International Journal of Computer Science and Applications,
©Technomathematics Research Foundation
Vol. 15, No. 1, pp. 1 – 15, 2018

AUDIO WATERMARKING SYSTEM RESISTANT TO REMOVAL ATTACKS BY DEREVERBERATION

VALERY KORZHIK, VASILY ALEKSEEV

The Bonch-Bruevich Saint-Petersburg State University of Telecommunication, Saint Petersburg, Russia
val-korzhik@yandex.ru , i@vasay.ru

GUILLERMO MORALES-LUNA

Computer Science, CINVESTAV-IPN, Mexico City, Mexico
gmorales@cs.cinvestav.mx
http://delta.cs.cinvestav.mx/~gmorales/

We consider a digital audio watermarking system based on reverberation. Extraction procedure uses a correlation receiver in cepstral area. Blind dereverberation attack as the most powerful method to remove the embedding information is presented both theoretically and experimentally including also addition noise and desynchronization attack. For a protection against such attacks are used the salient points and an addition of short pulses placed in appropriated time interval.

Our experiments with the modificated embedding algorithm showed that a quality of cover audio signal occurs good and desynchronization attack is useless.

Application of MP3 compression to WM-ed audio files results to something increasing of the error bit rates but they still occur acceptable and can be decreased by the use of error correction codes.

Keywords: Audio watermarking; reverberation ; cepstrum ; salient points ; dereverberation attack.

1. Introduction

It is well known that a technology of digital watermarking (WM) is the most effective approach to provide copyright protection for digital media products. Examples of such products are digital audio and video works. In the current paper we consider audio works (first of all musical files presented as digital signals in format `wav`). Such objects in which it is necessary to embed an additional information are called in this paper in the sequel as*cover object* (*CO*). Dishonest users (pirates in another words) can try to remove the embedded WM without remarkable corruption of CO in hope of their illegal redistribution to other users. They could reach the desired result after some processing of the watermarked objects in such a way that legal users were unable to extract theWM correctly from redistributed copies and as a consequence theywill be unable to arrange a forensic consideration against pirates.

Owners of the products can try, on the contrary, to embed in CO such WM that cannot be removed without significant corruption of CO. On the other hand a significant corruption of CO results in their lower values at the market and a redistribution occurs useless.

Several well known embedding WM techniques have appeared for audio-signals, for the thing,*phase-shift-keying* (PSK) modulation [1] or WM system based on *echo hiding* (EH) (see [2]). But as it was shown in [2] and [3] both PSK and EH WM systems can easily be removed without significant degradation of CO.

The use of spread spectrum signals in the embedding procedures that are controlled by secret *stegokeys* seems to be very attractive. But more carefully consideration [4] shows that such signals suffer on a breaking by desynchronization attack.

At a single glance, the use of reverberation procedure with a secure pulse response of the reverberation filter, controlled by astegokey, is the best approach. In fact, on the one hand the use of a reverberation with filter pulse response close to *a room pulse response filter* provides a good quality of audio CO [4]. On the other hand, the use of complex pulse response form prevents a compensation of reverberation (making a dereverberation – in other words) that could be allow to remove the embedding.

But unfortunately, a changing of pulse response form on every bit interval results (as our experiments showed) in a significant corruption of CO. Therefore we propose some "intermediate" approach that is presented in Section 2. and 3.

But without some additional transforms that are discribed in Section 4. the WM system presented in Sections2. and 3. will yet suffer from blind dereverberation attack described in those sections also. The proposed modified WM system is presented in Section 4. Section 5. concludes the paper and presents some open problems for the future work.

2. Attack on WM system that is based on the embedding with a reverberation usage

Let us assume that the WM system uses some fixed (but sufficiently complex) reverberation *filter pulse response*$(h_{nb})_{n\in\{1,\ldots,N\}}$, $b\in\{0,1\}$, for all watermarking session, where *N* is the number of samples on every bit interval. In order to embed bits *b=0* or *b=1* it is used only fixed but different time delays with each filter corresponding to additional information.

Then digital WM-ed signal *Z(n)* on each bit interval can be presented as follows:

$$\forall n\in\{1,\ldots,N\}, b\in\{0,1\}\text{:}\ Z(n) = S(n) * h_{nb} \qquad (1)$$

where $(S_n)_{n\in\{1,\ldots,N\}}$is the input audio signal (CO),$(h_{nb})_{n\in\{1,\ldots,N\}}$ is thefilter pulse response depending on the embedding bit *b*,"*" is theoperation of convolution, and*N* is the number of samples on each bit interval.

Applying cepstrum transform to both sides of (1) we get

$$\forall n\in\{1,\ldots,N\}, b\in\{0,1\}\text{:}\ \tilde{Z}(n) = \tilde{S}(n) + \tilde{h}_{nb} \qquad (2)$$

where "~" means cepstrum transform, calculated as follows [5]:

$$C(x) = \tilde{x}(n) = \tfrac{1}{N}\textstyle\sum_{k=0}^{N-1}(\log(x'(k) + j\,\Theta(k))e^{\frac{2\pi jnk}{N}} \qquad (3)$$

where

$$x'(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\, e^{\frac{2\pi jnk}{N}}$$

and $x'(k)_{k\in\{0,\ldots,N-1\}}$is the signal amplitude, and $\Theta(k)_{k\in\{0,\ldots,N-1\}}$ is the signal phase.

