Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

A high-capacity watermarking technique for audio signals based on MDCT-domain quantization

Jonathan Pinel (1), Laurent Girin (1), Cléo Baras (1) and Mathieu Parvaix (1)

(1) Grenoble Lab. of Images, Speech, Signal, and Automation - GIPSA-lab, Grenoble, France

PACS: 43.75.Zz, 43.66.Dc, 43.60.Hj.

ABSTRACT

Watermarking is a technique that consists in hiding/embedding binary information within a signal in an imperceptibly way, meaning in the present context of audio signals that the mark is inaudible. Watermarking was first used for the protection of digital contents as part of the DRM (Digital Rights Management). In this context of secured applications, important efforts were devoted to ensure robustness of watermarks against pirate attacks aiming at neutralizing it rather than improving the quantity of watermarked information; the bitrate was usually within the range of tens of bits per second bps for audio signals. Nowadays, audio watermarking can be used for other kinds of applications, and in particular for metadata transmission. However, bitrates are usually still quite low, although such applications require extended bitrates balanced with lower robustness. In this study we propose a high-capacity watermarking technique for audio signals. This technique is suitable for many uncompressed audio signals, more particularly for 16-bit Pulse Coded Modulation (PCM) signals as widely used in audio-CD and wav formats. The proposed technique is based on the application of the Quantization Index Modulation (QIM) technique on the MDCT (Modified Discrete Cosine Transform) coefficients of the signal. The underlying basic principle is that, if those coefficients can be significantly modified by quantization in audio compression schemes such as MPEG MP3/AAC without quality impairments, they can also be modified to embed watermark codes. Following audio compression principles, a psychoacoustic model (PAM) is used at the watermark embedder to take into consideration the behavior of the human auditory system and match the inaudibility constraint. The PAM is used to estimate an optimal watermarking capacity for each sub-band of each MDCT frame. The resulting capacity values are transmitted as (watermarked) side-information to the decoder (so that the decoder can retrieve the usefull watermarked information in the corresponding sub-band). For this aim, specific fixed capacities are allotted in the higher sub-band of the spectrum. With this technique, maximal bitrates of about 250kbps per audio channel can be reached (depending on the audio content), at the expense of robustness: the system can be used for “non-secure” applications where the signal suffers any attack other than quantization for uncompressed format conversion. For instance, we use this technique in a watermark-informed source separation system presented at the same congress.

INTRODUCTION

Digital watermarking is an area of signal processing that consists in imperceptibly embedding binary information in a digital media. Watermarking techniques appeared in the early 90’s [1][2][3] and were first mainly used as a security tool, in the context of the Digital Rights Managements (DRM). Indeed, due to the increasing use of compression techniques and the development of the Internet, digital piracy has been largely spreading. In the audio context, security watermarking usually consists in embedding small-size information (like a digital signature) in the audio signal. Therefore, the difficulty for security watermarking has not been the bitrate, which can be quite low, but it has rather been the robustness to possible attacks aiming at neutralizing the watermark.

For several years, watermarking tends to be used in new kinds of applications besides digital security: watermarking is no longer necessarily used to hide an information related to the content owners/users; it can also be used as a mean to transmit information useful for users [4] (e.g. “enriched-content” databases). For example, in [5] [6], watermarking is used to embed into a mixture audio signal metadata that guide the separation of the source signals composing the (watermarked) audio mix. In many of these potential applications, the main point is less the robustness to attacks (since the user as no interest in damaging the watermark) than the bitrate: the goal is to imperceptibly embed the maximum amount of information in the audio signal.

