

An Audio Watermarking Technique That Is Robust Against Random Cropping

Author(s): Wei Li and Xiangyang Xue

Source: *Computer Music Journal*, Winter, 2003, Vol. 27, No. 4 (Winter, 2003), pp. 58-68

Published by: The MIT Press

Stable URL: <https://www.jstor.org/stable/3681901>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



The MIT Press is collaborating with JSTOR to digitize, preserve and extend access to *Computer Music Journal*

JSTOR

An Audio Watermarking Technique That Is Robust Against Random Cropping

Recent years have seen increased Internet traffic and the proliferation of digital audio distribution in the form of MP3 files. As such, copyright protection of digital audio works is becoming increasingly important. As a complement to conventional encryption techniques, watermarking provides powerful tools for copyright protection and has become a very active research area in recent years.

Audio watermarking is a technique that embeds information with specific meaning into the host media without interference to the quality of the original work. Proposed applications of audio watermarking include copyright protection, annotation, authentication, broadcast monitoring, and tamper proofing. Depending on the application, the inserted watermark data can include copyright information, serial numbers, text (e.g., the name of the composer and title of the work), a small image, or even a small clip of audio. The watermark is hidden in host media and is usually imperceptible to humans. The watermark must also be able to withstand signal-processing manipulations.

In general, a good audio watermarking scheme should possess the following properties:

- The watermark must be embedded into the host media rather than stored in a file header or a separated file, or else it can be removed or modified easily.
- The watermark should not introduce any perceptible artifacts that affect the quality of the original signal (i.e., perceptual transparency).
- The watermark should not be detected without prior knowledge of the watermark sequence. (To ensure the security, a secret key must be used for the embedding and detection process.)
- The watermark should be robust against common signal processing techniques such as lossy

compression, filtering, resampling, noise addition, etc.

- The computational cost of embedding and detection should be low enough for real-time processing.
- The watermark should be self-clocking to ensure its recovery when facing malicious cropping or time-scale modification.
- Under most circumstances, the watermark should be detected without resorting to the original audio, which is often very difficult to find in an open environment like the Internet.
- Finally, the watermarking algorithm should be public; that is, the security depends on the secret key but not the secrecy of the algorithm.

It is very difficult to design a watermarking system that meets all these requirements. Some properties such as robustness, transparency, and data capacity conflict with each other. The goal is achieving the best trade-off among them.

In this article, we propose a novel audio watermarking scheme based on statistical features in the wavelet domain. In each audio frame, the mean value of the wavelet coefficients at the coarsest approximation subband is calculated as the statistical feature, which is supposed to be invariant to a wide range of attacks. The basic idea is to embed watermark data by transforming this feature to a given positive or negative number, one bit per frame. The experimental results demonstrate that this algorithm is robust to common audio signal processing such as MP3 compression, low-pass filtering, equalization, echo addition, resampling, and noise addition. Furthermore, it also shows certain robustness to synchronization attacks like random cropping and time-scale modifications. The watermarked audio has very high perceptual quality and is indistinguishable from the original signal. A blind watermark detection technique without resorting to the

original signal is developed to identify the embedded watermark under various types of attacks. To ensure the security of the watermark, a random chaotic sequence is employed in the process of embedding and detection.

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 most existing watermarking algorithms based on the classical spread-spectrum technique. This new approach is more robust when the original signal is not available, and it is very insensitive to the change of synchronization structure.

The concept of random cropping in audio watermarking is quite different from that used in most of the image-watermarking literature, where the cropped image is usually restored to the original size of standard test images before watermark detection. With audio watermarking, however, it is very difficult to know the exact length of the original audio and the positions where audio samples are cropped. As a result, the watermark detection must be performed on the shortened audio, which is much more difficult than image watermarking. To the best of our knowledge, although some audio watermarking methods have been developed (Bassia, Pitas, and Nikolaidis 2001; Kirovski and Malvar 2001; Veen et al. 2001; Muntean et al. 2002; Ricardo 1999; Seok, Hong, and Kim 2002; Swanson et al. 1998), most of them are vulnerable to random cropping such as "jittering," and very few researchers have performed and published sufficient experiments involving this malicious attack.

In communications systems, error coding the information bit sequence may help achieve more efficient and reliable transmission. Because watermarking can be viewed as a communications system, it is natural to consider using error correction codes to improve the detection precision, capacity, and robustness. Repetition code is the simplest one in error-correction codes, and it outperforms block codes like BCH code at conditions of high channel-error rate (greater than 10%). Because the error probability of a watermark channel is usually much higher than that of the conventional communications channel, we apply five times ($5 \times$)

repetition codes to significantly improve the detection performance.

