AUDIO FINGERPRINTING BASED ON NORMALIZED SPECTRAL SUBBAND CENTROIDS

Jin S. Seo, Minho Jin, Sunil Lee, Dalwon Jang, Seungjae Lee, Chang D. Yoo

Dept. of EECS, Div. of EE, KAIST,
373-1 Guseong Dong, Yuseong Gu, Daejeon 305-701, Korea
{jsseo, jinmho, sunillee, dal1, SeungjaeLee}@kaist.ac.kr, cdyoo@ee.kaist.ac.kr

ABSTRACT

For multimedia fingerprinting, it is crucial to extract relevant features that allow direct access to the distinguishing characteristics of a multimedia object. Features used for fingerprinting directly relate to the performance of the entire fingerprinting system. This paper proposes a novel audio fingerprinting method based on normalized spectral subband centroids. Spectral subband centroid is selected due to its resilience against equalization, compression, and noise addition. Both the reliability and the robustness issues in the fingerprinting system are addressed. The experimental results show that the proposed method is not only reliable but also robust against various audio processing steps including MP3 compression, equalization, random start, time-scale modification, and linear speed change.

1. INTRODUCTION

Protection, management, and indexing of digital contents are becoming more prominent with the increasing popularity of electronic commerce and on-line services. As one of the efficient solutions to these problems, fingerprinting is receiving increased attention. The goal of fingerprinting is to provide fast and reliable methods for content identification [1]. Promising applications [2] of multimedia fingerprinting are filtering for file-sharing services, automated monitoring for broadcasting stations, audio recognition through mobile network, and automated indexing of large-scale multimedia archives.

Similar to a human fingerprint used for identifying an individual, an audio fingerprint is used for recognizing audio. Fingerprints are perceptual features or short summaries of a multimedia object. This concept is an analogy with cryptographic hash function that maps data with arbitrary length to a bit sequence that consists of small and fixed number of bits [3]. Although cryptographic hashing is a proven method in message encryption and authentication, it is not possible to directly apply it to multimedia fingerprinting. Cryptographic hash functions are bit sensitive: an alteration of a single bit in the content will result in a completely different hash value. This renders cryptographic hash functions not applicable to multimedia objects that often undergo various manipulations including compression, enhancement, speed change, and analog-to-digital conversion during distribution. The modified version of the audio should have the same or similar fingerprints with the original audio. Various requirements on fingerprinting are summarized in [4]. In general, the fingerprinting function needs to have the following properties.

- **Robustness** (Invariance under perceptual similarity): the fingerprints resulting from degraded versions of an audio should result in the same or at least similar fingerprints with respect to the fingerprint of the original audio.
- **Pairwise independence** (Collision free): if two audios are perceptually different, the fingerprints from two audios should be considerably different.
- **Database search efficiency**: for the practical applications with a large-scale fingerprint database (DB), fast DB search is essential.

In this paper, we presented a new audio fingerprinting method based on the *normalized spectral subband centroid (SSC)*. Fingerprint matching is performed using the square of the Euclidean distance. By modelling the normalized SSC as a stationary process, the threshold for reliable fingerprint matching is obtained. SSC is originally proposed for speech recognition [5] and has shown recognition performance comparable to the widely-used cepstral features especially with noisy speech [6]. Moreover, SSC is resilient against the equalization of audio spectrum since it is the first-order normalized moment of the subband spectrum. By the comparative test, it was experimentally verified that SSC outperforms other widely-used features, such as tonality and MFCC, in the context of audio recognition. Experiments show that the proposed audio fingerprinting method satisfies the main requirements of fingerprinting.

This paper is organized as follows. Section 2 describes the extraction procedure of audio fingerprints used in the proposed method. Section 3 provides evaluation results on the performance of the proposed fingerprint. Finally, Section 4 concludes the paper.

