A High-Rate Data Hiding Technique for Uncompressed Audio Signals

JONATHAN PINEL, LAURENT GIRIN, AND
(Jonathan.Pinel@gipsa-lab.grenoble-inp.fr) (Laurent.Girin@gipsa-lab.grenoble-inp.fr)

CLÉO BARAS
(Cleo.Baras@gipsa-lab.grenoble-inp.fr)

GIPSA-Lab/University of Grenoble

In this paper we propose a high-rate data hiding technique for audio signals suitable for non-secure applications that require a large bit rate but no particular robustness to attacks. More particularly, the proposed technique is suitable for enriched-content applications involving uncompressed PCM audio signals, as used in audio-CD and .wav formats. It applies the Quantization Index Modulation (QIM) technique on the Modified Discrete Cosine Transform (MDCT) or Integer MDCT (IntMDCT) coefficients of the signal. The basic principle is that if these coefficients can be significantly modified by quantization in perceptual audio compression with very moderate quality impairments, they can also be modified to embed data. Following audio compression principles, a Psychoacoustic Model (PAM) is used at the embedding stage to consider the properties of the human auditory system and match the inaudibility constraint. The PAM is used to estimate the number of bits to be embedded in each MDCT coefficient for each frame. The resulting set of values is transmitted to the decoder as a minor part of the total embedded side-information. For this aim, a specific fixed embedding space is allocated in the high frequencies of the spectrum. With this technique, simulations on real audio signals show that bit rates of about 250 kbps per audio channel can be reached (depending on the audio content).

1 INTRODUCTION

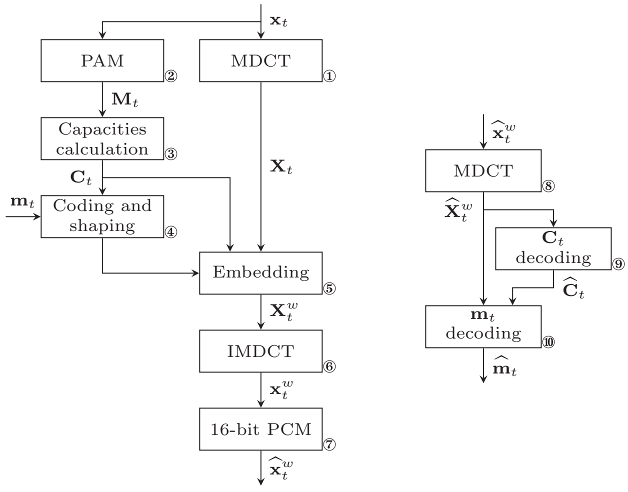
Data hiding consists in imperceptibly embedding information in digital media. Theoretical fundamentals can be found in [7], and the first papers and applications dedicated to audio signals were developed in the 1990s [2, 8]. In its beginning, data hiding for audio signals was mainly used for the Digital Rights Management (DRM). The embedded data were usually copyrights or information on the author or the owner of the audio content (in this context data hiding is often referred to as *watermarking*, and the embedded data is the *watermark*). For such applications, the size of the embedded data is relatively small, and a crucial issue is the robustness of the watermark to malicious processes (referred to as attacks) that aim at removing or modifying it [1, 18]. Therefore, research has long been (and still is) focused on enhancing the security and robustness of the data hiding techniques, at the price of limited embedding bit rate.

Data hiding is now used for non-secure applications as well [5]. For example, in [25] watermarking is used to transmit information that is used for the restoration of coding artifacts on the host signal. "Enriched-content" applications can use data hiding as a means to transmit side-information to the user, in order to provide additional interaction with the media. In this context the specifications of data hiding are different from security applications. Here, a high embedding rate is generally required to provide substantial interactive features. Therefore, the technical issue is usually to maximize the embedding bit rate under the double constraint of imperceptibility and robustness. Yet robustness is here to be taken in the weak sense because the user has no reason to impair the embedded data, since this would result in losing the enriching features. Therefore, robustness is generally limited to compliance with signal representation in a given format or robustness to transmission errors. In this paper we focus on high-rate data hiding for uncompressed audio signals (i.e., 44.1 kHz 16-bit PCM samples, such as audio-CD, .wav, .aiff, .flac formats), with potential application to enriched-content music processing. For example, the so-called Informed Source Separation techniques developed in [19, 20, 22] use embedded data to ease the separation of the different musical instruments and voices that form a music signal. In the present study the embedding constraints are inaudibility and robustness to time-domain PCM quantization (so that the embedded host signal can be stored or transmitted with usual uncompressed formats).

