ICICS-PCM 2003
15-18 December 2003
Singapore

Localized Robust Audio Watermarking in Regions of Interest

W Li; X Y Xue; X Q Li

Department of Computer Science and Engineering
University of Fudan, Shanghai 200433, P. R. China
E-mail: weili_fd@yahoo.com

Abstract

In this paper, a novel localized robust audio watermarking scheme is proposed. The basic idea is to embed watermark in selected high energy regions, that is, in regions of interest (ROI). By virtue of localization and ROI, the embedded watermark is expected to escape the damages caused by audio signal processing, random cropping and time scale modification etc, because these high energy local regions usually correspond to music edge like note attack, transition or percussion instruments like drum, which represent the music rhythm or tempo and are very important to human auditory perception. Taking advantage of localization and ROI, together with global redundancy, this method shows strong robustness against common audio signal processing, time domain synchronization attacks, and most distortions introduced in Stirmark for Audio.

Keywords: localized watermarking, ROI, random cropping, time scale modification

1. Introduction

Synchronization is a serious problem to any watermarking scheme, especially to audio watermarking scenario. Audio processing such as random cropping and time scale modification cause displacement between embedding and detection in the time domain and is hence difficult for watermark to survive.

Generally speaking, synchronization problem can be alleviated by the following methods: exhaustive search [1], synchronization pattern [2], invariant watermark [3], and implicit synchronization [4].

Time scale modification is a serious attack to audio watermarking, very few algorithms can effectively resist this kind of synchronization attack. According to the SDMI (Secured Digital Music Initiative) Phase-II robustness test requirement [5], a practical audio watermarking scheme should be able to withstand time scale modification up to ±4%. In the literature, several existing algorithms aimed at solving this problem. Mansour et al. [6] proposed to embed watermark data by changing the relative length of the middle segment between two successive maximum and minimum of the smoothed waveform, the performance highly depends on the selection of the threshold, and it is a delicate work to find an appropriate threshold. In [7], Mansour et al. proposed another algorithm for embedding data into audio signals by changing the interval lengths between salient points in the signal, the extrema of the wavelet coefficients of the envelope are adopted as salient points. The proposed algorithm is robust to MP3 compression, low pass filtering, and can be made robust to time scaling modification by using adaptive quantization steps. The errors are primarily due to thresholding problems. For modification scales lower than 0.92 or higher than 1.08, the bandwidth of the envelope filter as well as the coarsest decomposition scale should be changed accordingly. Tachibana et al. [1] introduced an audio watermarking method that is robust against random stretching up to ±4%. The embedding algorithm calculates and manipulates the magnitudes of segmented areas in the time-frequency plane of the content using short-term DFTs. The detection algorithm correlates the magnitudes with a pseudo-random array that corresponds to two-dimensional areas in the time-frequency plane. Tachibana et al. [8] further improved the performance up to ±8% by using multiple pseudo-random arrays, each of which is stretched assuming a certain amount of distortion. Since most of the detection process for the multiple arrays is shared, the additional computational cost is limited.

The above mentioned methods share one common problem, that is, they all highly depend on adjusting some parameters like threshold or some assumed factors, this makes them difficult to be applied in different kinds of music. In this paper, we present a novel localized robust audio watermarking method aiming at combating audio signal processing and the synchronization problems caused by random cropping and time scale modification. The basic idea is to embed watermark in selected high energy regions, that is, in regions of interest (ROI). High energy regions, which generally represent music transition or sound of percussion instruments like drum, tambourine and castanet, are closely related to the rhythm information and are very important to human auditory perception, they usually draw more attention to listeners than other mild sections. In order to maintain high auditory quality, such regions have to be left unchanged or altered very little under different kinds of modification. Moreover, watermark embedded in local areas shows natural resistance to random cropping. Since random cropping occurred at the ROI regions will degrade the audio quality, pirates usually crop some less important parts outside of these important regions, thus it will not

make any threat to the watermark at all. Therefore, by embedding the watermark in these relatively safe regions, we can expect the watermark to elude all kinds of attacks, especially those time domain synchronization attacks.

