A Novel Scheme for Merging Digital Audio Watermarking and Authentication

Nedeljko Cvejić, Tapio Seppänen
MediaTeam Oulu, Information Processing Laboratory
University of Oulu
Oulu, Finland
{cvejcic, tapio}@ee.oulu.fi

Abstract—We present a novel scheme that is able to combine digital watermarking and content authentication of digital audio. The embedding of additional data is performed in discrete wavelet domain. Watermark embedding is done by frequency hopping method, while the additional authentication data is hidden using the LSB modulation. The perceptual transparency is achieved using the frequency masking property of the HAS. The scheme obtains high robustness against standard watermark attacks and localizes accurately tampered parts of the audio clip.

Keywords—audio watermarking, data hiding, digital rights management

I. INTRODUCTION

Broadband communication networks and multimedia data available in digital format opened many challenges and opportunities for innovation. Versatile and simple-to-use software and decreasing prices of digital devices have made possible for consumers from all around the world to create and exchange multimedia data. Broadband Internet connections and near error-free transmission of data facilitate people to distribute large multimedia files and make identical digital copies of them. Perfect reproduction in digital domain has promoted the protection of intellectual ownership and the prevention of unauthorized tampering of multimedia data to become important technological and research issue.

Multimedia data hiding techniques have developed a strong basis for digital watermarking and steganography area with a growing number of applications like digital rights management, covert communications, hiding executables etc. In the past few years, several algorithms for the embedding and extraction of watermarks in audio sequences have been presented. All of the developed algorithms take advantage of the properties of the human auditory system (HAS) in order to embed a watermark into host signal in a perceptually transparent manner. A broad range of embedding techniques goes from simple least significant bit (LSB) scheme to the various spread spectrum methods.

Most of the developed audio watermarking algorithms perform well in the digital copyright protection applications, due to high robustness against watermark attacks. However, in some applications, in addition to a robust watermark there is a need for checking the authenticity of the watermarked audio. Applications for digital audio authentication can be found in many areas; e.g. sound recording of criminal events (authentic recording of legally essential conversation could lead to progress in criminal cases), broadcasting (tampered audio clip could be used for manipulating public opinion) and military intelligence (authentication allows the military to authenticate whether received audio does come from the legitimate source).

Therefore, the watermarking system should be able to perform the content authentication and, in addition, check whether the watermarked audio was tampered prior to watermark extraction. In the case of tampering, the embedded watermark should be declared invalid after any, even slightest, modification of the watermarked audio clip.

II. METHOD

The simplest visualization of the requirements of information hiding in digital audio is so called magic triangle. Inaudibility, robustness to attacks, and the watermark data rate are in the corners of the magic triangle. This model is convenient for a visual representation of the required trade-offs between the capacity of the watermark data and the robustness to certain watermark attacks, while keeping the perceptual quality of the watermarked audio at an acceptable level. It is not possible to attain high robustness to signal modifications and high data rate of the embedded watermark at the same time. Therefore, if a high robustness is required from the watermarking algorithm, the bit rate of the embedded watermark will be low and vice versa, high bit rate watermarks are usually very fragile in the presence of signal modifications. However, there are some applications that do not require that the embedded watermark has a high robustness against signal modifications. In these applications, the embedded data is expected to have a high data rate and to be detected and decoded using a blind detection algorithm. While the robustness against intentional attacks is usually not required, signal processing modifications, like noise addition, should not affect the covert communications [1]. To qualify as steganography applications, the algorithms have to attain statistical invisibility as well.

The proposed scheme utilizes spread spectrum (SS) technique and LSB modulation in discrete wavelet domain. The wavelet domain provides information about temporal domain and frequency spectrum simultaneously. In addition, it is very suitable for frequency analysis due to its multiresolutional properties that provide access both to the
most significant parts and details of signals spectrum [2]. Therefore, we are able to make easily the trade-off between the amount of the embedded information and perceptual distortion caused by information hiding, by handling subbands with different levels of power and perceptual significance.

