FUSING DIGITAL AUDIO WATERMARKING AND AUTHENTICATION IN DIVERSE SIGNAL DOMAINS

Nedeljko Cvejic, Tapio Seppänen

MediaTeam Oulu, Information Processing Laboratory, University of Oulu
P.O. Box 4500, 4STOINF, 90014 University of Oulu, Finland
phone: + 358 8 553 2797, fax: + 358 8 553 2534, email: cvejic@ee.oulu.fi
web: www.mediateam.oulu.fi

ABSTRACT

A novel scheme that is able to merge digital watermarking and content authentication of digital audio is presented in this paper. The embedding of additional data is performed in different signal domains. Watermark embedding is made by frequency hopping method in the Fourier domain, while the additional authentication data is hidden using the LSB modulation in the wavelet domain. 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

1. INTRODUCTION

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 if audio does come from a 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.

2. 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 [1]. 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 [2]. To qualify as steganography applications, the algorithms have to attain statistical invisibility as well.

The proposed scheme utilizes spread spectrum (SS) technique [3] and LSB modulation in discrete wavelet domain. Samples of the host audio sequence (mono signal, sampling frequency 44.1 kHz, 16 bits/sample) are forwarded to the SYNC module (Figure 1). In the SYNC module, host audio is divided into blocks used for data hiding and blocks used for watermark extraction synchronization. Data hiding blocks have a fixed length L, while synchronization blocks have a length chosen randomly from the interval $[L_1, L_2]$. Therefore, between each two consecutive data hiding blocks, there is one synchronization frame with variable length. In each synchronization frame, a perceptually shaped PN sequence is added to the host signal in time domain. Perceptual weighting of the added PN sequence is performed by prefiltering, using a filter with the frequency characteristic similar to the threshold in quiet curve of the HAS.

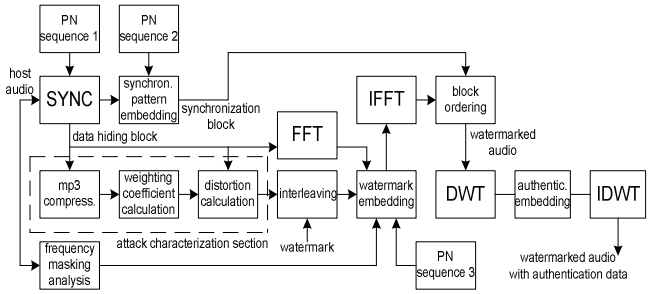

Figure 1: Watermark/authentication embedding scheme

Spreading gain of the embedded PN sequence is controlled through limits of the synchronization block length L_1 and L_2. Spreading gain must be large enough to make the detection of the cross-correlation peak reliable on the extraction side. Random length of the synchronization block improves the security of the algorithm, as it makes it very difficult for an adversary to find the exact position of the data hiding block inside a watermarked audio sequence.

The data hiding 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. Besides detection desynchronization attack, the most malicious attacks for the contemporary audio watermarking algorithms are MPEG compression and low pass (LP) filtering [3]. In a previously developed scheme [4], we use both MPEG compression and LP filtering attack characterization in order to find the subset of FFT coefficients least affected by these fading-like distortions. However, experimental tests showed that the characterization section selects similar subsets of FFT coefficients even if we leave out the LP filtering module, as MPEG compression has an inherently embedded LP filter [5].

Each data hiding block undergoes MPEG compression (48 kbps bit rate, mono). Distortion measure D for the ratio of the original magnitude of an FFT coefficient C_i and magnitude of the same FFT coefficient after the simulated attack C_i', is calculated during a predefined time interval T:

$$D = \sum_{i=1}^{N} a_i D_i$$

where

$$D_i = \frac{(C_i - C_i')^2}{C_i^2} \text{ and } a_i = \frac{\log(i+1)}{i}, \qquad i = 1,..., N$$

Coefficients a_i are introduced because experiments showed that modification of the FFT magnitudes at the lower frequencies introduces more perceptual distortion, as they contain more signal energy. The expression for a_i is derived from experimental data. Other models for weighting coefficients have been tested, with similar results; however, the presented test results were obtained using the above expression.

The algorithm selects a subband corresponding to 100 consecutive FFT coefficients (of 1024 coefficients in total) with the smallest cumulative value of D, with the constraint that the first 50 FFT coefficients are not considered, as their modification causes significant perceptual distortion. Identity of the first coefficient in the subband of coefficients that will be used for data embedding is binary encoded and submitted to the watermarking embedding module. In the first time interval of the audio sequence T, the position of the first coefficient of the subband is fixed. It does not compromise the security of the watermarking scheme, as two other secret keys, length of synchronization frame and hopping pattern inside the subband of coefficients are not known to the adversary. 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. Time interval T is chosen as a trade-off between two conflicting requirements. The first requirement is to get precise information about distortion of the magnitude of FFT coefficients at the particular time instant and the second one is decreasing the portion of the position information bits in the unified data stream.