2. Motivation and Embedding Regions Selection

Since the main purpose of this paper is to combat time scale modification, it is necessary to know something about the time scale modification algorithm, and see why watermark embedded in high energy ROI such as drum sections can be hoped to elude this challenging attack.

2.1 TSM attack and countermeasure

Recently developed TSM algorithms are usually performed on the harmonic components and residual components separately [10]. The harmonic portion is time-scaled by demodulating each harmonic component to DC, interpolating and decimating the DC signal, and remodulating each component back to its original frequency. The residual portion, which can be further separated into transient (edges) and noise components in the wavelet domain, is time-scaled by preserving edges and relative distances between the edges while time-scaling the stationary noise components between the edges. The edges are related to attacks of musical notes, transitions, or non-harmonic instruments such as castanets, drums and other percussive instruments. Such information may be related to temporal aspects of a music signal such as tempo and timbre. Special care must be taken when manipulating the time-scale of the residual component. First, it is important to preserve the shape or slope of the attacks (edges). If the slope is not preserved, the instruments tend to sound dull because the high frequency information is lost. Second, it is important to preserve the relative distances between the edges while maintaining synchronization with the harmonic component, because this contains the information relative to tempo [9].

Based on the above knowledge, we know that TSM algorithms stretch audio signals only in regions where there is minimum transient information and strive to preserve music edge. If we embed watermark in regions representing music edge, it is possible to elude time scale modification without delicately adjusting parameters like thresholds or predefined scale factors. In Figure 1, we can observe that although the absolute time domain positions of those local regions with high energy have some change after time scaling up to ±5%, the shape of them does not change a lot. Thus, by defining such high energy regions as regions of interest (ROI) and embed the watermark in these areas, it is reasonable to believe that the watermark will be safe under time scale modification attacks to some extent.

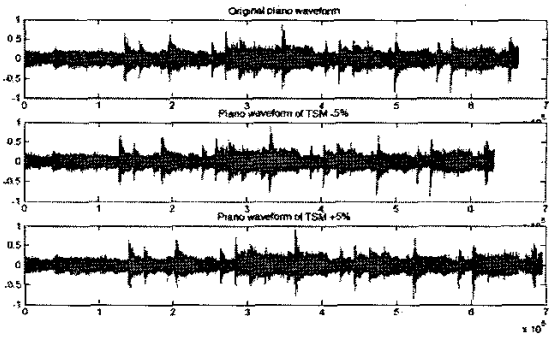

Figure 1. The waveform of the original and the ±5% time scaled piano waveform

2.2 Localization and random cropping

Random cropping is another serious synchronization attack to audio watermarking, it causes displacement between the embedding window and the detection window, thus makes the detector failed. Based on the discussion above, the method to embed watermark in regions of interest is by nature a kind of localized scheme, since the watermark is embedded only in some high energy regions corresponding to music edge, while not in the whole audio. Such localized watermarking scheme possesses natural resistance against random cropping, because pirates usually cut those trivial audio clips while preserve important clips including the ROIs in our case, for the purpose of keeping the value of the piratical music.

2.3 Embedding Regions Selection

ROI is the most important factor in localized watermarking schemes, because it indicates the area where the watermark bits are embedded, therefore, it must be designed to be able to withstand common audio signal processing attacks like lossy compression and synchronization attacks like random cropping and time scale modifications. If ROI regions are incorrectly identified, the detection is bound to be failed. It is our observation that short time frames around all the strong local energy peaks can serve as good regions of interest to embed and detect watermark. In experiment, we first smooth the waveform by applying denoising technique, then detect all the strong peaks by calculating gradient and slope. The adopted frame length is 4096 samples, which is approximately 0.1s long, corresponding to a single musical note or a sound of drum, under the condition of 44,100 Hz sampling rate.

