Feature Extraction for Audio Fingerprinting Using Wavelet Transform

Shang-Lin Hsieh
Dept. of CSE
Tatung University, Taiwan
slhsieh@ttu.edu.tw

Hsing-Chih Wang
Dept. of CSE
Tatung University, Taiwan
forsean@pchome.com.tw

Abstract

This paper proposes a novel feature extraction scheme for audio fingerprinting using discrete wavelet transform (DWT). The proposed scheme reduces the granularity, i.e., the minimal length of audio, needed for identification in an audio fingerprinting system. The scheme first decomposes the video frame into sub-bands by DWT. Then, it calculates statistical values from the DWT coefficients and extracts features according to the statistical data. Finally, it uses the features to construct the fingerprint. The proposed scheme has two advantages: (1) it needs smaller fingerprint granularity than other previous work; (2) it is not only reliable but also robust against various signal degradations according to the experimental results.

Keywords: audio fingerprinting, discrete wavelet transform, feature extraction

1. Introduction

Non-invasive techniques have received much interest in recent years. They are used in lots of applications such as retrieval, recognition and authentication of digital contents. The non-invasive techniques are performed without modifying the original signal but only analyzing it. Among them, fingerprinting is the most important application that provides a fast and reliable method for content identification.

In an audio content, audio fingerprinting extracts some identifiable features, i.e., the fingerprint, from a piece of audio and stores it in a database. When the system is presented with an unidentified piece of audio, its fingerprint is extracted and matched against those stored in the database. Using fingerprints and matching algorithms, distorted versions of a recording can still be identified as the same audio signal [7] [9] [10].

Haitsma and Kalker proposed five main parameters of an audio fingerprinting system in [3] [4]. They are:

- **Robustness:** determining how severely an audio clip can be processed before it cannot be recognized anymore.
- **Reliability:** expressing the probability that an audio is incorrectly identified.
- **Fingerprint size:** giving the number of bits of a fingerprint.
- **Granularity:** indicating the minimal length of audio needed for identification.
- **Search speed and scalability:** representing how fast a fingerprint can be found in a large fingerprint database.

The five parameters have influence on each other. For example, the increase of robustness might lead to the increase of search speed.

In this paper, we propose an novel feature extraction scheme to find a robust and reliable fingerprint for an audio fingerprinting system. We specifically address the problem of reducing the granularity required by an audio fingerprinting system.

This paper is organized as follows. The next section describes some previous work for fingerprint extraction and discusses their granularities. Section 3 describes the discrete wavelet transform adopted by the proposed scheme. Section 4 details the feature extraction phases of the scheme. Section 5 provides the evaluation results on the performance using the fingerprint extracted by the scheme. Finally, Section 6 draws the conclusion.

2. Related Work

This section describes the feature extraction methods of some previous work and discuss their granularities.

In the audio hashing approaches [3] [5] [11], the input signal is transformed into frequency domain using Fast Fourier Transform (FFT). Audio features are represented as the spectrum of Fourier transform.

Haitsma and Kalker [3][5] proposed a fingerprint extraction method which extracts 32-bit sub-fingerprints for every interval of 11.6 milliseconds. A fingerprint block consists of 256 subsequent sub-fingerprints, corresponding to about 3 seconds.

Mapelli and Lancini [2] used a hamming window, which includes a certain number of frames, to generate a hash block that is the smallest identifiable piece of a song. The size of the window and hence the size of the block correspond to about 3 seconds.

Lancini et al. [11] split the original song into shorter parts called frames. 64 frames form a hash window, which is the minimum identifiable shot. A shot corresponds to about 2 seconds.

Furthermore, Lu [1] proposed an audio fingerprinting method based on analyzing time-frequency localization of signals for audio recognition. Audio features are represented as the spectrum of continuous wavelet transform (CWT). The fingerprint block of the method corresponds to 3 seconds.

Table 1. Granularities of the related work.

Feature extraction scheme	Granularity
Haitsma and Kalker [3][5]	3 seconds
Mapelli and Lancini [2]	3 seconds
Lancini et al. [11]	2 seconds
Lu [1]	3 seconds
J.S. Seo et al.[6]	9.845 seconds

In addition, J.S. Seo et al. [6] presented an audio fingerprinting method based on the normalized spectral sub-band centroid (SSC). Fingerprint matching is performed using the square of the Euclidean distance. A fingerprint block in the method corresponds to 9.845 seconds.

