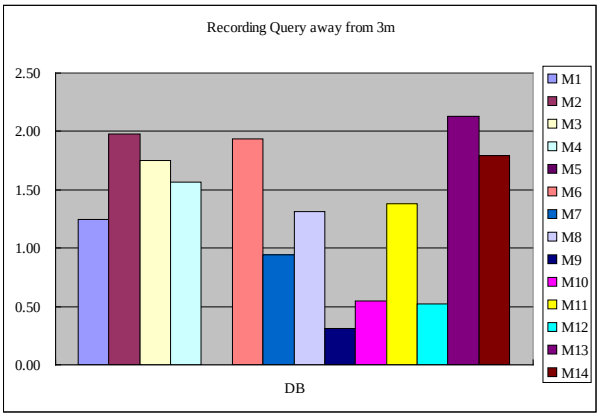
Fig. 5 shows the performance of the various methods for different recording distances from the loudspeaker and no specific music genre. Fig. 5 (A) and (B) show that M2 has the best performance for distances of 1 and 2 m. Fig. 5 (C) shows that M13 has the best performance for a distance of 3 m. However, in this case, the performance of M2 is lower only by 0.15.



(A) Recording distance of 1m.



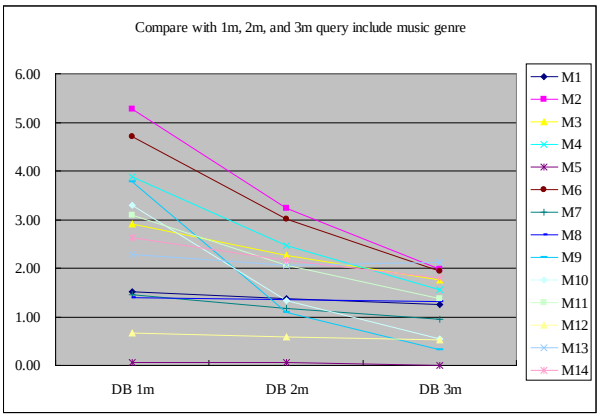
(B) Recording distance of 2m.



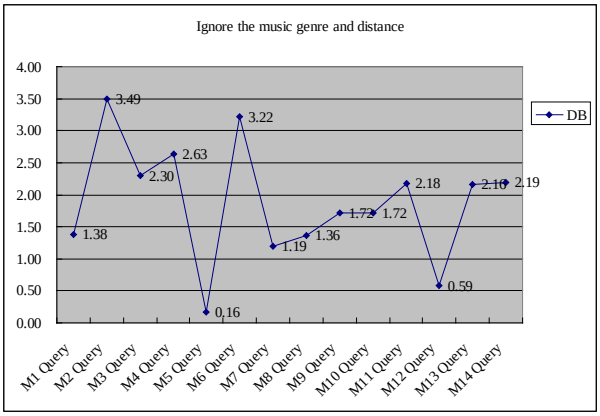
(C) Recording distance of 3m.

Fig. 5. Results for various recording distances.

Fig. 6 shows the performance of the various methods for different and irrespective of the recording distances and irrespective of the music genre. Fig. 6 (A) shows that M2 has the best performance at each recording distance irrespective of the music genre. Fig. 6 (B) shows that M2 has the best performance irrespective of the recording distance and the music genre; in other words, M2 has the best pre-retrieval result with the LUT. Overall, Fig. 5 and 6, show that the M2 method has the best pre-retrieval rate with the LUT.



(A) Results for various recording distances irrespective of the music genre.



(B) Results irrespective of the recording distance and music genre.

Fig. 6. Overall results.

5. CONCLUSION

In this study, we investigated various pre-processing methods for a hash-based MIR system. We performed experiments based on the Philips method for each pre-processing method and without filtering [2]. All methods were tested at various recording distances from a loudspeaker both for three specific music genres and irrespective of the music genre. The pre-retrieval rates of these methods were compared, and the method having the best performance was determined.

The obtained results show that the DNR method has the best pre-retrieval rate for a search query. For developing an MIR system, it would be best to use an unfiltered DB in combination with the DNR method when recording query music in a noisy environment using a mobile device.

In the future, first, we intend to compare each pre-processing method for various DBs and queries. Second, we will record queries from greater distances. Third, we will record queries from additional music genres.

6. 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.

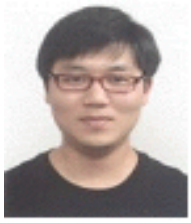
[2] J. Haitsma, and T. Kalker, “A Highly Robust Audio Fingerprinting System,” *Proc. Of the 3rd Int. Symposium on Music Information Retrieval*, 2002, pp.144-148.

[3] 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, Feb. 2010, pp. 156-160.

[4] S. W. Smith, *The Scientist and Engineer’s Guide to Digital Signal Processing*, CA: California Technical Publishing, San Diego, 2006.

[5] R. S. Burwen, “A Dynamic Noise Filter for Mastering,” *Audio Magazine*, Jun, 1972.
[6] LM1984, *Dynamic Noise Reduction System DNR*, National Semiconductor, Apr, 2002.
[7] M. K. Mihcak and R. Venkatesan, “A Perceptual Audio Hashing Algorithm: A Tool for Robust Audio Identifcation and Information Hiding,” LNCS, vol. 2137, 2001, pp. 51-65.
[8] http://en.wikipedia.org/wiki/Low-pass_filter
[9] http://forum.recordingreview.com/f8/dynamic-range-compression-normalization-31861/
[10] http://en.Wikipedia.org/wiki/Dynamic_range_compression

Dae-Jin Kim
He received the B.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 the Tech Lab, DirectMedia. His main research interests include multimedia information retrieval, codec, fingerprinting, watermark, video cartooning, and multimedia system.



Ddeo-Ol-Ra Koo
He received the B.S. in multimedia from Namseoul university, Korea in 2008 and also M.S in advanced imaging science from Chung-Ang university, in 2010. Since then, he has been with the Tech Lab, DirectMedia. His main research interests include multimedia information retrieval, fingerprinting, augmented reality, and human computer interaction.