

Robust Audio Watermarking for Copyright Protection

Chung-Ping Wu, Po-Chyi Su and C.-C. Jay Kuo
Media Fair, Inc., 1055 Corporate Center Dr., Ste 580
Monterey Park, CA 91754

and

Department of Electrical Engineering-Systems
University of Southern California, Los Angeles, CA 90089-2564

E-mail: {chungpin, pochyisu, cckuo}@sipi.usc.edu

ABSTRACT

A digital audio watermarking scheme of low complexity is proposed in this research as an effective way to deter users from misusing or illegally distributing audio data. Previous work on audio watermarking has primarily focused on the inaudibility of the embedded watermark and its robustness against attacks such as compression and noise. In this research, special attention is paid to the synchronization attack caused by casual audio editing or malicious random cropping, which is a low-cost yet effective attack to watermarking algorithms developed before. The proposed scheme is based on audio content analysis and watermark embedding in the Fourier transform domain. A blind watermark detection technique is developed to identify the embedded watermark under various types of attacks.

Keywords: digital watermark, blind watermark detection, audio content analysis, synchronization attack, human auditory system, malicious cropping attack

1. INTRODUCTION

Among all digital media, audio plays a very important role in E-commerce. That is, the fast growth of the Internet and the maturity of audio compression techniques enable the promising market of on-line music distribution. However, since the digital technology allows lossless data duplication, illegal copying and distribution would be much easier than before. This concern does make musical creators and distributors hesitant to step into this market quickly. Recently, researchers have investigated the embedding of inaudible information into digital audio data for the purpose of copyright protection and/or ownership verification. The hidden information is known as the digital watermark. The demand for a mature audio watermarking scheme is clearly high.

In order for the embedded watermark to effectively protect the copyright of the digital audio data, it has been generally agreed¹⁻⁷ that a good watermarking scheme should satisfy the following properties:

- The embedded watermark should not produce audible distortion to the sound quality of the original audio.
- The computation required by watermark embedding and detection should be low. The complexity of watermark detection should be especially low to facilitate its integration into consumer electronic products.
- Watermark detection should be done without referencing the original audio data. This property is known as blind detection.
- The watermark should be undetectable without prior knowledge of the embedded watermark sequence. This property prevents attackers from reversing the embedding process to remove the watermark.
- The embedded watermark should be robust against common signal processing attacks such as filtering, resampling and compression.
- The watermark should survive malicious attacks such as random cropping and noise adding. However, severe attacks that produce annoying noise can be ignored for the survival test.

Several techniques for digital audio watermarking were proposed before. One method is to modify the value of Fourier transform phase coefficients.¹ Another method is to embed decaying echoes with different delay times.² A third technique is to embed the watermark in the scale factors during the MPEG compression process.⁶ There is also a class of algorithms that embed spread spectrum noise in the time domain as the watermark.^{1,3-5}

Although these methods have their own features and properties, they share one common problem. That is, they are vulnerable to the synchronization attack in watermark detection. This problem could be resulted from casual audio editing such as cropping unwanted audio segments or intentional attacks such as randomly deleting or adding samples to watermarked audio data. This *random sample cropping attack* is very effective in interfering with the watermark detection process with respect to the algorithms mentioned above. This attack has a very low computational complexity. Besides, when done correctly, it would not introduce annoying noise to the underlying audio signals. One might argue that such a skillful attack could only be done by a few professionals and not by the majority of consumers. However, once a watermarking method is widely in use, it is almost certain that some professionals would produce and distribute attacking apparatuses so that a majority of common users would be able to perform the skillful attack. One method to solve the synchronization problem was proposed in,⁵ where an exhaustive search algorithm was used and the original audio signal was required. Consequently, its computational complexity is too high. Furthermore, it can only handle the casual editing attack, but not the random sample cropping attack.

In this research, we propose a low complexity solution to the synchronization problem caused by both casual and malicious attacks. The solution is composed of a salient point extraction technique and a Fourier transform domain watermark embedding procedure. Salient point extraction through audio content analysis is done during both watermark embedding and detection processes so that synchronization is regained at each salient point. The extraction algorithm is designed such that salient points remain stable after distortion. The Fourier transform domain watermark embedding and detection is adopted since the frequency domain information is less effected by sample cropping in the time domain.

One common characteristic among existing audio watermarking algorithms is that their watermark is embedded throughout the entire audio signal. However, this may not be the most efficient way to embed and detect watermarks. For a skilled attacker, different amount of attack could be applied to different segments of the audio signal to avoid introducing annoying noise. For example, randomly cropping (deleting) one sample out of every 100 samples in high energy tonal segments of audio signals would produce noticeable noise, but the effect of doing so in low energy segments would be inaudible. Thus, watermarks embedded in highly-attackable areas will face heavier attack and are more likely to be destroyed. The second major contribution of this work is the introduction of “attack-sensitive regions” via audio content analysis. If the watermark is only embedded in attack-sensitive regions where little attack could be applied, the computational complexity of both watermark embedding and detection could be reduced. Although we incorporate the concept of attack-sensitive regions into our own watermark embedding method here, it is our belief that other watermark embedding algorithms will benefit the same concept as well.