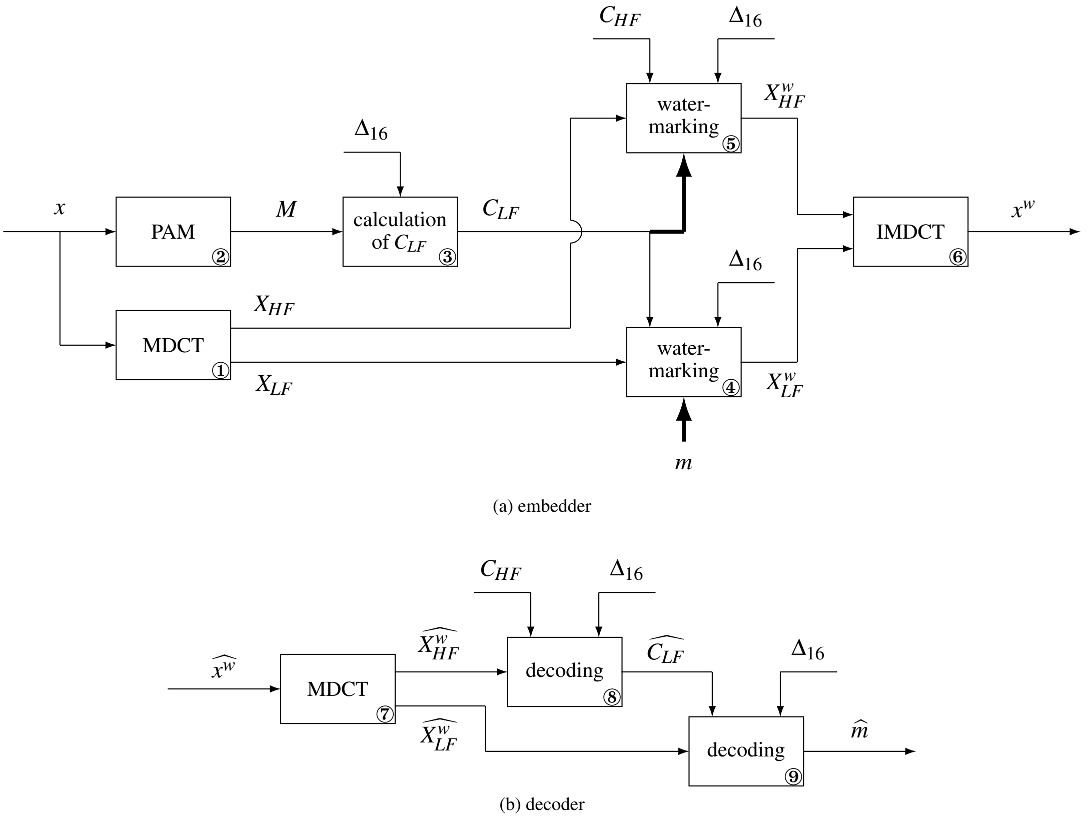
In this paper, we consider such an application scenario and we present a high-capacity watermarking technique developed for audio signals in raw (PCM) format (for instance, 44.1kHz-sampling frequency, 16-bit PCM samples), which goal is to maximize the embedding bitrate under inaudibility constraint.

This paper is organized as follows: the first section is a general overview of the watermarking system and the second section is a more detailed presentation of the main blocks of the system. Results are presented in the third section and the last section concludes this article.

GENERAL OVERVIEW OF THE WATERMARKING SYSTEM

In this section we quickly present a general overview of the watermarking system, focusing on its main principles. Each functional block will be further detailed in the next section. The system consists of two main blocks (figure 1):

- an embedder used to insert the watermark into the signal (figure 1a) and



(a) embedder

(b) decoder

Figure 1: Embedder and decoder diagrams of the proposed high-capacity audio watermarking system. The notation $\widehat{\cdot}$ means that the watermarked signal samples have been quantized by PCM quantization, and derived features may be different from the same features derived from the unquantized watermarked signal. Thick arrows indicate information embedded by watermarking.

- a (blind) decoder used to recover the watermark from the watermarked signal (figure 1b), the original signal being supposed unknown from the decoding part.

The watermarking technique is performed in the Time-Frequency (TF) domain. Therefore, at the embedder, the time-domain input signal[1] x is first transformed (block ①) in the TF plan using the Modified Discrete Cosine Transform (MDCT). The MDCT [7] is a real-valued frame-wise TF transform that is widely used in audio processing, and that will be described in more details in the next section. Note that the resulting MDCT coefficients are defined for each frame (indexed by t) and each frequency channel (or bin, indexed by f). Since the watermarking process is performed independently for each frame, we describe it at the frame level, and we omit the index t when denoting processes inside a frame.

Basically, the watermarking process consists in quantizing the MDCT coefficients X (block ④) using a specific set of quantizers, following the Quantization Index Modulation technique described in [8] (see Section QIM). Once the MDCT coefficients are watermarked, the inverse MDCT (block ⑥) is applied on the watermarked coefficients X^w to obtain the watermarked signal x^w. The key point of the proposed embedding strategy is that, a PsychoAcoustic Model (PAM) inspired from the MPEG-AAC standard (block ②) provides a masking threshold $M(f)$ that enables to calculate the watermark embedding capacity $C(f)$ for each frequency bin f and for each frame t (block ③), *i.e.* the maximum size of the binary code to be embedded in that frequency bin under inaudibility constraint. It is very important to note that the watermarking capacity is a crucial parameter in the proposed watermarking process: it not only characterizes the amount of transmitted information, but it also completely determines the parametrization of the QIM technique used to embed and decode this information (see Section QIM). In other words, it simultaneously determines how much information can be embedded and how to embed and retrieve it. Consequently, the capacity values $C(f)$ must be known at the decoder. In the proposed system, since only the watermarked signal is transmitted, they can be 1) reestimated from the transmitted signal or 2) transmitted as side-information with the watermarked signal. Series of preliminary experiments revealed that the first solution is not a trivial task: when high-capacity is targeted, as is the case here, the overall watermarking process modifies the signal in such a way that, at the receiver, the computation of $C(f)$ by applying the PAM to the transmitted signal generally provides wrong estimations of $C(f)$. Therefore, we rather consider the second solution and we propose to transmit the capacity values $C(f)$ as a part of the watermark itself (with some tricky “self-decoding” procedure), as explained in the sequel.