In reality, relation (2) is only an approximation for a finite signal. The accuracy of expression (2) depends on the number of zeros added tothe finite signal. If the number of added zeros is sufficiently large, then relation (2) holds with small errors. The advantage of presentation (2) against the presentation (1) consists in a fact that such transform allows to use awell known algorithm of the optimal receiver [6] if the interference$(S_n)_{n\in\{1,\ldots,N\}}$ can be approximated by awhite Gaussian noise.

Then extraction algorithm for such WM system will be the well known *correlation receiver*

$$b = \underset{b\in\{0,1\}}{\text{Arg max}} \sum_{n=1}^{N} \tilde{Z}(n)\widetilde{h_b}(n) \tag{4}$$

We can see from equation (4) that it is necessary to know exactly the initial and the end points of every bit interval in order to extract the embedded information correctly. If the special synchronizing signals are inserted into audio signal it results firstly in distortions to the original signal and secondly they can be easily removed by an attackers. In order to avoid of these defects there was used a method proposed in [7] that executed with a notion of so called *salient points* (SP). In this case we do insert nothing special synchronizing signals but extract SP from the raw audio via content analysis.

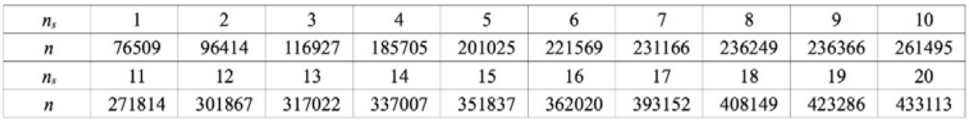
In line with [7] theextraction of salient points is provided by the following 7-th step procedures:

1. The audio signal is filtered to remove low and high frequency components to which human ears are not sensitive.
2. For each samples $Z(n)$, the total energy of r samples before $Z(n)$ and r samples after $Z(n)$ are calculated separately:
$$E_{before}(n) = \sum_{i=-r}^{-1} Z^2(n+i) \text{ and} E_{after}(n) = \sum_{i=0}^{r-1} Z^2(n+i).$$
3. The ratio of these two energy values is calculated for each $Z(n)$, $n = 1,\ldots, N$
$$ratio(n) = \frac{E_{after}(n)}{E_{before}(n)}$$
4. If $ratio(n)>T_1$ and $E_{after}(n)>T_2$, where T_1 and T_2 are some chosen thresholds, then $Z(n)$ is labeled as an energy fast-climbing point.
5. The energy of fast-climbing point appears in groups usually. Points that are separated by less than T_3 samples (where T_3 is some chosen integer threshold) are merged into one larger group.
6. Within each group, the sample with the largest $ratio(n)$ is labeled as one salient point.
7. If the salient point was derived from a group where the largest $ratio(n)$ times the number of samples in the group is less than some chosen threshold T_4 , then such salient point is deleted.

Next the embedding procedure on each bit interval of the length N samples is performed just after extraction of salient point except the case when the embedding of several bits can be done one by one in consecutive manner.

At the receiving side salient points are detected following the steps 1-7 above and hence a synchronizing of bit interval can be easily recovered even after desynchronization attack.

Table 1. The number of samples *n* corresponding to the number of salient points n_s.



n_s	1	2	3	4	5	6	7	8	9	10
n	76509	96414	116927	185705	201025	221569	231166	236249	236366	261495
n_s	11	12	13	14	15	16	17	18	19	20
n	271814	301867	317022	337007	351837	362020	393152	408149	423286	433113

In Table 1 thereare presented the numbers of samples *n* corresponding to the numbers n_s of silent points. We obtain it by simulation of typical audio files and following to the procedure 1-7 above.

We can see that salient points are found sufficiently frequently. This means that desynchronization attac can be practically cancel.

Let us assume that an attacker tries to remove WM and is able to estimate somehow filter cepstrum pulse responses for b=0and for b=1 as $\widetilde{h_0{}'(n)}$and$\widetilde{h_1{}'(n)}$ on each of bits intervals.

Then theattack intended to remove theWM could be

$$\widetilde{Z_a}(n) = C^{-1}\left(\tilde{Z}(n) - \widetilde{h_b}'(n)\right) \tag{5}$$

where $C^{-1}$is inverse to cepstrum transform *C* in (3).(We do not consider the hardness to perform transform C^{-1} in a favor of the attacker.)

It is worth to note that an operation to remove a reverberation from audio signal is called as *blind dereverberation*. Such problem was investigated in many papers, e.g.[8],[9],[10] and others. But the goal of such signal transforms is to make audio signal free from additional reverberation interference that may occur by natural manner.

In our case it is not enough to make audio signal free from reverberation “by ear”. We require to make impossible to extract WM from dereverberated signal even with the use at optimal receiver. Moreover in paper [10] for a dereverberation removal were used multiple microphones placed on some distances one against another. Of course such approach cannot be used in our scenario.

