한국음향학회지 제31권 제4호 pp. 260~265 (2012)
The Journal of the Acoustical Society of Korea Vol.31, No.4 (2012)

A Robust Audio Fingerprinting Method Based on Segmentation Boundaries

Jin Soo Seo

Dept. of Electrical Engineering, Gangneung-Wonju National University
(Received February 6, 2012; accepted March 6, 2012)

ABSTRACT: A robust audio fingerprinting method is presented based on segmentation boundaries. In order to obtain robustness against linear speed changes, fingerprint extraction and matching are synchronized with the segmentation boundaries. Experimental results show that the proposed method is also robust against other common audio processing steps including low bit-rate compression, equalization, and time-scale modification.
Key words: Audio fingerprinting, Content identification, Speed change, Segmentation, Matching
ASK subject classification: Acoustic Signal Processing (1.7)

I. Introduction

Fingerprints are perceptual features or short summaries of a multimedia content. Similar to a human fingerprint which has been used for identifying an individual, a multimedia fingerprint is used for recognizing a multimedia clip. A fingerprinting system must make allowances for some modification of content while distinguishing one content clip from another [1-4]. Promising applications 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.

This paper proposes an audio fingerprinting method that is robust to linear speed changes. The robustness against speed changes is essential since speed changes are a common phenomenon in many contexts where audio fingerprinting is of importance, for example in radio broadcast, and easy to impose with modern computers [5]. However, few results have been reported that explicitly consider this problem in the context of audio recognition. In [5], the shift invariance of autocorrelation function is utilized to extract speed-change resilient fingerprints from audio spectrum. This, however, can not provide explicit temporal synchronization, which makes the method only robust against speech changes up to 6 %.

Fig. 1 shows a 10-second clip of the original and speed-changed (10 % speed-up) audio. Although the start positions of the two audio clips are the same, there is a significant temporal mismatch, which renders the conventional fingerprint matching methods, based on fixed-length segments, fail to identify the two clips as the same. This motivates us to combine segmentation with fingerprinting. Temporal synchronization can be achieved from the segmentation boundaries. Moreover, by extracting a fingerprint from a segment, the number of fingerprints reduces, which is conducive for fingerprint storage and search. However, the segmentation boundaries extracted from the original and the severely distorted audio may be different. To mitigate the limited repeatability, we propose a local invariant matching method in which only small numbers of correct boundaries are needed

***Corresponding author:** Jin Soo Seo (jsseo@gwnu.ac.kr)
Department of Electrical Engineering, Gangneung-Wonju National University, Gangneung, Rep. of Korea
(Tel: 033-640-2428)

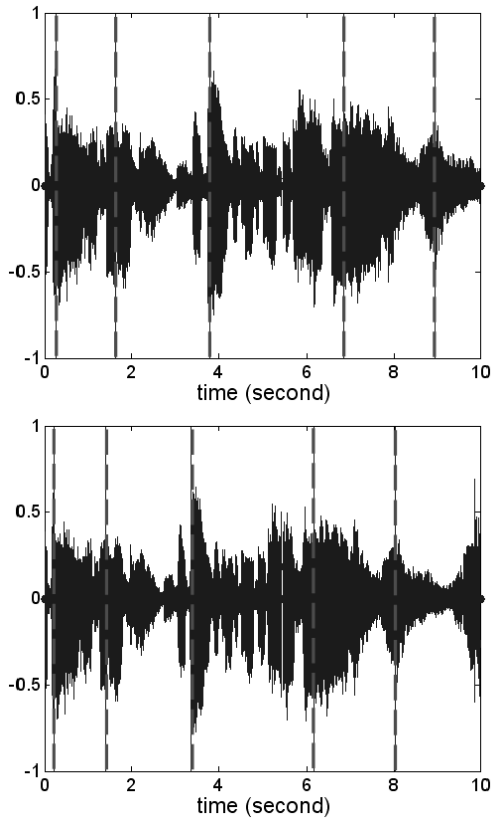

Fig. 1. A 10-second clip of audio signal (top) and its 10% speed-changed version (bottom). The vertical dashed lines represent segmentation boundaries extracted using the method presented in Section 2.1.

for claiming successful matching.

