Robust Audio Fingerprinting Method Using Prominent Peak Pair Based on Modulated Complex Lapped Transform

Hyoung-Gook Kim and Jin Young Kim

The robustness of an audio fingerprinting system in an actual noisy environment is a major challenge for audio-based content identification. This paper proposes a high-performance audio fingerprint extraction method for use in portable consumer devices. In the proposed method, a salient audio peak-pair fingerprint, based on a modulated complex lapped transform, improves the accuracy of the audio fingerprinting system in actual noisy environments with low computational complexity. Experimental results confirm that the proposed method is quite robust in different noise conditions and achieves promising preliminary accuracy results.

Keywords: Audio fingerprinting, modulated complex lapped transform, content identification, robust matching.

Manuscript received Dec. 26, 2013; revised May 7, 2014; accepted July 4, 2014.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A2007601). And this work was supported by Information Technology Research Center (NIPA-2014-H0301-14-1019), and Kwangwoon Research Grant 2013.

Hyoung-Gook Kim (corresponding author, hkim@kw.ac.kr) is with the Department of Wireless Communication Engineering, Kwangwoon University, Seoul, Rep. of Korea.

Jin Young Kim (beyondi@jnu.ac.kr) is with the Department of Electronic and Computer Engineering, Chonnam National University, Gwangju, Rep. of Korea.

I. Introduction

Audio fingerprinting techniques are used for successfully performing content-based audio identification even when audio signals are distorted. The audio fingerprints have been mainly applied to two identification usage modes: natural audio fingerprinting [1] and artificial audio fingerprinting [2]–[3]. Natural audio fingerprints are typically invariant characteristics of an audio signal, while artificial audio fingerprints are information embedded in an audio signal (typically following a watermarking strategy) to identify the origin of the audio signal.

In this paper, we focus on an acoustic audio fingerprinting system to detect natural-audio fingerprints.

Since smart consumer devices have become more ubiquitous, several applications of audio fingerprinting have been installed in mobile devices. Common uses include query-by-example music or advertisement identification [4]–[5], broadcast monitoring [6], copyright detection, filtering for file sharing, and automatic audio-based content library organization [7].

A successful audio fingerprinting system needs to satisfy several practical requirements [8]. First, it should be able to identify corrupted audio clips in spite of degradations caused by various noisy environments or distance from the audio source. Second, it should be able to identify audio clips that are only a few seconds long. Finally, it should be computationally efficient — both in calculating the fingerprints and in searching for the best match in the database.

Various methods [9] have been proposed to satisfy the aforementioned practical requirements. Among the various algorithms, the system developed by Wang [10] has become

commercially successful. The robust hash algorithm proposed by Haitsma and others [7] is also a well-studied content-based music identification or retrieval technique.

In Wang's method, each audio track is analyzed using the short-time Fourier transform (STFT) to find local prominent peaks concentrated in frequency. These peaks are formed into pairs within a target area, which is parameterized by the frequencies of the peaks and the times in-between them. These values are quantized to give a relatively large number of distinct landmark hashes. To identify a query, it is similarly converted into landmarks. Then, the database is queried to find all the reference tracks that share landmarks with the queries and to find the relative time differences between where they occur in the query and where they occur in the reference tracks. Its weakness is that it is not suited for pitch-shifted or time-stretched audio; the likes of which frequently occur in the context of broadcasting monitoring.

Based on the idea of Wang' algorithm, Pan and others [11] introduced a local energy centroid for generating an audio fingerprint, while a real-time peak-discovering method was proposed by Jiang and others [12] for audio fingerprinting.

In Haitsma's method, an audio track is segmented into overlapping frames, and then the spectrum of each, in each frame, is logarithmically divided into 33 sub-bands. Finally, the fingerprints are determined by the relationship of the energy in adjacent sub-bands. To compensate the distorted sub-fingerprints, the query for the database lookup is expanded into hash values within the Hamming distance of a one-bit error from the original sub-fingerprint, resulting in 33-times more lookup time for audio identification. Its drawbacks are that the amount of information is relatively large and that there is poor performance with a low signal-to-noise ratio (SNR) [11].

Park and others introduced frequency-temporal filtering for a robust fingerprinting scheme in an actual noisy environment based on Haitsma's algorithm and showed that a frequency-temporal filtering combination achieves robustness to channel and background noise in music identification [13]. Son and others [14] proposed a masking, generated by predominant pitch estimation, on each sub-fingerprint of Haitsma's robust hashing algorithm. Based on the ideas of both Wang and Haitsma, masked audio spectral keypoints for robust audio fingerprinting was proposed by Anguera and others [15].

To improve the accuracy of the audio fingerprinting system, a fingerprint should both capture and characterize the essence of the audio content. In this paper, based on the idea of Wang's method, a novel audio fingerprint extraction method based on the modulated complex lapped transform (MCLT) [16] is proposed to improve the robustness of audio fingerprinting in an actual noisy environment for an audio-based content identification system and in the context of broadcasting monitoring.

The contributions of this paper are as follows: (a) MCLT-based spectral peaks are estimated and provided to preserve the majority of the sound's peaks more effectively than STFT-based spectral peaks; (b) using emphasis filtering, the spectral peaks in the high-frequency bin are enhanced for generating robust peak pairs against attenuation distortions; (c) to obtain salient peak pairs against different types of noise and at different distances from the audio source, a dynamic peak-picking threshold based on linear interpolation was used; (d) the proposed algorithm improves the robustness of the audio fingerprinting in various real environments; and (e) it is computationally efficient, delivers high identification accuracy (in spite of a short query), and is suitable for use in any practical mobile phone.