Related Work on Digital–Audio Watermarking

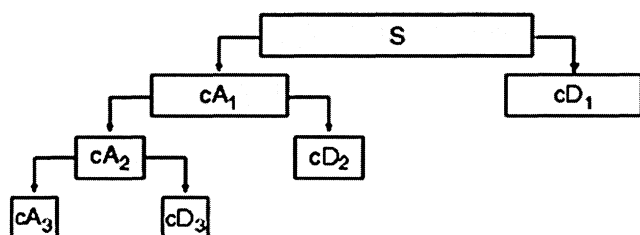
Several techniques for audio watermarking have been proposed. In general, they can be classified into time-domain, frequency-domain, and compression-domain algorithms.

Time-domain algorithms insert the watermark into the audio signal directly. Compared with frequency-domain algorithms, they are relatively easy to implement and require less computational cost, but they are not very robust to some audio signal processing such as compression and filtering. The simplest time-domain method is the so-called LSB scheme. Bassia, Pitas, and Nikolaidis (2001) proposed a blind-detection watermarking algorithm that takes advantage of audio masking. Mansour and Tewfik (2001) employed spline interpolation to change the relative length of the middle segment between two successive maxima and minima to embed data. In the echo-hiding technique (Guhl, Lu, and Bender 1997), information is embedded into an audio signal by introducing echoes with two different delays to the original audio.

Most audio watermarking methods operate in the frequency domain. Tilki and Beex (1996) proposed a method of hiding information in television broadcasts to accompany audio. In this technique, the Discrete Fourier Transform (DFT) coefficients of the middle frequencies (2.4–6.4 kHz) are replaced by the digital signature. Bender et al. (1996) proposed a method of phase coding.

Since the initial paper of Tirkel et al. (1993), spread-spectrum techniques borrowed from communications theory have been widely applied to watermarking systems. The watermark is first multiplied with a high-speed pseudorandom sequence and then added to the original audio signal. The audio masking of HAS is employed in Boney, Tew, and Hamdy (1996); Neubauer, Herre, and Brandenburg (1998); and Swanson et al. (1998) to ensure inaudibility. The work of Cox et al. (1996) is regarded as the most representative of the application of spread-spectrum techniques in digital watermark-

Figure 1. The tree structure of wavelet decomposition.



ing. Ricardo (1999) implemented a robust watermarking system by combining spread-spectrum theory with a psychoacoustic model. Lu, Liao, and Chen (2000) proposed a hybrid watermarking scheme by inserting two complementary modulated watermarks, a robust one and a “fragile” one, into the original audio. In addition to these articles, recent developments are addressed by Mansour and Tewfik (2001); Muntean et al. (2002); Ryuki et al. (2001); Ryuki (2002, 2003); and Seok, Hong, and Kim (2002).

Petitcolas from Cambridge University proposed a watermarking scheme called “MP3Stego” that embeds the watermark into MP3 files during the compression process. (See www.cl.cam.ac.uk/~fapp2/steganography/mp3stego for details.) However, this method suffers from poor robustness against decoding and re-encoding. Qiao and Nahrstedt (1999) presented two watermarking methods to embed the watermark directly into the MPEG audio bitstreams by modifying the scale factors and the time-domain samples. Xu and Zhu (2000) and Xu, Zhu, and Feng (2001) proposed embedding the watermark in a partially compressed domain of audio adaptively based on the multiple-bit-hopping method. The watermark is highly correlated to the audio content and can be embedded very quickly.

An Invariant Watermark

The Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT), which provides a compact representation of a signal in both time and frequency, is a relatively recent and computationally efficient technique for analyzing non-stationary signals like audio. It was developed

as an alternative to the Short-Time Fourier Transform (STFT). Instead of providing uniform time resolution for all frequencies like the STFT, the DWT provides greater time resolution and less frequency resolution for high frequencies while providing greater frequency resolution and less time resolution for low frequencies. In that respect, it behaves just like the human ear, which exhibits similar time–frequency resolution characteristics.

The one-dimensional DWT decomposes a signal into two parts: high frequencies and low frequencies. The discontinuity components of the signal are primarily confined to the high-frequency part. The low-frequency part is decomposed again into two parts of high and low frequencies. The number of decompositions in the above process is usually determined by the application and the length of the original signal. The data obtained from the above decomposition are called DWT coefficients, and the original signal can be reconstructed precisely from these coefficients. This reconstruction process is called the inverse DWT, or IDWT.

The DWT analysis can be performed using a fast, pyramidal algorithm related to multirate filterbanks. First, the time-domain signal is successively filtered using a high-pass filter and then a low-pass filter according to Equations 1 and 2:

$$y_{high}[k] = \sum x[n]g[2k - n] \quad (1)$$

$$y_{low}[k] = \sum_n x[n]h[2k - n] \quad (2)$$

where $y_{high}[k]$ and $y_{low}[k]$ are the outputs of the high-pass and low-pass filters, respectively, after downsampling by a factor of two. The signal is finally decomposed into a coarsest approximation signal and a series of detail signals, as shown in Figure 1, with each subband containing half the number of samples of its parent frequency subband (Mallat 1989).

Owing to the downsampling, the number of resulting wavelet coefficients is the same as the number of input points. A variety of different wavelet families has been proposed in the literature. In our experiment, the wavelet bases of “haar” and “db4” are used, and they show similar results.