2. PROPOSED AUDIO FINGERPRINTING METHOD

An overview of the proposed fingerprinting method is shown in Fig. 1. First, an input audio is converted to mono and downsampled to 11025 Hz. The downsampled signal is windowed by Hamming window (typically 371.5 ms) with 50% overlap and transformed into the frequency domain using FFT. The obtained audio spectrum is divided into 16 critical bands (from 300 Hz to 5300 Hz) [7]. For each critical band, normalized frequency centroid is calculated. The *frequency centroids* of the 16 critical bands are used as a *fingerprint* of the audio frame. A *fingerprint block*, composed of M fingerprints (typically $M = 53$ and consequently 848 centroids), from an audio block (typically 9.845 sec) is used for fingerprint matching. The details of the proposed method are explained in the next subsections.

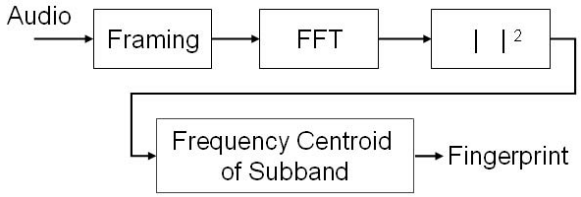

Fig. 1. Overview of the proposed audio fingerprint extraction

2.1. Fingerprint Based on Normalized Spectral Subband Centroids

The frequency centroid of an audio has been widely used for audio content analysis [8][9]. The frequency centroid has been found to be related to the human sensation of the brightness of a sound [10]. Paliwal used SSC [5] as features for speech recognition. The SSC has shown recognition performance comparable to the widely-used cepstral features especially with noisy speech [6]. The subband moment of order ν at the i-th subband of an audio spectrum $P[k,m]$ is defined as

$$M_i^{\nu}[m] = \sum_{k=CB_i+1}^{CB_{i+1}} k^{\nu} P[k,m] \tag{1}$$

where k, m and CB_i denote the frequency bin, the frame index, and the frequency boundary of the i-th critical band respectively. The SSC is the first-order normalized moment given as follows [6]:

$$C_i[m] = \frac{M_i^1[m]}{M_i^0[m]}. \tag{2}$$

Through the above normalization, the SSC is resilient against the equalization of the audio spectrum. Since the range of $C_i[m]$ is different at each critical band, it is normalized as follows:

$$NC_i[m] = \frac{C_i[m] - (CB_i + CB_{i+1})/2}{CB_{i+1} - CB_i}. \tag{3}$$

Then the normalized SSC, $NC_i[m]$, has a range between -0.5 and 0.5 regardless of the critical bands. The normalized SSC is used as a fingerprint.

2.2. Fingerprint Matching

In the fingerprint matching, the audios are declared similar if the distance between their fingerprints is below a certain threshold T. The problem could be formulated as the following hypothesis testing using the fingerprinting function $H(\cdot)$ and distance measure $D(\cdot,\cdot)$:

- L_0: Two audios A and A' are from the same audio if the distance $D(H(A), H(A'))$ is below a threshold T.
- L_1: Two audios A and A' are from the different audio if the distance $D(H(A), H(A'))$ is above a threshold T.

For the selection of threshold T, the false alarm rate P_{FA} and the false rejection rate P_{FR} are considered. The false alarm rate P_{FA} is the probability to declare different audios as similar. The false rejection rate P_{FR} is the probability to declare an audio and its processed versions as dissimilar. In practice, P_{FR} is difficult to analyze since there are plenty of audio processing steps of which the exact characteristics are not known. Thus it is common to deal with only P_{FA} for the selection of threshold T.

2.2.1. Fingerprint Modelling

The problem of fingerprint matching is approached by assuming the SSC as a stationary process. We note that similar analysis has been performed for watermark detection in [11]. Let x be the normalized (values between -0.5 and 0.5) SSCs of an audio block (9.845 sec). We further normalize x by the mean m_x and the variance σ_x^2 of x as follows:

$$p[n] = \frac{x[n] - m_x}{\sigma_x} \tag{4}$$

where $n = 1, 2, \ldots, N$ and N is the number of SSCs in an audio block (typically $N = 16 \times 53 = 848$ for 9.845 sec). Thus p is the random process with zero mean and unit variance. By simplifying the stochastic model of the normalized SSC as the first-order autocorrelation, we obtain the following expressions:

$$\begin{aligned} R[k] = E\Big[p[n]p[n+k]\Big] &= a^{|k|}, \\ Q[k] = E\Big[p^2[n]p^2[n+k]\Big] &= 1 + (\mu_4 - 1)b^{|k|} \end{aligned} \tag{5}$$

where $\mu_k = E[p^k[n]]$, and a and b are the measures of the correlation of SSC. Through the normalization by the mean and the variance, $\mu_1 = 0$ and $\mu_2 = 1$. Fig. 2 shows that the autocorrelation obtained from the audio data follows the first-order model reasonably well. Experiments reveal that the values of a, b, and μ_4 are typically 0.59, 0.44, and 3.0 respectively.