Let us estimate the probability of error P (incorrect bit b extraction) for WM system owner using the decision rule (4) after dereverberation attack:

$$\forall n \in \{1, \ldots, N\}: \quad \widetilde{Z_a}(n) = \tilde{Z}(n) - \widetilde{h_b}'(n).$$

It is easy to see from (2), (4) and (5) that even for opposite signals$\widetilde{h_0{}'(n)}$and $\widetilde{h_1{}'(n)}$:

$$P = \Pr(1|0) = \Pr\left\{\xi \leq -\sum\nolimits_n \left(\widetilde{h_0}(k) - \widetilde{h_0}'(k)\right)\widetilde{h_0}(k)\right\} \tag{6}$$

After a changing of variables we get from (6)

$$P = \frac{1}{\sqrt{2\pi\sigma^2 A}} \int_{-\infty}^{\tilde{A}} \exp\left(-\frac{x^2}{2\sigma^2 A}\right) dx \tag{7}$$

where

$$\tilde{A} = \sum_n \left(\widetilde{h_0}(n) - \widetilde{h_0}'(n)\right)\widetilde{h_0}(n) \quad , \quad A = \sum_n \widetilde{h_0}^2(n) \,\text{and}\sigma^2 = \mathrm{Var}\big(\tilde{S}\big).$$

(We note that relation (7) is true if a random variableξ is a zero mean Gaussian sequence with variance$\sigma^2 A$.)It is easy to prove that

$$\tilde{A} = \sum_n \left(\widetilde{h_0}(k) - \widetilde{h_0}'(k)\right)\widetilde{h_0}(k) = A(1-\eta) \qquad (8)$$

where

$$\eta = \frac{1}{A}\sum_n \widetilde{h_0}'(n)\widetilde{h_0}(n).$$

Substituting (8) into (7) we get after simple transform

$$P = 1 - F\left(\sqrt{\frac{A(1-\eta)}{\sigma^2}}\right) \qquad (9)$$

where

$$F\colon x \longmapsto F(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) dt.$$

(If thesignals $\widetilde{h_0'(n)}$ and $\widetilde{h_1'(n)}$are not opposite then the equality (9) holds as upper bound, in favor of attacker.)We see from (9) that if η=0, that is, an estimation $\widetilde{h_0'(n)}$is bad, then theattack occurs inefficiently. But if η=1 it results in *p*=1/2, which means a "*break of the legal WM channel*". Then theestimation attack is effective because it removes the WM embedding completely.

In Fig 1. there are presented the dependencies of the legal user bit error probability calculated by (9) against ofparameter η for different parameters A/σ^2 .

We can see from dependences presented in Fig. 1 that in order to provide high efficiency of attack it is necessary to get parameter η close to the value 0.8. Hence an attacker should correctly estimate filter pulse responses of legal user. We note first all that such problem cannot be solved by total exhaustion of all possible filter pulse response wave forms. In fact, thetypical length of "room pulse" that keeps a good quality of musical file after embedding is about 180 samples. Assuming that pulse response amplitude is at most about 0.2 with respect to audio signal amplitude, we get for a total numbers of quantization levels 65536 for format wav, the number of acceptable levels for pulse response will be about 13107. It results in a set of all possible pulse response wave forms about $1.4*10^{741}$, that is untractable value for exhaustion attack.

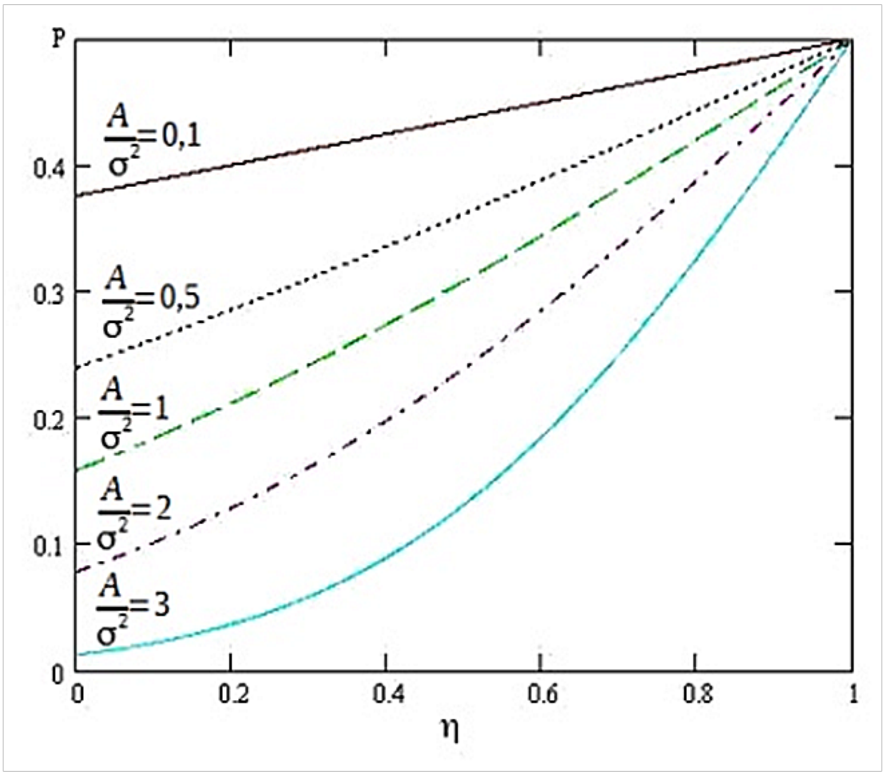


Fig 1. The dependences ofthe legal user bits error probabilities against of η for different A/σ^2

