Chaos Based Audio Watermarking with MPEG Psychoacoustic Model I
A. Giovanardi, G. Mazzini, M. Tomassetti
University of Ferrara, Italy

Abstract

A method to implement digital audio watermarking in the frequency domain is presented. The embedding is performed by using chaos based sequences to increase the signature hiding properties. The proposed scheme takes into account the MPEG psychoacoustic model I to improve the robustness against the MP3 compression. Some tests have been performed to verify the system robustness against the MP3 compression and also against the signal cropping, re-sampling, re-quantization and filtering.

1 Introduction

Researches on digital audio and video signals have permitted to design systems with low distortion and noise, by increasing the digital source diffusion on the worldwide market. The negative counterpart is the facility to copy audio and video materials and to tamper with them with digital files.

As a result, performers, studios, distributors and retailers need a reliable, tamper-proof and permanent audio watermarking solution ([1] [2], [3], [4]) to embed inaudible (or not visible) and indelible information, to protect and track inclusive content. In particular, to meet today’s copyright protection requirements, electronic watermarking techniques, applied to audio signals, must satisfy four basic requirements [5]: the watermark should be inaudible, i.e., the sound quality must not be significantly corrupted; indelible, i.e., the watermark should not be removed from the audio signal; robust, i.e., the watermark should be resistant to the main digital manipulations; invisible, i.e., the watermark presence should not be easily verified in order to prevent its removal. Besides, multiple watermarks should be supported to track the ownership passing or the movement of the proprietary source. Finally, the watermark detection should not require the copy of the original audio track.

This paper proposes a new scheme to protect audio signal copyright taking into consideration the above mentioned requirements. An often used watermarking algorithm is the Patchwork approach [1]. A similar approach which is working in the time domain has been discussed by [2]. In [3] a frequency domain approach is shown, where the watermark detection requires the copy of original audio tracks. Other methods act in frequency domain, e.g., [4]. The novelty of our work, with respect to [1], [2], [3], [4], is that it integrates different aspects: the watermark detection does not require the copy of the original track, the watermark is applied in the frequency domain, it is based on chaotic sequences and the embedding algorithm is based on the MPEG psychoacoustic model I, to improve robustness against MP3 compression.

2 MPEG Psychoacoustic Model I

The human auditory system [6] can detect sounds with frequencies between 20Hz and 20kHz and acts as a frequency analyzer. Thus, it can be modeled as a set of 32 bandpass filters with bandwidths increasing with the frequency. This 32 bands are usually known as "critical bands": if a faint tone lies in the critical band of a louder tone, the faint one results to be masked. However, also a temporal masking is present, i.e., to hear a faint tone following a louder one, a give amount of time must be wait.

The MPEG audio compression algorithms are based on these masking effects: perceptually irrelevant informations are removed to increase the data compression. Let us briefly review the basic steps involved in the psychoacoustic model I algorithm. Audio data time alignment: a frame structure, with a given number of samples, depending on the MP3 layers is created; frequency domain representation: the audio samples are converted to their frequency domain representation, by using a Fast Fourier Transform (FFT); tonal and non-tonal components identification: the spectral values are divided into tonal and non-tonal components; the model identifies tonal components on the basis of the local peaks of the audio power spectrum, then it sums all the remaining components (non-tonal) to obtain a single value for each critical band; masked components removal: by means of an empirical masking threshold, chosen close to the lower bound of the sound audibility, the masked components are decimated; global masking threshold choice: for each band a global masking threshold is chosen, on the basis of the remaining tonal and non-tonal components; signal-to-mask ratio (SMR) evaluation: the SMR, defined as the ratio between the signal energy within a given sub-band and the minimum masking threshold for that band, is computed. This value is used by the MPEG encoder to decide how many bits allocate in each band: the greater is the SMR, the more bits are allocated.