DWT Statistical Feature Extraction

In audio analysis and classification, the extracted wavelet coefficients provide a compact representation that shows the energy distribution of the signal in both time and frequency. To further reduce the dimensionality of the extracted feature vectors, statistics over a set of wavelet coefficients can also be used to represent the statistical characteristics of the “texture” or the “music surface” of audio pieces. In general, the following statistics can be extracted as statistical wavelet features (Tzanetakis, Essl, and Cook 2001): the mean of the absolute value of the coefficients in each subband, which provides information about the frequency distribution of the audio signal; the standard deviation of the coefficients in each subband, which provides information about the changes of the frequency distribution; and ratios of the mean values between adjacent subbands, which also provide information about the frequency distribution.

Invariant Watermark

In this research, for the convenience of watermark embedding, we adopt the mean of the coefficient values rather than the absolute coefficient values at the coarsest approximation subband as the statistical feature. Because these statistical features are calculated from the wavelet coefficients at the coarsest approximation subband (which represents the perceptually most significant low frequency components of the audio signal), they tend to be relatively stable under common signal processing techniques. Moreover, owing to the high relevance between adjacent audio samples and small frames of audio, random cropping of a small clip of audio will not change this statistical feature much, although individual coefficients may experience a large change. In this way, the statistical feature can also be relatively invariant to short time-domain random cropping. Therefore, this statistical feature of the wavelet coefficients at the coarsest approximation subband serves as a good candidate for watermark embedding. Here, we attempt to find a kind of feature that is insensitive to most common

signal processing and malicious random cropping attacks, as described in Cox, Matthew, and Bloom (2001). Several other schemes robust against synchronization attacks including exhaustive search, synchronization, autocorrelation, and implicit synchronization are also discussed in this book. Until now, although many different approaches have been investigated, resisting temporal distortions (i.e., synchronization attacks) such as random cropping, delay, and scaling still remains one of the most difficult problems in watermarking research.

To validate the efficiency of the above assumptions, we have performed many experiments on different kinds of music and instruments including pop, rock, saxophone, piano, violin, guitar, electronic organ, etc. Experimental results show that for most audio frames, the statistical mean changes only slightly after undergoing attacks.

Watermark Embedding and Detection

As mentioned, most existing watermarking methods share one common problem: they are vulnerable to the synchronization attacks. This problem could result from audio editing such as cropping unwanted audio segments or intentional attacks such as random deleting or adding samples to watermarked audio data. This random-sample cropping attack is very effective in interfering with most watermark-detection processes based on the spread-spectrum technique. It exhibits very low computational complexity and does not introduce noise to the underlying audio signal when done correctly (Wu, Su, and Kuo 2000).

As discussed earlier, we select the means of wavelet coefficients at the coarsest approximation component of audio signals as the statistical feature in which to embed data, because they experience much less variance after most signal processing manipulations and casual or malicious cropping attacks than individual wavelet coefficients and original samples in the time domain. Motivated by this idea and our experimental observation of this attack-invariant feature, we propose a novel way of embedding and detecting data described in detail as follows.

Figure 2. The original watermark: a binary logo image.



Embedding Algorithm

The input audio signal is first segmented into overlapping frames. Given a sampling frequency of 44.1 kHz, the frames are chosen to be 2,048 samples each with 75% (i.e., 1,536 samples) overlap between every two adjacent frames. Each frame is Hamming-windowed by the function $w(i) = 0.54 - 0.46 \cdot \cos(2\pi i/256)$ to minimize the Gibbs phenomenon and avoid generating clicking sounds at the borders of adjacent frames. Note that the frame size is a trade-off between perceptual transparency (small frame sizes) and detection reliability (large frame sizes). Experimental results demonstrate a good compromise in our selection of window size and overlap.

In the second step, a three-level wavelet decomposition is performed for each audio frame with the "db4" wavelet basis (although we use the "haar" wavelet basis on some audio sequences), and then the mean of all the wavelet coefficients at the coarsest approximation subband (i.e., the "ca3" level) is calculated.

In the third step, we use a 24×24 -pixel binary logo image shown in Figure 2 in our experiment. It is transformed into a one-dimensional sequence of ones and zeros as follows:

$$W = \{w(i); w(i) \in \{1, 0\}, 1 \leq i \leq 24 \times 24\} \quad (3)$$

Most other existing watermarking algorithms use a pseudorandom sequence as the watermark. Most other existing watermarking algorithms use a pseudorandom sequence as the watermark (which is not so intuitive as a logo image), and the correlation detection greatly depends on the selection of threshold. Next, each bit of the watermark data is mapped into an antipodal sequence using BPSK modulation ($1 \rightarrow -1; 0 \rightarrow +1$) and repeated five times according to the following equations:

$$w'(i) = 1 - 2w(i), 1 \leq i \leq 24^2 \quad (4)$$

$$w'(k) = w'(i), 5i - 4 \leq k \leq 5i, 1 \leq i \leq 24^2 \quad (5)$$

$$W' = \{w(k); w'(k) \in \{+1, -1\}, 1 \leq k \leq 24^2 \times 5\} \quad (6)$$

Finally, $w'(k)$ is embedded into the corresponding audio frame in the following way:

$$\begin{cases} \text{if } w'(k) = 1, x'(k, j) = x(k, j) - m(k) + \alpha \\ \text{if } w'(k) = -1, x'(k, j) = x(k, j) - m(k) - \alpha \end{cases} \quad (7)$$

where $x(k, j)$ and $x'(k, j)$ are the original and modified j th wavelet coefficient at the "ca3" level in the k th frame, respectively, $m(k)$ is the mean of the wavelet coefficients at the "ca3" level in the k th frame, and α is a small number on the same order of magnitude as $m(k)$. Furthermore, α is adjusted so as not to introduce any audible artifacts into the watermarked audio.

For the fourth step, the inverse discrete wavelet transform (IDWT) is applied to $x'(k, j)$, the modified wavelet coefficients in each frame, to transform them back to the time domain. Finally, steps 2 through 4 are repeated until all the watermark bits are embedded. Finally, all the modified frames are merged together to form the entire audio signal in the time domain.

Blind Watermark Detection

The detection algorithm is straightforward and "blind," that is, it does not need the original audio signal or the original watermark. For each segmented frame, if the mean of the wavelet coefficients at the coarsest approximation subband (i.e., "ca3") is greater than zero, a bit of 1 is extracted, whereas if the mean is less than zero, a bit of -1 is extracted. This process is repeated until all embedded bits are detected as follows:

$$w'(k) = \text{sign}(\text{mean}[\text{ca3}(k)]), 1 \leq k \leq 24^2 \times 5 \quad (8)$$

Then, the watermark bits are determined based on the majority rule and BPSK demodulation:

$$w'(i) = \text{sign} \left(\sum_{k=5 \cdot (i-1)+1}^{k=5 \cdot i} w'(k) \right), 1 \leq i \leq 24^2 \quad (9)$$

$$w(i) = \frac{(1 - w'(i))}{2}, \quad 1 \leq i \leq 24^2 \quad (10)$$

Finally, all the detected watermark bits $w(i)$ are rearranged to form the binary watermark image.

Synchronization is a serious problem in any watermarking scheme, especially with audio. For example, adding or removing one sample out of every 100 (a jittering attack) introduces no audible distortion but makes the detector of most audio watermarking schemes invalid (Li and Yu 2000a, 2000b). In contrast to conventional spread-spectrum watermarking strategies, which greatly depend on the correct alignment between the test watermarked signal and the signature signal, this method has much less synchronization sensitivity and is much more robust.

Security

Security is of great importance to watermarking schemes. If the embedding process is totally transparent, an attack might be able to modify the embedded data. To further increase the security of our scheme, an encryption/decryption technique can be employed. In this article, the hybrid chaotic dynamical system (Kennedy and Kolumban 2000) is adopted to produce random chaotic sequences according to Equation 11.

$$y = \begin{cases} 1 - 2x^2 & -1 \leq x < -0.5 \\ 1 - \frac{1}{2}(-2x)^{1.2} & -0.5 \leq x < 0 \\ 1 - 2x & 0 \leq x \leq 0.5 \\ -(2x - 1)^{0.7} & 0.5 < x \leq 1 \end{cases} \quad (11)$$

The generated random sequence greatly depends on the selection of the initial secret key and has a good autocorrelation and zero mean. Taking advantage of this chaotic sequence, the watermark data are first encrypted before embedding. Even if the embedding process is completely transparent to attackers, they can only detect the encrypted watermark data. Without the correct secret key in detection, these detected data will be incomprehensible. In this way, the ability to resist illegal detection is improved significantly.

Experimental Results

The algorithm was applied to a set of audio signals including pop music, rock music, and samples of a saxophone, piano, electronic organ, guitar, and violin. Each monaural selection had a duration of 45 sec and was recorded at 16 bits/sample with sampling rate of 44.1 kHz. The watermark in our experiment was a 24×24 -pixel binary logo image, as shown in Figure 2. The waveform of the original and watermarked piano music is shown in Figure 3.

Listening Test

To evaluate the watermarked audio quality, we performed an informal subjective listening test by using the so-called Subjective Diff-Grades (SDG). The meaning of each score in the SDG test is shown in Table 1.

Ten listeners were provided with the original and watermarked audio, and they were asked to classify the difference in terms of the SDG scale. The results of the subjective quality evaluations were averaged and tabulated in Table 2. It is apparent that all the average SDG scores are zero or very close to zero, which means that the watermarked audio and the original are perceptually undistinguishable.