In the data hiding literature, when security and robustness are not the main concerns, the highest bit rates are obtained for data hiding techniques based on quantization. For example, in [9] and [10], Cvejic and Seppänen use the Least Significant Bit (LSB) scheme, on either the temporal samples of the signals with bit rates around 170 kbps per channel (kbps/c), or on the coefficients of a wavelet transform with bit rates up to 400 kbps/c. In these works the inaudibility constraint is not clearly defined and thus not entirely exploited. To maximize the embedding bit rate while sticking as closely as possible to the inaudibility constraint, the properties of the human hearing system must be better taken into account. This involves the use of a Psychoacoustic Model (PAM). Since PAM are generally described in the frequency domain, it seems relevant to perform the embedding on the coefficients of a Time-Frequency (TF) transform of the signal, such as the Discrete Fourier Transform (DFT) or the Modified Discrete Cosine Transform (MDCT). In fact, the combination of quantization, TF transform, and PAM is actually the basis of most perceptual audio coding (PAC) systems [3, 21]. For example, in MPEG 2 Advanced Audio Coding (MPEG2-AAC) [15], the MDCT is first applied on the signal and the MDCT coefficients are then quantized with limited binary resources while the quantization error is shaped below the masking threshold provided by the MPEG2-AAC PAM. Such general scheme can be adapted to data embedding: host audio signals are also transformed into the MDCT domain, but the quantization stage is used to embed binary information instead of coding the host signal (i.e., the coefficients are modified according to the information to be embedded). The PAM is used to control the embedding error instead of the coding error. Finally the embedded signal, obtained by inverse MDCT, consists of time-domain PCM samples instead of a compressed bit stream.

This principle has already been implemented in [14]. In this study an LSB embedding scheme is applied on the Integer MDCT (IntMDCT) coefficients of the signal. The IntMDCT is an integer-valued approximation of the MDCT. The number of bits used for the LSB scheme is controlled by a PAM that is grossly estimated from the *lead bits* of the short-term spectrum. This is to ensure that the PAM can be exactly recalculated at the decoder to derive the corresponding LSB decoding. However this limits the accuracy of the PAM and may thus limit either the inaudibility or the embedding bit rate, or both, depending on the tuning of the system. With this approach and a basic PAM, embedding bit rates around 140 kbps/c are reported.

In the present study we propose a new high-rate data hiding technique also inspired from PAC principles. We use the MDCT or the IntMDCT transform, and the resulting coefficients are quantized using the Quantization Index Modulation (QIM) scheme [6], which is more general than LSB quantization. We use an accurate PAM directly inspired from the MPEG2-AAC standard, and, more importantly, we derive an embedding scheme that does not need recalculation of the PAM at the decoder. Instead, the time-varying and frequency-varying parameters of the quantization process are transmitted as a minor part of the embedded information within a "subchannel" with fixed parameters. This results in a very computationally efficient decoder and also enables to fully exploit the PAM-based embedding capacity of the TF representation, leading to bit rates up to 350 kbps/c (depending on the musical content). Synchronization issues will be considered: two specific cases relevant for the proposed system will be detailed. However the system is not designed for robustness to malicious attacks, to most processing techniques that affect the signal samples, and obviously to audio compression. Thus those issues will not be discussed.



(a) Embedder. (b) Decoder.

Fig. 1. Embedder (a) and decoder (b) diagrams of the proposed high-rate audio data hiding system. $\mathbf{x}_t$ is a frame of the host audio signal and $\mathbf{m}_t$ is the extra information to be embedded into $\mathbf{x}_t$. $\mathbf{M}_t$ is the masking threshold (output of the PAM) and $\mathbf{C}_t$ are the capacities. The notation $.^w$ indicates an embedded signal and the notation $\widehat{.}$ indicates samples modified by PCM quantization.

This paper is organized as follows: Sec. 2 is a general overview of the system and Sec. 3 is a more detailed technical presentation. Results and comparison with state-of-the-art data hiding system [10] (in terms of embedding bit rate and audio quality) are then presented in Sec. 4. Section 5 concludes this article.

2 GENERAL OVERVIEW OF THE DATA HIDING SYSTEM

In this section we provide a general overview of the proposed data hiding system focusing on the main principles. The functional blocks will be further detailed in Sec. 3. The system consists of two main blocks (see Fig. 1):

- An *embedder* used to embed the data into the host signal $\mathbf{x}$ in an imperceptible manner (Fig. 1a);
- A *decoder* used to recover the data from the embedded host signal $\widehat{\mathbf{x}}^w$ (Fig. 1b); the decoder is *blind* in the sense that the original signal is assumed to be unknown from the decoding part.