The Discrete Wavelet Transform (DWT) decomposes the signal into low-pass and high pass components subsampled by two; the inverse transform performs the reconstruction. We decided to make use of the simplest quadrature mirror filter - Haar filter. The Haar wavelet has the shortest support among all orthogonal wavelets, and it is the only quadrature mirror filter that has a finite impulse response [2]. FIR filters can be designed to be linear phase filters, what is important from the point of view of the perceptual transparency, as the linear phase filters delay the input signal, but do not distort its phase. In addition, Haar filter is computationally simple to implement, as on most DSP processors, the FIR calculation can be done by looping a single instruction. This property gives the opportunity for real time applications of the proposed algorithm.

The embedding algorithm is given in Figure 1. After wavelet decomposition of a block of 512 samples of host audio, using Haar filter and decomposition depth of five steps, algorithm produces 512 wavelet coefficients. Simultaneously, the same block is forwarded to the attack characterization section of the embedding scheme (Figure 1). The attack characterization section has the purpose of analyzing the signal for the watermark removal attacks with different signal processing methods [3]. Besides detection desynchronization attack, the most malicious attacks for SS audio watermarking algorithms are MPEG compression and LP filtering. Therefore, each data-hiding block undergoes mp3 compression (48 kbps bit rate) and LP filtering (cut-off frequency 8 kHz, stop band attenuation 80dB). Distortion measure D for the ratio of the original magnitude of an wavelet coefficient C_i and magnitude of the same wavelet coefficient after the simulates attack C_i', is calculated during a predefined time interval T:

\[ D = \sum_{i=1}^{N} a_i D_i, \quad D_i = \frac{(C_i - C_i')^2}{C_i^2} \quad \text{and} \quad a_i = \log(i+1) \]

for \( i=1,2,...,N \). Coefficients \( a_i \) are introduced because experiments showed that modification of the wavelet coefficients at the lower frequencies introduces more perceptual distortion, as they contain more signal energy. The \( a_i \) expression is derived from experimental data. Other models for weighting coefficients have been tested, with similar results; however, results in Section III are obtained using the expression above. Subsequently, weights \( a_i \) improve perceptual transparency of the algorithm, allowing less distortion in the frequency subbands of the higher sensitivity of the HAS.

Algorithm selects a subset corresponding to 128 consecutive wavelet coefficients (of 512 coefficients in total) with the least distorted magnitudes, with the constraint that the lowest 50 wavelet coefficients are not considered, as their modification causes significant perceptual artifacts. Identity of the first coefficient in the subset of coefficients that will be used for data embedding is binary encoded and submitted to the watermarking embedding module. At the embedding module, the binary coded identity of the position of the first coefficient is inserted along with watermark bits into single bit stream and embedded into data hiding blocks with N-fold repetition during time interval T.

![FIGURE I AUDIO WATERMARK/AUTHENTICATION EMBEDDING SCHEME](image)

Watermark embedding is performed by frequency hopping method [4], presented in Figure II. Thus, a secret key is used to select two wavelet coefficients from the subset least affected by modeled attacks. The mean value of the magnitudes of all the coefficients in the subset is calculated and assigned to the two mapped coefficients' magnitudes. The magnitude of the coefficient at the lower frequency is then increased by K decibels (dB) and the value of the second coefficient is decreased by the same value, if bit 1 is to be embedded. The opposite arrangement is done if bit 0 is signaled. The value K is chosen to be equal to distance from the mean value of the magnitudes of the subset to the frequency masking threshold, derived from the frequency masking property of HAS.

![FIGURE II FREQUENCY HOPPING METHOD USED IN WATERMARK EMBEDDING](image)

The authentication signature embedding is performed by LSB modulation of the remaining 384 wavelet coefficients [5], presented in Figure III. The wavelet coefficients are scaled using the maximum value inside the given subband and converted to binary arrays in the two's complement. A predetermined number of the LSBs are thereafter replaced with bits of information that represent authentication signature for the host audio. Coefficients are then converted and scaled back to the original order of magnitude and inverse DWT is performed. Data hiding in the LSB of the wavelet coefficients is practicable due to the near perfect reconstruction properties of the filter bank.