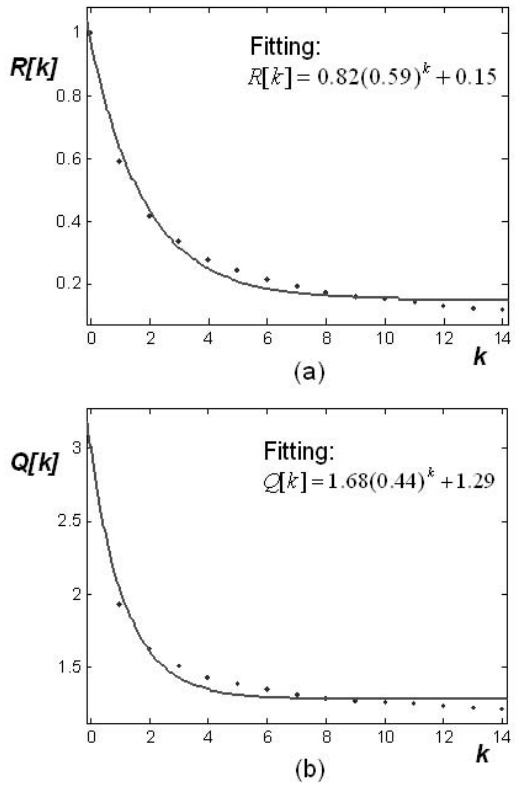

Fig. 2. Experimental results of (a) $R[k]$ versus correlation lag k (b) $Q[k]$ versus correlation lag k

2.2.2. Reliability Analysis

The square of the Euclidean distance measure D is used for fingerprint matching as follows:

$$D = \frac{1}{N}\sum_{n=1}^{N}(p[n]-q[n])^2 \tag{6}$$

where p and q are the normalized SSCs from the different audio blocks. By the central limit theorem, the distance measure D has a normal distribution if N is sufficiently large and the contributions in the sums are sufficiently independent [11]. The mean $E[D]$ of the distance measure D is given as

$$\begin{aligned} E[D] &= \frac{1}{N}E\Big[\sum_{n=1}^{N}(p[n]-q[n])^2\Big] \\ &= \frac{1}{N}\Big(\sum_{n=1}^{N}E[p^2[n]]+\sum_{n=1}^{N}E[q^2[n]] \\ &\quad -2\sum_{n=1}^{N}E[p[n]]E[q[n]]\Big) \\ &= 2\mu_2+0=2. \end{aligned} \tag{7}$$

The variance σ_D^2 of the distance measure D is expressed as

$$\sigma_D^2 = E[D^2]-(E[D])^2. \tag{8}$$

The mean of D^2 is given as follows:

$$\begin{aligned} E[D^2] &= \frac{1}{N^2}E\Big[\Big(\sum_{n=1}^{N}p^2[n]+\sum_{n=1}^{N}q^2[n]-2\sum_{n=1}^{N}p[n]q[n]\Big)^2\Big] \\ &= 2+(2\mu_4+4)/N \\ &\quad +\frac{4}{N^2}\sum_{k=1}^{N-1}(N-k)[1+(\mu_4-1)b^k+2a^{2k}]. \end{aligned} \tag{9}$$

Using the typical values of a, b, and μ_4, the standard deviation of the distance measure is given as $\sigma_D = 0.1479$. Through the normal approximation of the distance measure $N(2,\ \sigma_D^2)$, the false alarm rate P_{FA} is given as follows:

$$\begin{aligned} P_{FA} &= \int_{-\infty}^{T}\frac{1}{\sqrt{2\pi}\sigma_D}\exp\Big[\frac{-(x-2)^2}{2\sigma_D^2}\Big]dx \\ &= \frac{1}{2}\mathrm{erfc}\Big(\frac{2-T}{\sqrt{2}\sigma_D}\Big). \end{aligned} \tag{10}$$

For a certain value of P_{FA}, the threshold T for D can be determined. In the experiments we use $T = 0.8$. Then we arrive at a very low false alarm probability of $\mathrm{erfc}(5.7387)/2 = 2.414\times10^{-16}$.

