Audio Fingerprinting Based on Normalized Spectral Subband Moments

Jin S. Seo, *Associate Member, IEEE*, Minho Jin, Sunil Lee, *Student Member, IEEE*, Dalwon Jang, *Student Member, IEEE*, Seungjae Lee, and Chang D. Yoo, *Member, IEEE*

***Abstract*—The performance of a fingerprinting system, which is often measured in terms of reliability and robustness, is directly related to the features that the system uses. In this letter, we present a new audio-fingerprinting method based on the normalized spectral subband moments. A threshold used to reliably determine a fingerprint match is obtained by modeling the features as a stationary process. The robustness of the normalized moments was evaluated experimentally and compared with that of the spectral flatness measure. Among the considered subband features, the first-order normalized moment showed the best performance for fingerprinting.**

***Index Terms*—Content identification, fingerprinting, robust hashing, robust matching, spectral subband moment.**

I. Introduction

FINGERPRINTS are short summaries of multimedia content. The aim of fingerprinting (also known as content identification) is to provide fast and reliable means for protection, management, and indexing of multimedia contents. Similar to a human fingerprint that has been used for identifying an individual, an audio fingerprint is used for recognizing an audio clip. Promising applications [1] 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. These applications have boosted the interest in multimedia fingerprinting, which has also led to a number of audio fingerprinting methods [1]–[4]. A review of audio-fingerprinting methods is given in [5]. In practice, multimedia contents are liable to degradations due to compression, enhancement, noise addition, and analog-to-digital conversion. Thus, an audio-fingerprinting system must be able to allow for some modification of an audio while distinguishing one audio clip from another. In general, the fingerprinting function needs to have the following properties [6], [7].

- **Robustness** (invariance under perceptual similarity): The fingerprints of a degraded audio clip should be similar to the fingerprints of the original audio clip.
- **Pairwise independence** (collision free): Two audio clips, which are perceptually different, must have different fingerprints.
- **Database search efficiency**: The structure of fingerprints must be conducive to fast database (DB) search.

For fingerprinting, extracting features that allow direct access to the distinguishing information is crucial. In this letter, the *normalized spectral subband moments* are investigated. The normalized moments are selected due to their resilience against local spectral modifications, such as equalization. Among the various normalized moments, the first-order normalized moment (spectral subband centroid) has been used in speech recognition [8] and is known to give recognition performance comparable to the widely used cepstral features [9]. Fingerprint matching is performed using the square of the Euclidean distance measure for fast calculation and mathematical tractability for analysis. By modeling the normalized moments as a stationary process, a threshold that can be used to reliably determine a fingerprint match is obtained. In practice, there is a tradeoff between *robustness* and *pairwise independence* in selecting the threshold. The performance of the normalized moments for fingerprinting was experimentally compared with that of the spectral flatness measure (SFM), which is another subband feature successfully employed in audio fingerprinting [2].

This letter is organized as follows. Section II describes the extraction and the matching of audio fingerprints based on the spectral subband moments. Section III evaluates the proposed fingerprinting method. Finally, Section IV concludes the paper.

II. Proposed Audio-Fingerprinting Method

As shown in Fig. 1(a), a fingerprinting system used for identification is generally made up of three steps: fingerprint extraction, DB search, and fingerprint matching. Query fingerprints are extracted from an audio clip that is to be identified. Then the candidates for the query fingerprints are obtained by the nearest neighbor DB search. For fast DB search, an efficient indexing structure is needed. For our system, the k-d tree, known for its simplicity and reasonably high hit-ratio, is used. A survey of the various DB search methods can be found in [10]. As a

Manuscript received August 22, 2005; revised October 31, 2005. This work was supported in part by the Ministry of Information and Communications, Korea, under the Information Technology Research Center (ITRC) Support Program, in part by Grant R01-2003-000-10829-0 from the Basic Research Program of the Korea Science and Engineering Foundation, and tin part by he Brain Korea 21 Project, the School of Information Technology, KAIST, in 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mauro Barni.