As already mentioned in the introduction, due to the requirement of a high embedding bit rate, the data hiding system is based on a quantization technique. However, directly quantizing the time-domain samples of the host signal quickly leads to a deterioration of the audio quality when the bit rate increases. Therefore, at the coder, the time-domain input signal $\mathbf{x}$ is first transformed in the time-frequency (TF) domain using the MDCT or the IntMDCT[1] (Block ⑤). The MDCT is a real-valued frame-wise TF transform widely used in audio processing. Note that boldfaced variables denote vectors or matrices. Subscript $_t$ denotes frame index and f denotes frequency bin. For example if $\mathbf{x}$ is a single channel time-domain signal, $\mathbf{x}_t$ is the t^{th} frame of this signal, $\mathbf{x}_t(n)$ represents the n^{th} sample of frame t, and $\mathbf{X}_t(f)$ is the f-th coefficient of the MDCT transform of frame t.

Basically, the embedding process consists in quantizing each MDCT coefficient $\mathbf{X}_t(f)$ (Block ⑤) using a specific set of quantizers $\mathcal{S}(\mathbf{C}_t(f))$, following the QIM technique described in [6] (see Sec. 3). Once the MDCT coefficients are embedded, the signal is reverted back in the time-domain using the inverse MDCT (IMDCT; Block ⑥). Finally, the embedded time-domain signal is converted using PCM coding (Block ⑦).

As mentioned in the introduction, the key point of the proposed method is that for each frame t, a PAM (Block ②) provides a masking threshold $\mathbf{M}_t$ used to calculate the embedding capacity vector $\mathbf{C}_t$ (Block ③), i.e., the maximum size of the binary code to be embedded into each TF coefficient under inaudibility constraint. It is very important to note that the embedding capacities $\mathbf{C}_t(f)$ are crucial parameters in the proposed data hiding technique: they not only characterize the amount of embedded information, but they also completely determine the configuration of the QIM technique that is used to embed and retrieve this information (see Sec. 3). In other words, the embedding capacities $\mathbf{C}_t(f)$ determine at the same time *how much* information is embedded (in $\mathbf{X}_t(f)$) and *how* it is embedded and retrieved. Consequently, the vector of capacity values $\mathbf{C}_t$ must be known at the decoder. In the proposed system, data hiding is the only way of transmitting information. Therefore, those capacities $\mathbf{C}_t(f)$ have either to be estimated from the transmitted signal at the decoder, or to be transmitted within the host signal $\mathbf{x}$, as a part of the embedded data themselves. A series of preliminary experiments have revealed that the first solution is not a trivial task: when high bit rates are targeted (around hundreds of kbps/c), the overall data hiding process modifies the host signal $\mathbf{x}$ in such a way that the recalculation of the capacities $\mathbf{C}_t(f)$ by applying the PAM to the transmitted signal $\widehat{\mathbf{x}}_t^w$ generally provides wrong $\widehat{\mathbf{C}}_t(f)$ values. To overcome this problem the *lead bits* principle can be used [14] to ensure an identical output of the PAM at the embedder and the decoder but at the cost of a reduced embedding bit rate and a less accurate PAM. Therefore, we rather consider the embedding of the $\mathbf{C}_t(f)$ values and we propose the following process to overcome those difficulties.

At the embedder, the capacities $\mathbf{C}_t(f)$ are maximized under inaudibility and robustness constraints for each TF bin. This is the core of the proposed method that will be detailed in Sec. 3.4. A small part of the available payload located in the high frequencies of the spectrum is then used to embed the values of the resulting capacities $\mathbf{C}_t(f)$ that totally configure the data hiding process. The embedding location of those $\mathbf{C}_t(f)$ values is fixed and independent of the frame t to ensure blind decoding. The remaining payload is used to embed the "useful" information $\mathbf{m}_t$. Note that in the following, the set of $\mathbf{C}_t(f)$ values (plus potential error correction codes and synchronization data, see Sec. 3.6) is referred to as the *side-information*.

The decoding process is a simple inversion of the embedding chain. At the decoder, the embedded signal $\widehat{\mathbf{x}}_t^w$ is first transformed in the TF domain (Block ⑧). The embedding location of the side-information being fixed and known at the decoder, the decoded $\widehat{\mathbf{C}}_t(f)$ values are extracted (Block ⑨). This information is then used to decode the "useful" information $\mathbf{m}_t$ embedded in the frame (Block ⑩).

Finally, it can be worth noticing a particularity of this data hiding system: the length N of the MDCT frame can be chosen among several values (however once chosen this length is fixed for the whole process). This is motivated by two reasons: first, this length N is a parameter that is likely to change the system performance (in terms of embedding rate and audio quality), and thus it will be tested as such in Sec. 4. Second, this system can be used jointly with applications that use the MDCT transform, hence the interest of having the same frame length for the application and the data hiding system to optimize the computational load.