If we assume that filter pulse response$\widetilde{h_0(n)}$and $\widetilde{h_1(n)}$differ only by a fixed and known delay N_0 then the attacker could find both bit intervals I_0 corresponding to b=0 and I_1 corresponding to b=1. Next it is possible to average separately all cepstrum corresponding wave forms in order to get an approximation of cepstrum pulse response as follows:

$$\begin{aligned}\widetilde{h_0(n)} &\sim \frac{1}{L}\left[\sum_{i\in I_0}\tilde{Z}(n)+\sum_{i\in I_1}T_{N_0}\big(\tilde{Z}(n)\big)\right] \\ &= \frac{1}{L}\sum_{i\in I_0}\widetilde{S_\iota}(n)+\widetilde{h_0(n)}\end{aligned} \tag{10}$$

Using the last relation (10) and definition of η in (8) we get

$$\eta = 1-\frac{\sum_n \widetilde{h_0}(n)\frac{1}{L}\sum_{i\in\{1,\dots,L\}}\widetilde{S_\iota}(n)}{\sum_n \widetilde{h_0}^2(n)} = 1-\varepsilon \tag{11}$$

Let us find the variance of random value ε assuming that $\mathrm{Var}\big(\widetilde{S_i}(n)\big)=\sigma^2$and that samples of cepstrum $\widetilde{S_\iota}(n)$are *i.i.d.* random values. Then we can write

$$\begin{aligned} \mathrm{Var}(\varepsilon) &= \frac{\mathrm{Var}\left(\sum_n \widetilde{h_0}^2(n)\frac{1}{L}\sum_{i\in\{1,\dots,L\}}\widetilde{S_i}(n)\right)}{\sum_n \widetilde{h_0}^2(n)} \\ &= \frac{\sigma^2}{L\sum_n \widetilde{h_0}^2(n)} \\ &= \frac{\sigma^2}{LA} \end{aligned} \tag{12}$$

Next we can use the relation (12) for known cepstrum pulse response$\widetilde{h_0(n)}$ and known parameters L andσ^2in order to get that the parameter $\eta$is at most$3\sqrt{\mathrm{Var}(\varepsilon)}$ with probability 0.997.

Example:

Let us assume$\frac{A}{\sigma^2} = 0.5$, L=360, then we get by (12),$\mathrm{Var}(\varepsilon) < 0.0006$ and the parameter η is at least 0.93. with probability 0.997. Then we can see from Fig.1 that for the estimation attack presented above the extracted bit error for legal user occurs close to 0.5 and hence this attack be very effective.

But one gap in the description of the estimation attack is the fact that so far it is unknown how an attacker could be able to find all bit intervals belonging separately to embedding of bits.

Since the forms of filter pulse responses are constant for different bit interval (in line with our previous assumption) and differ only in a fixed delay, it results in the same condition for corresponding cepstrums. Thus we get that if for a pair of bit intervals i and j corresponding to them bits$b$and$\tilde{b}$are equal to one another that is$b = \tilde{b}$we get the following crosscorrelation for corresponding cepstrum wave form$\tilde{Z}_i(n)$and$\tilde{Z}_j(n)$:

$$\begin{aligned} \Lambda &= \frac{1}{N}\sum_{n=1}^{N}\tilde{Z}_i(n)\tilde{Z}_j(n) \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(\tilde{S}_i(n) + \widetilde{h_b}(n)\right)\left(\tilde{S}_j(n) + \widetilde{h_b}(n)\right) \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(\tilde{S}_i(n)\tilde{S}_j(n) + \tilde{S}_i(n)\widetilde{h_b}(n) + \widetilde{h_b}(n)\tilde{S}_j(n) + \widetilde{h_b}(n)^2\right) \end{aligned} \tag{13}$$

where L is the total number of bit intervals.

For the case of different embedding on i-th and j-th bit intervals, that is$b \neq \tilde{b}$, we get

$$\begin{aligned}\Lambda' &= \frac{1}{N}\sum_{n=1}^{N}\tilde{Z}_i(n)\tilde{Z}_j(n)\\ &= \frac{1}{N}\sum_{n=1}^{N}\left(\tilde{S}_i(n)+\widetilde{h_b}(n)\right)\left(\tilde{S}_j(n)+\widetilde{h_{\tilde{b}}}(n)\right)\\ &= \frac{1}{N}\sum_{n=1}^{N}\left(\tilde{S}_i(n)\tilde{S}_j(n)+\tilde{S}_i(n)\widetilde{h_{\tilde{b}}}(n)+\widetilde{h_b}(n)\tilde{S}_j(n)+\widetilde{h_b}(n)\widetilde{h_{\tilde{b}}}(n)\right)\end{aligned} \tag{14}$$

Comparing equations (13) and (14) we can conclude that in the first caseΛ is more on average than in the second case.