3. EXPERIMENTAL RESULTS

3.1. Performance of the proposed method

The pairwise independence and the robustness of the proposed method are evaluated. The proposed method is tested using the fingerprint DB with 8000 songs that include various genres, such as classic, jazz, pop, rock, and hiphop. Database search and fingerprint matching are performed every 9.845 sec of the input query audio. Many algorithms have been proposed for DB search [12], and among them, k-d tree algorithm [13] is used in this paper. The k-d tree outputs the positions of the nearest neighborhoods of the input query audio as candidate positions, and fingerprint matching is performed on these candidate positions for verification.

Pairwise independence is tested with 100,000 randomly selected pairs of audio blocks. Fig. 3 shows the histogram of the measured distances between the chosen pairs. All the measured distances were in the range between 1.1 and 2.9. The histogram of the measured distance shows that the proposed fingerprints follow the stochastic model in Section 2.2 fairly well. The mean of the measured distances was 1.9781 which is close to 2.0, and the standard deviation of it was 0.1569 which is also close to 0.1479 obtained from the model in Section 2.2. This result shows that the proposed method is approximately pairwise independent.

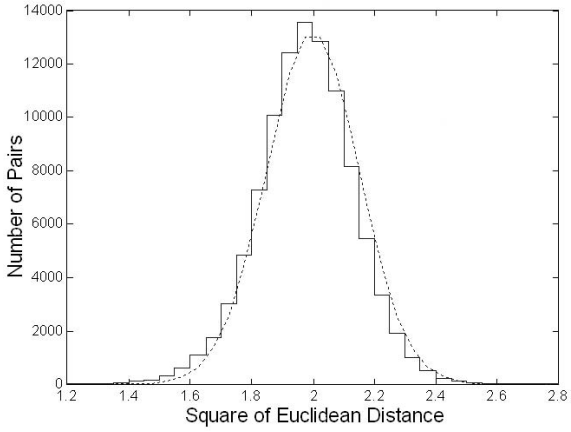

Fig. 3. Histogram of the square of the Euclidean distance between the fingerprints from different audio blocks

To test robustness of the proposed method, the original audios were subjected to various kinds of audio processing steps, including MP3 compression, equalization, random start, time-scale modification, and linear speed change, and their respective fingerprint blocks were extracted (see [2] for a detailed description of the processing steps). Mean, standard deviation, and false rejection rate of the measured distance between the original and the processed audio fingerprints are shown in Table 1 for randomly selected 5000 audio blocks in the 8000 song DB. For all the tested audio processing steps, the measured distance was below the threshold ($T = 0.8$). The table shows that the proposed method is robust against the common audio processing steps.

3.2. Comparison of the proposed method with other features

The robustness of the proposed normalized SSC is compared with that of the other audio features. Among the various audio features [8], tonality of subbands [14] and Mel Frequency Cepstral Coefficients (MFCC) are selected due to its popularity in audio and speech recognition respectively. We tested the features using the DB with 100 audios. Audio features, normalized SSC, tonality and MFCC, are generated from the 100 audios. The Euclidean distance is used as a measure of the distance between the original and the processed audios for all three features. For each frame of a processed audio, we extract features (16 dimensions), search the

Table 1. Mean, standard deviation (Std) and false rejection rate (with threshold $T = 0.8$) of the measured distance for different kinds of audio degradations

Processing	**Mean**	**Std**	P_{FR}
MP3 compression (32 kbps)	0.2297	0.0654	0.0
Equalization (3 dB)	0.0243	0.0091	0.0
Random start (worst case 92.9 ms shift)	0.2567	0.0704	0.0
Time scale (+4%)	0.3165	0.0903	0.0
Time scale (-4%)	0.3236	0.0916	0.0
Linear speed (+1%)	0.2715	0.0922	0.0
Linear speed (-1%)	0.2873	0.0961	0.0

DB exhaustively, and find the DB position with the minimum Euclidean distance. If the DB position with the minimum Euclidean distance is corresponding to the input processed audio frame, it is assumed that the input processed audio frame is correctly identified. Fig. 4 shows the probability of the correct identification for the three features. The probability stands for the robustness of the feature space against the audio processing steps. The result shows that the normalized SSC is more robust than other features for audio recognition.