After MDCT transform, the MDCT coefficients are separated into a “low-frequency” part (denoted $_{LF}$ on figure 1 and hereafter) and a “high-frequency” part (denoted $_{HF}$). Actually, the low-frequency region constitutes the main part of the spectrum, and the high-frequency region is limited to the few highest frequency bins (this point is detailed later). High frequencies are used to embed the values of the capacities $C_{LF}(f)$ which parametrize the watermark embedder on the low-frequency (*i.e.* main band) region. To do this, we chose to fix (*i.e.* set independent of frame index t and signal content) the capacities $C_{HF}(f)$ used to watermark the high-frequency coefficients, exploiting the fact that in this frequency region the power of audio signals is generally well below the absolute threshold of hearing. Those fixed $C_{HF}(f)$ capacities are known at the decoder. Therefore, at the decoder, the received watermarked sig-

[1] the format of the input samples is not a matter here: it can be usual audio PCM (16/20/24-bit) or floating-point values).

nal $\widehat{x^w}$ is first transformed in the time-frequency domain (block ⑦) and the resulting MDCT coefficients are separated into low and high frequencies subvectors, the same way as was done at the embedder. The capacities $C_{HF}(f)$ of the high frequencies being fixed, the information watermarked in this zone is first extracted and decoded (block ⑧) to obtain the decoded values $\widehat{C_{LF}}(f)$. This latter information is then used to decode the information $\widehat{m}$ in the low frequency region (block ⑨) which is the "main" or "useful" information.

In summary, the proposed system is characterized by two entangled watermarking processes :

- the first one is the watermarking of the "useful" information m in the low frequencies parametrized by capacities $C_{LF}(f)$ (block ④) and
- the second one is the watermarking of the $C_{LF}(f)$ values in the high frequencies using fixed capacities $C_{HF}(f)$ (block ⑤).

A detailed setting of those parameters will be described in the next section. The watermarking technique used in both cases is chosen to be the QIM, although two different techniques could be used, one for each watermarking zone.

Finally, it can be noted that a particularity of this watermarking system is that the length N of the MDCT frames can be adjusted. This variability is interesting for two reasons: first, the length N can likely change the system performance (in terms of watermarking bitrate); thus, the system will be tested with respect to this parameter. Second, the proposed watermarking system can be used jointly with applications using MDCT transformation; hence, the length N of MDCT frames used for watermarking can be set equal to the one used for the application, so to optimize computational load.

DETAILED PRESENTATION

In this section we describe in more details the blocks and the techniques composing the watermarking system.

Time-frequency transformation

Due to the high bitrate requirement, the watermarking strategy is based on quantization techniques (see the next section). However, quantizing directly the time-domain signal samples leads to quickly damaging the signal quality while obtaining low embedding capacity. That is why, as already mentioned in the previous section, the watermarking is processed on the time-frequency coefficients of the signal. In the present study, the choice of the Modified Discrete Cosine Transform (MDCT at blocks ① and ⑦, and inverse MDCT at block ⑥) was guided by the following points [7]:

- MDCT coefficients are real (when opposed to complex Discrete Fourier Transform coefficients)
- MDCT coefficients are particularly robust to quantization (and more generally to any manipulation affecting their values)
- the MDCT is an overlapping transform, with particularly good behavior regarding block effects (see the "perfect reconstruction" or "time-domain aliasing cancellation" property in [7]).

Those reasons notably explain why the MDCT is used in many audio applications, particularly in audio compression (AAC, AC3, Vorbis...).

Technically, the MDCT applied to a given frame of N time-domain samples $x(n)$ (N being even) is given by:

$$X_M(f) = \frac{2}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) w_a(n) \cos\left(\frac{2\pi}{N}(n+n_0)\left(f+\frac{1}{2}\right)\right), \tag{1}$$

with $f \in \left[0, \frac{N}{2}-1\right]$, $n_0 = \frac{1}{2}+\frac{N}{4}$ and w_a is the analysis window with duration N. The inverse transform (IMDCT) is given by:

$$y(n) = w_s(n) \frac{2}{\sqrt{N}} \sum_{f=0}^{\frac{N}{2}-1} X_M(f) \cos\left(\frac{2\pi}{N}(n+n_0)\left(f+\frac{1}{2}\right)\right), \tag{2}$$

with $n \in [0, N-1]$ and w_s is the synthesis window (also with duration N).

Since the MDCT transform of N temporal samples only gives $\frac{N}{2}$ coefficients, it is not (strictly speaking) an invertible transformation on a single frame basis. However using an overlap of 50% between successive frames and ensuring some conditions on the analysis and synthesis windows, perfect reconstruction of the time signal from unmodified MDCT coefficients can be achieved (as well as acceptable reconstruction of the time signal from watermarked MDCT coefficients in our case). More information about the MDCT can be found in [7].