Therefore we can select a threshold and take a decision that *i*-th and *j*-th interval belong to the bit intervals$b = \tilde{b}$ if the threshold is exceeded and otherwise to different bit intervals $b \neq \tilde{b}$. Thus it is possible to find the sets $I_0$and $I_1$for a calculation in (10)

However the last question arises – how can an attacker find filter pulse response but not filtrum cepstrum pulse response by (10)? It has been proved in [11] that for small embedding amplitude it is possible to take into account only the first term in Tailor series for thecepstrum expansion of signal in (2). This means that the last equation can be rewritten as

$$\tilde{Z}(n) = \tilde{S}(n) + \lambda h_b{}'(n) \tag{15}$$

where λ is some scale coefficient, $h_b{}'(n)$ is thealready filter pulse response but not cepstrum pulse response.

Expression (15) asserts that if an attacker has estimation of cepstrum pulse response correctly he (she) wouldbe able to find pulse response after a specification (may be even by exhaustive trial) of the coefficient λ.

After all calculation with filter pulse responses an attacker, with the knowledge of bits embedding an each bit interval, be manage to apply inverse filter pulse response and as a consequence to remove all embedded information.

However, in a theoretical investigation above were suggested some model for cover objects that can be not exactly valid in practice. Therefore in the next Section we investigate the proposed attack experimentally. In section 4. we modify the embedding scheme in such a way to be resistant against a dereverberation attack.

3. Experimental investigation of the proposed dereverberation attack

We select the filter pulse response (FPR) for both embedding bits *b*=0 and *b*=1 shown in Fig. 2. Delays chosen for embedding are 30 and 25 samples for bits zero and one, respectively. Cepstrums of these FPR are shown in Fig. 3. These figures confirm thestatements given before that firstly cepstrum delays coincide with FPR delays and secondly, that cepstrum wave forms copy FPR wave forms.

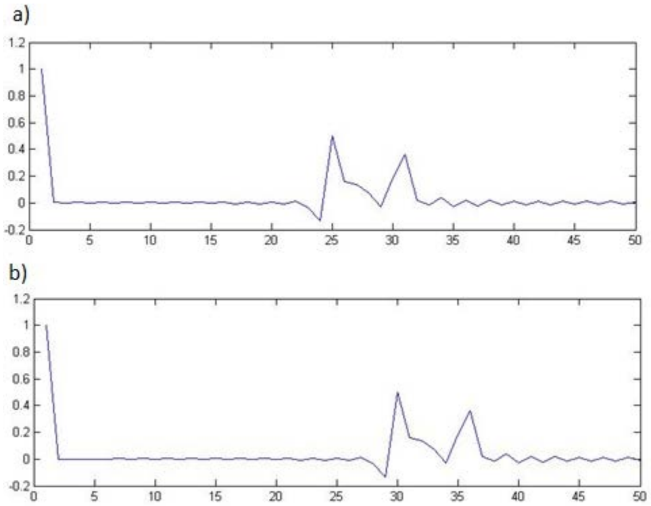


Fig. 2 Filter pulse responses: a) for bit “1”, b) for bit “0”

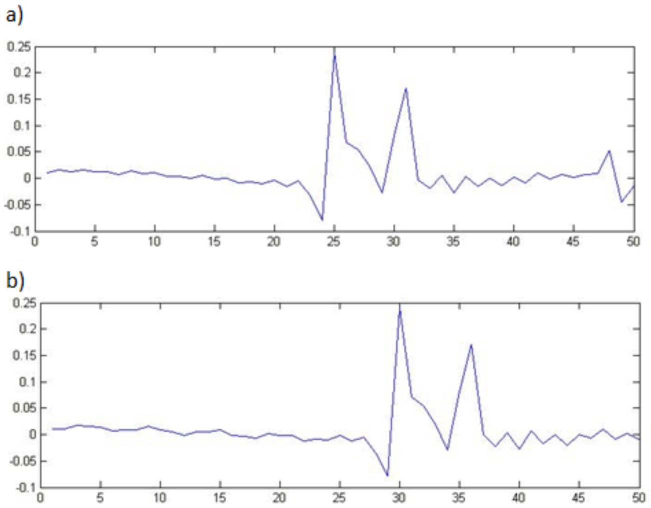


Fig. 3 Filter cepstrum pulse responses: a) for bit “1”, b) for bit “0”

They were found all bit intervals corresponding to bit b=0 and b=1 with the use of crosscorrelation Λ, and Λ' by eq. (13), (14) respectively.

In Fig. 4 it is presented the FCPR averaged in line with formula (10). We can see from this figure that a form of FCPR copies a form of FPR with accuracy of some scale factor.

Our experiment showed that even 3 bit intervals are sufficiently in order to estimate FPR correctly. But on the other hand 3-4 bits is too small values for a useful WM embedding.

If an attacker is able to find the scale factor then the wave form of FPR can be easily estimated (see Fig. 5).