This paper is organized as follows. Section II describes the proposed method. Section III discusses the experimental results. Finally, Section IV presents the conclusion.

II. Proposed Robust Audio Fingerprint Generation in Portable Consumer Devices

The two key components of the proposed fingerprinting system are the fingerprint server and one or more fingerprint clients. A fingerprint client, such as a portable consumer device, captures an audio clip that is a few seconds long and then extracts a robust fingerprint based on an MCLT peak pair and submits it to the fingerprint server. The extracted fingerprint is then used to query the fingerprint database at the server and is compared with the stored fingerprints. If a match is found, then the resulting track identifier is retrieved from the server database.

For robust fingerprint extraction against noise and distortion, we propose to use MCLT peak pairs. As shown in Fig. 1, the robust MCLT peak pair–based fingerprint extraction method

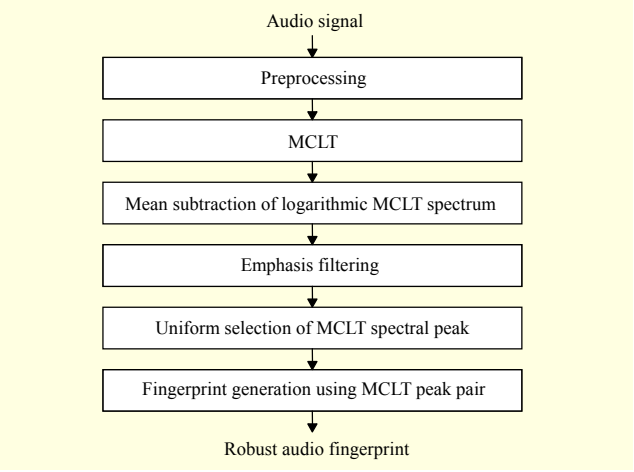

Fig. 1. Block diagram of robust audio fingerprint extraction.

is composed of six main blocks.

First, a stereo audio signal, captured by a user's mobile phone, is converted into mono and then downsampled to 16 kHz. The converted signal is divided into overlapping frames by the application of a Hanning window function (each of which contains 512 overlapped samples). To find the spectral peaks, an MCLT is then applied to each frame (1,024 samples). A log spectrum is generated by taking the log modulus of each MCLT coefficient. From the logarithmic MCLT spectrum, a frequency-time averaged MCLT spectrum is calculated and subtracted, thus yielding a normalized logarithmic MCLT spectrum. To increase the local spectral peaks of high frequencies against attenuation distortion, an emphasis filter is applied to each normalized logarithmic MCLT spectrum. The emphasis-filtered MCLT spectral peaks are fed into a uniform selection step, where the salient peaks are selected by applying appreciative forward and backward filtering using a dynamic peak-picking threshold. In a local target area of the frequency-time plane, nearby salient MCLT peaks are combined into a pair or landmark. Landmarks are 3-tuples, using start frequency, frequency difference, and time difference of the pair of peaks, and are converted into hashes with a 32-bit value. The robust fingerprint generated in consumer devices is submitted to the fingerprint server for content-based identification.

1. Time-to-MCLT and Its Logarithmic Mean Subtraction

First, the audio signal, $s(n)$, is segmented into a Hanning-windowed overlapping frame and analyzed using the MCLT (which has a complex-valued portion based on a discrete Fourier transform), and is given by

$$S_{\mathrm{MCLT}}(k, l) = \left| jV(k, l) + V(k+1, l) \right|, \quad (1)$$

using

$$V(k, l) = b(k, l) \cdot U(k, l), \quad (2)$$

$$U(k, l) = \sqrt{\frac{1}{2N}} \sum_{n=0}^{2N-1} x(n + lM) \cdot h(n) \cdot \exp\left(\frac{-j2\pi kn}{N} \right), \quad (3)$$

$$b(k, l) = W_8(2k+1, l) \cdot W_{4K}(k, l), \quad (4)$$

and

$$W_T(r, l) = \exp\left(\frac{-j2\pi r}{T} \right), \quad (5)$$

where k is the frequency bin index, l is the time frame index, h is an analysis window of size N, and M is the framing step. Also, $U(k, l)$ is a $2N$-point fast Fourier transform (FFT) with an orthonormal basis function of the input block $s(n)$. This means that the MCLT coefficients can be generated by computing the FFT of $s(n)$ to obtain $U(k, l)$ and by carrying out the operations with factor $b(k, l)$. Unlike orthogonal transforms, such as FFT, using $b(k, l)$ and complex-valued transform coefficients, the MCLT has a significant overlap in its frequency response for the basis functions and provides twice-frequency resolution. Therefore, the MCLT has approximate shift invariance properties [17]. The spectral peaks detected by the MCLT preserve the majority of the original sound's peaks more effectively than the STFT-based spectral peaks, against different distortions caused by additive noise, additive echo, and coding artifacts; a sufficient number of peak pairs can, therefore, be identified as coming from the same reference track. Figure 2 illustrates the spectrogram of the MCLT compared to FFT.