3. Embedding Strategy

a) First, all magnitude peaks of the watermarked audio waveform are calculated. Let *iPeakNum* be the number of all detected peaks, then the number of embedding regions *ROINum* is calculated as follows, to ensure its being odd when applying the majority rule in detection.

$$ROINum = iPeakNum + (iPeakNum \% 2 - 1) \quad (1)$$

b) After determining all the watermark embedding regions, Fast Fourier Transformation is performed to each region, AC FFT coefficients from 1kHz to 6kHz are selected as the dataset for watermark embedding.

c) The watermark adopted in our experiment is a 64-bit pseudorandom number sequence W, denoted by (2), it is mapped into an antipodal sequence W' before embedding using BPSK modulation ($1 \rightarrow -1$, $0 \rightarrow +1$) according to (3), for the convenience of applying majority rule in detection. Experimental results show that a 64-bit watermark can maintain high audio perception quality, while a 128-bit or bigger watermark will introduce annoying distortion, that is, exceeding the watermark capacity of some 4096-sample embedding regions.

$$W = \{w(i) \mid w(i) \in \{1,0\},\ 1 \le i \le 64\} \quad (2)$$

$$W' = \{w'(i) \mid w'(i) = 1 - 2 * w(i),\ w'(i) \in \{+1,-1\},\ 1 \le i \le 64\} \quad (3)$$

d) Each watermark bit, $w'(k)$, is repeatedly embedded into all the selected ROI regions by exchanging the corresponding AC FFT coefficient pair according to (4)

$$\begin{aligned}
&for\ \ l = 1 : ROINum \\
&\quad for\ \ k = 1 : 64 \\
&\quad\quad flag = ROIFFTR(off + 2*k - 1) < ROIFFTR(off + 2*k) \\
&\quad\quad \begin{cases} if\ \ w'(k) = 1\ \ and\ \ flag = 1 \\ \quad exchange\ the\ absolute\ value \\ if\ \ w'(k) = -1\ \ and\ \ flag = 0 \\ \quad exchange\ the\ absolute\ value \end{cases} \\
&\quad end \\
&end
\end{aligned} \quad (4)$$

where $ROIFFTR(off+2*k-1)$ and $ROIFFTR(off+2*k)$ are the AC FFT coefficients at the low-middle frequency band ranging from 1kHz to 6kHz, off is a user defined offset. Because most of these coefficients are in the same order of magnitude, exchanging them while preserving the biggest low frequency (<1kHz) coefficients will not introduce annoying auditory quality distortion.

e) Inverse Fast Fourier Transformation (IFFT) is applied to the modified AC FFT coefficients in each ROI region to transform them back to the waveform in the time domain.

4. Detection Strategy

a) First, the same method with embedding is used to determine all watermark detection regions. Let $iPeakNum1$ be the number of calculated local high energy peaks, then the number of detection regions $ROINum1$ can be calculated as (5), to ensure its being odd when applying the majority rule in detection. Note that the number of detection regions ($ROINum1$) may be different from that of embedding regions ($ROINum$), since it is usually changed more or less after undergoing all kinds of distortions such as audio signal processing or time domain synchronization attacks.

$$ROINum1 = iPeakNum1 + (iPeakNum1 \% 2 - 1) \quad (5)$$

b) Next, Fast Fourier Transform is performed to each ROI region, obtaining a series of AC FFT coefficients for watermark detection.

c) The embedded watermark bits in each region are extracted based on the following rule (6), then the BPSK modulated antipodal watermark bits are determined based on the majority rule according to (7), since it is equal to global redundancy to embed the same watermark into all embedding regions.