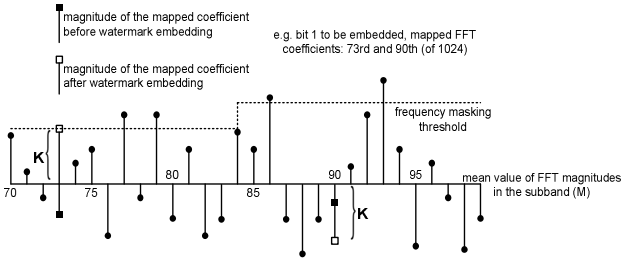

Figure 2: Frequency hopping method used during watermark embedding

Data embedding is performed by frequency hopping method, as shown in Figure 2. Thus, a secret key is used to select, from the subband selected as above, two FFT coefficients least affected by modelled attacks. The mean value of the magnitudes of all the coefficients in the subband is calculated and assigned to the two selected coefficients' magnitudes. If bit 1 is to be embedded, the magnitude of the coefficient at the lower frequency is increased by K decibels (dB) and the value of the second coefficient is decreased by the same value, while keeping the phases unchanged. The opposite arrangement is done if bit 0 is signalled. Increase in the magnitude of a randomly chosen FFT coefficient introduces narrowband noise, perceptually similar to the sound of tone dialling used in digital telephony. Listening tests proved that HAS is very sensitive to that noise, especially if the host audio has a dominant low frequency spectral structure and a small dynamics range. Therefore, the value K cannot be larger than the distance from the mean value of the magnitudes of the subband to frequency masking threshold (in dB), in order to keep the introduced distortion below just noticeable distortion (JND) value. Frequency masking threshold is obtained from the frequency masking property of HAS, and calculated for a given data hiding block after two mapped coefficients' magnitudes are set to the mean value of the subband. The model is derived from Psychoacoustic Model 1, which is used to control quantization step and bit allocation during MPEG audio coding. Controlled distortion of the perceptual quality of the host audio sequence provides perceptual transparency of the watermarked signal, while keeping robustness of the embedded watermark at a high level. After

the additional data bit has been embedded, the block is transformed back to time domain and inserted between two synchronization frames.

The authentication signature embedding is performed by LSB modulation of the data hiding block's (Figure 1) samples in wavelet domain, presented in Figure 3. Data hiding in the LSBs of the wavelet coefficients is practicable due to the near perfect reconstruction properties of the filterbank. The DWT decomposes the signal into low-pass and high pass components subsampled by two; the inverse transform performs the reconstruction.

We decided to use the simplest quadrature mirror filter - Haar filter. The Haar basis is obtained with a multiresolution of piecewise constant functions [6]. The scaling function is equal to one. 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 [6]. Signal decomposition into the low-pass and high pass part of the spectrum is performed in five successive steps. After subband decomposition of 512 samples of host audio, using the Haar filter and decomposition depth of five steps, algorithm produces 512 wavelet coefficients. All 512 wavelet coefficients are then scaled using the maximum value inside the given subband and converted to binary arrays in the two's complement [7]. A fixed number of the LSBs are thereupon replaced with bits of authentication data. Coefficients are then converted and scaled back to the original order of magnitude and an inverse transformation is made.

The order of hidden data extraction is opposite to the order in which the data was embedded into the host audio. In the beginning the authentication information is decoded and after that the watermark is detected and decoded, if the watermarked audio is determined not to be tampered. Before the extraction process the samples of the watermarked audio are first checked for synchronization. Mean removed cross-correlation, between synchronization block and the same prefiltered PN sequence as the one used during watermark embedding, is calculated and peak of the cross-correlation values is detected. If a desynchronization attack (e.g. time shifting or time offset caused by MPEG compression) has been introduced, correlation peak is displaced from the expected position by the amount equal to the shift in time domain. If a time shift is noticed, the following data hiding block is shifted for the same number of samples after which the extraction process from the data hiding block begins. If a time scaling attack is performed, the correlation peak is decreased for a random value, depending on the place where the samples of watermarked audio were deleted or additional samples inserted. However, the parameters of the synchronization block (L_1, L_2, T), set during the experiments enable reliable detection of the correct position of the data hiding block, if the scaling factor is in range [-3%,+3%]. Further increase/decrease of length in the watermarked audio significantly decreases performance of the watermarking extraction scheme.

As described above, the authentication data hiding operation was performed on the LSBs is a subset of wavelet coefficients chosen by a secret key. The authentication extraction process straightforwardly retrieves the watermark by reading the value of these bits. Therefore, the decoder needs all the samples of the watermarked audio that were used during the authentication embedding process.

Using the same key-based hopping pattern as on the embedding side, the detector reads the magnitude F_{k1} (in dB) of the first (lower frequency) FFT coefficient. The same operation is repeated for the second FFT coefficient with the magnitude F_{k2}. The detection value D is calculated as the difference between values F_{k1} and F_{k1}:

$$D = F_{k1} - F_{k2}$$

The sign of D determines the value of the extracted bit; positive value is mapped to bit 1, otherwise bit 0 is extracted. If the watermarked audio is not processed and modified, the value of D is equal to 2K; the trade-off between robustness of watermark extraction and perceptual transparency is clear. After time interval T, a new subband of FFT coefficients is selected, using the extracted information about the position of the first coefficient of the subband.