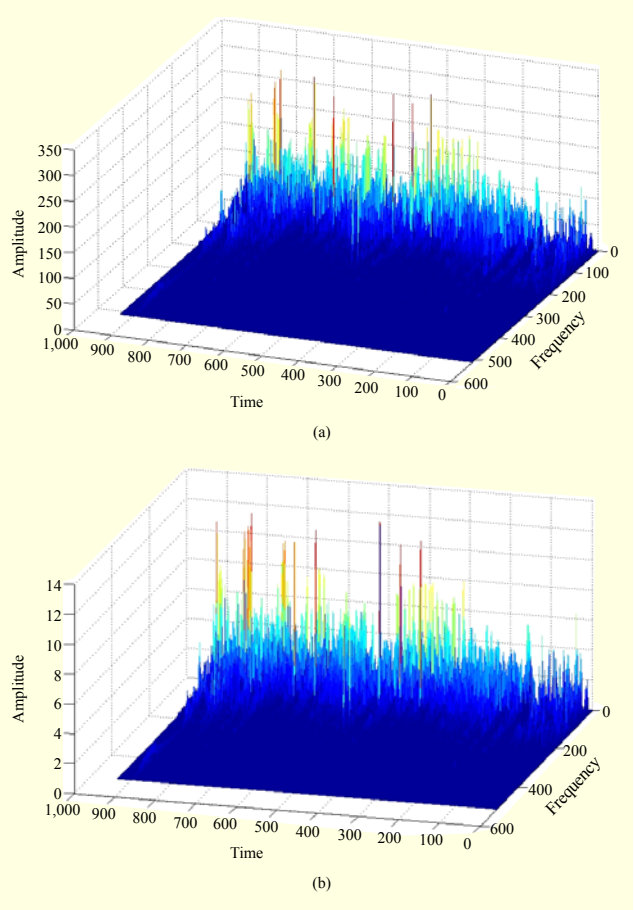

Fig. 2. Spectrogram of MCLT compared to FFT: (a) spectrogram of FFT and (b) spectrogram of MCLT.

Using a comparison of the searched minimum value of the previous frame and the MCLT spectrum, the minimum $S_{\mathrm{min}}(k, l)$ of the local energy is searched for in all frames. From the searched minimum values, the maximum of $S_{\mathrm{min}}(k, l)$ in each frequency bin is obtained by the following:

$$S_{\max}(k, l) = \max\left\{ S_{\mathrm{MCLT}}(k, l), S_{\min}(k-1, l) \right\}, \; 0 \le k < K. \quad (6)$$

Then, a logarithmic operation of each MCLT coefficient is performed by comparison of the MCLT spectrum and $S_{max}(k, l)$, and is given by

$$\begin{aligned} &\text{If} && S_{\text{MCLT}}(k, l) > S_{\max}(k, l), \\ &\text{then} && S_{\log}(k, l) = \log_{10}\left(S_{\text{MCLT}}(k, l)\right); \\ &\text{else} && S_{\log}(k, l) = \log_{10}\left(S_{\max}(k, l)\right). \end{aligned} \tag{7}$$

From the logarithmic MCLT spectrum $S_{log}(k, l)$, the mean S_{mean} of the MCLT spectrum is estimated and subtracted in every frame to minimize the ripple in the low and high ends of the logarithmic MCLT spectrum.

$$S_{\text{norm}}(k, l) = S_{\log}(k, l) - S_{\text{mean}}, \tag{8}$$

where

$$S_{\text{mean}} = \frac{1}{K \cdot L} \sum_{l=0}^{L-1} \sum_{k=0}^{K-1} S_{\log}(k, l). \tag{9}$$

2. Emphasis Filtering and Uniform Selection of Salient MCLT Spectral Peaks

An emphasis filter is applied to the normalized logarithmic MCLT spectrum by

$$S_{\text{F}}(k,l) = \sum_{j=-Q}^{Q} F(k, j) \cdot S_{\text{norm}}(k, l-j). \tag{10}$$

Generally, high-frequency components of the audio signal are susceptible to the noise effect. The emphasis filter is designed to increase the magnitude or spectral peaks of high frequencies with respect to the magnitude of other (usually lower) frequencies. The emphasis filtering is performed across successive frames and frequency bins. This improves the overall SNR of the normalized logarithmic MCLT spectrum by minimizing the adverse effects of phenomena such as attenuation distortions, which do not degrade subjective sound quality. In the present implementation, the band-specific non-causal FIR with $Q = 10$ is used.

An emphasis-filtered time-frequency point, $S_F(k, l)$, is a candidate peak if it has a higher energy content than all its neighbors in a region centered around the point. Candidate peaks are chosen according to a density criterion to assure that the time-frequency strip for the audio file has reasonably uniform coverage. The peaks in each time-frequency locality are also chosen according to amplitude, with the justification that the highest amplitude peaks are most likely to survive the distortions.

Figure 3 depicts the uniform selection of MCLT spectral peaks to obtain noise-robust salient peaks. The noise-robust salient peak extraction procedure based on forward and backward filtering using dynamic peak-picking threshold is as follows.