3 DETAILED PRESENTATION

In this section we describe more precisely the main blocks or techniques composing the data hiding system. Section 3.1 presents the MDCT and IntMDCT transforms, Sec. 3.2 presents the QIM embedding technique, and Sec. 3.3 presents the PAM. In Sec. 3.4 we describe the core of the proposed method, which is the calculation, encoding and embedding of the capacities. In Sec. 3.5 we present how to easily control the embedding bit rate, and finally in Sec. 3.6 we address synchronization issues.

3.1 Time-Frequency Transformation

3.1.1 MDCT

The MDCT is a very popular transform for audio processing. In the present study the choice of the MDCT was guided by several points:

- The MDCT is a transform with 50% overlap, which shows good behavior against block effect (often heard as "clicks" in audio signals).
- The MDCT coefficients are real-valued, as opposed to complex coefficients for the DFT: it is easier to perform a quantization-based embedding on a single real value than on a pair of real/imaginary or modulus/phase values.

[1] Those transforms will be briefly described in Sec. 3.1.1. The differences resulting from each choice will be discussed in Secs. 3.1.1 and 4. When there is no need to differentiate between the two transforms, the term MDCT is assumed to represent any of the two.

- Most importantly, the MDCT possesses the Time-Domain Aliasing Cancellation (TDAC) property. This means that, after modification of the coefficients in a given frame t by data embedding, transforming to the time-domain (Block ⑥) and back to the MDCT domain (Blocks ⑧) will yield the same modified coefficients on frame t and also will not affect the adjacent frames. In fact this is true only in absence of PCM quantization noise (Block ⑦), and in the present study the PCM quantization will be the only source of potential error to be accounted for (see Sec. 3.4).

Technically, the MDCT coefficients of a given frame t of N samples (N being even) of the host signal $\mathbf{x}$ is given for each $f \in [0, \frac{N}{2} - 1]$ by:

$$\mathbf{X}_t(f) = \frac{2}{\sqrt{N}} \sum_{n=0}^{N-1} \mathbf{x}_t(n)\mathbf{w}(n) \cos\left(\frac{2\pi}{N} n' f'\right), \tag{1}$$

where $\mathbf{w}$ is the analysis window, $n' = n + \frac{N}{4} + \frac{1}{2}$, and $f' = f + \frac{1}{2}$. The inverse transformation of the same frame is given for each $n \in [0, N-1]$ by:

$$\widetilde{\mathbf{x}}_t(n) = \frac{2}{\sqrt{N}} \mathbf{w}(n) \sum_{f=0}^{\frac{N}{2}-1} \mathbf{X}_t(f) \cos\left(\frac{2\pi}{N} n' f'\right). \tag{2}$$

Note that $\widetilde{\mathbf{x}}_t \neq \mathbf{x}_t$: the signal is perfectly reconstructed only after the overlap-add if $\mathbf{w}$ satisfies the Princen-Bradley conditions [24]:

$$\forall n \in \left[0, \frac{N}{2} - 1\right] \begin{cases} \mathbf{w}^2(n) + \mathbf{w}^2\left(n + \frac{N}{2}\right) = 1 \\ \mathbf{w}(n) = \mathbf{w}(N-1-n) \end{cases}. \tag{3}$$

In the present study we use a Kaiser–Bessel Derived window, which satisfies these conditions.

3.1.2 IntMDCT

The disadvantage of using the MDCT is that the 16-bit PCM quantization (Block ⑦) introduces a noise on the decoded MDCT coefficients (see Sec. 3.4), leading to possibly wrong decoded values for the embedded data $\mathbf{m}$. To get rid of this problem, an integer-valued transform can be used, i.e., a bijection from $\mathbb{Z}^N$ to $\mathbb{Z}^N$. We thus consider the IntMDCT which is an integer-to-integer approximation of the MDCT. One of the possible ways for building such an integer approximation is the following [13]: the first step is to decompose the transform matrix in a product of matrices that can be either permutation matrices or block diagonal matrices with each block consisting of:

- A 1-by-1 matrix 1 or -1, or
- A 2-by-2 Givens rotation $R(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$.