Watermarking

In this section, we present the core of the watermarking technique itself. We first present the general principle of the watermarking technique for a given capacity $C(f)$. Then we explain in details how the capacities $C(f)$ are determined as a result of an optimization problem under constraints. As already mentioned in the general overview, the capacities for the low frequencies are calculated from the masking threshold $M(f)$ (block ③), and the capacities for the high frequencies are fixed regarding the absolute threshold of hearing (block ⑤). The calculation of the masking threshold will be detailed in a further section. Then, practical settings of the capacities will be derived.

QIM principles

The considered watermarking technique is the quantization-based technique called Quantization Index Modulation (QIM) introduced in [8] applied to the MDCT coefficients. The principle is the following: for each MDCT coefficient at frequency bin f, a set of $2^{C(f)}$ quantizers $\{\mathcal{Q}_c\}_{c=0\dots 2^{C(f)}-1}$ with intertwined quantization levels is defined (see figure 2) in such a way that:

- Each quantizer represents a $C(f)$-bit binary code so that the watermarking capacity of this particular time-frequency coefficient is $C(f)$
- The $2^{C(f)}$ quantizers are uniform and the intertwining is regular in order to obtain a good compromise in watermarking performance (good audio quality while low decoding error rate).

Watermarking the $C(f)$-bit code c on a given MDCT coefficient $X_M(f)$ is done by quantizing $X_M(f)$ with the quantizer $\mathcal{Q}_c$ associated to (*i.e* indexed by) the code c to transmit. In other words, the MDCT coefficient $X_M(f)$ is replaced with its code-indexed quantized value:

$$X_M^w(f) = \mathcal{Q}_c\left(X_M(f)\right). \tag{3}$$

At the decoder, the transmitted code is recovered by:

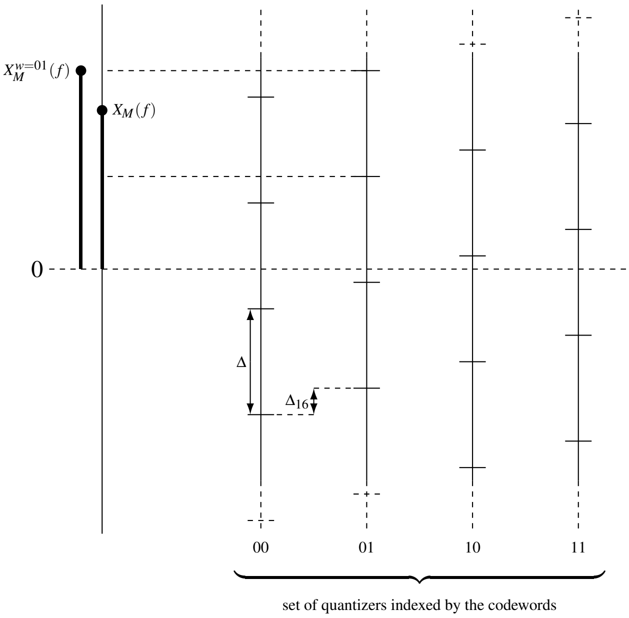


Figure 2: Setting of the QIM quantizers. For example, if one wants to watermark the binary code 01 in the MDCT coefficient $X_M(f)$, one just quantizes it in $X_M^{w=01}(f)$ on the quantizer indexed by 01.

1. comparing the transmitted (quantized) MDCT coefficient $\widehat{X_M^w(f)}$ (potentially corrupted by transmission noise) with the quantization levels of the $2^{C(f)}$ quantizers (assumed to be available at both the embedder and the decoder), and
2. selecting the quantizer $\mathscr{Q}_{\hat{c}}$ to which belongs the quantization level the closest to the transmitted MDCT coefficient $\widehat{X_M^w(f)}$. The decoded code $\hat{c}$ is finally obtained as the index of the selected quantizer $\mathscr{Q}_{\hat{c}}$.

Obviously, the whole binary message m to transmitfor the considered application scenario has to be previously split and spread across the different MDCT coefficients according to the local capacity values, so that each MDCT coefficient carries a small part of the complete message. Conversely, the decoded elementary messages have to be concatenated to recover the complete message $\hat{m}$.

Computation and choice of the capacities $C(f)$

The capacities computation is determined by two related constraints:

- first, the watermarking process must be robust to the 16-bit PCM conversion of the watermarked audio signal; in other words, the quantization of the original MDCT coefficients $X_M(f)$ at block ④ (resp. block ⑤) of the embedder and the quantization of the transmitted MDCT coefficients $\widehat{X_M^w(f)}$ at block ⑨ (resp. block ⑧) of the decoder must provide the same result. Only a low error probability p_e on the decoding will be tolerated, and
- second, the obtained watermark must be inaudible.

Since the goal of our contribution is to maximize the embedding rate, the capacities computation can be formulated as an optimization problem under a double constraint. As detailed below, the first constraint imposes the quantization step $\Delta(f)$ of the quantizers to be lower that a certain bound, fixed to achieve robustness to PCM quantization. Conversely, the inaudibility constraint induces an upper bound on the number of quantizers, hence a corresponding upper bound on the individual (MDCT coefficient-wise) capacity.