II. Proposed Audio Fingerprinting Method

2.1 Fingerprint Extraction Based on Segmentation

Any suitable segmentation and fingerprint extraction method can be utilized for the proposed fingerprinting method. In this paper we use the segmentation algorithm [6] based on the spectral difference, which has shown superior performance in a comparative test [6] and does not require any prior training, and the audio fingerprinting method based

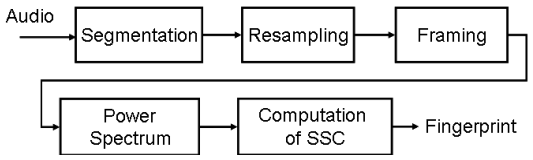

Fig. 2. Overview of proposed fingerprint extraction method.

on the normalized spectral subband centroid (NSSC), which was originally proposed for speech recognition and has shown superior performance for audio fingerprinting [2].

An overview of the proposed fingerprinting method is shown in Fig. 2. First, an input audio is converted to mono and downsampled to 11025 Hz. The downsampled signal is windowed by the Hanning window of length 2048 with 75 % overlap, and each windowed frame is transformed into the Fourier domain. Let $S[n,k]$ be the short-time Fourier transform (STFT) of an audio signal at frequency bin k of the n-th frame. The magnitude and the phase of $S[n,k]$ are respectively denoted by $R[n,k]$ and $\phi[n,k]$. The spectral difference in complex domain [6] is given as the difference between the current STFT and the estimated (or predicted) STFT as shown below:

$$\Gamma[n,k] = \| S[n,k] - R[n-1,k]e^{\hat{\phi}[n,k]} \| \quad (1)$$

where $\|x\|$ denotes the magnitude of complex number x, and $\hat{\phi}[n,k]$ is principal argument of $(2\phi[n\text{-}1,k]\text{-}\phi[n\text{-}2,k])$. Summing the above measures across all k, we can construct a frame-by-frame novelty function η as

$$\eta[n] = \sum_k \Gamma[n,k] \quad (2)$$

A peak-picking algorithm is applied to the novelty function to locate the segmentation boundaries. Prior to applying the peak-picking algorithm, the novelty function η is lowpass-filtered (typically, Gaussian filter with zero mean and standard deviation 5.0).

The lowpass filtering makes the resulting segmentation boundaries more reliable and robust against various audio processing steps. As in [6], we use the median-based peak-picking algorithm. The resulting segmentation boundaries do not change significantly after various processing steps including speed changes.

The length of each audio segment is normalized into a certain length (typically 8192) by resampling. We divide the resampled audio into two non-overlapping frames. The spectrum of each frame is divided into M critical bands [7] (typically M=16 from 300 Hz to 5300 Hz, which is known to be relevant to human perception and robust [2]), and at each critical band, the NSSC is calculated. Let $P[n, k]$ be the short-time power spectrum of an audio signal at frequency bin k of the n-th frame. The centroid of the n-th subband audio spectrum is defined as

$$C[n,m]=\frac{\sum_{k=B[m]+1}^{B[m+1]} kP[n,k]}{\sum_{k=B[m]+1}^{B[m+1]} P[n,k]} \quad (3)$$

where $B[m]$ denotes the frequency boundary of the m-th critical band. Through the above normalization, the subband centroids are resilient against the equalization of the audio spectrum [8]. By normalizing the the range of $C[n, m]$, NSSC of the m-th critical band is given by [8]

$$NSSC[n,m]=\frac{C[n,m]-(B[m]+B[m+1])/2}{B[m+1]-B[m]} \quad (4)$$

Then NSSC has a range between -0.5 and 0.5 regardless of the critical bands.

As we divide a segment into two frames, we obtain $2M$ NSSCs from a segment. The NSSCs of each segment, denoted by x, are further normalized by its mean and standard deviation (denoted by m_x and σ_x respectively) as follows:

$$p[n]=\frac{x[n]-m_x}{\sigma_x} \quad (5)$$

where n=1,2, ... , $2M$. Finally the $p[n]$ is used as the fingerprint of a segment.

2.2 Local Invariant Fingerprint Matching

Conventionally the fingerprint matching is performed on a fingerprint block (fingerprints from N consecutive audio segments) [1-4] since a fingerprint from one segment does not contain enough information to identify an audio. However, the segmentation boundaries obtained from a distorted audio may be different from those from the original audio. If one segmentation boundary is missing or added, the segments after the segmentation boundary are out of synchronization as shown in Fig. 3. Thus we propose a novel local fingerprint matching method.