The granularities of the related work mentioned above are summarized in Table 1. It shows the granularities of the related work are all greater than 2 seconds. The proposed scheme reduces the granularity to less than 2 seconds.

3. Discrete Wavelet Transform

The proposed scheme adopts the one dimension (1D) DWT. There is now great interest in using wavelet transform for feature extraction in automatic recognition applications. Wavelet transform has taken the place of short time Fourier transform (STFT) to extract sub-band energy features. Furthermore, it has also been used in the decorrelation process of features in place of discrete cosine transform (DCT) [8].

Figure 1 shows the decomposition process of two-level DWT. First, it decomposes an audio excerpt into one low-pass sub-band L1 and one

high-pass sub-band H1. Then, it decomposes the L1 sub-band into two sub-bands, L2 and H2. Therefore, there are three sub-bands (L2, H2 and H1) after the two-level DWT decomposition. Generally, the energy is concentrated in the low frequency sub-band L2. Hence, the proposed scheme generates features from the coefficients in the L2 sub-band.

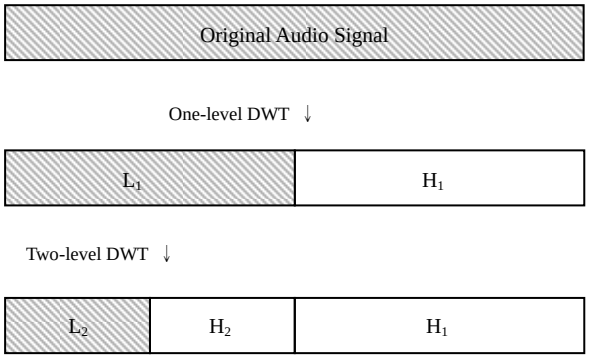


Figure 1 (a) Two-level DWT decomposition of an audio signal.

4. Fingerprint Extraction Scheme

This section describes the proposed fingerprint extraction scheme. The scheme, shown in Figure 2, takes into account the main phases of an extraction algorithm according to Cano *et al.* [9].

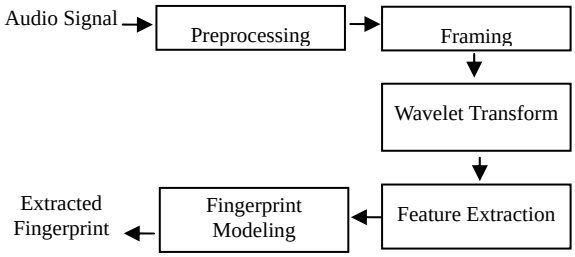


Figure 2 Fingerprint extraction scheme.

The functions of the five major phases in the proposed scheme are described as below.

- **Preprocessing**: converting the audio signal to a certain format.
- **Framing**: dividing the audio signal into overlapping frames.
- **Wavelet Transform**: decomposing a frame into sub-bands and obtaining the coefficients in the L2 sub-band.
- **Feature Extraction**: calculating statistical values from the coefficients and extracting features according to the statistical data.
- **Fingerprint Modeling:** constructing the fingerprint block by the features.

The following subsections describe each phases in detail.

4.1 Preprocessing

The first phase digitizes the audio signal and converts it to an identical format: 16-bit monophonic signal sampled at 44.1 kHz. An audio might suffer various manipulations, e.g., amplitude change, resolution change, resampling, filtering, perceptual audio coding, noise addition, etc. Hence, the phase is necessary for the following phase to work properly. Furthermore, it can improve the efficiency of the algorithm and obtain a better model of the audio signal [2] [11].

4.2 Framing

The phase divides the audio signal into *overlapping* frames. The length of an overlapping frame is 0.37 seconds (16384 samples) with an overlap factor of 63/64. Following the approach in [3] [5], a 32-bit sub-fingerprint is extracted from one frame. A fingerprint block corresponding to about 1.8 seconds of audio is used as the basic unit for identification. In other words, the granularity of the proposed scheme requires is 1.8 seconds.