By combining techniques of salient point extraction, attack-sensitive region identification, and Fourier transform domain watermark embedding and detection, we propose a complete audio watermark embedding and detection system for copyright protection. This system satisfies all desired properties of watermark design described earlier. Furthermore, it has a very low computational complexity, and is robust to casual and intentional synchronization attacks.

This paper is organized as follows. The procedures of salient point extraction and attack-sensitive region identification through audio content analysis are described in Section 2. The Fourier domain watermark embedding is presented in Section 3. The blind watermark detection algorithm is detailed in Section 4. Experimental results and their analysis are given in Section 5. Finally, concluding remarks are provided in Section 6.

2. AUDIO CONTENT ANALYSIS FOR WATERMARKING

In our system, audio content analysis is performed for the purposes of salient point extraction and attack-sensitive region identification. Salient points in an audio signal allow watermark detection to resynchronize at these locations. Synchronization by salient points has far less complexity than exhaustive search and makes blind watermark detection possible. It should be noted that we do not insert salient points, but extract them from the raw audio via content analysis. This approach has two advantages over explicitly embedding synchronization signals. One is that our

content analysis approach does not introduce any distortion to the original audio signal since we do not add anything to it. The other is that the explicitly added synchronization signal is more likely to be taken out by attackers.

A good salient point extraction method should produce approximately the same set of salient points from audio signals before and after attacks such as audio compression, low-pass filtering and noise adding. To achieve this, we extract salient points based on audio features that are sensitive to human ears. In this way, if an attacker wants to destroy these salient points, he/she would have to alter these features and produce noticeable distortions. We choose the energy variation as the main feature for salient point extraction because the associated computational cost is low and alterations in this feature would be audible. Salient points are extracted as locations where the audio signal energy is fast climbing to a peak value.

The procedure of attack-sensitive region identification aims at decreasing the watermark embedding and detection complexity. Thus, it is important that the identification process itself does not require too much computation. In this work, we integrate attack-sensitive region identification process into the salient point extraction process so that almost no extra computation is needed for attack-sensitive region identification. The attack that we are mainly concerned with is the random sample cropping attack. The corresponding attack-sensitive regions, as described in Section 1, is the high energy tonal region. Since salient points chosen with our algorithm are located at positions where the audio signal energy is fast climbing to a peak, the region following each salient point would contain high energy. We simply define this region as the attack-sensitive region, so that no additional computation is needed.

Consider an original audio signal $x(n)$ of N samples, i.e. $n = 1, \dots, N$. The procedure of audio content analysis is stated as follows.

1. The audio signal is first band-pass filtered to remove low and high frequency components to which human ears are not sensitive.
2. For each sample $x(n)$ in the audio signal, the total energy of r samples directly before $x(n)$ and r samples directly after $x(n)$ are separately calculated.

$$\begin{cases} E_{before}(n) = \sum_{i=-r}^{-1} x^2(n+i), \\ E_{after}(n) = \sum_{i=0}^{r-1} x^2(n+i) \end{cases} \quad (1)$$

3. The ratio of these two energy values is calculated for $x(n)$, $n = 1, \dots, N$, i.e.

$$ratio(n) = \frac{E_{after}(n)}{E_{before}(n)} \quad (2)$$

4. If $ratio(n) > T_1$ and $E_{after}(n) > T_2$, then $x(n)$ is labeled as an energy fast-climbing point.
5. The energy fast-climbing points usually appear in groups. Points that are separated by less than T_3 samples are merged into one larger group.
6. Within each group, the sample with the largest $ratio(n)$ is labeled as one salient point.
7. If the salient point is derived from a group where the largest $ratio(n)$ times the number of samples in the group is less than a threshold T_4 , this salient point is deleted.
8. The first 2^p samples following each salient point are selected as the attack-sensitive region.

Step 1 in the above procedure prevents attackers from adding inaudible large energy signals to interfere with our salient point extraction process. Steps 2 to 6 search for points where the energy is climbing fastest. It is our observation that this energy climbing feature is very difficult to disturb without introducing audible distortion. Thus, it would produce stable salient points. The condition $E_{after}(n) > T_2$ in Step 4 prevents $x(n)$ from being labeled as an energy fast-climbing point just because $E_{before}(n)$ is nearly zero. Steps 5 to 7 improve the salient point stability by pre-merging and pre-deleting groups that might merge or disappear when small distortions are introduced by using compression or the addition of noise. Also note that Step 8 does not require any additional computation.