In the proposed method, we regard each segment of an audio as an independent information source for the final decision whether the two audio clips are from the same audio or not. A fingerprint block (fingerprints of N consecutive segments) is used for fingerprint matching. By querying the reference-fingerprint database (DB), which contains all the fingerprints of audios that we want to identify, with each fingerprint in the block, we obtain the candidate matched DB positions corresponding to the block. If

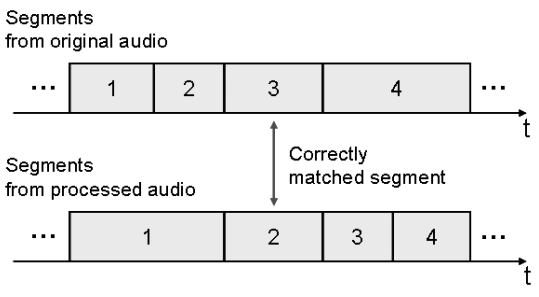

Fig. 3. An example of extracted segmentation boundaries from the original (top) and the processed (bottom) audio. Each rectangular block represents a segment.

the same DB position is matched more than ν times out of the N fingerprints in a fingerprint block, the query audio block is claimed to be the audio in the DB corresponding to the matched DB position.

For the DB query of each fingerprint, the square of the Euclidean distance measure D is used as follows:

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

where p and q are the fingerprints from the different audio segments. If D is smaller than a threshold T, it is hypothesized that the two audio segments p and q are from the same audio segment. The selection of the threshold T is usually performed based on the false alarm rate P_{FA} which is the probability to declare different audios as similar. By the central limit theorem, the distance measure D has a normal distribution if M is sufficiently large and the contributions in the sums are sufficiently independent [1]. Through the normal approximation $N(m_D, \sigma_D^2)$ of the distance measure D, the false alarm rate $P_{FA\text{-}seg}$ of individual matching is given as follows:

$$\begin{aligned}P_{FA-seg}&=\int_{-\infty}^{T}\frac{1}{\sqrt{2\pi}\,\sigma_D}exp\left[\frac{-(x-m_D)^2}{2\sigma_D^2}\right]dx\\&=\frac{1}{2}erfc\left(\frac{m_D-T}{\sqrt{2}\,\sigma_D}\right)\end{aligned} \tag{7}$$

By assuming that the fingerprint p is independent and identically distributed random process with zero mean and unit variance from (5), the mean and the standard deviation of D can be estimated as $m_D = 2$ and $\sigma_D = 0.4677$ from the typical value of $E[p^4[n]] = 2.5$. The details of the estimation were covered at [2].

By viewing each fingerprint as independent information sources, we claim that the two fingerprint blocks are from the same audio if the same DB position is matched more than ν times. Then the false alarm probability of an audio block $P_{FA\text{-}blk}$ is given by

$$P_{FA-blk}=\sum_{i=\nu}^{N}\binom{N}{i}P_{FA-seg}^{i}(1-P_{FA-seg})^{(N-i)} \tag{8}$$

For a certain value of $P_{FA\text{-}blk}$, the threshold T for D can be determined from (7) and (8).

III. Experimental Results

The proposed method was tested using the fingerprint DB generated from 1000 songs that include various genres, such as classic, jazz, pop, rock, and hiphop. DB search and fingerprint matching were performed using a fingerprint extracted from each segment of the input query audio. The local invariant fingerprint matching method in Section 2.2 is used to finally verify the matching results.

To test the robustness of the proposed method, the original audios were subjected to various kinds of audio processing steps, including MP3 compression, equalization, linear speed change, and time-scale modification, and their respective fingerprint blocks were extracted (see [1] for a detailed description of the processing steps). For the two different values of ν, the false rejection rates of the proposed fingerprint matching method are shown in Table 1 for randomly selected 2000 audio blocks in the 1000 song DB. In

Table 1. Probability (%) of correct fingerprint matching for (ν = 4, T = 0.254) and (ν = 5, T = 0.448).