4.3 Wavelet Transform

The scheme adopts the one-dimensional (1D) Haar DWT to decompose a frame into three sub-bands (L2, H2, and H1) as mentioned in Section 3. The L2 sub-band is selected for feature extraction.

Figure 3 briefly shows an example of the two-level decomposition. Figure 3 (a) shows a frame of an audio signal with 16384 samples. Figure 3 (b) shows the one-level DWT decomposition of an audio signal. There are 8192 coefficients in the L1 sub-band and the H1 sub-band, respectively. The two-level DWT decomposes the L1 sub-band into L2 and H2 sub-bands and each contains 4096 coefficients. They are shown in Figure 3(c).

4.4 Feature Extraction

The phase employs three statistical values generated from the coefficients in the L2 sub-band in order to find robust features. The three statistical values (P_k , k = 1, 2, 3) are:

1. Mean of all coefficients(P_1).
2. Standard deviation of all coefficients(P_2).
3. Mean of the coefficients greater than the third quartile(P_3), i.e., the mean of the last 25% of the coefficients in ascending order.

The following lists the detailed steps of the feature extraction.

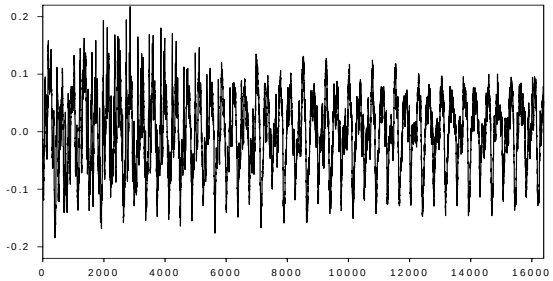
*Step*1. Calculate the three statistical values of the sub-band, B_{p_1} , B_{p_2} , and B_{p_3} .

Step2. Segment L2 sub-band into 32 *non-overlapping* sections.

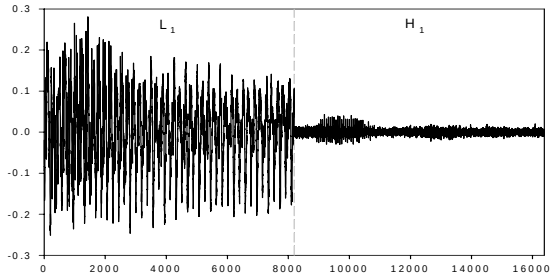
Step3. Calculate the three statistical values of each section, S_{i,p_1} , S_{i,p_2} , and S_{i,p_3} , where i is section number, and i = 1, 2, …, 32.

Step4. Compare the three statistical values between the section and sub-band, and then construct the feature block (Ft_{i,p_k}) according to Eq. 1.

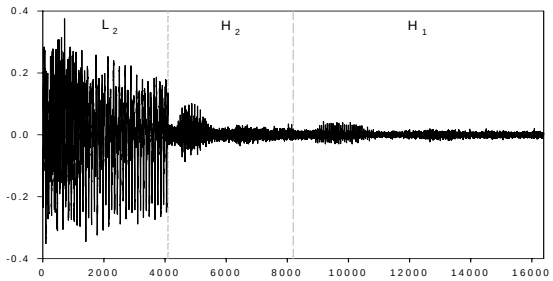
$$Ft_{i,p_k} = \begin{cases} 1, & if\ S_{i,p_k} \geq B_{p_k} \\ 0, & otherwise \end{cases}, \qquad (1)$$



(a)



(b)



(c)

Figure 3 (a) A frame of an audio signal, (b) one-level DWT decomposition of the audio signal, (c) two-level DWT decomposition of the audio signal.

4.5 Fingerprint Modeling

The voting method is used to build the sub-fingerprint. Every sub-fingerprint bit (sF_i) is determined by the corresponding feature block bits, Ft_{i,p_1} , Ft_{i,p_2} and Ft_{i,p_3} , according the following equation.

$$sF_i = \begin{cases} 1, & if\ Ft_{i,p_1} + Ft_{i,p_2} + Ft_{i,p_3} \geq 2 \\ 0, & if\ Ft_{i,p_1} + Ft_{i,p_2} + Ft_{i,p_3} < 2 \end{cases} \qquad (2)$$

Finally, according to the frame order, the fingerprint block is constructed by appending the sub-fingerprints.