Robustness constraint to PCM quantization: Although the proposed system is not designed to achieve robustness to attacks, it must be robust to the PCM quantization. In the present study, we consider the 16-bit PCM, since it is a common storage format for uncompressed audio signals (*e.g.* it is used in wav and audio-CD data).

To model the effects of the time-domain PCM on the MDCT coefficients, let us assume the realistic hypothesis that the noise $b(n) = x^w(n) - \widehat{x^w(n)}$ added to the signal samples by the PCM has independent values from one sample to another. According to the Central Limit Theorem, the MDCT coefficients $B(f)$ of the noise b at frequency bin f follows a normal distribution. Moreover, it can be proved using the normalized equation (1) of the MDCT, that the variance $\sigma^2_{MDCT}(f)$ of the noise coefficient $B(f)$ is equal to the variance σ^2_{16} of the noise $b(n)$ in the time domain and thus is independent of the frequency index f. In summary, we have:

$$\begin{gathered}\forall f \in \left[0, \frac{N}{2} - 1\right], B(f) \sim \mathscr{N}\left(0, \sigma^2_{16}\right), \\ \text{with } \sigma^2_{16} = \sigma^2_{MDCT}(f) = \frac{\left(2^{-15}\right)^2}{12}\end{gathered} \tag{4}$$

Since the PCM effects can be modeled as an additive white Gaussian noise in the MDCT domain, the minimum distance Δ_{16} that can be tolerated between two watermarked values (see figure 2) to achieve a predefined decoding error probability p_e can be computed (see [8]) as:

$$\Delta_{16} = 2\sqrt{2}\sigma_{16}\mathrm{erf}^{-1}(1 - p_e), \tag{5}$$

where $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ is the common error function.

Finally, due to the intertwined property, the quantization step $\Delta(f)$ of the quantizers for the MDCT coefficient $X_M(f)$ indexed by f is then given by:

$$\Delta(f) = \Delta_{16} 2^{C(f)}. \tag{6}$$

Inaudibility constraint: The inaudibility constraint is guided by the masking threshold $M(f)$ provided by the psychoacoustic model. Specifically, the power of the watermarking error (that is the watermark itself) has to remain under the masking threshold whatever the watermark is. As the watermarking strategy is a simple quantization, the worst case watermark is equal to half the quantization step $\Delta(f)$, which is directly related to $C(f)$ through equation (6). Thus, assuming the watermark power well estimated using MDCT coefficients, the inaudibility constraint can be written as:

$$\left(\frac{\Delta(f)}{2}\right)^2 < M(f), \tag{7}$$

for each frequency bin f.

Choice of the capacities: As explained previously, the capacities choice depends on the considered frequency zone: low-frequencies are involved in the watermarking of useful information m parametrized by adaptive capacities $C_{LF}(f)$ whereas high-frequencies support the watermarking of $C_{LF}(f)$ values parametrized by fixed capacities $C_{HF}(f)$.

In the high-frequency zone, the capacities $C_{HF}(f)$ being fixed, the yielded watermark must always be inaudible, whatever the frame and signal content. This particular constraint can be satisfied when only the absolute threshold of hearing is considered. Indeed,

- the absolute threshold of hearing is raising rapidly as a function of frequency in the high frequency region, and
- the power of audio signals is generally very low at such high frequencies.

Therefore, the power of audio signals is generally significantly lower than the masking threshold $M(f)$ in the high frequency region, and a reasonable amount of information can be embedded there without noticeable effects. In the present case, this information is the values of the capacities in the low frequency region, and as shown below, it can actually be set at a reasonable (even low) amount.

For the low frequencies, combining the two formulas (6) and (7) to reach the maximum bitrate for frequency index f, the adapted capacity $C_{LF}(f)$ is given by:

$$C_{LF}(f) = \left\lfloor \frac{1}{2}\log_2\left(\frac{M(f)}{\Delta_{16}^2}\right) + 1 \right\rfloor. \tag{8}$$

Experiments on real audio signals show that the resulting values are always lower than 15[2]. Those values can be coded with 4-bit codewords (from 0 to 15). Unfortunately, embedding the high-frequency zone with as many 4-bit codewords as there are frequency bins in the low frequency-zone is clearly impossible (or this would require the high-frequency zone to be nearly the same size as the low-frequency zone, which is clearly not what we want). For this reason, we choose to define watermarking bands as groups of adjacent frequency bins and to allocate one capacity per band instead of one capacity per frequency bin: the capacity is identical for each MDCT coefficient within a band (similarly to coding bands used in compression: for each coding band, only one quantizer is used). The number of low-frequency capacities values $C_{LF}(f)$ to be watermarked in the high-frequency zone can then be significantly reduced. In order to achieve both the inaudibility constraint and a small high-frequency zone, the capacities $C_{HF}(f)$ (that parametrize the watermark of the $C_{LF}(f)$) are fixed to 1 or 2 bits per MDCT coefficient[3].