J. S. Seo, M. Jin, S. Lee, D. Jang, and C. D. Yoo are with the Division of Electrical Engineering, Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea (e-mail: jsseo@kaist.ac.kr; jinmho@kaist.ac.kr; sunillee@kaist.ac.kr; dal1@kaist.ac.kr; cdyoo@ee.kaist.ac.kr).

S. Lee is with the Digital Contents Research Division, Electronics and Telecommunications Research Institute, Daejeon 305-700, Korea (e-mail: seungjlee@etri.re.kr).

Digital Object Identifier 10.1109/LSP.2005.863678

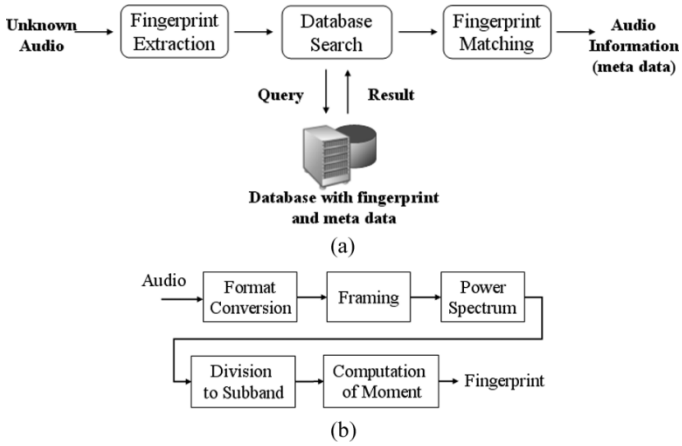

Fig. 1. Overview of (a) audio fingerprinting system used for identification and (b) fingerprint extraction from spectral subband moments.

final step, the fingerprint matching is performed only on these candidates [1].

The focus of this letter is on the fingerprint extraction and matching. The fingerprint extraction consists of five steps, as shown in Fig. 1(b): 1) an input audio clip, which may have come from a number of different formats, is converted to a single unified format (mono and sampling frequency 11 025 Hz); 2) the converted audio signal is split into overlapping segments (called frames) of length L with $P\%$ overlap (in our system, L is 371.5 ms and $P = 50$); 3) each frame is windowed by a Hamming window of length L and transformed into the frequency domain; 4) the spectrum of each frame is divided into M critical bands [11] (typically $M = 16$ from 300 to 5300 Hz, which is known to be relevant to human perception [1] and robust [9]); and 5) at each critical band, a subband feature is calculated. The subband features of the M critical bands are used as a *fingerprint* of the frame. As an M-dimensional fingerprint from a frame does not contain enough information to identify the whole audio, a *fingerprint block*, which is composed of N consecutive fingerprints (in our system, $N = 27$ or 54 for 5–10 s), is used for fingerprint matching [1]. The details of the proposed method are explained in the next subsections.

A. Normalized Spectral Subband Moments

Fingerprinting should be based on robust and perceptually relevant features to meet the requirements mentioned in Section I. Although the moment features have been widely used for image and audio content analysis [12], [13], few studies have applied them to audio fingerprinting. Our prime objective is to examine the performance of the normalized spectral subband moments for audio fingerprinting. Let $P[n,k]$ be the short-time power spectrum of an audio signal at frequency bin k of the nth frame. Then the νth-order moment of the mth subband audio spectrum is defined as

$$\zeta_\nu[n,m] = \sum_{k=C[m]+1}^{C[m+1]} k^\nu P[n,k] \tag{1}$$

where $C[m]$ denotes the frequency boundary of the mth critical band. Then the first- and second-order normalized moment η_1 and η_2 are

$$\eta_1[n,m] = \frac{\zeta_1[n,m]}{\zeta_0[n,m]}$$

$$\eta_2[n,m] = \left[\sum_{k=C[m]+1}^{C[m+1]} (k-\eta_1[n,m])^2 \frac{P[n,k]}{\zeta_0[n,m]}\right]^{\frac{1}{2}} = \left[\frac{\zeta_2[n,m]}{\zeta_0[n,m]} - (\eta_1[n,m])^2\right]^{\frac{1}{2}}. \tag{2}$$

The normalized moments are resilient against local spectral modifications, such as equalization. Among other subband features, the SFM was successfully employed in audio recognition [2]. The SFM is defined as

$$\mathrm{SFM}[n,m] = \frac{\left[\prod_{k=C[m]+1}^{C[m+1]} P[n,k]\right]^{\frac{1}{\kappa_m}}}{\frac{\zeta_0[n,m]}{\kappa_m}} \tag{3}$$

where $\kappa_m = C[m+1] - C[m]$. This letter evaluates and compares the three subband features η_1, η_2, and SFM in the context of fingerprinting.