$$\begin{aligned}
&for\ \ m = 1 : ROINum1 \\
&\quad for\ \ n = 1 : 64 \\
&\quad\quad flag = FFTR(2*n - 1 + off) > FFTR(2*n + off) \\
&\quad\quad \begin{cases} if\ \ flag = 1\ \ then\ \ w'(m,n) = 1 \\ if\ \ flag = 0\ \ then\ \ w'(m,n) = -1 \end{cases} \\
&\quad end \\
&end
\end{aligned} \quad (6)$$

$$w(n) = sign\left(\sum_{m=1}^{m = ROINum\ 1} w'(m,n) \right) \quad 1 \le n \le 64,\ 1 \le m \le ROINum1 \quad (7)$$

where m is the m-th embedding region, n means the n-th watermark bit embedded in the m-th region, and $ROINum1$ is the number of all detection regions.

d) Finally, BPSK demodulation is used to obtain the original watermark bits:

$$w(i) = (1 - w'(i))\ /2 \quad 1 \le i \le 64 \quad (8)$$

5. Experimental Results

The algorithm was applied to a set of audio signals including pop, saxophone, rock, piano, and electronic organ (15s, mono, 16 bits/sample, 44.1kHz). The waveform of the original and the watermarked piano music is shown in Figure 2, with the signal noise rate (SNR) of 33.5 dB, which is rather high to show that little apparent distortions have been introduced.

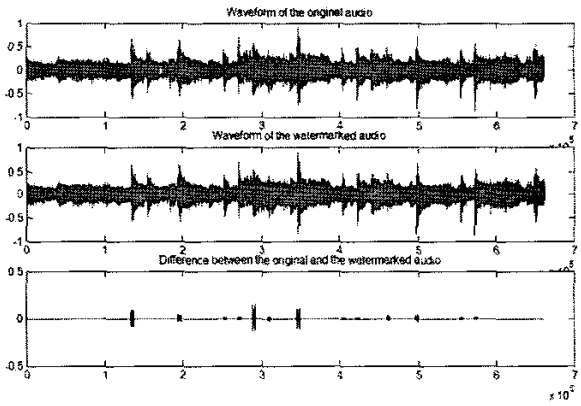

Figure 2. (a) The original piano waveform (b) The watermarked piano waveform (c) The difference between (a) and (b).

5.1 Robustness Test

The experimental conditions and robustness test results under common audio signal processing, random cropping, time scale modification and Stirmark for Audio are listed in Table 1-4.

From table 1 it can been seen that this algorithm is very robust to high strength audio signal processing, for example, it can resist MP3 compression up to 32kbps (22:1), low pass filtering with the cutoff frequency of 4kHz, noise addition that can be heard clearly by everybody, resample, echo, denoise etc.

Table 1. RCDR (Ratio of Correctly Detected Regions), sim, BER of piano under audio signal processing

Attack Type	RCDR	Sim	BER
UnAtacked	12/15	1	0%
MP3 (32kbps)	2/15	0.9276	7.81%
MP3 (48kbps)	2/15	0.9398	6.25%
MP3 (64kbps)	7/15	0.9706	3.13%
MP3 (96kbps)	12/15	1	0%
MP3 (128kbps)	12/15	1	0%
Low pass (4kHz)	9/13	1	0%
Low pass (8kHz)	12/15	1	0%
Resample (44100 Hz-> 16000 Hz ->44100 Hz)	10/15	1	0%
Resample (44100 Hz -> 22050 Hz ->44100 Hz)	12/17	1	0%
Echo (100ms, 40%)	8/15	0.9852	1.56%
Noise (audible)	4/13	0.9852	1.56%
Denoise (Hiss Removal)	2/11	0.8986	10.94%

Table 2 shows strong robustness to random cropping, as long as one or more embedding regions are not cropped, the detection will succeed. In our experiment, even 10000 samples are cropped at each of 8 randomly selected positions, it does not make any affection to the watermark detection.