3 Watermark Embedding

The watermark embedding is performed by dividing the audio track into blocks, and by superimposing the same watermarking sequence on each block. So, the sequence casting is realized more times along the audio content. In particular, the main steps to make the embedding are: the audio file is decoded to get the audio sample stream; the audio sample stream is divided into blocks and sub-blocks: each block contains N sub-blocks, where the sub-block length is equal to the FFT block size, L; therefore, each block length is NL. The FFT is applied to each sub-block with resulting vector of length L/2, due to the FFT symmetry; a chaotic sequence sized NL/2 is generated; the chaotic sequence is superimposed to each block, with size NL/2, containing the samples in the frequency domain.

The signature sequence to be hidden can be chosen in many ways, but watermarking schemes are often using pseudo-random (PN) number generators. In this work we are interested to use chaotic sequence generators, giving easy control on the chaotic trace statistical behavior in term of expected value, auto- and cross-correlation, due to that some studies in the literature are showing better performance of chaotic watermarking with respect to the PN one.

The chaotic watermarking signal for audio sources can be obtained by the recursive application of a suitable one-dimensional discrete-time dynamical system, i.e., a chaotic map [7] [8]. Let
us refer to a particular class of maps [9], called Piecewise Affine Markov Maps (PWAM), characterized by a $M : X \to X$, where $X = [0, 1]$, with $x_{k+1} = M(x_k)$ and assume that $n + 1$ points $0 = a_0 < a_1 < \ldots < a_n = 1$ exist defining the intervals $X_j = [a_{j-1}, a_j]$, for $j = 1, \ldots, n$, such that, for any couples of indices $j$ and $k$, either $X_k \subseteq M(X_j)$ or $X_k \cap M(X_j) = \emptyset$.

A particular class of PWAM are the so called $(n, t)$-tailed shift maps [10], where $n, t$ are integer such that $n$ is even and $t < n/2$:

$$M(x) = \begin{cases} (n - t)x & \text{if } 0 \leq x < \frac{n - t}{n} \\ (t - x) & \text{if } \frac{n - t}{n} \leq x < n \\ 0 & \text{otherwise} \end{cases}$$

The sequences that we have superimposed to the audio content in the frequency domain have been obtained by iterating the $(n, t)$-tailed shift maps, where the first iteration element $x_0$ is the seed of the sequence and it represents the watermarking key. Before the casting, the generic chaotic sequence $x_k$ has been processed by means of the function $Q : M \to [-1, +1]$, where $Q(x) = 2x - 1$, in order to obtain a sequence with null mean. With this, the final chaotic sequence is $y_k = Q(x_k)$. Furthermore, we have considered just $NL/4$ different values of the final chaotic sequences, $y_k$, with $k = 0, \ldots, NL/4 - 1$, and put them into the even position of the audio frequency block with natural sign, and into the odd one with the opposite sign, i.e., $y_k = (-1)^k y_{[k/2]}$, with $k = 0, \ldots, NL/2 - 1$. This procedure permits to have lower cross-correlations, as verified by means of experimental tests. Let us underline that this mapping does not change the watermarking sequence auto- and cross-correlation statistics, considering that the same law has been used in the detection unit.

In order to minimize the sound distortion, let us observe that audio spectrum components which are less audible by human hearing, can be more affected by the watermark cast with respect to components more audible. So, we have shaped the audio spectrum components which are less audible by human natural sign, and into the odd one with the opposite sign, i.e., those spectrum components which are less audible by the human old curve, hearing. In particular, we have considered the absolute threshold curve, function of the frequency, studied in [11] and reported in Figure 1. Then we have identified a simple first degree polynomial function, based on $T(f)$, i.e., the minimum audibility threshold for the human auditory system, function of the frequency, studied in [11] and reported in Figure 1. Then we have identified a simple first degree polynomial function, based on $T(f)$, such that the shaped quantized chaotic sequences in each block, $\tilde{z}_k$, with $k = 0, \ldots, NL/2 - 1$, results:

$$\tilde{z}_k = y_k \left\{ (1 - S) + \frac{S \left( T \left[ L \frac{f_k}{f_M} \right] - T_m \right)}{T_M - T_m} \right\}$$

where $S$ indicates the shaping strength, $f_M$ is the maximum frequency of the absolute threshold curve, $T_M$ and $T_m$ represent the maximum and minimum values of such curve, respectively.