B. Fingerprint Matching

In the fingerprint matching, two audio clips 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 audio clips A and A' are from the same audio if the distance $D(H(A), H(A'))$ is below a threshold T.
- L_1: Two audio clips 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 T, there is a tradeoff between the false alarm rate P_{FA} and the false rejection rate P_{FR}. The false alarm rate P_{FA} is the probability to declare different audio clips 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 select a threshold T of minimizing P_{FR} subject to a fixed P_{FA}.

1) Fingerprint Modeling: The problem of fingerprint matching is approached by assuming the subband feature as a realization of a stationary process. We note that a similar analysis has been performed for watermark detection in [14]. Let $x[n,m]$ be the subband features of an audio clip $(1 \le n \le N, 1 \le m \le M)$. The mean and the standard deviation of the mth subband features $x[\cdot,m]$ are denoted by $\mu_x[m]$ and $\sigma_x[m]$, respectively. We further normalize $x[n,m]$ using the mean and the standard deviation as follows:

$$p[n,m] = \frac{x[n,m] - \mu_x[m]}{\sigma_x[m]} \tag{4}$$

so that p is a random process with zero mean and unit variance. By simplifying the stochastic model of the subband feature with

a first-order autocorrelation, the following expressions are obtained:

$$\begin{aligned} R[k_1,k_2] &= E\left[p[n,m]p[n+k_1,m+k_2]\right] \\ &= \alpha_1^{|k_1|}\alpha_2^{|k_2|} \\ Q[k_1,k_2] &= E\left[p^2[n,m]p^2[n+k_1,m+k_2]\right] \\ &= 1+(\mu_4-1)\beta_1^{|k_1|}\beta_2^{|k_2|} \end{aligned} \tag{5}$$

where $\mu_k = E[p^k[n,m]]$ as well as (α_1,β_1) and (α_2,β_2) represent a measure of the correlation in the direction of the time and the frequency, respectively. By the normalization, $\mu_1 = 0$ and $\mu_2 = 1$. To determine the value of parameters in (5), the autocorrelations $R[k1,k2]$ and $Q[k1,k2]$ were estimated from 50 000 clips and fitted into the first-order model using cftool in Matlab. From the fitting result, the value of $(\alpha_1,\alpha_2,\beta_1,\beta_2,\mu_4)$ was determined as (0.54, 0.00, 0.29, 0.06, 2.87), (0.45, 0.00, 0.21, 0.00, 2.76), and (0.43, 0.37, 0.17, 0.30, 3.12) for η_1, η_2, and SFM, respectively.

2) Reliability Analysis: Fast and mathematically tractable fingerprint matching can be achieved by using the square of the Euclidean distance measure as follows:

$$D = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\left(p[n,m]-q[n,m]\right)^2 \tag{6}$$

where p and q are the subband features of the different audio clips. By the central limit theorem, the distance measure D has a normal distribution if NM is sufficiently large and the contributions in the sums are sufficiently independent [14]. Assuming that the two fingerprints p and q are independent, the mean $E[D]$ of the distance measure D is given as

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

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

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

where $E[D^2]$ can be calculated as in (7) using the first-order stochastic model in (5). The detailed derivation is available at http://mmp.kaist.ac.kr/~jsseo/fingerprint.html. Using typical values of $(\alpha_1,\alpha_2,\beta_1,\beta_2,\mu_4)$, the standard deviation of the distance measure σ_D for $N = 54$ is 0.098, 0.094, and 0.11 for η_1, η_2, and SFM, respectively. By approximating the distance measure as the normal distribution $N(2,\sigma_D^2)$, the false-alarm rate P_{FA} is given as follows:

$$P_{FA} = \int_{-\infty}^{T}\frac{1}{\sqrt{2\pi}\sigma_D}\exp\left[\frac{-(x-2)^2}{2\sigma_D^2}\right]dx = \frac{1}{2}\mathrm{erfc}\left(\frac{2-T}{\sqrt{2}\sigma_D}\right). \tag{9}$$