There are four thresholds, i.e. T_1 to T_4 , in the procedure stated above. From experiments, we observe that thresholds T_3 and T_4 can be set to 30 and 100, respectively, to obtain good results while T_1 and T_2 should be

adaptively adjusted. Threshold T_1 effects the total number and the behavior of the extracted salient points. When T_1 is high, the number of salient points is small, and they tend to be very stable against compression and noise attacks. However, too few salient points gives us few places to embed watermarks. Empirically speaking, we need at least 30 salient points to ensure successful watermark detection. Since it is desirable to detect a watermark by using as short as an audio segment of 15 sec, we need on the average 2 salient points per second for the audio data. Thus, salient point extraction is performed several times by using different values of T_1 , and the result that is closest to 2 salient points per second is selected. Threshold T_2 is adaptively selected to give an rough estimate of the average energy of a 5 second neighborhood.

Since the total number of salient points is by several orders less than the total number of samples, it is clear that Steps 5 to 8 demands a relatively lower computation cost than Steps 1 to 4. In Step 2, $E_{before}(n)$ and $E_{after}(n)$ can be calculated in constant time via

$$\begin{cases} E_{before}(n) = E_{before}(n-1) - x^2(n-r-1) + x^2(n-1) \\ E_{after}(n) = E_{after}(n-1) - x^2(n-1) + x^2(n+r-1) \end{cases} \quad (3)$$

Thus, the complexity of Step 2 is $O(N)$. Note also that the complexity of Steps 1, 3 and 4 are apparently $O(N)$. The total complexity of the proposed algorithm for audio content analysis is $O(N)$ (i.e. it is linear to the size of the original audio data).

3. FOURIER TRANSFORM DOMAIN WATERMARK EMBEDDING

Although salient points are selected to be as stable as possible, it is difficult to get exactly the same salient points after some audio processing such as compression. A certain amount of displacement in the location of salient points is common and should be tolerated. If we embed and detect watermark in the time domain, it is obvious that even a small amount of displacement would have a problem since embedding and detection cannot be synchronized. However, this problem is alleviated by considering the magnitude coefficients of the discrete Fourier transform.

This property is illustrated in Figure 1, where $a(i)$, $i = 1, \dots, 2^p$, is the watermarked region. The watermark is embedded in $|A(k)|$, $k = 1, \dots, 2^p$, where $A(k)$ is the discrete Fourier transform coefficient of $a(i)$. Suppose that the salient point is displaced in the detection process, and the watermarked region is mistaken to be another region $b(i)$, $1 \leq i \leq 2^p$.

However, it is a well known property that if $c(i)$ is formed by moving the right-most part of $a(i)$ to the left-most part, then $c(i)$ and $a(i)$ have identical discrete Fourier transform magnitude coefficients, i.e.

$$|C(k)| = |A(k)|, \quad k = 1, \dots, 2^p, \quad (4)$$

Let us denote the difference between $b(i)$ and $c(i)$ with

$$d(i) = c(i) - b(i), \quad i = 1, \dots, 2^p. \quad (5)$$

Then, we have

$$\begin{aligned} |B(k)| &\approx |C(k)| + |D(k)| \\ &= |A(k)| + |D(k)|, \quad k = 1, \dots, 2^p \end{aligned} \quad (6)$$

Thus, from (6), we see that the error caused by the displaced salient point is $|D(k)|$. There is no disastrous mis-synchronization effect in the frequency domain. When the displacement amount is small relative to the window size, the energy in $|D(k)|$ is small.

In order for the embedded watermark to be inaudible, it is common to utilize the temporal and frequency masking effects of the human auditory system (HAS).^{1,4,5} Temporal masking refers to the effect that weaker signals immediately before and after a stronger signal may be inaudible while frequency masking refers to the effect that when two signals occur simultaneously and are close together in the frequency, the stronger signal may make the weaker one inaudible.

Since our watermark is only embedded in attack-sensitive regions, which have a high energy value, the temporal masking effect is used. That is, the weak-energy watermark is masked by the high energy audio samples in these regions. To take the advantage of the frequency masking effect as well, the proposed scheme only embeds the