After the shaping rule application, the $\tilde{z}_k$ sequence energy $E_{\tilde{z}} = \sum_k z^2_{\tilde{z}}$ has been forced to a referring value, $E_{\tilde{z}}$, in order to enhance the reliability of the watermarking detection process, as discussed in the following. The final embedding normalized sequence results $z_k = \bar{z}_k \sqrt{E_{\tilde{z}}/E_{\tilde{z}}}$. The $E_{\tilde{z}}$ value is chosen such that $z_k \ll 1$, condition guaranteeing a good detection, as reported in Section 4. An example of the normalized shaped chaotic sequences evolution is reported in Figure 2.

The casting of the watermark is made by using a spread spectrum technique, as in [12] and [13]. A possible implementation of this technique is obtained by slightly modifying the magnitude of each frequency tone in each frequency sub-block (in this case none modification is imposed to the tone phase).

Let us call $R_k$ and $W_k$ the original FFT sample module and the watermarked sample module, respectively. Then: $W_k = R_k(1 + z_k)$.

Note that the amplitude of the casting is modulated by mean of the parameter $E_{\tilde{z}}$. Another possible casting rule has been investigated: $W_k = R_k + z_k$, which is characterized by a simple additive process. Unfortunately, this possible rule suffers particularly of the problem of magnitude clipping, i.e., the crop of possible negative amplitudes, due to the superimposition of negative watermark values. In particular the negative clipping destroys both the audibility and the watermarking detection capability, so this technique has not further investigated. Finally, let us observe that, since $R_k$ and $z_k$ are independent and the sequence $z_k$ has null mean, it follows that:

$$\sum_k z_k R_k \approx 0 \quad \text{or} \quad \sum_k z_k \ln R_k \approx 0 \quad (1)$$

### 4 Watermark Detection

The classical watermarking detection system, independently on the domain in which the watermarking is applied, is based on a correlation scheme. Thus, the trivial way to implement a detection consists to perform a simple correlation between the watermarking sequence and the investigated audio stream. By implementing this strategy the correlation is:

$$\sum_k \tilde{z}_k W_k = \sum_k z_k R_k + \sum_k z^2_k R_k \approx \sum_k z^2_k R_k$$

where the approximation follows the observation in (1) and where $k = 0, \ldots, NL/2 - 1$ (i.e., correlation is computed on each block).

Observe that $y_k$ and the sample $R_k$ are independent each other and that this property holds also when the $z_k$ sequence is considered, due to the independence of the absolute threshold curve from $R_k$. With this, the previous correlation should be approximated to $\sum_k \tilde{z}_k^2 \sum_i R_i$. Even if the term $\sum_k \tilde{z}_k^2$ is well known, the term $\sum_i R_i$ depends on the particular audio stream. So, it should be measured or approximatively evaluated at the receiver from the watermarking stream, with some possible errors.
practice, the optimal detection should be based on the knowl-
dge of the original audio stream.

Another problem with this approach is related to the potential high
dynamic of \( R_k \); in the correlation computation the terms
with higher energy are strongly considered with respect to the
others, so the watermarking sequence statistical properties are
broken by the particular \( R_k \) stream.

To reduce the two above cited problems, we propose to use a
logarithmic correlation index, as follows:

\[ C = \frac{1}{E_0} \sum_k z_k \ln W_k \]

By observing that we have chosen \( E_0 \) such that \( z_k \ll 1 \) and re-
calling (1), it is possible to consider the approximation \( \ln W_k =
\ln[R_k(1 + z_k)] = \ln R_k + \ln(1 + z_k) \approx \ln R_k + z_k \) that gives
\( C \approx 1 \) if the watermarking is present and it is vanishing oth-
erwise. In such a way we have both neglected the problem of the
knowledge of the original audio track and of the potential high
dynamic of \( R_k \). Note that the variances of the auto- and cross-
correlations decrease, independently on the audio stream con-
sidered, by increasing the watermarking sequence length. So,
to enhance this effect we have averaged the correlation on a certain
number of available blocks, \( N_B \), each of dimension \( N L/2 \), by
obtaining \( \overline{C} \).