There is now to determine the number and the size of the watermarking bands. We have chosen two types of partition for the watermarking bands:

- a partition with 25 bands following the Bark scale (*i.e.* approximate logarithmic distribution). The last band of this 25-band log. partition being very large with respect to a frequency bin (whatever the MDCT frame length N), the high-frequency zone consists of the 50 last frequency bins and each of them is watermarked with 2 bits per coefficient, offering a total amount of 100 bits, which is appropriated to encode 25 $C_{LF}(f)$ values with 4-bit codewords.
- a partition with 32 bands equally distributed (quite alike the MPEG-1 filterbank [9]). For this linear 32-band partition, the high-frequency zone must be adapted to the frame length value N in order to maximize the embedding capacity in the main band. Therefore, the high-frequency zone is composed of the n_{HF} last bands, with n_{HF} the smallest integer so that watermarking those n_{HF} bands with maximum 2 bits per MDCT coefficient gives enough bits to encode the $C_{LF}(f)$ values of the 32-n_{HF} first bands. For example, for frames of length $N = 2048$, the high-frequency zone consists of the $n_{HF} = 2$ last watermarking bands, *i.e.* $\frac{N/2}{32} \times 2 = 64$ frequency bins, offering 128 bits to embed the 30 capacity values $C_{LF}(f)$.

Psychoacoustic model

The psychoacoustic model used in our system (block ②) is directly inspired from the psychoacoustic model of the MPEG-AAC standard [10][11], with some adaptations allowing the user to adjust the frame length N. As written earlier, the output of the PAM is a masking threshold $M(f)$ defined for each watermarking band (similarly to the masking threshold for coding band in AAC). $M(f)$ represents the maximum power of the watermarking error (coding error in AAC) that can be added to the audio signal while ensuring inaudibility.

PAM computation is carried out in the time-frequency domain, however the transform used for the PAM calculation is not the MDCT used or the watermarking but the classical Fast Fourier Transform (FFT). In broad outline, a first masking curve is computed as the convolution of the DFT power spectrum of the signal and a spreading function that models elementary frequency masking phenomenons. This curve is then adjusted according to the signal tonality[4], and is combined with the absolute threshold of hearing. Next, some pre-echo control is applied, resulting in the DFT masking threshold (see figure 3 for an example) and a Signal-to-Mask Ratio (SMR) is calculated as the difference between the DFT spectrum and the DFT masking threshold. The MDCT masking threshold $M(f)$ is finally simply computed as the ratio between the MDCT power spectrum and the SMR.

Usually, in audio compression, the masking threshold is used in the bit allocation procedure to manage the distribution of the quantization error over the frequency range with respect to local psychoacoustical phenomenons; the mean quantization error power is not only controlled by $M(f)$ but also by the coding bit budget. Regarding the watermarking scenario, this last constraint can be converted into a scaling factor α (multiplying the masking threshold before its combination with the absolute threshold of hearing) and controlling the overall inaudibility of the watermark. Moreover, since the masking threshold determines the watermarking capacities, the scaling factor α permits to adjust the embedding bitrate to the need before the calculation of the capacities. Thus, with this flexible PAM, it is possible to trade bitrate against audio quality and vice versa.

Note: an important characteristic of the AAC psychoacoustic model is that all the intermediate parameters that step in the masking threshold calculation are not defined for each frequency bin f but for "partitions". In AAC, the partitions are approximately equal to the minimum between a third of a critical band (in Bark scale, see [12]) and a frequency bin, in order to achieve good quality. As for the coding bands, the partitions are arbitrarily fixed (their configuration does not change

[2] It can be noted that this maximal value of 15 bits for a single coefficient is a very high capacity; it is comparable to the number of bits necessary for accurate coding of time-domain samples. However, as detailed in the results section, all MDCT coefficients cannot carry such a large amount of watermarked information.

[3] the margin seems to be quite high here; we plan to experiment greater capacities in future works, in order to improve the overall bitrate of our watermarking system.

[4] It is important to note that the main reason why the PAM of the AAC works with the FFT and not the MDCT is because the phase information given by the FFT can be used to estimate the tonality of the signal in a better way than it is possible with the MDCT.

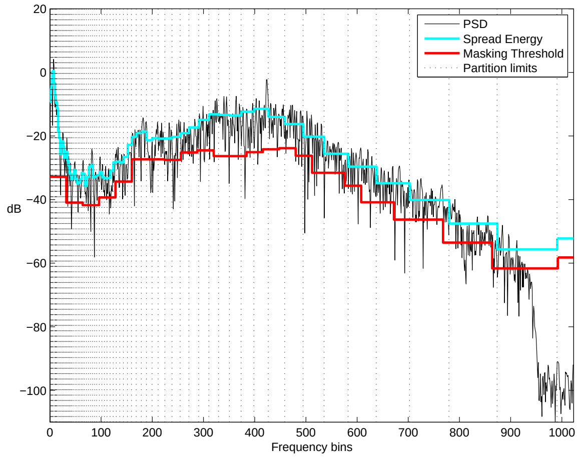


Figure 3: Example quantities of the psychoacoustic model for a frame length of N=2048.