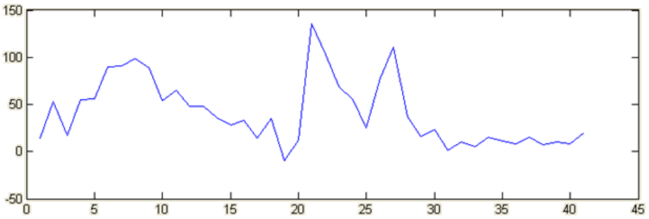


Fig. 4 Averaged FCPR in line with eq. (10)

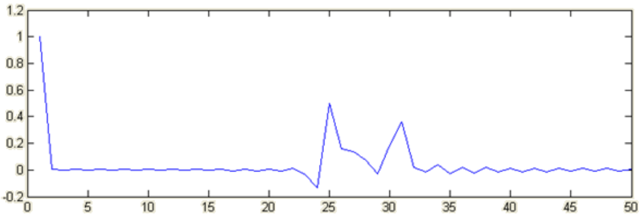


Fig. 5 Estimation of FPR after a selection of scale factor.

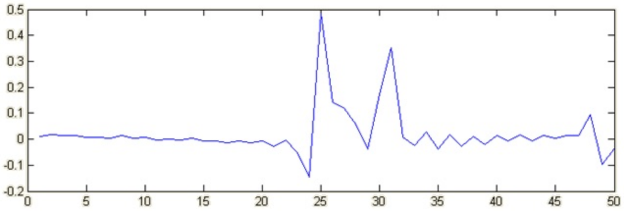


Fig. 6 FCPR calculated from FPR (Fig. 5)

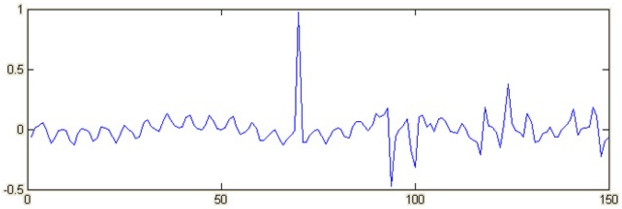


Fig. 7 FPR for dereverberation attack.

Next, the dereverberation attack can be performed in the following steps:

1. For known FPR (Fig. 5) calculate FCPR (see Fig. 6)
2. Inverse with respect to zero the wave form of FCPR
3. Find FPR for the attack filter computing inverse cepstrum transform from FPR. The result is presented in Fig. 7

(In a similar manner inverse FPR can be calculated for the embedding of bit *b*=1).

4. Apply inverse filters to the embedded bits “0” and “1” which has been found before in corresponding bit intervals.
5. Use a transition function between bit intervals with linear form that is necessary to keep high quality of audio signal after dereverberation procedure.

In Table 2 there are presented the extracted bit error probabilities before and after dereverberation attack under different parameters of WM system. The wave forms of FPR were presented in Fig 2. They have finite length equal to 180 samples.

We can see from this table that before attack the proposed WM system is working acceptably but after dereverberation attack the bit error probability is close to 50%, that is similar to “break of channel”. (We note that a fact that sometimes the probability exceeded 50% owing incorrect estimation of scale factor. But it does not affect on a final conclusion).

We note also that as it was showed before the dereverberation attack cannot be simplified if we short reasonably the number of the embedded bits.

4. Modification of WM system to be resistant against a dereverberation attack

In order to protect WM system against the proposed above dereverberation attack it is necessary to provide an impossibility for an attacker to separate 0-bit intervals from 1-bit intervals.

In fact, if an attacker does not know which of bit intervals correspond to embedding bit “1” and which ones to the bit “0”, then replacing expression (10) to a summation over all bit intervals as follows

$$\Lambda'' = \frac{1}{L}\sum_n \widetilde{Z(n)} \tag{16}$$

we get a large corruption of FPR wave form in comparison with original one.

Table 2. The extracted bit error probabilities before and after dereverberation attack for different system parameters and different audio files.

Name of music files and their duration	Delays of WM signal		The length of bit intervals (in number of samples)	The number of the embedded bits	Bit error rate before attack in %	Bit error rate after attack in%
	«1»	«0»				
Vysocki ”Song of Boxer” (fragment 20 sec)	25	29	4000	142	4,5%	72%
Vysocki ”Song of Boxer” (fragment 20 sec)	25	29	6000	94	0%	78%
Vysocki ”Song of Boxer” (fragment 20 sec)	15	19	6000	94	17%	63%
Yuta, «Jealosy» (fragment 29 sec)	25	29	10000	55	2%	48%
Yuta, «Jealosy» (fragment 29 sec)	25	29	5000	113	1%	57%
Yuta, «Jealosy» (fragment 29 sec)	20	24	5000	113	7%	61%

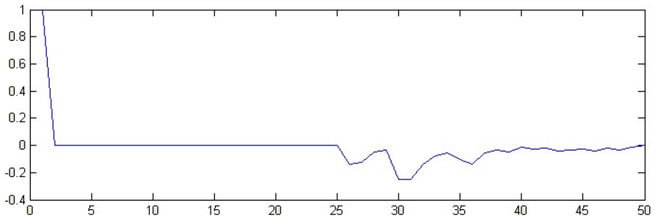


Fig. 8 FPR wave form obtaining by (16)

In Fig. 8 the FPR wave form is presented after such “total averaging”. We can see that the FPR in Fig. 8 has no similarity with the original FPR wave form (see Fig. 2) and hence an attacker will be unable to arrange a dereverberation attack (In fact we have checked that the use of such FPR in dereverberation attack cannot result even in a remarkable increasing of the extracted bit error probabilities for legal users).