A permutation is directly a bijection form $\mathbb{Z}^N$ to $\mathbb{Z}^N$, so the integer approximation problem comes down to the integer approximation of the Givens rotations. If $\theta = k\pi/2 (k \in \mathbb{Z})$, the Givens rotation is a bijection from $\mathbb{Z}^2$ to $\mathbb{Z}^2$. Otherwise, denoting $c = \cos\theta$ and $s = \sin\theta$ the following factorization in *lifting steps* [11] can be done:

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{c-1}{s} & 1 \end{pmatrix} \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{c-1}{s} & 1 \end{pmatrix}. \tag{4}$$

If we note $l_a = \left(\begin{smallmatrix} 1 & 0 \\ a & 1 \end{smallmatrix}\right)$ and $.^T$ the matrix transposition, then we have $R(\theta) = l_{\frac{c-1}{s}} \cdot l_s^T \cdot l_{\frac{c-1}{s}}$. l_a corresponds to an operator:

$$\begin{aligned} L_a : \mathbb{R}^2 &\longrightarrow \mathbb{R}^2 \\ (x, y) &\longrightarrow (x, y + ax) \end{aligned} \tag{5}$$

The last part for building the integer approximation is to approximate operators L_a by the operators:

$$\begin{aligned} \mathrm{Int}L_a : \mathbb{Z}^2 &\longrightarrow \mathbb{Z}^2 \\ (x, y) &\longrightarrow (x, [y + ax]) \end{aligned} \tag{6}$$

where [.] denotes the rounding operation. Also notice that if we note Int$R(\theta)$ the integer approximation of $R(\theta)$ then we have:

$$R(\theta)^{-1} = R(-\theta) \tag{7}$$

$$\mathrm{Int}R(\theta)^{-1} = \mathrm{Int}R(-\theta), \tag{8}$$

which means that the IntIMDCT will be the inverse of the IntMDCT, resulting in a coherent framework.

Applying this process directly on the MDCT matrix (i.e., the matrix used to compute $\mathbf{X}_t$ from $\mathbf{x}_t$) is not possible, since this matrix is not square (N/2-by-N). However it can be shown that the whole MDCT transform process is the cascading of two operations [13]: windowing with overlap and DCT4. As the windowing operation and the DCT4 are orthogonal transforms, the corresponding matrices can be decomposed as explained above. The decomposition of the windowing matrix is straightforward, whereas for the DCT4 we use the decomposition developed in [27].

3.2 Embedding Technique: QIM

The Quantization Index Modulation (QIM) is a quantization-based embedding technique introduced in [6]. The scalar version of the technique is used here (embedding at Blocks ④ and ⑤, and decoding at Blocks ⑨ and ⑩), which means that each MDCT coefficient $\mathbf{X}_t(f)$ is modified by the QIM independently from the others.

The embedding principle is the following. If $\mathbf{X}_t(f)$ is the MDCT coefficient that has to be processed with capacity $\mathbf{C}_t(f)$, then a unique set $\mathcal{S}(\mathbf{C}_t(f))$ of $2^{\mathbf{C}_t(f)}$ quantizers $\{\mathcal{Q}_c\}_{0\le c\le 2^{\mathbf{C}_t(f)}-1}$ is defined with a fixed arbitrary rule. This implies that for a given value $\mathbf{C}_t(f)$ the set generated at the decoder is the same as the one generated at the embedder. The quantization levels of the different quantizers are intertwined (see Fig. 2) and each quantizer is indexed by a $\mathbf{C}_t(f)$-bit codeword c. Note that the quantizers are uniform, the indexation follows the Gray code, and the intertwining is regular to simplify the implementation and minimize the Bit Error Rate (BER). Embedding the codeword c into the MDCT coefficient $\mathbf{X}_t(f)$ is simply made by quantizing $\mathbf{X}_t(f)$ with the quantizer $\mathcal{Q}_c$ indexed by c (see Fig. 2 for an example). In other words, the MDCT

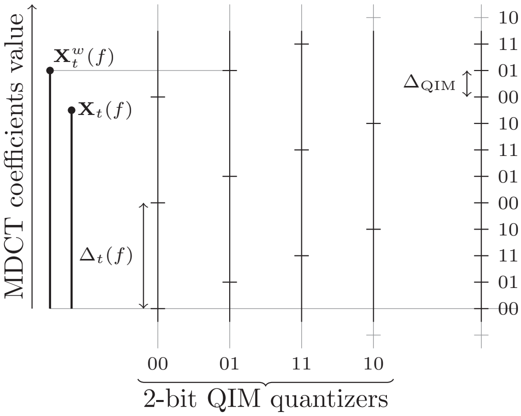


Fig. 2. Set of QIM quantizers $\mathcal{S}(\mathbf{C}_t(f))$ for $\mathbf{C}_t(f) = 2$. The 2-bit Gray codes that index the quantizers correspond to the elementary messages that can be embedded into a MDCT coefficient $\mathbf{X}_t(f)$. For example, the binary code 01 is embedded into $\mathbf{X}_t(f)$ by quantizing it to $\mathbf{X}_t^w(f)$ using the quantizer indexed by 01. The levels of the 4 quantizers are gathered on a single equivalent grid on the right.