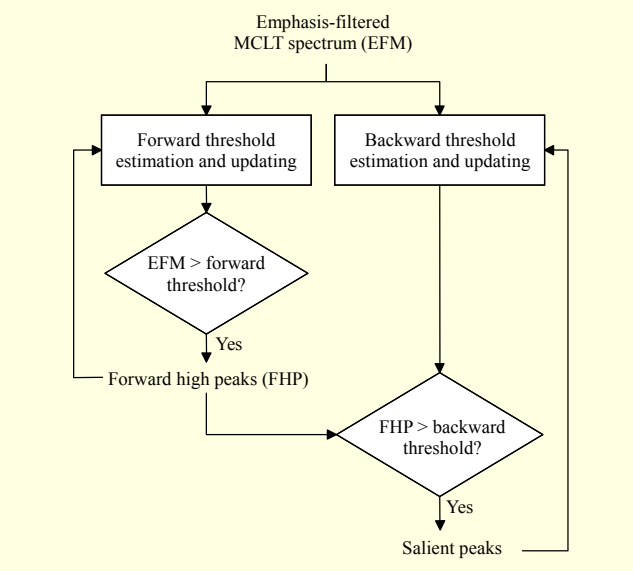

Fig. 3. Block diagram of uniform selection of MCLT spectral peak using forward and backward filtering.

Step 1. Initial forward threshold computing. A frame-wise comparison of the emphasis-filtered MCLT spectral peak and the maximum value of the previous frame within J frames yields the maximum value for each frequency bin, given by

$$F_{\max}(k, 0) = \max\left\{S_{\text{F}}(k, l), F_{\max}(k, l-1)\right\}, \; 0 \le l \le J. \tag{11}$$

Using the basic assumption that a transition from a positive to a negative slope occurs, transited high peaks, $F_s(k, l)$, are selected from $F_{max}(k, l)$. Each transited peak is linearly interpolated and used as an initial forward threshold, $T_f(k, 0)$, for the first frame.

$$\begin{aligned} &\text{If} && \left(F_{\text{S}}(k_{\text{pre}}, l) > F_{\text{S}}(k, l)\right), \\ &\text{then} && T_{\text{f}}(k, 0) = F_{\text{S}}(k_{\text{pre}}, 0) - \frac{F_{\text{S}}(k_{\text{pre}}, 0) - F_{\text{S}}(k, 0)}{k - k_{\text{pre}}} k`; \\ &\text{else} && T_{\text{f}}(k, 0) = F_{\text{S}}(k_{\text{pre}}, 0) + \frac{F_{\text{S}}(k, 0) - F_{\text{S}}(k_{\text{pre}}, 0)}{k - k_{\text{pre}}} k`, \end{aligned} \tag{12}$$

where $k`$ ($k_{pre} < k` < k$) is the frequency index between k_{pre} and k, and k_{pre} is a previous frequency index where the transited peaks are found.

Step 2. Forward high peak selection and dynamic peak-picking threshold update. All peaks of $S_F(k, l)$ higher than the forward threshold $T_f(k, l)$ are stored in a set of (k, l) tuples named FHPs.

$$S_{\text{F}}(k,l) = \begin{cases} \text{FHP} & \text{if } S_{\text{F}}(k,l) > T_{\text{f}}(k,l), \\ \text{non-FHP} & \text{otherwise.} \end{cases} \tag{13}$$

If the FHP is selected among $S_F(k, l)$, then the FHP is

http://dx.doi.org/10.4218/etrij.14.0113.1405

represented as $P_f(k, l)$ and the peak-picking threshold is updated by raising the previous threshold with the linear interpolation of all new peaks.

$$\text{If} \quad \left(P_f\left(k_{pre}, l\right) > P_f\left(k, l\right)\right),$$

$$\text{then} \quad T_{fup}\left(k, l\right) = P_f\left(k_{pre}, l\right) - \frac{P_f\left(k_{pre}, l\right) - P_f\left(k, l\right)}{k - k_{pre}} k`; \quad (14)$$

$$\text{else} \quad T_{fup}\left(k, l\right) = P_f\left(k_{pre}, l\right) + \frac{P_f\left(k, l\right) - P_f\left(k_{pre}, l\right)}{k - k_{pre}} k`.$$

The new peak-picking threshold for the next frame is dynamically obtained by comparing the previous threshold attenuated with a decay factor, $d(k, l)$, with the updated forward peak-picking threshold $T_{fup}(k, l)$.

$$T_f\left(k, l\right) = \max\left(d\left(k, l-1\right) \cdot T_f\left(k, l-1\right), T_{fup}\left(k, l\right)\right), \quad (15)$$

using the following decay factor:

$$d\left(k, l-1\right) = \exp\left(\lambda \cdot \left|\frac{T_f\left(k, l-1\right) - m}{\sigma}\right|\right), \quad -1 \le \lambda \le 1, \quad (16)$$

where m and σ represent the mean and variance of the frequency bands of each threshold value, respectively. The threshold using the decay factor helps to extract more salient peak-pairs for comparing the query fingerprint with the original fingerprints.

Figure 4 illustrates examples of steps 1 and 2. The upper red line in Fig. 4 represents the dynamic peak-picking threshold, while the lower red line represents the updated dynamic peak-picking threshold. The green line represents the forward higher peaks, which are higher than the dynamic peak-picking threshold.

Step 3. Steps 2 to 3 are repeated for $l = l + 1$ until all frames are processed.