Using the same hopping key-based pattern as on the embedding side, the detector reads the magnitude of the first wavelet coefficient and calculates the difference between it and the mean value of all the magnitudes of the coefficients in the sub band. The same operation is repeated for the wavelet coefficient on the higher frequency. At the end, the detection value is calculated as the difference between the values. The
sign of the difference determines the value of the extracted bit. After time interval T, a new subset is selected using the

![Wavelet Coefficients Diagram](image)

FIGURE III SIGNAL DECOMPOSITION PRIOR TO LSB EMBEDDING

extracted information about the position of the first coefficient of the subset.

III. EXPERIMENTAL RESULTS

Subjective quality evaluation of the watermarking method has been done by listening tests involving ten persons. A total number of eight audio pieces were used as test signals, of 10 s duration each. The audio excerpts were selected so that they represent a broad range of music genres, i.e. audio clips with different dynamic and spectral characteristics. In the first part of the test, participants listened to the original and the watermarked audio sequences and were asked to report dissimilarities between the two signals, using a 5-point impairment scale: (5: imperceptible, 4: perceptible but not annoying, 3: slightly annoying, 2: annoying 1: very annoying). Table I presents results of the test, the lowest and the highest values from the impairment scale and average MOS for given audio excerpt. In the second part, test participants were repeatedly presented with the original and watermarked audio clips and were asked to determine the watermarked one.

<table>
<thead>
<tr>
<th>file name</th>
<th>SNR</th>
<th>discrimination</th>
<th>MOS range</th>
<th>aver. MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>lovet</td>
<td>31.0</td>
<td>52%</td>
<td>5</td>
<td>4.6</td>
</tr>
<tr>
<td>ritenour</td>
<td>28.2</td>
<td>49%</td>
<td>4-5</td>
<td>4.8</td>
</tr>
<tr>
<td>yoyoma</td>
<td>29.2</td>
<td>50%</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>titanic</td>
<td>28.1</td>
<td>52%</td>
<td>4-5</td>
<td>4.8</td>
</tr>
<tr>
<td>yanni</td>
<td>31.7</td>
<td>47%</td>
<td>5</td>
<td>4.8</td>
</tr>
<tr>
<td>joecocker</td>
<td>28.4</td>
<td>49%</td>
<td>4-5</td>
<td>4.5</td>
</tr>
<tr>
<td>abba</td>
<td>28.4</td>
<td>52%</td>
<td>4-5</td>
<td>4.2</td>
</tr>
<tr>
<td>eurhythmics</td>
<td>29.4</td>
<td>48%</td>
<td>4-5</td>
<td>4.1</td>
</tr>
</tbody>
</table>

Total average MOS 4.56

Experimental results are presented also in Table I, values near to 50% show that the two audio clips (original audio sequence and watermarked audio signal) cannot be discriminated. The following parameters were used during watermark embedding: time interval T=1s and number of repetitions N=1.

The audio clips were compressed to MPEG layer-3 files, at a rate of 48 kb/s using Syntrillium’s commercial mp3 coder based on software implementation licensed from the Fraunhofer IIS. The extraction results after the employed compression are presented in Table II. The detection performance of the algorithm was also tested against common signal processing attacks [6]:

1. All-pass filtering using system function: H(z)=0.81z^2 - 1.64z + 1 / (z^2 - 1.64z + 0.81)
2. Echo-addition (delay 100ms, decay 50%)
3. Band-pass filtering using a second order Butterworth filter with cut-off frequencies 100 Hz and 3000 Hz
4. Amplitude compression (8.9:1 for A>29dB, 1.73:1 for –46dB<A<29dB and 1.16:1 for A<-46dB)
5. Equalization (6-band equalizer, signal suppressed or amplified by 6 dB in each band)
6. Noise addition (with uniform white noise. Maximum magnitude of 200 quantization steps)
7. Time-scale modification of –3% or +3%, where the pitch remains unaffected.
8. Subsequent D/A and D/A conversion using standard analogue tape recorder
9. Resampling (consisting of subsequent down and up sampling to 22.05 kHz and 44.1 kHz, respectively)

Watermark detection results after the attacks described above are shown in Table II.