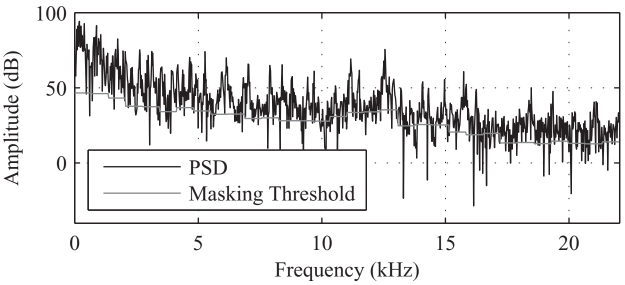


Fig. 3. Example of a masking threshold given by the PAM with frame length $N = 2048$.

coefficient $\mathbf{X}_t(f)$ is replaced by its closest code-indexed quantized value: $\mathbf{X}_t^w(f) = \mathcal{Q}_c\left(\mathbf{X}_t(f)\right)$.

The decoding principle is also very simple: if the capacity $\mathbf{C}_t(f)$ is known at the decoder, the set of quantizers $\mathcal{S}(\mathbf{C}_t(f))$ is generated (and is the same as the one generated at the embedder). Then, the quantizer $\mathcal{Q}_c$ with the quantization level that is the closest to the received embedded MDCT coefficient $\widehat{\mathbf{X}}_t^w(f)$ is selected, and the decoded message is the index c of the selected quantizer.

Obviously if one wants to transmit a large binary message $\mathbf{m}$, it has to be previously split into sub-messages $\mathbf{m}_t$ that are embedded into the corresponding frame. In each frame, $\mathbf{m}_t$ has to be spread across the different MDCT coefficients according to the local capacity values (Block ④), 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.

3.3 Psychoacoustic Model (PAM)

The PAM used in our system (Block ②) is directly inspired from the PAM of the MPEG2-AAC standard [15], with some adaptations allowing the user to adjust the frame length N. The output of the PAM is a masking threshold $\mathbf{M}_t$, which represents the maximum power of the quantization error that can be introduced while ensuring inaudibility. The calculations are made in the time-frequency domain, however the transform used for the computations inside the PAM is not the MDCT but the Discrete Fourier Transform (DFT). The main computations consist first in a convolution of the DFT power spectrum of the host signal with a spreading function that models elementary frequency masking phenomenons to obtain a first masking curve. This curve is then adjusted according to the tonality of the signal[2], and the absolute threshold of hearing is integrated. After that, some pre-echo control is applied, resulting in the DFT masking threshold (see Fig. 3 for an example). From the DFT spectrum and the DFT masking threshold a Signal-to-Mask Ratio (SMR) is computed (for each frequency bin f). This SMR is then used to obtain the MDCT masking threshold $\mathbf{M}_t$ (by simply computing the ratio between the MDCT power spectrum coefficients and the SMR coefficients). This masking threshold $\mathbf{M}_t$ is then used to shape the embedding noise (under this curve) so that it remains inaudible. Note that in order to control the embedding rate, it is possible to adjust the masking threshold $\mathbf{M}_t$ by translating it by a factor α (in dB) (see Sec. 3.5).

An important characteristic of the MPEG2-AAC PAM is that all the intermediate parameters used in the masking threshold calculation are not defined for each frequency bin f but for "partitions." In MPEG2-AAC, the partitions are approximately equal to the minimum between a third of a Bark-scale critical band [29] and a frequency bin in order to achieve good quality. The MPEG2/4-AAC standard uses different window lengths (e.g., 2048 and 256 time-samples for long windows and short windows respectively in MPEG2-AAC), and the corresponding partition limits are saved in tables. In order to ensure the adaptability of our system to different window lengths N, an algorithm computing the partitions for a given length N has been developed (eligible values for N being powers of 2). This algorithm simply computes the partitions limits starting from frequency bin 0 and choosing for each partition the size (in number of frequency bins) that is the closest to a third of a critical band (using the analytical expression for the conversion Bark/Hertz given in [26]).