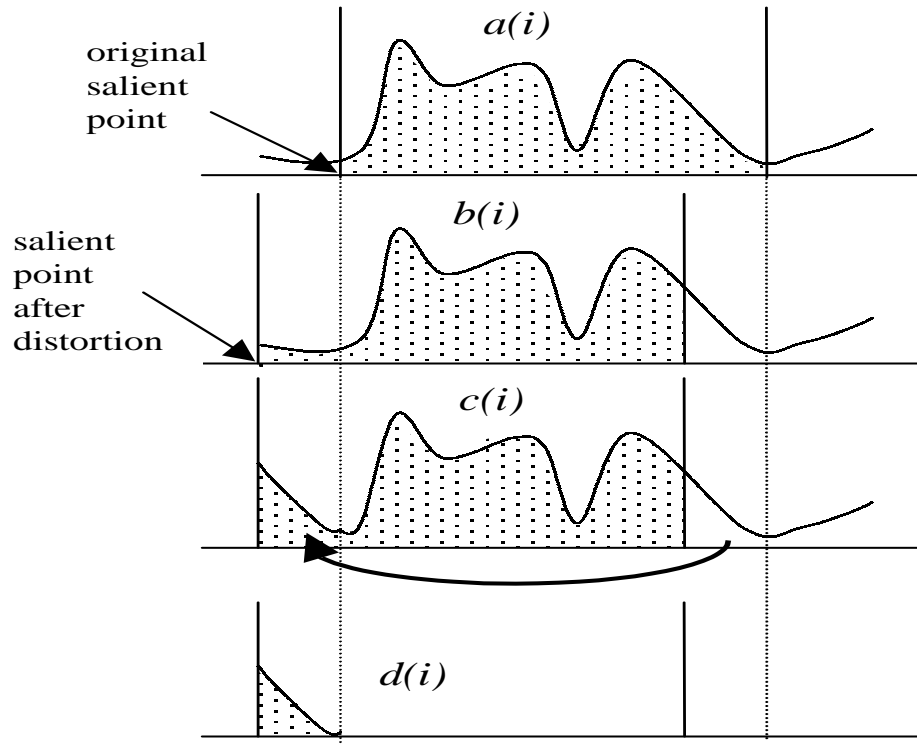


Figure 1. The effect of salient point displacement on the discrete Fourier transform domain watermarking.

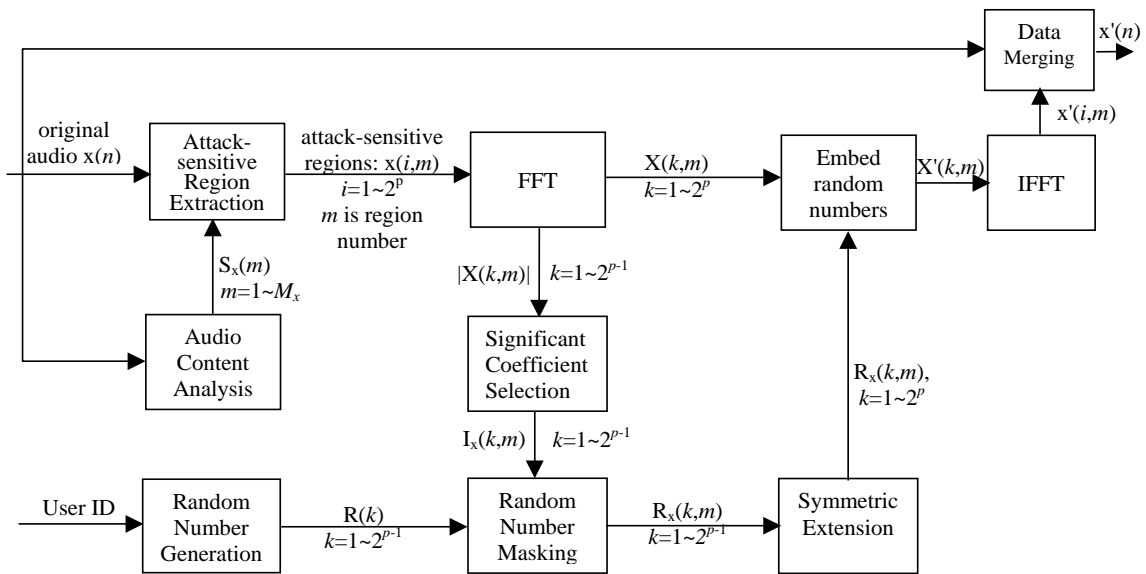


Figure 2. The block diagram of watermark embedding in the discrete Fourier transform domain.

watermark signal in the magnitude of the discrete Fourier transform coefficients that have large values. We select the q largest DFT magnitude coefficients in each attack-sensitive region for watermark insertion.

The block diagram of the DFT domain watermark embedding is shown in Fig. 2. The watermark $R(k)$ is a sequence of 2^{p-1} random numbers generated with user's ID as the seed. Each random number is independent, and has the zero mean and the unity variance. Meanwhile, the original audio is analyzed as described in Section 2 to identify salient points $S_x(m)$. Then, 2^p audio samples following each salient point in the original audio are extracted as the attack-sensitive region denoted by $x(i, m)$. The discrete Fourier transform is performed on $x(i, m)$ via FFT to result in $X(k, m)$, which stands for the k th DFT coefficient in the m th region. The first half of FFT coefficients in each region are examined for its q coefficients with the largest magnitudes. They are used as the position for watermark insertion. The reason to omit the second half of the DFT coefficients is that these two halves have symmetric magnitudes. The result of this selection process is recorded in $I_x(k, m)$, which is set to 1 if $|X(k, m)|$ was selected, and otherwise set to 0. For random number masking, $I_x(k, m)$ is used as a binary mask to filter out random numbers corresponding to DFT magnitude coefficients with small values. This is done with the equation

$$R_x(k, m) = I_x(k, m)R(k), \quad k = 1 \sim 2^{p-1}. \quad (7)$$

With symmetric extension, the watermark sequence is extended via

$$\begin{cases} R_x(k, m) = 0, & k = 2^{p-1} + 1 \\ R_x(k, m) = R_x(2^p - k + 2, m), & k = (2^{p-1} + 2) \sim 2^p \end{cases} \quad (8)$$

Then, $R_x(k, m)$ is embedded into $X(k, m)$ via

$$X'(k, m) = X(k, m)[1 + \alpha R_x(k, m)], \quad k = 1 \sim 2^p, \quad (9)$$

where α is a constant usually less than 0.2. Finally, data merging is performed by replacing signals in the attack-sensitive region in the original audio with $x'(i, m)$.