Table 2. RCDR, sim, BER of piano under random cropping and jittering

Attack Type	RCDR	Sim	BER
Crop1 (10000*8)	11/13	1	0%
Jittering (1/1000)	3/15	0.9118	9.38%
Jittering (1/1500)	5/15	1	0%
Jittering (1/2000)	7/15	0.9856	1.56%

Pitch-invariant time scale modification is a challenging problem in audio watermarking, it can be viewed as a special form of random cropping, removing or adding some parts of audio signal while preserving the pitch. In our test dataset, the algorithm shows strong robustness to this attack up to at least ±10%, exceeding the ±4% standard requested in the SDMI phase-II proposal. Based on the introduction in section 2.1, this is mainly due to the relative invariance of the high energy regions under such attacks. The test results of piano under time scale modification from -20% to +20% are tabulated in table 3 (— means that watermark detections in all embedding regions are failed).

Table 3. RCDR, sim, BER of piano under time scale modification

Attack Type	RCDR	Sim	BER
TSM-1%	10/13	1	0%
TSM-2%	9/15	1	0%
TSM-3%	10/15	1	0%
TSM-4%	8/11	1	0%
TSM-5%	5/15	1	0%
TSM-6%	7/13	1	0%
TSM-7%	6/15	1	0%
TSM-8%	6/15	1	0%
TSM-9%	7/15	1	0%
TSM-10%	5/15	1	0%
TSM-11%	2/15	0.8956	10.94%
TSM-12%	5/13	1	0%
TSM-13%	4/15	1	0%
TSM-14%	4/13	0.9852	1.56%
TSM-15%	1/15	0.9701	3.13%
TSM-16%	1/11	0.8359	17.19%
TSM-17%	0/15	—	—
TSM-18%	2/11	0.9412	6.25%
TSM-19%	0/13	—	—
TSM-20%	3/15	0.9276	7.81%
TSM+1%	10/15	1	0%
TSM+2%	12/15	1	0%
TSM+3%	8/15	1	0%
TSM+4%	10/15	1	0%
TSM+5%	9/15	1	0%
TSM+6%	9/17	1	0%
TSM+7%	9/15	1	0%
TSM+8%	8/17	1	0%
TSM+9%	8/15	1	0%
TSM+10%	5/15	1	0%
TSM+11%	6/15	1	0%
TSM+12%	7/17	1	0%
TSM+13%	3/15	1	0%
TSM+14%	4/17	1	0%
TSM+15%	10/15	1	0%
TSM+16%	1/13	0.8788	14.06%
TSM+17%	3/17	1	0%
TSM+18%	1/17	0.8658	14.06%
TSM+19%	4/13	1	0%
TSM+20%	0/11	—	—

Stirmark for Audio is a standard robustness evaluation tool for audio watermarking technique. All operations are performed by default parameter except that the MP3 compression bit rate is changed to 32kbps. From table 4, we can see that most results are satisfactory. In the cases of failure, the auditory quality is also distorted severely.

Table 4. RCDR, sim, BER of piano under Stirmark for Audio

Attack Type	RCDR	Sim	BER
write_addbrumm_100	12/15	1	0%
write_addbrumm_1100	11/13	1	0%
write_addbrumm_2100	10/13	1	0%
write_addbrumm_3100	8/15	1	0%
write_addbrumm_4100	6/15	0.9852	1.56%
write_addbrumm_5100	6/17	0.9706	3.13%
write_addbrumm_6100	6/17	0.9553	4.69%
write_addbrumm_7100	6/21	0.9566	4.69%
write_addbrumm_8100	6/25	0.9412	6.25%
write_addbrumm_9100	5/35	0.9304	7.81%
write_addbrumm_10100	4/37	0.9147	9.38%
write_addnoise_100	12/15	1	0%
write_addnoise_300	12/15	1	0%
write_addnoise_500	13/15	1	0%
write_addnoise_700	12/15	1	0%
write_addnoise_900	12/13	1	0%
write_addsinus.wav	13/15	1	0%
write_amplify	12/15	1	0%
write_compressor	8/15	1	0%
write_copysample	1/5	0.8933	10.94%
write_cutsamples	0/13	—	—
write_dynnoise	8/13	0.9856	1.56%
write_echo	2/19	0.8792	12.50%
write_exchange_30	12/15	1	0%
write_exchange_50	12/15	1	0%
write_exchange_70	12/15	1	0%
write_fft_hlpass	6/13	0.9566	4.69%
write_fft_invert	12/15	1	0%
write_fft_real_inverse	12/13	1	0%
write_fft_stat1	4/13	0.9852	1.56%
write_fft_test	4/13	0.9852	1.56%
write_flippsample	1/17	0.9276	7.81%
write_invert	12/15	1	0%
write_lsbzero	12/15	1	0%
write_normalize	12/15	1	0%
write_nothing	12/15	1	0%
write_original	12/15	1	0%
write_rc_highpass	6/15	0.9701	3.13%
write_rc_lowpass	12/13	1	0%
write_smooth2	12/13	1	0%
write_smooth	11/13	1	0%
write_stat1	12/13	1	0%
write_stat2	12/13	1	0%
write_zerocross	10/15	1	0%
write_zerolength	2/13	0.8658	14.06%
write_zeroremove	12/17	1	0%