Figure 4 shows an example of 256 subsequent 32-bit sub-fingerprints (i.e., a fingerprint block) extracted by the proposed scheme from a short excerpt of "*The moment*" by Kenny G. Figure 4(a) and Figure 4(b) show a fingerprint block from an original CD and the MP3 compressed (128Kbps) version of the same excerpt, respectively. Ideally, the two fingerprints should be identical. However, due to the compression, some of the bits are retrieved incorrectly. The bit errors, which are used as the similarity measure in our fingerprint scheme, are shown in black in Figure 4(c).

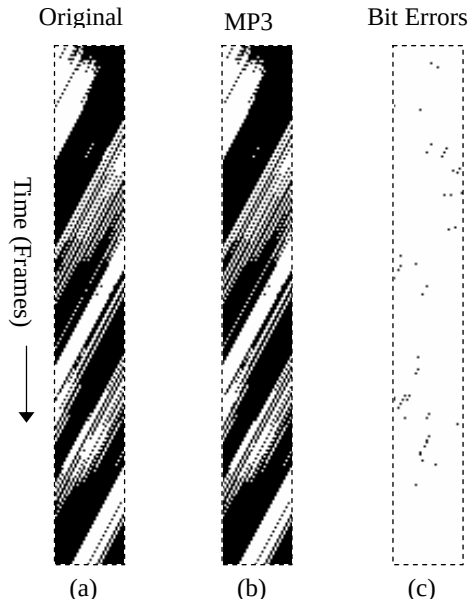


Figure 4 (a) The fingerprint block of the original music clip, (b) the fingerprint block of the compressed version, (c) the difference between (a) and (b) with the bit errors shown in black.

5. Experimental Results

Two experiments were conducted to show the robustness and discrimination of the proposed scheme. Bit Error Rate (BER) is used to assess the experimental effect. The results are obtained from the test of 5 different songs (16 bits monophonic signal sampled at 44.1 kHz): "*The moment*" by Kenny G, "*The 33rd corner*" by Yun-chang Dong, "*Now and forever*" by Richard Marx, "*Goodbye*" by Air Supply, and "*There You'll Be*" by Faith Hill. For each song, a short audio excerpt is selected to take part in the experiments.

5.1 Robustness Experiment

The first experiment uses the following signal degradations to prove the robustness.

- **MP3 Compression:** 192 Kbps, 128 Kbps, and 32 Kbps
- **Windows Media Audio (WMA) Compression:** 48 Kbps and 20 Kbps
- **Real Media Compression:** 20 Kbps
- **Echo addition:** 100ms delay and 40% decay
- **Noise addition:** White noise (SNR = 20 dB)
- **Equalization** A typical 10-band equalizer with the following settings [3]:

Freq.(Hz)	31	63	125	250	500	1k	2k	4k	8k	16k
Gain(dB)	-3	+3	-3	+3	-3	+3	-3	+3	-3	+3

- **Low-pass Filtering:** Cut-off frequencies above 4000Hz.
- **Resampling:** Sampling rate down to 22.05 kHz.
- **Requantization:** Resolution quantized to 8 bits.

The BERs between the fingerprint blocks of the original version and all degraded versions are shown in Table 2. All the resulting BERs are all below 0.11. Furthermore, most of the BERs of the perceptual audio coding (MP3, Real Audio, Windows Media Audio) are below 0.07. The experiment shows that the proposed scheme is robust.

Table 2. BERs after different kinds of signal degradations.