The computational complexity of the watermark embedding process depicted in Fig. 2 is dominated by audio content analysis and FFT. FFT of each attack-sensitive region demands computation of $O(2^p \log(2^p)) = O(p2^p)$. Thus, FFT of all attack-sensitive regions takes $O(pM_x2^p)$, where M_x is the total number of attack-sensitive regions extracted from the original audio. Note that M_x2^p is the total number of samples in attack-sensitive regions, which cover much less than one-tenth of all original samples. i.e. $M_x2^p < N/10$. With our experiments, p should be set to around 10 to get good results. Hence, we have $O(pM_x2^p) = O(10 \times N/10) = O(N)$. As shown in Section 2, the complexity of audio content analysis is also $O(N)$. To conclude, the whole watermark embedding process has a linear complexity.

4. BLIND WATERMARK DETECTION

The block diagram of the proposed blind watermark detection process is depicted in Fig. 3. Note that the structure of watermark detection before the correlation calculation stage is exactly the same as watermark embedding. In the correlation calculation stage, the average correlation coefficient between $|Y(k, m)|$ and $R_y(k, m)$ is calculated by

$$\text{average correlation coefficient} = \frac{1}{M_y} \sum_{m=1}^{M_y} \frac{\sum_{k=1}^{2^p} |Y(k, m)| \times R_y(k, m)}{\sqrt{\sum_{k=1}^{2^p} |Y(k, m)|^2} \times \sqrt{\sum_{k=1}^{2^p} R_y^2(k, m)}}, \quad (10)$$

where M_y is the total number of attack-sensitive regions extracted from $y(n)$. Then, the average correlation coefficient is compared to a threshold to decide whether there is watermark embedded.

It can be shown that the average correlation coefficient retrieved from a watermarked signal is always significantly greater than that of the original signal. In the ideal case, if no attack is applied to the watermarked audio, the input audio $y(n)$ to the watermark detection unit is identical to the output audio $x'(n)$ of Fig. 2. By assuming that audio-content analysis in Figs. 2 and 3 come up with exactly the same set of attack-sensitive regions, we have

$$S_x(m) = S_y(m), \quad (11)$$

$$Y(k, m) = X'(k, m), \quad (12)$$

$$R_y(k, m) = R_x(k, m). \quad (13)$$

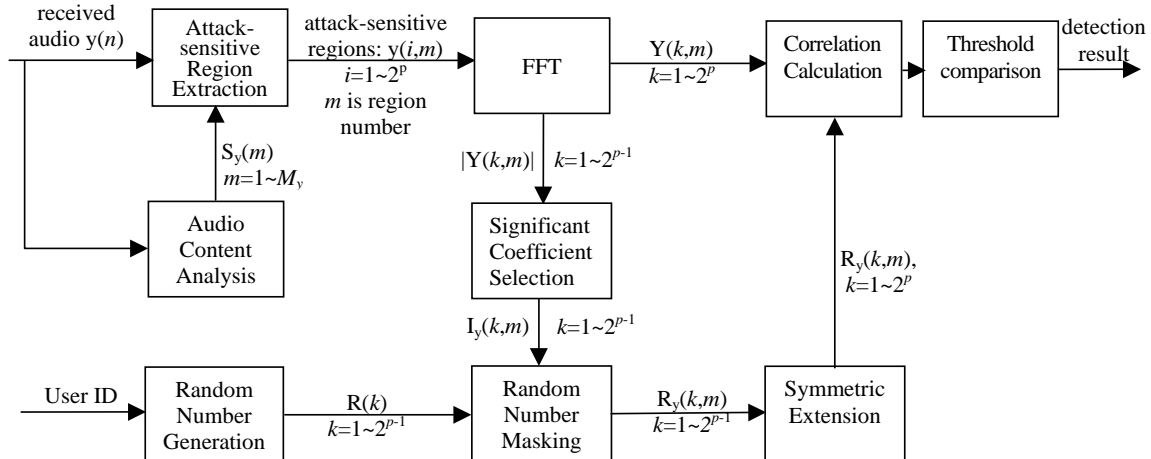


Figure 3. The block diagram of blind watermark detection.