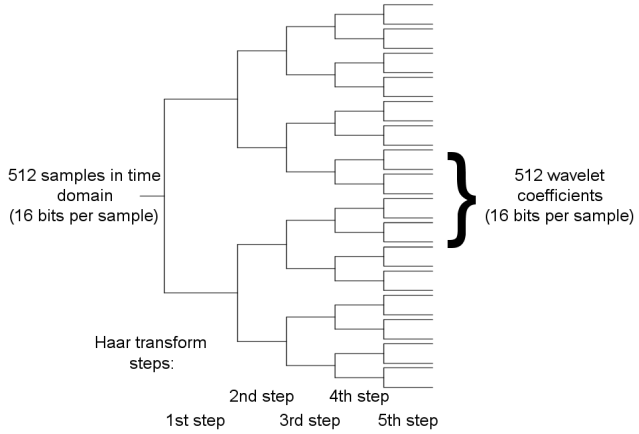

Figure 3: Signal decomposition prior to LSB embedding

3. 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 tests 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 present 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.

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.

TABLE I SNR, MEAN OPINION SCORES AND DISCRIMINATION OF ORIGINAL AND WATERMARKED AUDIO EXCERPTS

file name	SNR	discrimination	MOS range	Aver. MOS
Lovett	32.0	49%	5	5
Ritenour	29.1	52%	4-5	4.8
Yoyoman	29.9	51%	5	5
Titanic	28.9	48%	4-5	4.7
Yanni	32.5	51%	5	5
Joe Cocker	29.1	52%	4-5	4.6
Abba	29.2	52%	4-5	4.4
Eurythmics	30.0	48%	4-5	4.2
Total average MOS				4.71

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 [8]:

1. All-pass filtering using system function: H(z)=(0.81z2 - 1.64z + 1) / (z2 - 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 (8.91:1 for A>-29dB, 1.73:1 for –46dB<A<-29dB and 1:1.61 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.

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.

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.

TABLE II BIT ERROR RATE OF EXTRACTED WATERMARKS IN PRESENCE OF ATTACKS

Attack type /bit rate (bps)	Clip1	Clip2	Clip3
1. MPEG comp. (48 kbps)	$1.2\cdot10^{-2}$	$1.3\cdot10^{-2}$	$1.2\cdot10^{-2}$
2. Band pass filter	$6.1\cdot10^{-3}$	$4.6\cdot10^{-3}$	$4.1\cdot10^{-3}$
3. Resampling (44-22-44)	$7.1\cdot10^{-3}$	$8.4\cdot10^{-3}$	$8.8\cdot10^{-3}$
4. Amplitude compression	0	$1.1\cdot10^{-4}$	$1.9\cdot10^{-4}$
5. Echo addition	0	0	0
6. All-pass filtering	0	0	0
7. Equalization	0	0	$2.8\cdot10^{-4}$
8. Noise addition	0	0	$2.9\cdot10^{-4}$
9. Time scaling (+3%)	$1.8\cdot10^{-2}$	$1.7\cdot10^{-2}$	$1.9\cdot10^{-2}$
10. D/A–A/D conversion	$1.9\cdot10^{-4}$	$1.9\cdot10^{-4}$	$1.9\cdot10^{-4}$

TABLE III DETECTED PERCENTAGE OF THE TAMPERED SAMPLES

starting point/clip	Clip1	Clip2	Clip3
t=500000	100%	96%	97%
t=800000	100%	100%	98%

ACKNOWLEDGEMENTS

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 Tauno Tönning Scholarship.

REFERENCES

[1] Chou J, Ramchandran K, Ortega A "High capacity audio data hiding for noisy channels", *Proc. International Conference on Information Technology: Coding and Computing*, Las Vegas, NV, 2001, pp. 108–111.

[2] I. Cox, M. Miller and J. Bloom, "Digital Watermarking", Morgan Kaufmann Publishers, San Francisco, CA, 2003

[3] D. Kundur and D. Hatzinakos, "Diversity and attack characterization for improved robust watermarking", *IEEE Transactions on Signal Processing,* Vol. 29, No. 10, pp. 2383-2396.

[4] N. Cvejic and T. Seppänen "Audio watermarking using attack characterization", *Electronics Letters*, Vol. 13, No. 39, pp. 1020-1021.

[5] N. Cvejic and T. Seppänen "Spread spectrum audio watermarking using frequency hopping and attack characterization", *Signal Processing*, Vol. 84, No.1, pp. 207-213.

[6] Mallat S "Wavelet Tour of Signal Processing," Academic Press, San Diego, CA, 2001.

[7] N. Cvejic and T. Seppänen "A wavelet domain LSB insertion algorithm for audio steganography", *Proc. IEEE Digital Signal Processing Workshop*, Callaway Gardens, GA, pp. 53-55.

[8] J. Haitsma, M. van der Veen, T. Kalker, F. Bruekers. "Audio watermarking for monitoring and copy protection", *Proc. ACM Multimedia Workshop*, Marina Del Ray, CA, pp.119-122.