Step 4. Backward high peak extraction. To extract the noise-distorted robust salient peaks, steps 1 to 3 are repeated but from the last frame back and considering only (k, l) tuples already in FHP. This is referred to as "pruning".

$$P_f\left(k, l\right) = \begin{cases} \text{salient peak} & \text{if} \quad P_f\left(k, l\right) > T_b\left(k, l\right), \\ \text{non-salinet peak} & \text{otherwise}, \end{cases} \quad (17)$$

using

$$T_b\left(k, l\right) = \max\left(d\left(k, l-1\right) \cdot T_b\left(k, l-1\right), T_{bup}\left(k, l\right)\right). \quad (18)$$

$$\text{If} \quad \left(P_b\left(k_{pre}, l\right) > P_b\left(k, l\right)\right),$$

$$\text{then} \quad T_{bup}\left(k, l\right) = P_b\left(k_{pre}, l\right) - \frac{P_b\left(k_{pre}, l\right) - P_b\left(k, l\right)}{k - k_{pre}} k`; \quad (19)$$

$$\text{else} \quad T_{bup}\left(k, l\right) = P_b\left(k_{pre}, l\right) + \frac{P_b\left(k, l\right) - P_b\left(k_{pre}, l\right)}{k - k_{pre}} k`,$$

where $P_b(k, l)$ is the salient peak after backward filtering using the dynamic peak-picking threshold $T_b(k, l)$, and $T_{bup}(k, l)$ is the updated backward peak-picking threshold.

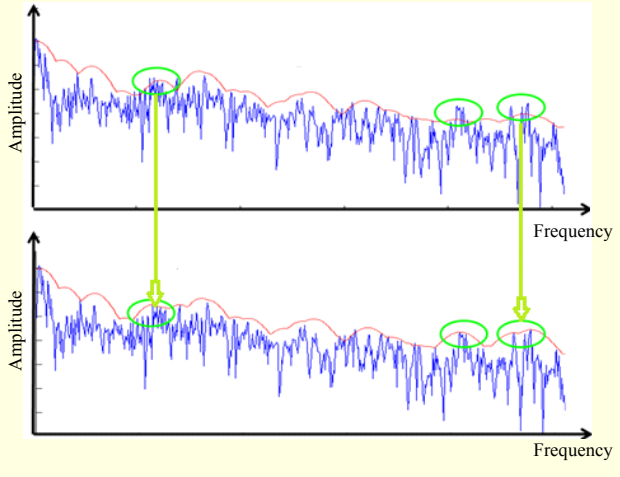

Fig. 4. Dynamic peak-picking threshold updating.

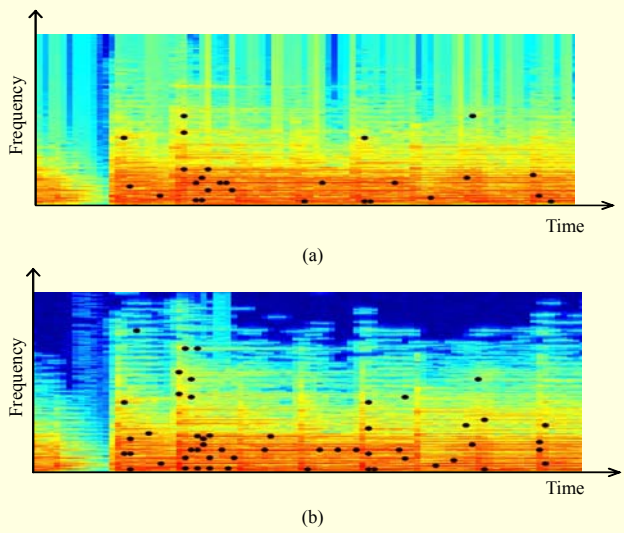

Fig. 5. Salient peaks of proposed MCLT-based method compared to Wang's approach: (a) salient peaks of Wang's approach and (b) salient peaks of proposed MCLT-based method.

Figure 5 presents the salient peaks (black dots) discovered by the proposed method compared to those of Wang's approach. As shown in Fig. 5, the proposed method provides more salient spectral peaks than Wang's approach so as to improve the accuracy of the identification.

3. Fingerprint Hashes Using MCLT Peak Pair

Fingerprint hashes are generated by associating the time-frequency information of $P_b(k, l)$ pairs. As in Wang's approach, we use pairs of peaks. However, a new type of hash is defined and used to improve the identification accuracy against time

stretching and pitch shifting. Each peak pair (called a landmark) is selected as an anchor point and paired with nearby landmarks from within a defined zone (including the pairing horizon in time and in frequency). Assuming that $P_b(k_a, l_a)$ is the anchor point and that it is paired with another landmark point $P_b(k_p, l_p)$, its hash, h, is obtain by

$$h = \left\{k_a\left(l_p - l_a\right),\ k_p - k_a,\ \left(k_p - k_a\right)\left(l_p - l_a\right)\right\} = \left(k_a \Delta l,\ \Delta k,\ \Delta k \Delta l\right). \quad (20)$$

All k (in frequency bins) and l (in frames) are integers with a fixed higher bound, so each landmark point generates a fixed number of pairs. The first component, $k_a\Delta l$, is a rough frequency location of the pair of peaks. The second component, Δk, is the spectral extent of the pair in the MCLT domain and provides good robust performance under time stretching or time-scale modification. Time-scale modification refers to the process of changing the speed or duration of an audio signal without affecting its pitch. Due to the vertical frequency axis, Δk is limited to a lower value than k_p. In addition, the component Δl is the time extent of the pair in the MCLT domain and provides good robust performance under pitch shifting. Pitch shifting refers to the process of changing the pitch of an audio signal without affecting its speed. Due to the horizontal time axis, Δl is limited to a lower value than l_p. As this hash takes into account relative time–frequency information (that is, the third component $\Delta k\Delta l$), it is robust to cropping, time stretching, and pitch shifting within the defined zone. If the time stretching and pitch shifting occur over the defined zone, then the identification is decreased.