Regarding the watermarking robustness to the MPEG audio
compression, let us observe that tones having the greater SMR
in a sub-block have allocated for their representation the largest
number of bits. So these tones have the higher probability to
be not corrupted by the MPEG compression. Thus, in order to
improve the watermarking robustness, the detection may take in-
account only sets of tones, in each sub-block, having the greater
SMR. In particular, to identify the fraction of tones se-
lected for the detection with respect to the total, a parameter \( P \),
representing the ratio between the selected tones and the number
of available tones \( L/2 \), has been introduced. When \( P = 1 \) the
psychoacoustic model is not considered. By considering \( P < 1 \)
and by decreasing \( P \) the robustness against the MPEG compres-
sion increases, i.e., the mean value of the correlations do not con-
siderably change with respect the case without attack, even if,
since the number of watermarking sequence samples is lower,
the variance increase (note that this trend can be neglected by
increase the number of blocks \( N_B \)).

Furthermore, if we consider \( P < 1 \), an intrinsic robustness to
a band-pass filter manipulation is acquired, due to that the de-
tection is performed only by considering components located at
medium frequencies, which are the more audible, following the
psychoacoustic model. Note that, in order to jointly threat the
cases \( P = 1 \) (without tone selection) and \( P < 1 \) (with tone
selection), the indexes of the sums present in the various mathe-
matical expressions are not explicitly specified.

To take a decision regarding the presence of the watermarking,
the correlation \( \overline{C} \) must be compared with a threshold, in order to
verify if the watermark is present \( (\overline{C} \approx 1) \) or not \( (\overline{C} \approx 0) \).
Let us note that \( \overline{C} \) should be null both in the case of absence of the
watermark selected to prove the ownership and in the case in which
the watermark selected is present but we prove to detect a
watermark not correct. The detection threshold should be in
the interval \([0,1]\), and its choice is depending on how the auto-
and cross- correlations are distributed around 1 and 0, respec-
tively. In particular this choice should minimize the probabilities
to have false alarm (i.e., in the audio track a wrong watermark
is detected) and false rejection (i.e., in the audio track the se-
lected watermark is not detected). In general, the pdf of the auto-
and cross- correlations show a Gaussian shape with mean values
around 1 and 0, respectively and variances depending on the wa-
termarking sequence statistical properties and on the audio track
characteristics.

Note that many types of manipulations, such as compression,
filtering, re-sampling or re-quantization, act shifting versus low
values the auto-correlation pdf and its expected value, and by
increasing the relative variance. Also the cross-correlation pdf
shows a slight shift versus left values and a variance increase. In
any way, the auto- and cross- correlation variances increase can be
reduced by increasing the watermarking sequence length (in
our case \( N_B \)).

By calling \( \hat{z}_k \) the watermarking sequence used for the detec-
tion, it follows that:

\[ \overline{C} = \frac{1}{N_B} \sum_{i=0}^{N_B-1} C_i = \frac{1}{N_B E_0} \sum_{i=0}^{N_B-1} \left( \sum_k \hat{z}_k \ln R_k + \sum_k \hat{z}_k z_k \right) \]

where \( C_i \) is the correlation on the block \( i \). If \( \hat{z}_k = z_k \) we expect
\( \overline{C} \approx 1 \), if \( \hat{z}_k \neq z_k \) \( \overline{C} \approx 0 \).

The knowledge of the auto- and cross- correlation pdf Gaus-
sian shape, with means and variances, permits to select the opti-
mal threshold value [3]. In particular, if we assume that the
Gaussian pdfs have means \( m_a \) and \( m_c \) and variances \( v_a \) and \( v_c \)
for cross- and auto- correlations, respectively, the false alarm and
false rejection probabilities, \( P_{fa} \) and \( P_{fr} \), respectively, result:

\[ P_{fa}(m_c,v_c,th) = \frac{1}{2 \sqrt{v_c}} \exp \left( \frac{th-m_c}{\sqrt{2v_c}} \right) \]

\[ P_{fr}(m_a,v_a,th) = \frac{1}{2 \sqrt{v_a}} \exp \left( \frac{m_a-th}{\sqrt{2v_a}} \right) \]

where \( th \) is the decision threshold. The optimal threshold can be
obtained by considering \( P_{fa} = P_{fr} \), having specified means
and variances.