In addition to the above-mentioned subjective listening test, the signal-to-noise ratio (SNR) between the original and the watermarked audio can serve as an objective measure:

$$SNR = 10 \cdot \log \left(\frac{\sum_{k=1}^d I^2(k)}{\sum_{k=1}^d [I(k) - I'(k)]^2} \right) \quad (12)$$

where $I(k)$ and $I'(k)$ are the sample values of the original and the watermarked audio, respectively, and d is the number of samples in the audio files. The calculated SNR of Figure 3 is 34.2 dB, which is rather high, showing that there is almost no perceptual difference between the original and the watermarked audio. In addition to SNR, the ITU objective quality measure system of Treurniet and Soulodre (2000) can also be used.

Figure 3. (a) The original piano waveform; (b) the watermarked piano waveform; (c) the difference between the original and the watermarked waveform.

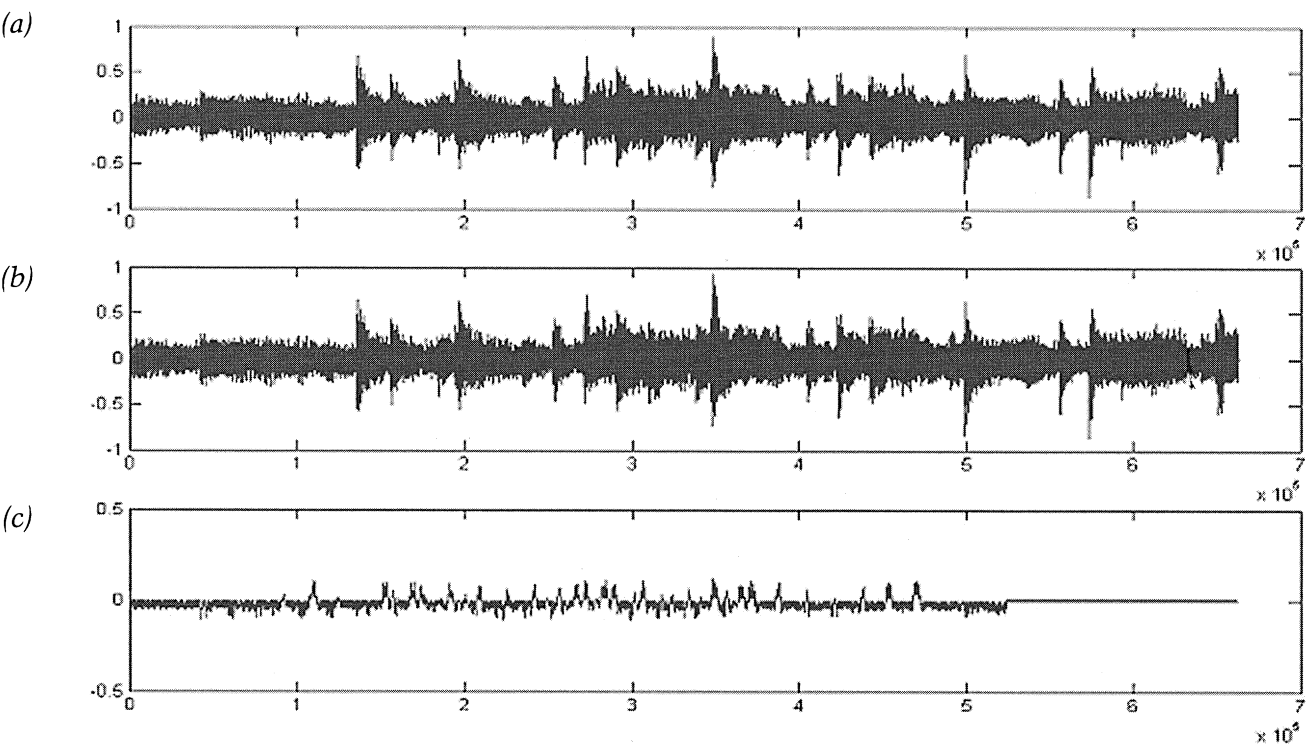


Table 1. Subjective Diff-Grades

SDG	Description
0.0	Imperceptible
-1.0	Perceptible, but not annoying
-2.0	Slightly annoying
-3.0	Annoying
-4.0	Very annoying

Robustness Test

To evaluate the performance of the proposed watermarking algorithm, we tested its robustness according to the SDMI (Secured Digital Music Initiative) Phase-I robustness test procedure. The audio editing and attacking tools adopted in the experiment were Cool Edit Pro 2.0 and GoldWave 4.26.

According to their influence on synchronization, attacks can be divided into two categories (Li and Yu 2000a, 2000b). Type I attacks include MPEG lossy compression, low-pass/band-pass filtering, ad-

ditive/multiplicative noises, resampling, echo addition, and equalization. These attacks may distort the perceptual quality but do not affect the synchronization structure. Type II attacks include random cropping, jittering, and time-scale warping. These attacks introduce very little distortion to the watermarked audio but destroy the synchronization needed by most existing audio watermarking algorithms.





