3.4 Computation of the Capacities

In the proposed system three sets of parameters have to be set: the capacities $\mathbf{C}_t(f)$, the step sizes of the QIM quantizers $\Delta_t(f)$, and the minimum distance between two different QIM quantizers levels Δ_{QIM} (see Fig. 2). However, due to the regular intertwining of the QIM quantizers, those parameters are linked by the fundamental relation:

$$\Delta_t(f) = 2^{\mathbf{C}_t(f)} \cdot \Delta_{\mathrm{QIM}} \tag{9}$$

[2] The main reason why the PAM of the MPEG2-AAC works with the DFT and not the MDCT is because the phase information given by the DFT can be used to estimate the tonality of the signal in a better way than it is possible with the MDCT.

and thus only two parameters have to be set. In order to set those parameters two constraints have to be taken into account:

- *Robustness*: the data hiding process must be robust to the PCM quantization of the host audio signal. In other words, the embedded data must remain decodable from MDCT coefficients corrupted by the time-domain PCM quantization.
- *Inaudibility*: the data-hiding process must not (or only very slightly) impair the audio quality of the host signal.

The problem is thus to optimize the embedding rate under these two constraints. The robustness constraint will set Δ_{QIM}, and we will see in the following that this parameter does not depend on *t* or *f*. The inaudibility constraint will then set the two remaining parameters.

3.4.1 Setting of Δ_{QIM} (Robustness)

Although the goal of the system is not the robustness to attacks, it must be robust to the PCM quantization of the time samples of the host signal **x**. In the present study we consider 16-bit PCM since it is a very usual format for uncompressed audio signals (e.g., it is used in audio-CD, .wav, .aiff, .flac). First, we need to know the effects of the time-domain PCM quantization of $\mathbf{x}^w$ on the TF coefficients $\mathbf{X}_t^w$. We consider the 16-bit PCM time-domain samples as integer values between -2^{15} and $2^{15}-1$. In the case of the IntMDCT there is no noise introduced by the 16-bit PCM quantization since the IntMDCT is an integer-to-integer mapping. Thus the only constraint is that the quantized IntMDCT coefficients $\mathbf{X}_t^w(f)$ remain integers, i.e.:

$$\Delta_{\mathrm{QIM}} = 1. \qquad \text{(IntMDCT)} \qquad (10)$$

For the MDCT case, we use the classical (and realistic) hypothesis that the quantization error $\mathbf{b}_t(n)$ introduced on the time-domain samples $\mathbf{x}_t^w(n)$ is an independent and identically distributed (i.i.d.) sequence, following a uniform distribution. Still considering the 16-bit PCM time-domain samples as integer values, the corresponding quantization step Δ_{PCM} is equal to 1. Let $\mathcal{U}(a,b)$ be the uniform distribution within $[a,b]$, then we have:

$$\forall t, \forall n \in [0, N-1], \mathbf{b}_t(n) \sim \mathcal{U}\left(-\frac{\Delta_{\mathrm{PCM}}}{2}, \frac{\Delta_{\mathrm{PCM}}}{2}\right). \qquad (11)$$

Using the Central Limit Theorem, it can be proven that the noise $\mathbf{B}_t(f)$ introduced on the MDCT coefficients $\mathbf{X}_t^w(f)$ follows a normal distribution (see Appendix) :

$$\forall t, \forall f \in \left[0, \frac{N}{2}-1\right], \mathbf{B}_t(f) \sim \mathcal{N}\left(0, \sigma^2_{\mathbf{B}_t(f)}\right). \qquad (12)$$

Moreover, when using the normalized version of the MDCT as is the case here, it can be easily shown that the variance $\sigma^2_{\mathbf{B}_t(f)}$ is equal to the variance of the PCM quantization noise in the time domain. This variance is thus independent of the frame *t* and the frequency index *f* (see Appendix):

$$\sigma^2_{\mathbf{B}_t(f)} = \sigma^2_{\mathrm{PCM}} = \sigma^2 = \frac{\Delta^2_{\mathrm{PCM}}}{12}. \qquad (13)$$

In summary, the effect of the time-domain PCM quantization on the MDCT coefficients can be modeled as an Additive White Gaussian Noise (AWGN). Thus on first approximation the minimum distance Δ_{QIM} between two levels of the set of quantizers $\mathcal{S}(\mathbf{C}_t(f))$ can be set to achieve an expected error ratio p_e:

$$\Delta_{\mathrm{QIM}} = 2 \cdot \sqrt{2\sigma^2}\,\mathrm{erf}^{-1}(1-p_e), \qquad \text{(MDCT)} \qquad (14)$$

with erf the usual error function:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \qquad (15)$$

This expected error ratio p_e is not exactly an expected BER, it is rather a Symbol Error Rate (SER), each symbol being the data embedded in one MDCT coefficient and thus of variable size. The BER should thus be quite lower than p_e. Comparisons between theoretical SER and BER and their estimated values will be discussed in Sec. 4.