Finally, the generated hash extracted in the mobile phone is used as a query for audio fingerprinting.

4. Building of Fingerprint Database and Identification

When building the fingerprint database at the server, a database index is created by a fingerprint hash, and a Track ID and time-frequency offset of the hash are stored according to the hash value to facilitate fast processing.

In retrieval or identification processing [9], the similarity searches of audio are performed in the fingerprinting domain. A query signal is fingerprinted in the user's mobile phone, and the resulting hashes are compared against the hashes stored in the database hash table. After all matching hashes are found, a candidate set, c_{set}, of match segments can be obtained by combining the Track ID stored in the database and the time–frequency offset of the hash in the query audio as follows:

$$c_{set} = \left(ID, \Delta k \Delta l\right). \quad (21)$$

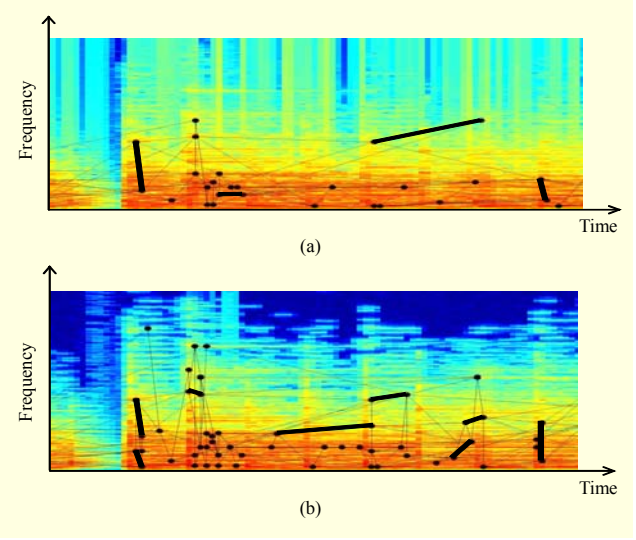

Fig. 6. Matching results of proposed MCLT-based method compared to Wang's approach: (a) matching results of Wang's approach and (b) matching results of proposed MCLT-based method.

The differences $\Delta k\Delta l$, between where the relative frequency–time extent occur in the query fingerprint and where they occur in the candidate set of fingerprints (references), are stored in the histogram (one histogram per reference). If the query frame corresponds to the reference m starting at time l, then its fingerprint will have more pairs in common with m's fingerprint than with any another reference fingerprint and will show a higher matching histogram than any other matching histogram. That is, the Track ID with the largest number of the same (or very similar) frequency–time extent is determined to be the matched result.

Figure 6 represents the matching results (black lines) of the proposed method compared to those of Wang's approach. Since the proposed method provides a larger matching number than Wang's approach, as shown in Fig. 6, the proposed method can obtain a higher identification accuracy than that of Wang's approach.

In addition, a statistical filter to remove the continuity and redundancy was applied to the matching process for improving the query and response accuracy from the database.

III. Experimental Results

In this subsection, the performance of the proposed MCLT peak-pair fingerprint extraction algorithm is evaluated. Additionally, the performance of the algorithm is compared with the modified implementations of four previous methods. Method 1 is an STFT-based peak-pair fingerprint extraction method proposed by Wang [10], while Method 2 is an audio fingerprint method using sub-fingerprint masking based on the

predominant pitch extraction [11]. Method 3 is an audio fingerprint extraction based on the masked audio spectral keypoints [15]. Method 4 is a local feature extraction from adaptively scaled patches of the time-chroma representation of the audio signal [18].

For experiments, three test database types were selected. Set I consists of a database of 7,000 songs from different genres such as pop, hip-hop, jazz, and classical. Set II is a database containing 4,000 TV advertisements with a total time amounting to 740 hours, and each advertisement ranges from 10 to 15 minutes in length. Set III comprises a database of 6,000 TV programs from various genres, such as drama, shows, comedy, and animation.

All of the audio data are stored in PCM format with mono, 16-bit depth, and 16 kHz sampling rate converted from real audio data in consideration of portable devices, such as mobile phones. Audio query clips with lengths of two, three, four, and five seconds were captured using a mobile phone, which was placed at 5 or 7 meters from a 2.1-channel loudspeaker connected to a TV or radio. Through a built-in fingerprint generation module in the mobile phone, the MCLT peak-pair fingerprint is extracted from the captured audio clip and submitted as a query to the matching module at the server site.

With the randomly created 3,000 queries, query sets are created by adding various types of noise of different levels. Five different types of noise (babble noise, moving car noise, white noise, street noise, and computer fan noise) have been artificially added to different portions of the database at SNRs ranging from clean to 12 dB, and 6 dB. The audio query is converted to a standard PCM format, which is sampled at 16 kHz and quantized with 16 bits in a mono-channel. Audio query data are captured from 1,000 randomly selected audio samples per set. Each audio sample is played 30 times at randomly set offsets.