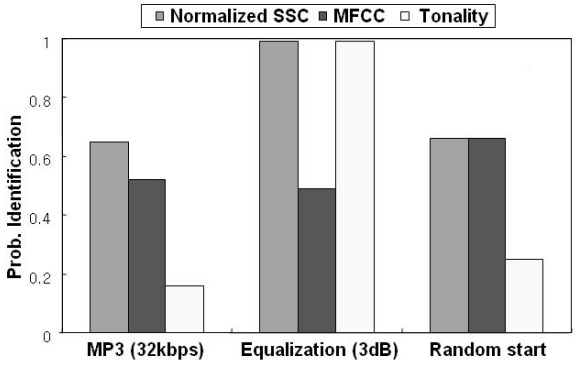

Fig. 4. Robustness of the features against audio processing steps

4. CONCLUSION

For a reliable fingerprinting system, the features should be both fairly discriminative and robust. In this paper, we presented a new audio fingerprinting method based on the normalized SSC. Fingerprint matching is performed using the square of the Euclidean distance. The problem of reliable fingerprint matching is approached by assuming a stationary process as a model for the normalized SSC. The stochastic model was experimentally verified. Experiments show that the normalized SSC is pairwise independent with different inputs and robust under quality preserving signal processing steps. In the comparative test, the normalized SSC outperformed other widely-used features, such as tonality and MFCC, in the context of audio fingerprinting.

5. ACKNOWLEDGMENTS

This work was supported in part by the Ministry of information & Communications, Korea, under the Information Technology Research Center (ITRC) Support Program, the grant No. R01-2003-000-10829-0 from the Basic Research Program of the Korea Science & Engineering Foundation, and the Brain Korea 21 Project, the school of information technology, KAIST in 2005.

6. REFERENCES

[1] T. Kalker, J.A. Haitsma, and J. Oostveen, "Issues with digital watermarking and perceptual hashing," in *Proc. SPIE 4518, Multimedia Systems and Applications IV*, Nov. 2001.

[2] J.A. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. International Conf. on Music Information Retrieval*, 2002.

[3] A. Menezes, P. Oorshot, and S. Vanstone, *Handbook of Applied Cryptography*, CRC press, 1997.

[4] J.S. Seo, J. Haitsma, T. Kalker, and C.D. Yoo, "A robust image fingerprinting system using the Radon transform," *Signal Processing: Image Communication*, vol. 19, pp. 325–339, 2004.

[5] K.K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. IEEE ICASSP*, 1998, pp. 617–620.

[6] J. Chen, Y. Huang, Q. Li, and K.K. Paliwal, "Recognition of noisy speech using dynamic spectral subband centroids," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 258–261, 2004.

[7] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, 1999.

[8] Y. Wang, Z. Liu, and J.-C. Huang, "Mutimedia content analysis using both audio and visual clues," *IEEE Signal Processing Mag.*, pp. 12–36, 2000.

[9] S.Z. Li, "Contenet-based audio classification and retrieval using the nearest feature line method," *IEEE trans. on Speech and Audio Processing*, vol. 8, no. 5, 2000.

[10] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia Mag.*, vol. 3, 1996.

[11] J.P. Linnartz, T. Kalker, G. Depovere, and R. Beuker, "A reliability model for the detection of electronic watermarks in digital images," in *Symposium on Communications and Vehicular Technology*, 1997.

[12] C. Bohm, S. Berchtold, and D. Keim, "Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases," *ACM Computing Surveys*, vol. 33, no. 3, pp. 322–373, 2001.

[13] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[14] J. Herre, E. Allamanche, and O. Hellumth, "Robust matching of audio signals using spectral flatness features," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 127–130.