Absolute Grade				ODG
	5,0	Imperceptible	0,0	
	4,9 to 4,0	Perceptible, but not annoying	-0,1 to -1	
	3,9 to 3,0	Slightly annoying	-1,1 to -2	
	2,9 to 2,0	Annoying	-2,1 to -3	
	1,9 to 1,0	Very annoying	-3,1 to -4	

Table 1: Meanings of the ODG.

the main calculations of the psychoacoustic model). The AAC standard using only two frame lengths (2048 and 256), the partitions are saved in tables. In order to ensure the adaptability of our system in regard to the frame length N, an algorithm computing the partitions for a given length N has been developed (eligibles values for N being powers of 2, in particular 512, 1024, 2048 and 4096). This algorithm simply calculates the partitions starting from the frequency bin 0 and chooses for each partition the size that is the closest to a third of a critical band (using the analytical expression of the conversion Bark/Hertz given in [13]) with the constraint that a partition's minimum size is 1 frequency bin.

RESULTS

In order to test the performance of our watermarking system, ten musical excerpts of 10-second duration and of different musical styles were used. For each test signal, we were interested in the audio quality (*i.e.* the (in)audibility of the watermark) and the watermarking/embedding rate. As already mentioned, tests were made for both watermarking band partitions, the one with 32 bands uniformly distributed (*lin*) and the other one with 25 bands following the Bark scale (*log*). The expected decoding error probability p_e was fixed to 10^{-6} and the scaling factor α was fixed to 1 (no modification of the masking threshold $M(f)$).

Inaudibility of the watermark

The audio quality of the signals was estimated using the Perceptual Evaluation of Audio Quality (PEAQ) algorithm [14]. This algorithm compares the original signal and the watermarked signal, and provides a comparative score, called Objective Difference Grade (ODG). Grades range from 0 for inaudible effect to -4 for severe degradation. More detailed meanings of the grades are given in table 1.

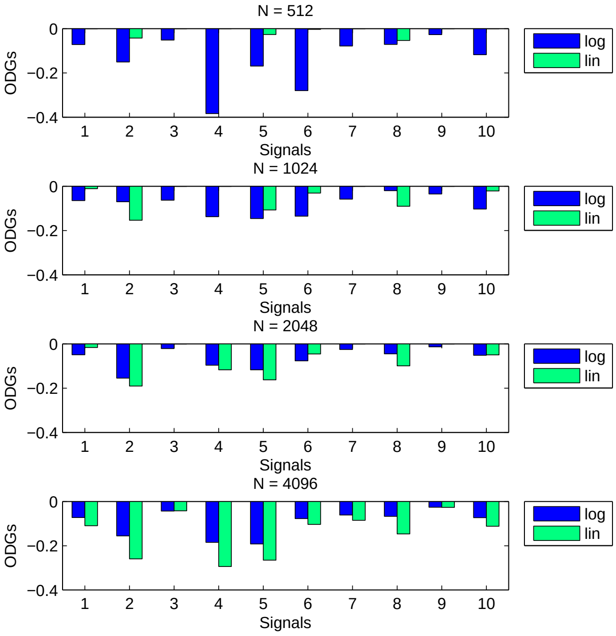


Figure 4: ODGs for each audio excerpt numbered from 1 to 10 for the two types of watermarking bands partition and for different values of the frame length N.

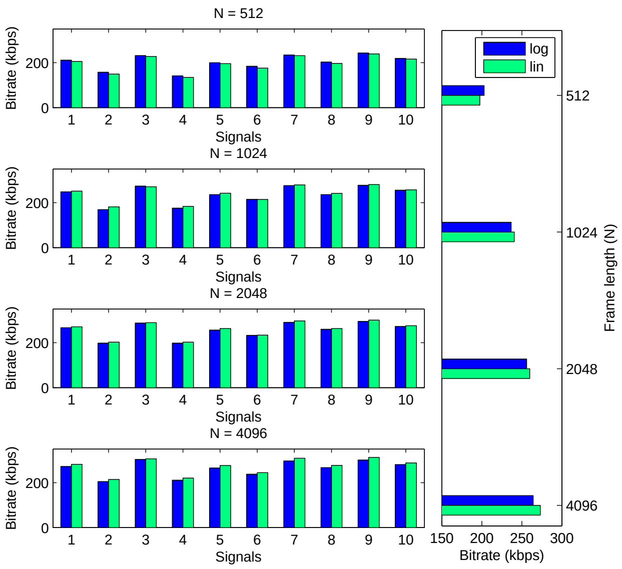
The ODG obtained for both watermarking band partitions and for different frame lengths are given figure 4. We can see that the grades are good for both schemes, since none is below -0.5, and even very good for frame length larger than 512, with minimum between -0.2 and -0.3. Hence the proposed watermarking system is shown to only slightly damage the signal quality. This is confirmed by informal listening tests that reveal undistinguishable quality between original and watermarked signals.