Table 1 depicts the experimental results of the four methods when a five-second-long query from Set I was used. MW, MS, MC, MX, and MCLT denote Method 1, Method 2, Method 3, Method 4, and the proposed method, respectively. The recognition results under the five different noisy environments are averaged for the evaluation. As shown in Table 1, the best recognition accuracy was 98.5% for query-by-example music identification, which was obtained with the proposed MCLT. The recognition rates of MW and MS were similar, but lower than those of MCLT. MX yields the lowest identification rate and provides worse results at SNR 0 dB.

Table 2 presents the results of the advertisement identification performed on a Set II database. The recognition accuracies for advertisement identification are not better than those of Table 1 for music identification, because some advertisements in Set II contain silent segments. The query was captured frequently from the silent segments and used for the matching. Also, the proposed MCLT yields better performance than MW, MS, MC, and MX.

Table 1. Comparative performance of four schemes with Set I.

SNR	Averaged recognition rate (%)				
	MCLT	MW [8]	MS [9]	MC [13]	MX [16]
Clean	98.5	95.5	96.6	94.8	93.5
12 dB	96.8	94.2	95.3	89.6	78.9
6 dB	93.6	83.4	84.5	77.5	63.8
0 dB	80.7	70.6	70.8	61.7	57.6
Total	92.4	85.9	86.8	80.9	73.5

Table 2. Comparative performance of four schemes with Set II.

SNR	Averaged recognition rate (%)				
	MCLT	MW [8]	MS [9]	MC [13]	MX [16]
Clean	95.5	93.6	94.5	93.5	92.6
12 dB	94.3	89.3	92.8	86.2	75.4
6 dB	91.7	80.6	81.7	74.6	60.2
0 dB	78.6	67.5	67.9	58.5	53.4
Total	90.0	82.8	84.2	78.2	70.4

Table 3. Comparative performance four schemes with Set III.

SNR	Averaged recognition rate (%)				
	MCLT	MW [8]	MS [9]	MC [13]	MX [16]
Clean	92.7	89.6	89.7	87.4	86.5
12 dB	91.5	86.7	89.5	82.4	76.8
6 dB	87.9	80.8	78.3	72.6	60.9
0 dB	75.5	66.5	65.5	55.6	51.7
Total	86.9	80.9	80.8	74.5	68.9

The results of the identification of TV (or radio) programs are shown in Table 3. Compared with the results in Tables 1 and 2, the results in Table 3 are not as high, because TV programs contain similar sound segments. The proposed MCLT method generally outperforms the other four methods in identification accuracy, especially in noisy environments.

Table 4 shows the recognition performance of the MCLT scheme for when the query length was changed. This result shows that the performance increases as the length of the query increases. Also, the proposed scheme shows satisfactory performance with four- and five-second-long queries, showing a recognition rate above 90%.

To cover a wide variety of robustness requirements for real-world application scenarios, Set I database was selected and the following signal modifications were carried out as a robustness test of the accuracy of the proposed audio fingerprinting system:

- Resampling: downsampling to 8 kHz and then upsampling back, upsampling to 32 kHz and then downsampling back.
- Equalization: gain –5 dB and 3 dB from 31 Hz to 16 kHz.
- Noise addition: SNR 6 dB, street noise.
- Echo addition: from 100 ms to 500 ms, 50% echo addition.
- Time stretching: from 70% to 150%. Only the pitch changes; the time duration remains unaffected.
- Pitch shifting: from –50% to +50%. Only the tempo changes; the pitch remains unaffected.

The identification results of the proposed method for different signal distortions are presented in Table 5 compared to those of the method based on peak discovery in two-direction scanning (TWS). The applied TWS method is modified from the contents of the reference paper and then implemented. As shown in the results in Table 5, the averaged identification rates of the proposed method are higher than those of TWT. In the proposed method, in addition to time stretching and pitch shifting, the averaged identification results are decreased from 98.5% to 87.9% for time stretching and 92.3% for pitch shifting.

Table 4. Performance evaluation according to query length.

SNR	Averaged recognition rate (%)			
	2 sec	3 sec	4 sec	5 sec
Clean	76.8	91.5	95.1	98.5
12 dB	71.5	90.7	94.3	96.8
6 dB	63.3	84.7	91.8	93.6
0 dB	55.6	76.5	81.7	80.7
Total	66.8	85.9	90.7	92.4

Table 5. Comparative performance under different distortions.

Types of distortions	Averaged recognition rate (%) using 5 s query	
	MCLT	TWS
Resampling	98.5	92.5
Equalization	98.5	92.5
Noise addition	92.8	78.7
Echo addition	98.5	84.6
Time-stretching	87.9	65.3
Pitch-shifting	92.3	67.4

IV. Conclusion

A salient audio peak-pair fingerprint extraction method, based on a modified spectrogram representation of the audio signal called the modulated complex lapped transform, has been proposed and evaluated. The proposed algorithm enhances Wang's fingerprint algorithm by generating local, stable salient peak-pair fingerprints based on MCLT; thus, it improves the accuracy of the audio fingerprinting system in the various real environments. The experimental results show that the proposed method has respectable results compared to other methods. However, the proposed method is clearly not robust enough with regards to time stretching and pitch shifting. An enhanced robust method is, therefore, needed. In future work, focus will be centered on the optimization of a more robust fingerprint extraction and search algorithm and on the combination of audio and visual fingerprints for a more robust content identification. The method will be applied to content security applications running on smart TVs and mobile phones.