Processing	ν = 4	ν = 5	Without Seg. [2]
Linear speed (+2 %)	100	100	82.35
Linear speed (+5 %)	98.95	99.90	0.00
Linear speed (+10 %)	91.65	95.20	0.00
Linear speed (-2 %)	100	100	75.70
Linear speed (-5 %)	98.25	99.30	0.00
Linear speed (-10 %)	83.30	89.15	0.00
Time scale (+2 %)	87.05	87.10	100
Time scale (-2 %)	97.25	96.85	100
MP3 compression (32 kbps)	98.20	98.60	99.65
Equalization (3 dB)	99.60	99.55	100

the experiments, we use M=16, N=25, and the matching threshold T corresponding to $P_{FA\text{-}blk}=10^{-12}$. The table shows that the proposed method is robust against the common audio processing steps as well as speed changes. Although the previous method [2], based on fixed-length segment, showed slightly better performance on compression and equalization, it was highly vulnerable to the linear speed changes. In practice, the weak robustness against speed changes is a pitfall in broadcast monitoring and could be easily utilized by attackers to avoid filtering in file-sharing system. However, some of the recently-published audio fingerprinting methods [9-10] still could not handle the linear speed changes properly and thus show robustness up to just 2% speed changes [9]. Specialized schemes [5,11,12], that can handle the linear speed changes, have been proposed. But those countermeasures to speed changes typically degrade robustness against other distortions as in the proposed method (Table 1). Thus future work is needed to develop a method which can maintain high robustness against all kinds of audio distortions including compression, equalization, filtering, speed changes, and time-scale modifications.

IV. Conclusion

In this paper, we show that the robustness of an audio fingerprinting against speed changes can be significantly improved by synchronizing fingerprint extraction and matching with the segmentation boundaries. The limited repeatability of the segmentation boundaries was overcome by the local invariant matching in which only small numbers of correct boundaries are needed for claiming successful matching. Experimental results show that the proposed method is highly robust against significant amount of speed changes and other common distortions including compression and equalization.

Acknowledgments

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 (2012012876).

Referance

1. J. Haitsma and T. Kalker, “A highly robust audio fingerprinting system,” *Proc. ISMIR 2002*, pp. 144-148, 2002.
2. J. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. Yoo, “Audio fingerprinting based on normalized spectral subband moments,” *IEEE Signal Processing Letters*, vol. 13, no. 4, pp. 209-212, Apr. 2006.
3. J. Seo and S. Lee, “Robust audio fingerprinting using compressed-domain features,” *Journal of Acoustical Society of Korea*, vol. 28, no. 4, pp. 375-382, May 2009.
4. P. Cano, E. Batlle, T. Kalker, and J. Haitsma, “A review of algorithms for audio fingerprinting,” *Proc. IEEE Workshop on Multimedia Signal Processing*, pp. 169-173, 2002.
5. J. Haitsma and T. Kalker, “Speed-change resistant audio fingerprinting using auto-correlation,” *Proc. IEEE ICASSP 2003*, pp. 728-731, 2003.
6. J. Bello, C. Duxbury, M. Davies, and M. Sandler, “On the use of phase and energy for musical onset detection in the complex domain,” *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 553-556, June 2004.
7. E. Zwicker and H. Fastl, Psychoacoustics: Facts and Models, *Springer-Verlag*, 1999.
8. J. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. Yoo, “Audio fingerprinting based on normalized spectral subband centroids,” *Proc. IEEE ICASSP 2005*, pp. 213-216, 2005.
9. M. Mohri, P. Moreno, and E. Weinstein, “Efficient and robust music identification with weighted finite-state transducers,” *IEEE Tr. Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 197-207, Jan. 2010.
10. N. Chen, H. Xiao, and W. Wan, “Audio hash function based on non-negative matrix factorisation of mel-frequency cepstral coefficients,” *IET Information Security*, vol. 5, no. 1, pp. 19-25, Mar. 2011.
11. E. Dupraz and G. Richard, “Robust frequency-based audio fingerprinting,” *Proc. IEEE ICASSP 2010*, pp. 281-284, 2010.
12. B. Zhu, W. Li, Z. Wang, and X. Xue, “A novel audio

fingerprinting method robust to time scale modification and pitch shifting," *Proc. ACM Multimedia 2010*, pp. 987-990, 2010.

Profile

▸ Jin Soo Seo

received the B.S., M.S., and Ph.D. degrees from Korea Advanced Institute of Science and Technology in 1998, 2000, and 2005 respectively, all in electrical engineering. While working toward Ph.D. degree, he was an adjunct research staff at Electronics and Telecommunications Research Institute (ETRI) in 2001 and a thesis trainee at Philips Research Eindhoven in 2002. He was a senior researcher at ETRI from 2006 to 2008. He joined the Department of Electrical Engineering at Gangneung-Wonju National University in 2008. His research interests are speech and audio processing, multimedia retrieval, and pattern recognition.