Watermarking bitrates

Watermarking bitrates obtained with the proposed system are presented in figure 5a for each tested frame length and each musical excerpt. The mean rates (averaged across all excerpts) are given in figure 5b. Results show that even if the rates are quite variable with the musical styles, they all are quite high, ranging from 135kbps for a jazz-rock track with N = 512 to 312kbps for a pop-music track with N = 4096. Average bitrates (across signals) higher than 200kbps are obtained for all frame lengths, and an average bitrate of 250kbps is obtained for N = 2048 (which is a standard length for most of the audio systems using MDCT; it is notably the length of the "long" frames in AAC). The coding rate of an audio channel in the usual 16-bit PCM format being 705kbps, the watermarking rate accounts for 30 to 35% of this coding budget, which is remarkable[5]. A marked increase of the rate can also be noticed when the length N increases. This can be related to the two following points:

- The psychoacoustic model being based on the frequency domain, the larger is the frequency resolution (*i.e.* the greater the frame length is), the better the results are. However, it is important to respect the signal dynamic, and thus to limit frame lengths (around 50ms).
- The number of bits in the high-frequency zone used to watermark the low-frequency capacities values is observed to be nearly invariant with the frame length N.

[5] Although this result may appear quite limited at first sight when compared with AAC transparency at 96-kbps coding rate, it must be taken into consideration that it is generally much more difficult to imperceptibly embed extra information within a signal than to remove information that was not audible at first hand.



(a) Bitrate per signal. (b) Mean bitrates.

Figure 5: Bitrates obtained for each signal and mean bitrates (function of the frame length N).

Therefore, when N increases (and so the total number of frames decreases), the total number of bit used in the high frequencies decreases as well, leading to gain rate for the “useful” information.

CONCLUSIONS AND PERSPECTIVES

The watermarking technique presented in this paper enables watermarking of audio signals in the PCM format at high capacity/bitrates (more than 200kbps for 16-bit PCM signals) while retaining a very good quality.

In further works, we will try to improve the system by:

- best considering the pre-echo phenomenon, as it is actually responsible of most of the bad ODGs for small frame lengths (512 essentially);
- improving the watermarking bands distribution to gain in bitrate and quality;
- adaptating the psychoacoustic model in order to recalculate the capacities at the decoder, and thus getting rid of the high-frequency zone. This would probably improve the bitrate consistently.

ACKNOWLEDGMENTS

These works are supported by the French National Research Agency (ANR) in the context of the DReaM project (ANR 09 CORD 006).

REFERENCES

[1] M. Costa. Writing on dirty paper. *IEEE Trans. Inform. Theory*, 29(3):439–441, 1983.

[2] L. Boney, T. Ahmed, and H. Khaled. Digital watermarks for audio signals. *Third IEEE Int. Conf. on Multimedia Computing and Systems*, pages 473–480, June 1996.

[3] I.J. Cox, M.L. Miller, and A.L. McKellips. Watermarking as communications with side information. *Proc. IEEE*, 87(7):1127–1141, 1999.

[4] B. Chen and C.-E.W. Sundberg. Digital audio broadcasting in the FM band by means of contiguous band insertion and precanceling techniques. *IEEE Trans. Commun.*, 48(10):1634–1637, 2000.

[5] M. Parvaix, L. Girin, and J.-M. Brossier. A watermarking-based method for single-channel audio source separation. In *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 101–104, Taipei, Taiwan, 2009.

[6] M. Parvaix and L. Girin. Informed source separation of underdetermined instantaneous stereo mixtures using source index embedding. In *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, Texas, 2010.

[7] J.P. Princen and A.B. Bradley. Analysis/synthesis filter bank design based on time domain aliasing cancellation. *IEEE Trans. Acoust., Speech, Signal Process.*, 64(5):1153–1161, 1986.

[8] B. Chen and G. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Inform. Theory*, 47(4):1423–1443, 2001.

[9] ISO/IEC MPEG. *IS11172-3 Information technology - Coding of moving pictures and associated audio for digital storage media at up to about 1.5Mbits/s, Part 3: Audio*. ISO, 1992.

[10] ISO/IEC. *ISO/IEC 13818-7:2004(E) Information technology - Generic coding of moving pictures and associated audio information - Part 7: Advanced Audio Coding (AAC)*.

[11] N. Moreau, O. Derien, S. Larbi, and M. Perreau Guimares. Le codeur MPEG-2 AAC expliqué aux traiteurs de signaux. *Ann. Télécommun.*, 55(9-10):442–461, 2000.

[12] E. Zwicker and U. Zwicker. *Psychoacoustics Facts and Models*. Springer-Verlag, 1990.

[13] H. Traunmüller. Analytical expressions for the tonotopic sensory scale. *J. Acoust. Soc. Am.*, 88:97–100, 1990.

[14] ITU. *ITU-R Recommendation BS.1387-1: Method for objective measurements of perceived audio quality (PEAQ)*, 2001.