References

[1] P. Cano et al., "A Review of Algorithms for Audio Fingerprinting," *IEEE Workshop Multimedia Signal Process*., St. Thomas, VI, USA, Dec. 9–11, 2002, pp. 169–173.

[2] J.J. Garcia-Hernandez, C. Feregrino-Uribe, and R. Cumplido, "Collusion-Resistant Audio Fingerprinting System in the Modulated Complex Lapped Transform Domain," *J., PLoS ONE*, vol. 8, no. 6, June 2013, pp. 1–15.

[3] A. Tirkel et al., "Collusion Resistant Fingerprinting of Digital Audio," *Proc. Int. Conf. Security Inf. Netw*., Sydney, Australia, Nov. 2011, pp. 5–12.

[4] W. Li, C. Xiao, and Y. Liu, "Low-Order Auditory Zernike Moment: A Novel Approach for Robust Music Identification in the Compressed Domain," *EURASIP J. Adv. Signal Process*., no. 132, Aug. 2013, pp. 1–15.

[5] A. Sinitsyn, "Duplicate Song Detection Using Audio Fingerprinting for Consumer Electronics Devices," *IEEE Int. Symp. Consum. Electron*., St. Petersburg, Russia, June 2006, pp. 1–6.

[6] J. Cerquides, "A Real Time Audio Fingerprinting System for Advertisement Tracking and Reporting in FM Radio," *Radioelektronika, Int. Conf*., Brno, Czech Republic, Apr. 24–25, 2007, pp. 1–4.

[7] J. Haitsma and T. Kalker, "A Highly Robust Audio Fingerprinting System," *Int. Conf. Music Inf. Retrieval*, Paris, France, Oct. 2002, pp. 107–115.

[8] Y. Liu, H.-S. Yun, and N.S. Kim, "Audio Fingerprinting Based on Multiple Hashing in DCT Domain," *IEEE Signal Process. Lett*., vol. 16, no. 6, June 2009, pp. 525–528.

[9] V. Chandrasekhar, M. Sharifi, and D. Ross, "Survey and Evaluation of Audio Fingerprinting Schemes for Mobile Query-by-Example Applications," *Int. Conf. Music Inf. Retrieval*, Miami, FL, USA, Oct. 2011, pp. 801–806.

[10] A. Wang, "An Industrial Strength Audio Search Algorithm," *Int. Conf. Music Inf. Retrieval*, Baltimore, MD, USA, Oct. 2003, pp. 7–13.

[11] X. Pan et al., "Audio Fingerprinting Based on Local Energy Centroid," *IET Int. Commun. Conf. Wireless Mobile Comput*., Shanghai, China, Nov. 14–16, 2011, pp. 351–354.

[12] T. Jiang et al., "A Real-Time Peak Discovering Method for Audio Fingerprinting," *Int. Conf. Internet Multimedia Comput. Service,* Huangshan, China, Aug. 2013, pp. 368–371.

[13] M. Park, H. Kim, and S. Yang, "Frequency-Temporal Filtering for a Robust Audio Fingerprinting Scheme in Real-Noise Environment," *ETRI J*., vol. 28, no. 4, Aug. 2006, pp. 509–512.

[14] W. Son et al., "Sub-fingerprint Masking for a Robust Audio Fingerprinting System in a Real-Noise Environment for Portable Consumer Devices," *IEEE Trans. Consum. Electron*, vol. 56, no. 1, Feb. 2010, pp. 156–160.

[15] X. Anguera, A. Garzon, and T. Adamek, "MASK: Robust Local Feature for Audio Fingerprinting," *Int. Conf. Multimedia Expo*, Melbourne, Australia, July 9–13, 2012, pp. 455–460.

[16] H. Malvar, "Fast Algorithm for the Modulated Complex Lapped Transform," *IEEE Signal Process. Lett*., vol. 10, no. 1, Jan. 2003, pp. 8–10.

[17] M.K. Mihcak and R. Venkatesan, "A Perceptual Audio Hashing Algorithm: A Tool for Robust Audio Identification and Information Hiding," *Workshop Inf. Hiding*, Pittsburgh, PA, USA, Apr. 25–27, 2001, pp. 51–65.

[18] M. Malekesmaeili and R.K. Ward, "A Novel Local Audio Fingerprinting Algorithm," *IEEE Int. Workshop Multimedia Signal Process*., Banff, Canada, Sept. 17–19, 2012, pp. 136–140.

Hyoung-Gook Kim received his Dipl-Ing degree in electrical engineering and his Dr-Ing degree in computer science from the Technical University of Berlin, Germany. From 1998 to 2005, he worked on mobile service robots at Daimler Benz and speech recognition at Siemens, Berlin, Germany. From 2005 to 2007, he was a project leader at the Samsung Advanced Institute of Technology, Suwon, Rep. of Korea. Since 2007, he has been a professor in the Department of Electronics Convergence Engineering, Kwangwoon University, Seoul, Rep. of Korea. His research interests include audio signal processing; audiovisual content indexing; and retrieval, speech enhancement, and robust speech recognition.

Jin Young Kim received his PhD degree in electronic engineering from Seoul National University, Seoul, Rep. of Korea. From 1993 to 1994, he worked on speech synthesis at Korea Telecom, Seoul, Rep. of Korea. Since 1995, he has been a professor in the Department of Electronics and Computer Engineering, Chonnam National University, Gwangju, Rep. of Korea. His research interests are speech synthesis; speech and speaker recognition; and audio-visual speech processing.