Processing	Song 1	Song 2	Song 3	Song 4	Song 5
MP3/192Kbps	0.00293	0.00354	0.01062	0.00464	0.00647
MP3/128Kbps	0.00586	0.00635	0.01525	0.00940	0.00794
MP3/32Kbps	0.02734	0.01526	0.02551	0.03552	0.02869
WMA/48Kbps	0.00732	0.01013	0.01867	0.04102	0.02771
WMA/20Kbps	0.02795	0.05347	0.04260	0.06116	0.10388
RealAudio	0.03821	0.09570	0.06213	0.06397	0.06714
Echo addition	0.01990	0.02710	0.02099	0.01660	0.02734
Noise addition	0.01502	0.00745	0.00842	0.00842	0.01172
Equalization	0.09900	0.05042	0.04406	0.05859	0.08374
Low-pass filter	0.02747	0.02820	0.03308	0.01819	0.02063
Resampling	0.00024	0.00012	0.00024	0.00159	0.00000
Requantization	0.00964	0.00208	0.00488	0.00464	0.00342

5.2 Discrimination Experiment

Every song should have a unique fingerprint. This experiment tests the difference of the fingerprint blocks between different songs. Table 3 lists the resulting BERs obtained from discriminations between different audio excerpts. The BERs all fall within the interval [0.44 0.56], which is reasonable because BERs are expected to approximate to 0.5 between two different audios [1]. The experiment shows that the proposed scheme is discriminative and hence reliable.

Table 3. BERs of the five different songs.

	Song 1	Song 2	Song 3	Song 4	Song 5
Song 1	0	0.551636	0.455566	0.474854	0.460815
Song 2	0.551636	0	0.495483	0.448853	0.498779
Song 3	0.455566	0.495483	0	0.484863	0.534058
Song 4	0.474854	0.448853	0.484863	0	0.524048
Song 5	0.460815	0.498779	0.534058	0.524048	0

6. Conclusions

In this paper, we presented a novel algorithm for audio feature extraction using the discrete wavelet transform for audio recognition. The robustness experimental results show the extracted features can resist the signal degradations such as MP3 compression, Windows Media Audio (WMA) compression, Real Media (RM) compression, echo addition, noise addition, equalization, low-pass filtering, resampling, and requantization. Furthermore, according to the discrimination experiment results, the extracted features possess high distinguishability. Therefore, the extracted features are robust and reliable. Moreover, a fingerprint block corresponding to 1.8 seconds of audio is used as the basic unit for identification. Consequently, the granularity of the proposed algorithm is smaller than those of the previous work as mentioned in Table 1.

References

[1] C. S. Lu, “Audio Fingerprinting based on Analyzing Time-Frequency Localization of Signals,” IEEE Workshop on Multimedia Signal Processing, pp. 174 – 177, Dec. 2002.

[2] F. Mapelli, R. Lancini, “Audio hashing technique for automatic song identification,” ITRE2003. International Conference on Information Technology: Research and Education, pp. 84 – 88, Aug. 2003.

[3] J. Haitsma, T. Kalker, “A Highly Robust Audio Fingerprinting System,” in Proc. ISMIR’02, 2002.

[4] J. Haitsma, T. Kalker, “Speed-change resistant audio fingerprinting using auto-correlation,” IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '03), Volume 4, pp. IV - 728-31, Apr. 2003.

[5] J. Haitsma, T. Kalker and J. Oostveen, "Robust audio hashing for content identification," in Proc. of the Content-Based Multimedia Indexing, Firenze, Italy, Sept. 2001.

[6] J.S. Seo et al. "Audio Fingerprinting Based on Normalized Spectral Subband Centroids," IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '05), Volume 3, pp. 213 – 216, Mar. 2005.

[7] L. Gomes, P. Cano, E. Gómez, M. Bonnet and E. Batlle, "Audio Watermarking and Fingerprinting: For Which Applications?," Journal of New Music Research, Vol. 32 .1, 2003.

[8] O. Farooq, S. Datta, "Wavelet-based denoising for robust feature extraction for speech recognition," Electronics Letters, Volume 39, Issue 1, pp. 163 – 165, Jan. 2003.

[9] P. Cano, E. Batlle, T. Kalker and J. Haibma, "A Review of Algorithms for Audio Fingerprinting," IEEE International workshop on MMSP, pp. 169 – 173, Dec. 2002.

[10] RIAA-IFPI. "Request for Information on Audio Fingerprinting Technologies," http://www.riaa.com/, July 2001.

[11] R. Lancini, F. Mapelli and R. Pezzano, "Audio Content Identification by using Perceptual Hashing," IEEE International Conference on Multimedia and Expo, 2004.