For a certain value of P_{FA}, the threshold T for D can be determined from (9). To examine the validity of the stochastic-model assumption and the normal approximation, we calculate the distance measure between fingerprints from randomly selected pairs of audio clips (more than 10^8 pairs were used in the calculation). Fig. 2 shows P_{FA} from the experiments and the model assumption. The figure shows that the fingerprints from the three features follow the stochastic-model assumption and the normal approximation fairly well. This result shows that the threshold T obtained from (9) can be used in practice with reasonable accuracy.

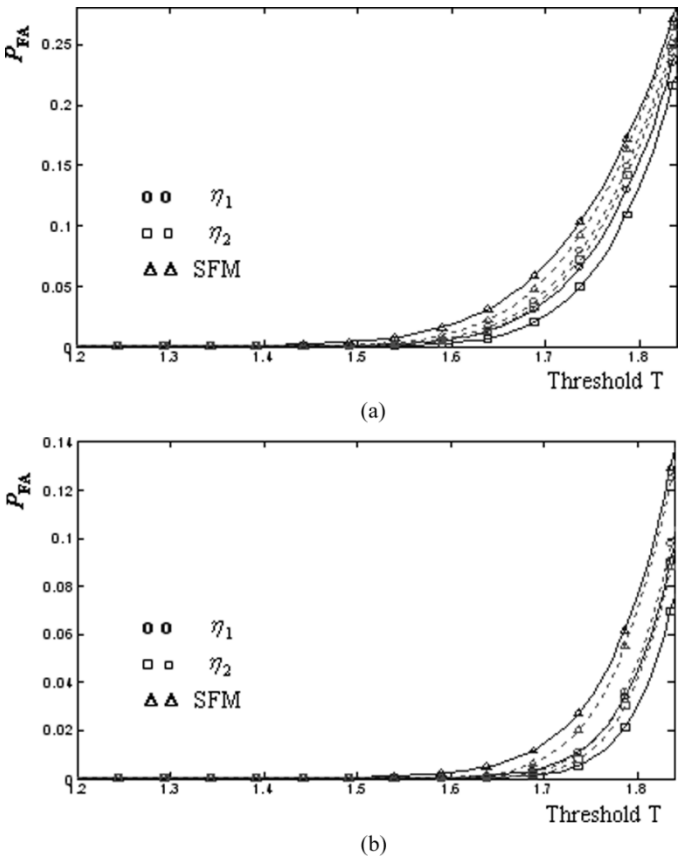

Fig. 2. P_{FA} versus the fingerprint matching threshold T; —— empirical value, −− theoretical value from (9). (a) $N = 27$. (b) $N = 54$.

III. Performance Evaluation

The performance of the normalized spectral subband moments was evaluated using the fingerprint DB generated from 8000 songs belonging to various genres, such as classic, jazz, pop, rock, and hiphop. The fingerprint was extracted from the 16 critical bands $(M = 16)$ between 300 and 5300 Hz. The fingerprint matching was performed using the fingerprints from 5- or 10-s audio clips ($N = 27$ or 54) after normalization, as in Section II-B1. As mentioned in Section II-B, fingerprint matching can be formulated as a hypothesis testing in which there are two types of errors: the false-alarm rate and the false-rejection rate. For a fair comparison between different features, the receiver operating characteristic (ROC) curve, which plots the false-rejection rate versus the false-alarm rate, was used. The ROC curve is obtained by measuring both error rates while varying the threshold used in the fingerprint matching. The false alarm rate was calculated from the original fingerprint DB, as in