In order to prevent a crosscorrelation attack (13), we propose to add to WM signal short pulses at the beginning of each of bit interval. (See Fig. 9 where additional pulse is presented on the 21-th samples of bit interval).

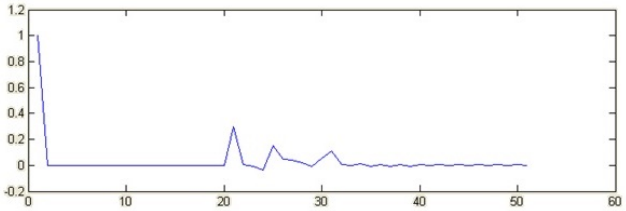


Fig. 9 Wave form of FPR with additional pulse on 21-th sample of bit interval.

Wave form of FCPR after such procedure is presented in Fig. 10

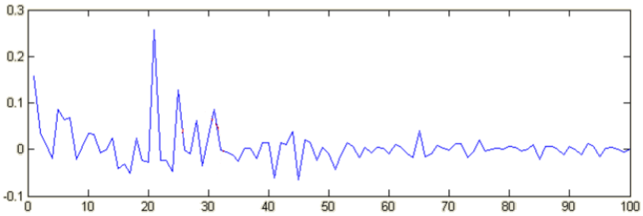


Fig. 10 Wave form of FCPR after insertion of short pulse.

We can see from this Figure that the wave form of FPR cannot be recognize there correctly as it was in reality (see Fig. 5). This fact can be explained by “a noising” of valid FCPR by cepstrum components of short pulses.

The use of crosscorrelation attack by (13), (14) results in an occurrence of single pulse independently on a coinciding or dis-coinciding of information bits corresponding to signal $\tilde{Z}_i(n)$ and $\tilde{Z}_j(n)$. (See Fig. 11 for a confirmation).

Table 3. The extracted bit error probabilities before and after mp3 attack for different system parameters

Name of music files and their duration	Delays of WM signal		The length of bit intervals (in number of samples)	The number of the embedded bits	Bit error rate before attack (in %)	Bit error rate after mp3 attack (in %)
	«1»	«0»				
Vysocki ”Song of Boxer” (fragment 20 sec)	25	29	4000	142	4,5%	5,5%
Vysocki ”Song of Boxer” (fragment 20 sec)	25	29	6000	94	0%	0%
Vysocki ”Song of Boxer” (fragment 20 sec)	15	19	6000	94	17%	23%
Yuta, «Jealosy» (fragment 29 sec)	25	29	10000	55	2%	3%
Yuta, «Jealosy» (fragment 29 sec)	25	29	5000	113	1%	1%
Yuta, «Jealosy» (fragment 29 sec)	20	24	5000	113	7%	15%

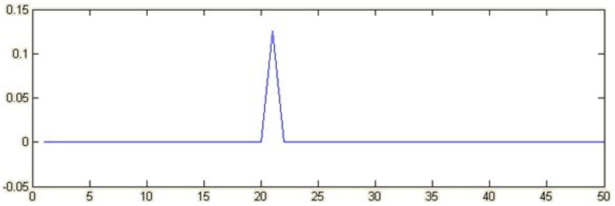


Fig. 11 Result of crosscorrelation computation with additional pulse on the 21-th sample

Thus we can conclude that a modification of reverberation-based WM system by additional pulses results a resistance of this system to a most power blind dereverberation attack.

We have tested the proposed WM system also with respect to audio signal quality after embedding. A group consisting of 5 experts has come into a conclusion that a quality of musical files after WM embedding keep practically the same as before embedding.

Another requirement to audio watermarking is its resistance against standard compression methods, likeMP3. MP3 is the most popular effective owing the following reason.
First of all it allows to short original audio file about 11 times.
In order to keep a good quality of audio signal MP3 algorithm uses spectral filtering in line with psychoacoustic model of human auditory system. Audio signal is divided into equal duration blocks each of them are packed in individual frame after processing. Spectral decomposition requires signal continuation that can be provided by joint processing of both previous and following frames. If in audio signal there are spectral components with small amplitude close to more intensive components, the first ones can be deleted owing of the property of frequency masking. It is possible also to replace two and more nearby peaks to one averaged peak. It is worth to nothing that on the contrary to JPEG compression high frequencies are not removed completely in MP3 algorithm but removed by selective manner in order to decrease information stream by spectral rarefaction. After spectral transform, it is used methods of source coding and a packing into frames. The compression degree can be vary in interval about 8-320 kbit/sec.
Justification of reverberation based WM embedding for MP3-wise file compression can be supported by fact that similar transform are very commonly for conventional music (not necessary WM-ed file) in order to provide a "decoration" of such musical files.
In Table 3 are presented bit error rates into extracted WM-ed information before and after MP3 compression of audio files with reverberation wise embedding.
We can see from this Table that the bit error rates slightly increase after MP3 compression but they are still acceptable and can be improved by error correction code application.