<table>
<thead>
<tr>
<th>ATTACK TYPE</th>
<th>BIT RATE (bps)</th>
<th>CLIP1</th>
<th>CLIP2</th>
<th>CLIP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. MPEG comp. (48 kbps)</td>
<td>1.1·10^2</td>
<td>1.2·10^2</td>
<td>1.1·10^2</td>
<td></td>
</tr>
<tr>
<td>2. Band pass filter</td>
<td>5.7·10^3</td>
<td>4.1·10^3</td>
<td>3.9·10^3</td>
<td></td>
</tr>
<tr>
<td>3. Resampling (44-22-44)</td>
<td>7.4·10^3</td>
<td>9.2·10^3</td>
<td>8.3·10^3</td>
<td></td>
</tr>
<tr>
<td>4. Amplitude compression</td>
<td>0</td>
<td>0</td>
<td>1.9·10^4</td>
<td></td>
</tr>
<tr>
<td>5. Echo addition</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6. All-pass filtering</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7. Equalization</td>
<td>0</td>
<td>0</td>
<td>3.8·10^4</td>
<td></td>
</tr>
<tr>
<td>8. Noise addition</td>
<td>0</td>
<td>0</td>
<td>3.8·10^4</td>
<td></td>
</tr>
<tr>
<td>9. Time scaling (+3%)</td>
<td>1.6·10^2</td>
<td>1.9·10^2</td>
<td>1.9·10^2</td>
<td></td>
</tr>
<tr>
<td>10. D/A–A/D conversion</td>
<td>3.8·10^4</td>
<td>3.8·10^4</td>
<td>3.8·10^4</td>
<td></td>
</tr>
</tbody>
</table>

The reason for poorer extraction capabilities after MPEG coding is that these compression techniques crop high frequency spectrum of the watermarked audio and quantize.
wavelet coefficients in other subbands. Time scaling or
detection desynchronization attack is always one of the most
malicious attacks on watermarking algorithms based on time
domain, but this algorithm showed a good performance after
that kind of attack as well.

<table>
<thead>
<tr>
<th>starting point/clip</th>
<th>Clip1</th>
<th>Clip2</th>
<th>Clip3</th>
<th>Clip4</th>
<th>Clip5</th>
</tr>
</thead>
<tbody>
<tr>
<td>t=500000</td>
<td>100%</td>
<td>96%</td>
<td>98%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>t=800000</td>
<td>100%</td>
<td>100%</td>
<td>98%</td>
<td>100%</td>
<td>98%</td>
</tr>
</tbody>
</table>

In content authentication tests the watermarked audio samples
were replaced by random samples from a selected starting
point. The detected percentage is shown in Table III. Any
number below 100% indicates that a part of audio has been
modified. After finding the incorrect authentication bit, the
detection system uses spatial information of the wavelet
coefficients to localize the modified content.

ACKNOWLEDGMENTS
The work is part of the Stego project, which is supported by
the Finnish National Technology Agency (TEKES), Nokia
and Yomi Solutions. The research topic has been supported by
Nokia Oyj Foundation and Taimo Tönnöng Research
Scholarship.

REFERENCES
improved robust watermarking”, IEEE Transactions on Signal
for high capacity audio steganography", Proc. IEEE Digital Signal
Processing Workshop, Callaway Gardens, GA, pp. 53-55.
watermarking for monitoring and copy protection", Proc. ACM
Multimedia Workshop, Marina Del Ray, CA, pp.119-122.