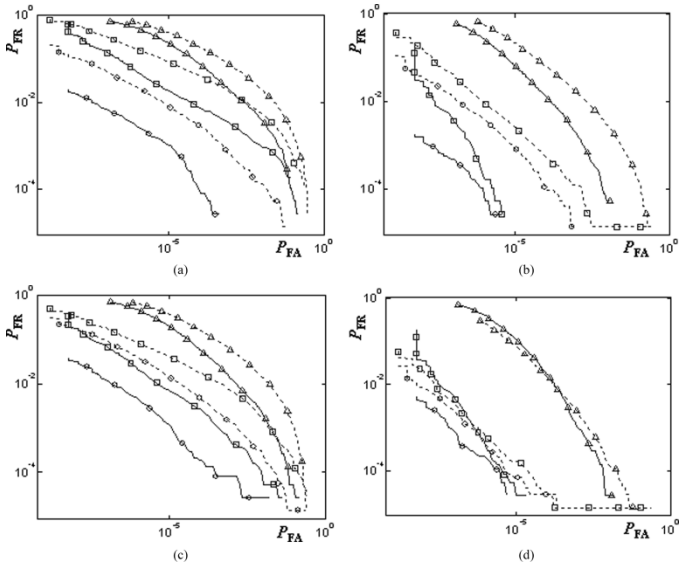

Fig. 3. ROC curves for four sets of distortions; $\circ\circ\,\eta_1$, $\square\square\,\eta_2$, $\triangle\triangle$ SFM, $--$ $N = 27$, —— $N = 54$. (a) Distortion set 1. (b) Distortion set 2. (c) Distortion set 3. (d) Distortion set 4.

Fig. 2. To calculate the false rejection rate, four sets of distortions were considered (using Cool Edit Pro 2.1 software). Distortion set 1 consists of a filter-emulating old time radio, pitch increase by 1%, and 92.9-ms delay. Distortion set 2 consists of a filter-emulating ambient metal room, pitch decrease by 1%, and 92.9-ms delay. Distortion set 3 consists of super loud and linear speed change by 1%. Distortion set 4 consists of a filter-emulating rich chamber and time-scale modification by 4%. Each original audio was subjected sequentially to all the distortions in a distortion set and then to common distortions that are 30-band classic equalization, 30-band pop equalization, and MP3 compression (128 kbps). By comparing the distance between the fingerprints from the original and the corresponding processed audio clips with the threshold, the false-rejection rate was obtained. The resulting ROC curves of the three features are shown in Fig. 3. The first-order normalized moment showed the lowest false-rejection rate for a given false-alarm rate (vice versa), and the SFM showed the worst performance among the three subband features. We note that the tendency also holds for most of the distortions that were not considered, such as other types of compression, equalization, and filtering.

IV. Conclusion

For a reliable fingerprinting system, the features should be both discriminative and robust. In this letter, the spectral subband features were investigated for audio fingerprinting. We considered three subband features: the first-order normalized moment, the second-order normalized moment, and the SFM. For the fast fingerprint matching, the square of the Euclidean distance measure was used. The problem of reliable fingerprint matching is approached by assuming the features as a realization of a stationary process. The assumed stochastic model was experimentally verified. The threshold used in the fingerprint matching is obtained from the analysis using the stochastic model. In the comparative test, the first-order normalized moment outperformed the other two subband features in the context of audio fingerprinting. Further work includes the comparison of the subband moments and the Euclidean distance measure with the other features and distance measures, respectively.

References

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

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

[3] C. Burges, J. Platt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 165–174, May 2003.

[4] S. Sukittanon, L. Atlas, and J. Pitton, "Modulation scale analysis for content identification," *IEEE Trans. Signal Process.*, vol. 52, no. 10, pp. 3023–3035, Oct. 2004.

[5] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of algorithms for audio fingerprinting," in *Proc. IEEE Workshop Multimedia Signal Processing*, 2002, pp. 169–173.

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

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

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

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

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

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

[12] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, Fall 1996.

[13] Y. Wang, Z. Liu, and J.-C. Huang, "Mutimedia content analysis using both audio and visual clues," *IEEE Signal Process. Mag.*, vol. 17, no. 6, pp. 12–36, Nov. 2000.

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