A successful digital watermark attack can make the watermark detection fail without affecting the audio quality too much. In general, to simulate the real operations a hacker may take, the experiment should distort the watermarked audio to an extent that is just objectionable to listeners. Stronger distortion will be of no use, because the audio file will no longer be of interest to hackers. If one algorithm can survive these critical attacks, it can also resist the attacks a pirate will perform.

The final calculated bit error rate (BER), the similarity between the extracted watermark and the original watermark, and the extracted binary water-

Table 2. The average SDG score of the subjective listening test

Music Type	Saxophone	Rock	Pop	Piano	Electronic Organ	Guitar	Violin
SDG Score	0	0	0	0	0	−0.1	−0.1

Table 3. Similarity, BER, and the extracted watermark image

No.	Type of Attacks	Original			Rep5		
		Sim	BER (%)	Image	Sim	BER (%)	Image
1	MP3	0.9662	5.73		1	0	
2	Resample	0.9116	14.93		1	0	
3	Low pass	0.9105	15.10		1	0	
4	Equalization	0.9424	9.72		1	0	
5	Noise	0.9119	14.93		1	0	
6	Echo	0.9260	12.85		0.9948	0.87	
7	Pitch Shift (+4%)	0.9550	6.89		1	0	
8	Pitch Shift (−4%)	0.9612	5.92		1	0	
9	Crop1 (5*100)	0.8931	18.06		0.9489	8.51	
10	Crop2 (10*100)	0.8908	18.40		0.9281	11.98	
11	Crop3 (10*500)	0.8381	27.26		0.8933	17.71	
12	Jittering (1/500)	0.9038	16.50		0.9256	11.80	
13	TSM (+3%)	0.8553	24.61		0.8924	17.88	
14	TSM (−3%)	0.8294	28.65		0.8395	24.56	

BER = bit error rate; Sim = similarity between the extracted watermark and the original watermark.

mark images are listed in the right half of Table 3. In all cases, it can be seen that with the help of repetition coding, the watermark image can be extracted and clearly identified. The detailed robustness test procedure is described as follows.

MP3 Compression

Lossy compression is a very common procedure to increase transmission and storage efficiency in multimedia applications. Some redundant information is thrown away during the compression process, thus creating a potential hazard for watermark detection. The audio samples we used were first compressed at the rate of 22:1 with a bit rate of 32 kbps and then decompressed into the wav format for watermark detection. Most MP3 music on the Internet is compressed at the bit rate

of 128 kbps, which is much higher than that used in our test. From Table 3, we can see that even under the 22:1 compression ratio, which is already beyond the typically acceptable level of most music listeners owing to the loss of signal quality, the watermark image can still be extracted without any error.

Resample

The audio clips to be tested are first downsampled to the sampling rate of 16 kHz, then upsampled to 44 kHz. The watermark detection is not affected by this operation, and the BER is zero.

Equalization

The “Bass Boost” preset of the audio editing tool GoldWave was used, which is a 7-band graphic

Li and Xue

equalizer. The 60-Hz, 150-Hz, and 400-Hz frequency bands were boosted by 6 dB, and the remaining bands at 1 kHz, 2.4 kHz, 6 kHz, and 15 kHz were set to 0 dB. Although equalization of course changes the frequency distribution, the detection still succeeds with a BER of zero.

Low-Pass Filtering

If the watermark is embedded in the frequency domain, low-pass filtering with a very low cutoff frequency could effectively eliminate the embedded watermark. However, because our watermark is embedded in the coarsest approximation subband, which corresponds to the perceptually significant low-frequency components of the audio, low-pass filtering with a cutoff frequency of 4 kHz used in our experiment has no effect on the detection, although it makes the audio sounds dull owing to the loss of high-frequency components.

Noise Addition

White noise with a constant level was added to the watermarked audio, but it does not affect watermark detection. The binary logo image can still be extracted exactly.

Echo Addition

An echo signal with a delay of 200 msec and a decay of 40% was added to the original audio signal. Here, the echo can be clearly perceived by listeners and becomes annoying. The similarity between the original and the extracted watermark is high enough, although there are a few bit errors.

Pitch Shifting

Tempo-preserved pitch shifting is a difficult attack for audio watermarking algorithms, because it causes frequency fluctuation. This problem could be caused by intentional attacks or unintentional side effects of analog editing. Our strategy is rather insensitive to frequency fluctuation, and the BER is kept zero even when the pitch is shifted up to $\pm 4\%$, which is very annoying to listeners. In fact, the watermark can be extracted without any error even if the pitch is shifted up to $\pm 10\%$.

Random Cropping and Jittering

Our approach is rather insensitive to synchronization structure. Even if several thousands of samples are cropped at different positions randomly, the binary watermark image can still be detected and identified.

Three random cropping tests were performed in our experiment. First, we cropped 100 samples at each of five randomly selected positions; next, 100 samples were cropped at each of ten randomly selected positions; finally, we cropped 500 samples at each of ten randomly selected positions. In these three tests, a total of 500, 1,000, and 5,000 samples were cropped, respectively, from the unattacked watermarked audio, and the length of the cropped watermarked audio was reduced accordingly. In the second and third cropped files, obvious discontinuity can be heard at the cropping positions.