6. Conclusion

In this paper, by embedding the watermark in the perceptually important localized regions of interest, we obtain high robustness against common audio signal processing and synchronization attacks. The selection of the ROI is the most crucial step in this algorithm. Our future work aims at finding better ROI to further improve the ROI stability under time scale modification and other audio signal processing attacks.

References

[1] R. Tachibana, S. Shimizu, T. Nakamura, and S. Kobayashi, "An audio watermarking method robust against time and frequency fluctuation," in SPIE Conf. on Security and Watermarking of Multimedia Contents III, San Jose, USA, January 2001, vol. 4314, pp. 104– 115.

[2] http://amath.kaist.ac.kr/research/01-11.pdf.

[3] W. Li, X.Y. Xue, "Audio Watermarking Based on Statistical Feature in Wavelet Domain", in Poster Track of the Twelfth International World Wide Web Conference (WWW2003). Budapest, Hungary, May 2003.

[4] C. P. Wu, P. C. Su, and C-C. J. Kuo, "Robust and efficient digital audio watermarking using audio content analysis," in SPIE Int. Conf. on Security and Watermarking of Multimedia Contents II, San Jose, USA, January 2000, vol. 3971, pp. 382–392.

[5]http://www.sdmi.org/download/FRWG00022401-Ph2_CFPv1.0.pdf, SDMI Phase II Screening Technology Version 1.0, Feb 2000.

[6] M. Mansour, A. Tewfik, "Time-Scale Invariant Audio Data Embedding". Proc. IEEE International Conference on Multimedia and Expo, ICME, 2001.

[7] M. Mansour and A. Tewfik, "Audio Watermarking by Time-Scale Modification", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Salt Lake City, May 2001.

[8] R. Tachibana, "Improving audio watermarking robustness using stretched patterns against geometric distortion," Proc. of the 3rd IEEE Pacific-Rim Conference on Multimedia (PCM2002), pp. 647-654.

[9] K. N. Hamdy, A. H. Tewfik, T. Chen, and S. Takagi, "Time-Scale Modification of Audio Signals with Combined Harmonic and Wavelet Representations," ICASSP-97, Munich, Germany.

[10] C. Duxbury, M. E. Davies and M. B. Sandler, "Separation of Transient Information in Musical Audio Using Multiresolution Analysis Techniques", the 4th International Workshop on Digital Audio Effects, DAFx01, Limerick, December 2001.

Wei Li is a Ph.D. candidate of Fudan University, P.R.China and the corresponding author, whose research interest includes audio watermarking and image processing etc.

Acknowledgement:

This work was supported in part by NSF of China under contract number 60003017, China 863 Projects under contract numbers 2001AA114120 and 2002AA103065, Local Government R&D Funding under contract numbers 01QD14013 and 015115044, National Nature Science Funds of China (10171017, 90204013).