Unfortunately, means and variances of the auto- and cross-
correlation pdfs are changing with the watermarking sequences,
with the audio stream considered and with the particular digit-
al manipulation applied to the audio file, while the threshold
\( th \) must be fixed a priori. So, \( th \) is in general fixed by con-
sidering only one watermarking sequence, a given audio stream
and without consider any attack [3]. To avoid dramatic effects
when attacks are present or when a detection with different se-
quences is tried, \( th \) should be selected by considering all possible
watermarking sequences and all possible digital manipulations,
independently from the audio sample. Note that experimental
tests have shown that the attacks act in a relevant way espe-
cially on the auto-correlation, while the cross- correlation is mi-
nus affected; its mean remains close to 0 and its variance does
not vary if the watermark sequence is applied to a long audio
stream. So, due to the low variability of the cross-correlation pdf,
a possible solution to make the \( th \) selection is to fix it more
close as possible to 0 than 1 (to counteract the shift of the auto-
correlation), having evaluated the average mean and variance of
the cross-correlation pdf in different cases, and having fixed a
given \( P_{fa} \). In particular, we estimate the average mean and vari-
ance of the cross-correlation pdf using a given set of watermark-
ing sequences (e.g., generated by a selected chaotic map with a
given number of digits to specify the seed), considering an ex-
tended but limited set of audio samples, and by considering a
set of attacks. On the basis of this average mean and variance
(depending on the watermark length), we set the \( P_{fa} \) and
obtain the relative \( th \). Having fixed \( th \), we can proceed to verify a
posteriori the actual values for \( P_{fa} \) and \( P_{fr} \).
Let us note that some investigations on the selected class of maps give value for the term $1/E_0 \sum_k \hat{s}_k \hat{s}_k$, into the cross-correlation formula (2), ranging around 0.1. Furthermore, by considering an extended set of audio samples and by selecting different kind of maps, we have found a bound value (in the worst case) for the $1/E_0 \sum_k \hat{s}_k \ln R_k$ ranging around 0.25. With this, the minimum value of the threshold $\hat{th}$ for a good detection is 0.35.

In order to make a good choice for the $n$ and $t$ parameters characterizing our chaotic maps, we should calculate the correlation between the selected sequence (correct sequence) and any other possible watermarking sequence (attack sequence). This procedure should be iterated on all possible correct sequences we consider. Then, the optimal $(n, t)$ couple is that minimizing the maximum correlation in the whole set calculated. Note that if we consider $n \in \{2, \ldots, 100\}$, $t \in \{1, \ldots, n/2\}$, and 1000 seeds uniformly distributed in $[0, 1]$, about 200 days of processing should be necessary (with the hardware system available). Consequently, we have performed a limited set of measurements and decided to use the couple $n = 11$, $t = 1$, that result to be a good choice compared with others investigated.

The proposed watermarking scheme support multiple watermark additions. Let us call $z_k^{(i)}$ the $i$-th watermarking sequence to be applied, then the watermarked audio stream after the application of different watermarking sequences results: $W_k = R_k \prod_i (1 + z_k^{(i)})$.

When the generic watermarking sequence $j$ is used in the detection, the correspondent approximated correlation, before the average on $N_B$ blocks, is:

$$C \approx \frac{1}{E_0} \left[ \sum_k z_k^{(j)} \ln R_k + \sum_i \sum_k z_k^{(j)} z_k^{(i)} \right]$$

which is vanishing if the index $i$ never assumes the $j$ value, due to the observation in (1) and due to the low cross-correlation property of chaos based sequences, when the rate of mixing [7] is large enough, as in the case of the selected maps. The same considerations give that if $i = j$, $C$ is close to 1.

5 Watermarking Robustness

To test the reliability and robustness of the presented watermarking algorithm we have randomly chosen various combinations of sequence seeds and audio contents and considered different kind of attacks, as reported in the following. All audio files are 60s long, sampled at 44.1 kHz (CD quality), coded with 16 bits stereo and the shaping strength is fixed to $S = 0.6$, value guaranteeing almost totally inaudible distortions for every type of audio contents [4]. We have considered $N = 8$ and $L = 512$, so that in each block in the frequency domain we have $NL/2 = 2048$ samples.