Consequently, the nominator in (10) can be estimated via

$$\begin{aligned}
 \sum_{k=1}^{2^p} |Y(k, m)| \times R_y(k, m) &= \sum_{k=1}^{2^p} |X(k, m)| \times [1 + \alpha R_y(k, m)] \times R_y(k, m) \\
 &= \sum_{k=1}^{2^p} |X(k, m)| \times R_y(k, m) + \alpha \sum_{k=1}^{2^p} |X(k, m)| \times R_y^2(k, m) \\
 &\approx 0 + \alpha \sum_{k=1}^{2^p} |X(k, m)| \times R_y^2(k, m)
 \end{aligned} \tag{14}$$

Since $|X(k, m)|R_y^2(k, m)$ is always greater than zero, the value of (14) is significantly greater than zero, which in turn makes (10) significantly greater than zero. On the other hand, if no watermark is embedded in the input audio $y(n)$, $|Y(k, m)|$ and $R_y(k, m)$ are uncorrelated so that the nominator of (10) is near zero. Thus, values of the average correlation coefficient retrieved from watermarked and unwatermarked signals are well separated.

In reality, attack-sensitive regions extracted in the watermark embedding and detection processes are not identical. Moreover, the result of significant coefficient selection in watermark embedding and detection are also slightly different. These differences make the average correlation coefficient smaller. The stronger an attack is applied to the watermarked audio signal, the more the average correlation coefficient decreases. However, in all cases of reasonable attacks, the average correlation coefficient retrieved from watermarked signals is always well above that of unwatermarked signals.

The final step in watermark detection is to compare the average correlation coefficient with a threshold. The criterion for selecting the threshold is to minimize the expected cost of detection errors. Note that the cost of miss (i.e. failure to detect when there is a watermark) is different from the cost of false alarm (i.e. claim a detection while there is no watermark). Although these costs vary in different applications, it is generally true that the cost of false alarm is much greater than the cost of miss. The false alarm rate should be extremely low because it undermines the credibility of the watermarking method to prove copyright ownership. In contrast, the constraint on the miss (or failure-to-detect) rate need not be so stringent, since the failure-to-detect rate of 1% or 10% might have a similar effect in scaring people away in illegally copying audio data. To conclude, the detection threshold should be set relatively high to ensure no false detection happens.

5. EXPERIMENTAL RESULTS

We test the inaudible and robust properties of the proposed watermarking scheme on three pieces of audio signals: flute and cello music, guitar music, and human vocal with the background music. All signals are sampled at a frequency of 22.05 kHz, and each piece is about 15 seconds long. Also, the FFT window size (i.e. 2^p) is chosen to be 1024.

salient point location extracted from original file	salient point location extracted from distorted file	salient point shift amount between two files	salient point location extracted from original file	salient point location extracted from distorted file	salient point shift amount between two files	salient point location extracted from original file	salient point location extracted from distorted file	salient point shift amount between two files
4401	4557	-156	138454	138392	62	343182	343309	-127
6581	6581	0	144478	144489	-11	347048	347233	-185
14196	none		145827	145823	4	351030	351003	27
14463	none		153485	153484	1	359383	359351	32
19464	19471	-7	185107	185056	49	382173	382186	-13
21092	21063	29	192565	192297	268	384255	384259	-4
28657	28651	6	216786	216784	2	389912	389914	-2
44152	44104	48	224510	224555	-45	391882	391884	-2
59635	59637	-2	232790	232808	-18	397653	397654	-1
91080	91126	-46	242895	242878	17	399407	399526	-119
94883	94879	4	264519	264518	1	406960	none	
98548	98545	3	271803	271803	0	422233	422234	-1
none	105946		273508	273507	1	426680	426682	-2
112475	112471	4	297097	297097	0	429936	none	
127941	127958	-17	304761	304760	1	437456	437473	-17
129319	129315	4	320039	320013	26	444820	444795	25
131028	131025	3	335700	335700	0	460640	460643	-3

Table 1. An example of the comparison between salient points extracted from original and processed audio files. The original audio is 15 sec of flute and cello music, and the processing is MPEG-1 Layer III compression with a bit rate 16 kbps. The rows printed in the bold type are regarded as failures, and the success rate in this example is 78.5%.

5.1. Audio content analysis

The effectiveness of the proposed audio content analysis is measured by its ability to extract the same set of salient points from audio signals before and after signal attack and/or processing. An example of the comparison between the salient points extracted from the original and processed files is shown in Table 1. The processing in this example is the MPEG Layer III compression with bit rate 16 kbps, which is a very high ratio compression scheme. As we can see from this example, almost every salient point is more or less shifted by a few points. However, as explained in Section 3, this does not cause a catastrophic effect on watermark detection. Empirically, a displacement of less than 50 points produces very little decrease to the average correlation coefficient in watermark detection. Therefore, it should be viewed as successful salient point extraction. Some salient points may disappear and some may be created after processing. However, again these phenomena only cause a marginal deterioration to detection results. The success rate of correctly extracting salient points from the three pieces of audio after processing are listed in Table 2. The vocal music with the music background has a lower success rate than the other two due to its smooth energy variation. According to our observations, a success rate of more than 60% is enough for correct watermark detection and all cases in Table 2 meet this criterion.

5.2. Watermark Embedding

The quality of the proposed watermarking method is evaluated by using the blind listening test. Listeners are presented with the original and watermarked audio without the knowledge of which one is watermarked. They are asked to tell which one has better sound quality. We do not use the question “whether any differences could be detected between the two audio signals”⁵ since people tend to imagine the difference while they actually cannot hear any. In fact, several listeners reported that audio signals were different when the same piece of audio clip was played twice.