5. Conclusion

In this paper audio WM system resistant to a remove attack is proposed. Embedding of WM in this system is performed by a reverberation of audio signal that is controlled by secret stegokey. The main advantage of reverberation – based watermarking system is its possibility to provide a high quality of audio signal after embedding. But there exist effective attack on such WM system known as blind dereverberation attack. We investigated this attack in detail and showed that in fact it is able to remove the embedding information without significant degradation of audio signal quality. Therefore we propose some modification of WM-based system and show that then such attack is useless.
Experimental investigation confirm our conclusion. All experiments presented above were performed with FPR amplitude 0.15 that provides embeddingrateabout 20 bits/sec. Of course this value is much less than 333 bit/sec that provides PSK modulation [15] but it requires using cover audio signal for WM extraction. Another embedding method based on masking of WM into inaudible tonesprovides bit rate 250 bit/sec [15]. But then WM can be removed after ordinary compression. Method of WM insertion into speech pause can be used only for speech but not for musical files.

On the over hand the embedding rate 20 bit/sec is as a rule sufficient one in order to provide copyright protection with additional needed information. But if it is necessary to increase the number of embedded bits, or we have more short length of musical clip then it is possible at the cost of FPR amplitude increase to 0.3. Then there may be reached the embedding rate 122 bits/sec but unfortunately it results in a slightly corrupted audio signal.

References

[1] M. Arnold, P. G. Baum, and W. Voeßing, “Information hiding,”, S. Katzenbeisser and A.-R. Sadeghi, Eds. Berlin, Heidelberg:Springer-Verlag, 2009, ch. A Phase Modulation Audio WatermarkingTechnique, pp. 102–116, DOI: 10.1007/978-3-642-04431-1_8. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04431-1_8

[2] V. I. Korzhik, G. Morales-Luna, and I. Fedyanin, “Audio watermarking based on echo hiding with zero error probability.” International Journal of Computer Science and Applications vol. 10, no. 1, pp. 1–10, 2013.

[3] V. Alekseyev, A. Grudinin, and V. Korzhik, “Design of robust audiowatermark system,” in Proceedings of the XI International Symposium on Problems of Redundancy in Information and Control Systems , Aug 2007, pp. 163–165.

[4] H. Liu and W. Zhang, “Overview of audio watermarking algorithmagainst synchronization attacks,” in Advances in Intelligent Systems Research: ICAITA-16, Aug 2016, DOI: 10.2991/icaita-16.2016.52.

[5] J. M. Arend and C. Pörschmann, “Audio watermarking of binaural room impulse responses,” in Audio Engineering Society Conference: 2016 AES International Conference on Headphone Technology, Aug 2016, DOI: 10.17743/aesconf.2016.978-1-942220-09-1. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18346

[6] [6] J. Proakis, Digital Communications, Fourth Edition. Mc Graw Hill, 2001.

[7] C.-P. Wu, P.-C. Su, and C.-C. J. Kuo, “Robust Audio Watermarking for Copyright Protection,” SPIE's 44th Annual Meeting Advanced Signal Processing Algorithms, Architectures, and Implementations, July 18-23, 1999.

[8] D. G. Childers, D. P. Skinner, and R. C. Kemerait, “The cepstrum: A guide to processing,” Proceedings of the IEEE, vol. 65, pp. 1428–1443, 1977, DOI: 10.1109/PROC.1977.10747.

[9] T. Nakatani, M. Miyoshi, and K. Kinoshita, “One microphone blind dereverberation based on quasi-periodicity of speech signals,” in Advances in Neural Information Processing Systems 16, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2003, p. None. [Online]. Available: http://books.nips.cc/papers/files/nips16/NIPS2003_SP06.pdf

[10] C. Evers, “Blind dereverberation of speech from moving and stationary speakers using sequential Monte Carlo methods,” Ph.D. dissertation, The University of Edinburgh (United Kingdom, 2010.

[11] H. Attias, J. Platt, A. Acero, and L. Deng, “Speech denoising and dereverberation using probabilistic models,” November 2000. [On-line]. Available: https://www.microsoft.com/en-us/research/publication/speech-denoising-and-dereverberation-using-probabilistic-models/

[12] N. Cvejic and T. Seppanen, Digital Audio Watermarking Techniques and Technologies: Applications and Benchmarks. Hershey, PA, USA: IGI Global, 2007, DOI: 10.4018/978-1-59904-513-9.

[13] G. Chardon, T. Nowakowski, J. de Rosny, and L. Daudet, “A blind dere-verberation method for narrowband source localization,” IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 5, pp. 815–824, Aug 2015, DOI: 10.1109/JSTSP.2015.2422673.

[14] K. Imoto and N. Ono, “Spatial cepstrum as a spatial feature using a distributed microphone array for acoustic scene analysis,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1335–1343, 2017, DOI: 10.1109/TASLP.2017.269059.

[15] M. A. Nematollahi, C. Vorakulpipat, H. G. Rosales, “Digital Watermarking Techniques and Trends”, Springer, 2016, ISBN 978-981-10-2095-7