Jittering is an evenly performed form of random cropping. We removed one sample out of every 500 samples in our jittering experiment.

Owing to the high relevance between adjacent audio samples or small blocks, random cropping and jittering will not seriously affect the mean of the wavelet coefficients at the coarsest approximation subband, and so the watermark can be extracted exactly.

Time-Scale Modification (TSM)

In our experiment, pitch-invariant time-scale modifications were also applied to the watermarked audio. However, our approach exhibits only moderate resistance to this type of attack. When the time-scale modification is not greater than $\pm 3\%$, the extracted binary image can still be identified, but the image gets increasingly difficult to recognize for larger time-scale modifications. Although our method's ability to resist TSM attacks is not very strong, it has already exceeded most other algorithms and is rather close to the $\pm 4\%$ performance index demanded by SDMI.

The pitch-invariant time speed change can also be viewed as a special form of random cropping; it removes or adds some parts of audio signal while preserving the pitch. Thus, for those audio frames whose data are greatly altered, the sign of the mean

of the wavelet coefficients at the coarsest approximation subband may change be inverted. In this case, the detected watermark bit will be incorrect.

From Table 3, we can see that after applying five-times ($5\times$) repetition codes, the overall detection performance is improved significantly. For non-synchronization attacks such as MP3 compression, noise addition, echo addition, equalization, low-pass filtering, and resampling, the BERs have been reduced to zero or very close to zero. As mentioned earlier, these attacks are somewhat objectionable to listeners, and stronger attacks are of no use to hackers owing to the excessive loss of auditory quality.

Because our algorithm can withstand the distortion attacks listed above, it will also be able to resist degradation performed by hackers on the Internet. To synchronization attacks such as random cropping and time-scale modifications, the use of repetition code also improves the detection performance greatly, the bit error rates are reduced, and the extracted image is much clearer.

Conclusion

From these analyses and experimental results, it can be seen that with the help of repetition coding this proposed algorithm is robust to common audio signal processing operations and rather insensitive to synchronization change. Finding a steady feature under all kinds of audio signal processing and synchronization attacks is a delicate work, and this kind of statistical feature does not always work in other transformation domains. In future work, we will try to find other features more suitable for audio watermarking. The impact of error-correction coding technique on signal processing and synchronization attacks will also be investigated. Based on the concept of Kutter, Bhattacharjee, and Ebrahimi (1999), this proposed approach can be viewed as a second-generation watermarking scheme.

Acknowledgments

This work was supported in part by National Science Foundation of China under contract

60003017, China 863 Projects under contracts 2001AA114120 and 2002AA103065, Local Government R&D Funding under contracts 01QD14013 and 015115044, and National Nature Science Funds of China (10171017, 90204013).

References

- Bassia, P., L. Pitas, and N. Nikolaidis. 2001. "Robust Audio Watermarking in the Time Domain." *IEEE Transactions on Multimedia* 3:232–242.
- Bender, W., et al. 1996. "Techniques for Data Hiding." *IBM Systems Journal* 35:313–336.
- Boney, L., A. H. Tew, and K. N. Hamdy. 1996. "Digital Watermarks for Audio Signals." *IEEE International Conference on Multimedia Computing and Systems*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 473–480.
- Cox, I. J., et al. 1996. "Secure Spread Spectrum Watermarking for Images, Audio and Video." *Proceedings of the International Conference on Image Processing*. Piscataway, New Jersey: IEEE Signal Processing Society, pp. 243–246.
- Cox, I. J., M. Matthew, and J. Bloom. 2001. *Digital Watermarking*. San Francisco: Morgan Kaufmann.
- Guhl, D., A. Lu, and W. Bender. 1997. "Echo Hiding." *Proceedings of the First International Workshop on Information Hiding*. Berlin: Springer, pp. 295–315.
- Kennedy, M. P., and G. Kolumban. 2000. "Digital Communication Using Chaos." *Signal Processing* 80:1307–1320.
- Kirovski, D., and H. S. Malvar. 2001. "Robust Spread-Spectrum Audio Watermarking." *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 1345–1348.
- Kutter, M., S. K. Bhattacharjee, and T. Ebrahimi. 1999. "Towards Second Generation Watermarking Schemes." *Proceedings of the 6th International Conference on Image Processing*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 320–323.
- Li, X., and H. H. Yu. 2000a. "Transparent and Robust Audio Data Hiding in the Cepstrum Domain." *Proceeding of the IEEE International Conference on Multimedia and Expo*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 397–402.
- Li, X., and H. H. Yu. 2000b. "Transparent and Robust Audio Data Hiding in the Subband Domain." *Proceedings of the International Conference on Information*