Nine people took the listening test, and the percentage of preferring the original audio to the watermarked audio is given in Table 3. The result shows that about one half of listeners preferred watermarked audio to the original. Therefore, no audible distortion is introduced by the embedded watermark.

Test Audio	Distortion by MPEG compression			Distortion by white noise		
	Success rate	Failure due to excessive shifting	Failure due to disappeared and inserted salient points	Success rate	Failure due to excessive shifting	Failure due to disappeared and inserted salient points
flute and cello music	78.5%	11.8%	9.7%	95.4%	2.3%	2.3%
guitar music	87.3%	10.6%	2.1%	91.4%	4.3%	4.3%
vocal with background music	69.5%	13.1%	17.4%	86.4%	4.5%	9.1%

Table 2. The success rate of correct salient point extraction after MPEG Layer III compression with bit rate 16 kbps and 10% white noise.

Test Audio	Original preferred to watermarked
flute and cello music	53.3%
guitar music	46.7%
vocal with background music	53.3%

Table 3. The blind listening test of watermarked audio pieces.

5.3. Blind Watermark Detection

We tested the robustness of the proposed blind watermark retrieval algorithm against several kinds of attacks, including additive noise, MPEG compression, random cropping, low pass filtering, and resampling. The quality of watermark detection is evaluated by the ratio between the correlation value obtained from the correct user ID and the largest correlation obtained from 1000 other random user IDs. An example of the detection results after various attacks is displayed in Figure 4. The audio used in this figure is the flute and cello music, and ID number 135 is the one used in watermark embedding. We can see from these figures that the correlation peak of the correct user ID still stands out of others after various kinds of attacks. Detailed detection results for the guitar music and the vocal music are similar to that in Figure 4 and therefore omitted. The ratio between the correlation value from the correct user ID and the largest correlation obtained from 1000 other random user IDs are summarized for the three test audio pieces in Table 4. Each kind of attack leads to a different amount of decrease in this peak ratio. However, in all cases experimented, the correlation peak of the correct user ID always stands out of the rest correlation peaks.

We have the following observations.

1. Additive white noise.

White noise with 10% of the power of the audio signal is added. Noise of this level is clearly audible, but only causes a moderate decrease in the peak ratio.

2. MPEG compression.

In multimedia applications, lossy compression is a very common procedure to increase transmission and storage efficiency. Some information is thrown away during the compression process, thus creating a potential hazard for watermark detection. To test the robustness of the proposed watermarking approach to lossy compression,

Attack	Flute and cello music	Guitar music	Vocal with background music
No attack	2.51	2.92	2.41
Additive noise	2.06	2.29	1.88
MPEG compression	1.93	2.49	1.66
Random cropping	2.03	2.36	1.95
Low pass filtering	1.87	2.17	1.90
Resampling	2.35	2.73	2.26

Table 4. The ratio between the correlation peak with the correct user ID and the largest correlation in 1000 random trials.