Figure 3: Auto- and cross- correlations of a watermarked audio signal.

Figure 4: Experimental auto-correlation pdf in the case of MP3 attack, having considered $P = 1$, and Gaussian distribution.

Figure 5: Experimental auto-correlation pdf in the case of MP3 attack, having considered $P = 0.6$, and Gaussian distribution.

To verify the watermarking reliability we have selected a set of 1000 seeds, by fixing between them the correct one; then, we have embedded the correct seed in the audio stream; and, finally, we have evaluated the auto- and cross- correlations.

The results are shown in Figure 3, where the correlation $C$, before the average procedure on $N_B$ blocks, is reported as a function of the seed. Observe the correlation peak, close to 1, for the correct seed ($x_0 = 0.5$) and the low correlation values for all other possible choices. This result has been obtained for a particular audio stream but no difference has been observed by changing the audio content. Furthermore, to show the performance improvement by averaging $C$ on several blocks, we have considered 3 different $N_B$ values: 32 (3 seconds of the audio stream), 54 (5 seconds), and 108 (10 seconds). For the 3 cases we have considered the same threshold $\hat{th} = 0.36$. The corresponding $P_{fa}$, decreasing with the increase of $N_B$, are: $0.77 \cdot 10^{-3}$, $2.05 \cdot 10^{-5}$ and $3.08 \cdot 10^{-9}$, respectively. The $P_{fr}$ are lesser than $10^{-12}$.

The robustness against MPEG layer 3 (MP3) compression has been tested by using a compression degree generating tracks of 128kbps. In Figure 4 the auto-correlation pdf relative to a watermarked audio track compressed and decompressed is reported, having set $P = 1$, i.e., without consider the psychoacoustic model. In this case we have considered only one block $N_B = 1$ (by increasing $N_B$ the performance increase). With the aim of comparison, the Gaussian distribution, characterized by the same average and variance parameters, is plot. Let us underline that the experimental pdf is very close to the Gaussian one, as previously discussed. Furthermore, the auto-correlation mean (0.45) is far from the expected value 1, implying a high probability of false rejection; the variance is 0.023. In this case, having fixed $\hat{th} = 0.36$, and with $N_B = 32$, 54, 108 we have $P_{fa} = 0.70 \cdot 10^{-3}$, $2.00 \cdot 10^{-5}$ and $3.00 \cdot 10^{-9}$, respectively, and $P_{fr} = 0.57 \cdot 10^{-5}$, $P_{fr} = 1.25 \cdot 10^{-5}$ and $P_{fr} = 1.20 \cdot 10^{-9}$. In Figure 5 the same study of Figure 4 (with $N_B = 1$) is performed, but considering the psychoacoustic model and fixing $P = 0.6$. The experimental pdf remains close to the Gaussian one (with same parameters), while the mean value is 1.1, implying a low probability of false rejection, even if the variance is higher (0.14). The variance decreases by increasing $N_B$. In this case, having fixed $\hat{th} = 0.36$ and by considering $N_B = 32$, 54, 108 the $P_{fa}$ are of the same order of the case with $P = 1$ but the $P_{fr}$ are lesser than $10^{-12}$, so a relevant performance improvement is present. All results show that, when the
A new watermarking scheme applied to digital audio streams, with the scope of copyright protection, has been proposed and tested. The embedding of the digital signature into the audio file has been performed in the frequency domain by integrating a shaping function derived from MPEG psychoacoustic model I, whose effect is to enhance the reliability of the watermarking detection, without degrading the audio quality. The watermarking signals have been generated by considering a class of PWAM chaotic maps, guaranteeing a high number of distinguishable watermarks.

The proposed watermarking technique, thanks to the psychoacoustic model I, shows high robustness against MP3 compression. Furthermore, this scheme guarantees high reliability in case of digital manipulations like re-sampling, re-quantization, band-pass filtering and cropping. Finally, this technique supports multiple watermarking and does not require the original signal during detection.

References