Li and Xue

- Technology: Computer and Communication*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, p. 74.
- Lu, C. S., H. Y. Liao, and L. H. Chen. 2000. "Multipurpose Audio Watermarking." *Proceedings of the 15th IAPR International Conference on Pattern Recognition*. Surrey, UK: International Association for Pattern Recognition, pp. 282–285.
- Mallat, S. G. 1989. "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11:674–693.
- Mansour, M., and A. Tewfik. 2001. "Time-Scale Invariant Audio Data Embedding." *Proceedings of the IEEE International Conference on Multimedia and Expo*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online at www.ece.umn.edu/users/mmansour/366.pdf.
- Muntean, T., E. Grivel, I. Naformita, and M. Najim. 2002. "Audio Digital Watermarking Based on Hybrid Spread Spectrum." *Proceeding of the IEEE Wedel Music*, pp. 150–156.
- Neubauer, C., J. Herre, and K. Brandenburg. 1998. "Continuous Steganographic Data Transmission Using Uncompressed Audio." *Proceedings of the Second International Workshop on Information Hiding*, pp. 208–217.
- Qiao, L. T., and N. Nahrstedt. 1999. "Non-Invertible Watermarking Methods for MPEG Encoded Audio." *SPIE Proceedings on Security and Watermarking of Multimedia Contents*. Bellingham, Washington: International Society for Optical Engineering, pp. 194–202.
- Ricardo, A. 1999. "Digital Watermarking of Audio Signals Using a Psychoacoustic Auditory Model and Spread Spectrum Theory." *Proceeding of the 107th Audio Engineering Society Conference*. New York: Audio Engineering Society, pp. 24–27.
- Ryuki, T. 2002. "Improving Audio Watermark Robustness Using Stretched Patterns against Geometric Distortion." *Proceedings of the 3rd IEEE Pacific-Rim Conference on Multimedia*. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 647–654.
- Ryuki, T. 2003. "Audio Watermarking for Live Performance." *Proceeding of the SPIE Conference on Security and Watermarking of Multimedia Contents*. Bellingham, Washington: International Society for Optical Engineering.
- Ryuki, T., et al. 2001. "An Audio Watermarking Method Robust against Time- and Frequency-Fluctuation." *Proceedings of the SPIE Conference on Security and Watermarking of Multimedia Contents*. Bellingham, Washington: International Society for Optical Engineering, pp. 104–115.
- Seok, J., J. Hong, and J. Kim. 2002. "A Novel Audio Watermarking Algorithm for Copyright Protection of Digital Audio." *ETRL Journal* 24:181–189.
- Swanson, M. D., et al. 1998. "Robust Audio Watermarking Using Perceptual Masking." *IEEE Transactions on Signal Processing* 66:337–355.
- Tilki, J. F., and A. A. Beex. 1996. "Encoding a Hidden Digital Signature onto an Audio Signal Using Psychoacoustic Masking." *Proceedings of the 7th International Conference on Signal Processing Applications and Technology*, pp. 476–480.
- Tirkel, A. Z., et al. 1993. "Electronic Watermark." *Proceedings of Digital Image Computing, Technology and Applications*, pp. 666–672.
- Treurniet, W. C., and G. A. Soulodre. 2000. "Evaluation of the ITU-R Objective Audio Quality Measurement Method." *Journal of the Audio Engineering Society* 48(3):164–173.
- Tzanetakis, G. G. Essl, and P. Cook. 2001. "Audio Analysis Using the Discrete Wavelet Transform." *Proceedings of the 2001 International Conference of Acoustics and Music: Theory and Applications*. Skiathos, Greece. Available online at www.cs.princeton.edu/~gessl/papers/amta2001.pdf.
- Veen, M., W. Oomen, F. Bruekers, J. Haitisma, T. Kalker, and A. N. Lemma. 2001. "Robust, Multi-Functional, and High Quality Audio Watermarking Technology." Convention paper of the 110th Convention Conference: Audio Engineering Society. Available online at www.wireless.per.nl/202/watermarks/research/papers/documents/aes2001audio.pdf.
- Wu, C.-P., P.-C. Su, and C. C. J. Kuo. 2000. "Robust and Efficient Digital Audio Watermarking Using Audio Content Analysis." *Proceedings of SPIE*. Bellingham, Washington: International Society for Optical Engineering, pp. 382–392.
- Xu, C., and Y. Zhu. 2000. "Content-Based Digital Watermarking for Compressed Audio." Paper presented at the 6th Conference on Content-Based Multimedia Information Access, 12–14 April, Paris, France. Available online at citeseer.nj.nec.com/cache/papers/cs/20368/http://zSzzSz133.23.229.11zSz~ysuzukizSzProceedingsallzSzRIO2000zSzWednesdayzSz33CP1.pdf/content-based-digital-watermarking.pdf.
- Xu, C., Y. Zhu, and D. D. Feng. 2001. "Digital Audio Watermarking Based on Multiple-Bit Hopping and Human Auditory System." *Proceedings of the ACM Multimedia*. New York: Association for Computing Machinery, pp. 568–571.