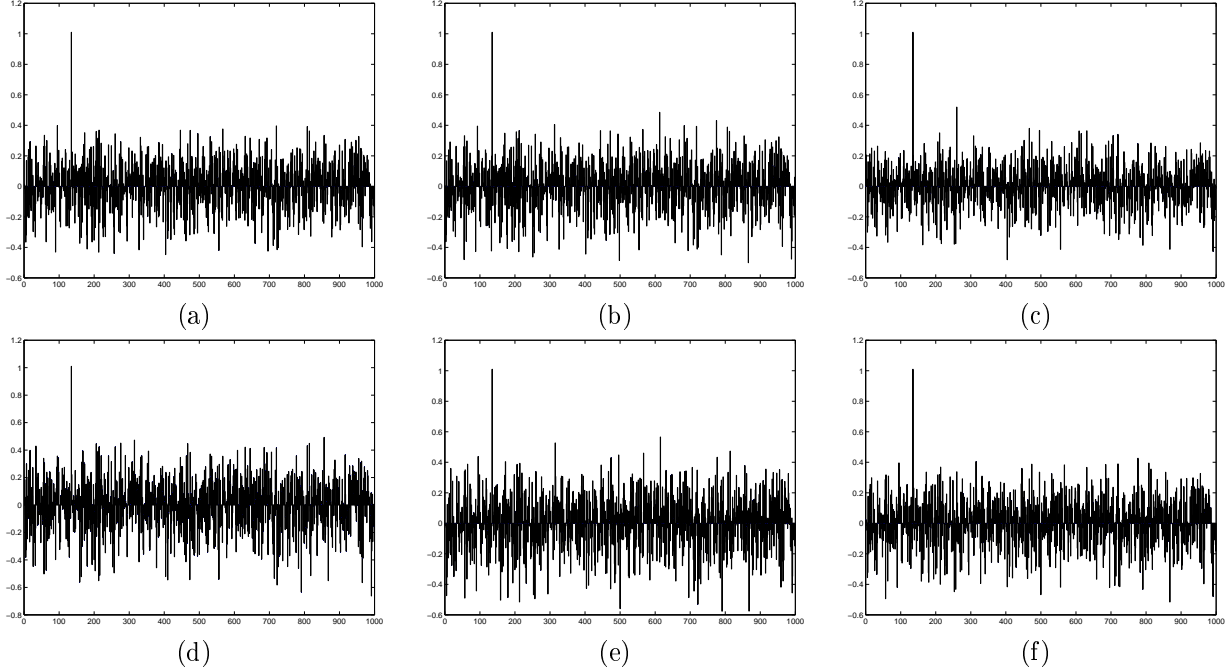


Figure 4. An example of detection results on watermarked audio after various attacks. The vertical axis is the normalized correlation while the horizontal axis is the user ID. The watermarked audio in this example is the flute and cello music, and the highest peak in each graph is the result of using the correct user ID with: (a) no distortion, (b) 10% white noise, (c) MPEG Layer III compression with a bit rate of 16 kbps, (d) randomly cropping one sample out of every 100 samples, (e) low pass filtering with a cutoff frequency of 3 kHz, and (f) down-sampling to 11 kHz and then re-sampled at 22.05 kHz

the watermarked audio signal is compressed and decompressed by MPEG layer III coder with a bit rate of 16 kbps, which corresponds to 22:1 compression. As shown in Table 4, this attack is more serious than others. However, the watermark can still be detected correctly.

3. Random cropping.

Randomly cropping one sample out of every 100 samples produces a disastrous synchronization problem for time-domain watermarking methods. However, the correlation peak ratio is only slightly decreased with the proposed method.

4. Low pass filtering.

With watermarks embedded in the frequency domain, low pass filtering with a very low cutoff frequency could effectively eliminate the embedded watermark. However, since our watermark is embedded in the frequency bands with the highest energy, filtering out the inserted watermark also greatly effects the sound quality. In our experiment, a low pass filter with a cutoff frequency of 3kHz is applied to watermarked audio signals. The loss of high frequency components is clearly audible, but the correlation peak ratio is only decreased around 25%.

5. Resampling.

Resampling is a common audio signal processing procedure. In this experiment, watermarked audio signals are down sampled to 11 kHz and then upsampled back to 22.05 kHz. The effect of resampling is audible, but there is almost no decrease in the peak ratios.

As shown in Table 4, the correlation peak ratios after various kinds of attacks are scattered between 1.5 ~ 2.5. These values could be increased if the watermark is embedded and retrieved everywhere in the audio signal, or if the original audio is used in watermark detection. However, the correlation ratio in Table 4 is already high enough for unambiguous watermark detection. The efficiency achieved by blind watermark detection and embedding in attack-sensitive regions only is very important for the practical use of audio watermarks.

6. CONCLUSION

The rapid growth of multimedia technologies facilitates the production and transmission of digital media data. It brings us not only opportunities but also challenges to copyright protection. An audio watermarking scheme which meets both the robustness and the low computational complexity requirements via audio content analysis was presented in this paper. The analysis identifies attack-sensitive regions which are suitable for watermark insertion, and provides consistent audio segmentation results before and after attacks. After audio content analysis, a watermark embedding scheme was developed in the Fourier transform domain that utilizes the temporal and frequency masking effects of the human auditory system. The embedded watermark is inaudible. The impact of malicious random cropping attack on audio watermarking was considered. The proposed watermarking solution, which combines the synchronizing feature of audio content analysis and the time-shift tolerance of Fourier domain watermarking, provides a low complexity solution to this kind of attack. There is more research to be done in the near future. It includes more experiments to derive an optimal decision threshold and the application of pitch-based audio content analysis to audio pieces which have a smooth energy variation.

7. ACKNOWLEDGEMENTS

This work was completely performed in Media Fair, Inc. when the first two authors were hired as student interns by the company. The support of Media Fair, Inc. of this research is highly appreciated.

REFERENCES

1. W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal*, vol. 35, no. 3-4, pp. 313-336, 1996.
2. D. Gruhl, A. Lu, and W. Bender, "Echo hiding," in *Info Hiding 96*, pp. 295-315, 1996.
3. I. J. Cox, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Processing*, vol. 6, no. 12, 1997.
4. I. Pitas and P. Bassia, "Robust audio watermarking in the time domain," in *EUSIPCO '98*, pp. 25-28, 1998.
5. M. Swanson, B. Zhu, A. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Processing Journal*, vol. 66, pp. 337-355, 1998.
6. J. Lacy, S. Quackenbush, A. Reibman, and J. Snyder, "Intellectual property protection systems and digital watermarking," *Journal of Optics Express*, vol. 3, pp. 478-484, 12 1998.
7. A. Piva, M. Barni, F. Bartolini, and V. Cappellini, "Dct-based watermark recovering without resorting to the uncorrupted original image," *ICIP*, vol. 1, pp. 520-523, 1997.