3.4.2 Calculation of $\mathbf{C}_t(f)$ (Inaudibility)

The inaudibility constraint is guided by the masking threshold $\mathbf{M}_t$ provided by the PAM. Specifically, the constraint is that the power of the embedding error in the worst case remains under the masking threshold $\mathbf{M}_t$. As the embedding is performed by quantization, the embedding error in the worst case is equal to half the quantization step $\Delta_t(f)$, which is directly related to $\mathbf{C}_t(f)$ through Eq. (9). Thus the inaudibility constraint in a given TF bin can be written as:

$$\left(\frac{\Delta_t(f)}{2}\right)^2 < \mathbf{M}_t(f). \qquad (16)$$

For a given frame *t*, we simply combine Eq. (9) and Eq. (16) to obtain for each $f \in \left[0, \frac{N}{2}-1\right]$:

$$\mathbf{C}_t(f) < \frac{1}{2}\log_2\left(\frac{\mathbf{M}_t(f)}{\Delta^2_{\mathrm{QIM}}}\right) + 1. \qquad (17)$$

Since the capacity per coefficient is an integer number of bits, and we want to maximize this capacity, we choose:

$$\mathbf{C}_t(f) = \left\lfloor \frac{1}{2}\log_2\left(\frac{\mathbf{M}_t(f)}{\Delta^2_{\mathrm{QIM}}}\right) + 1 \right\rfloor. \qquad (18)$$

where $\lfloor . \rfloor$ denotes the *floor* function. Recall that in the MDCT case, Δ_{QIM} is given by Eq. (14), whereas in the IntMDCT case $\Delta_{\mathrm{QIM}} = 1$. Experimentally, the resulting values are always lower than 15.[3] Thus those values are

[3] 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 PCM coding of time-domain samples. However, as detailed in the results section, all MDCT coefficients cannot carry such a large amount of embedded information.

coded with 4-bit codewords (from 0 to 15), in order to transmit them as side-information (Block ④).

3.4.3 Subband Processing

Embedding 4 bits of side-information per frequency bin is not appropriate as it would require 176.4 kpbs/c of embedding bit rate (44100 MDCT coefficients per second × 4 bits) lost for the "useful" information $\mathbf{m}$. For this reason, *embedding subbands* are defined as groups of adjacent frequency bins where the capacities $\mathbf{C}_t(f)$ are fixed to the same value.[4] The capacity value within each subband b, denoted $\widetilde{\mathbf{C}}_t(b)$, is given by applying Eq. (18) using the minimum value of the mask within the subband. Preliminary experiments have shown that equally spaced subbands give the best results (in particular when compared to log-scale subbands such as the Bark scale). To further simplify the implementation, a subband size of $N_b = 32$ bins was chosen: $\forall t, \forall b \in [0, N/64 - 1]$,

$$\widetilde{\mathbf{C}}_t(b) = \min_{f \in [bN_b, (b+1)N_b - 1]} \mathbf{C}_t(f). \tag{19}$$

In this case, the message $\mathbf{m}$ can be seen as a round number of 32-bit words, and each frame contains a round number of those words. This way the bit rate needed to transmit the capacities is reduced to about 5.5 kbps/c, which is reasonable given that the targeted embedding bit rates are around hundreds of kbps/c. This side-information is completed with error correcting codes and synchronization information (see Sec. 3.6), resulting in a total side-information bit rate of less than 10 kbps/c.

Now that the side-information is small enough to be embedded in the host signal in addition to the "useful" information $\mathbf{m}$, a fixed embedding "subchannel" must be chosen to embed it, so that it can be retrieved at the decoder without recalculating the PAM while remaining inaudible. This embedding subchannel dedicated to the side-information is chosen as the LSB of the QIM in the highest frequencies of each frame. This is possible for two reasons:

- Because the QIM quantizers are intertwined, the QIM enables hierarchical/scalable decoding. Indeed, if a coefficient is embedded with a capacity of $\mathbf{C}_t(f)$ bits, there is no need to know the value of $\mathbf{C}_t(f)$ to decode the C_{SI} LSB (assuming of course that $C_{\mathrm{SI}} \leq \mathbf{C}_t(f)$). This is illustrated in Fig. 4 for a 2-bit code and 1 LSB, and it can be easily generalized to larger code and LSB sizes.
- The absolute threshold of hearing is very high in the high frequency region, particularly at 44.1 kHz sampling frequency. This allows us to set the number of LSB dedicated to side-information embedding to up to 3 per MDCT coefficient, while ensuring inaudibility with a fair margin.

The exact configuration depends on the frame length N, but is arbitrarily fixed for each N value (number of embedding subbands for side-information embedding, and number of LSB used). For example, for $N = 2048$, the bit rate for the capacities is 5.5 kbps/c, the total side-information bit rate is 6.9 kbps/c corresponding to 160 bits per frame, hence the number of subbands concerned by the side-information embedding is 2, with respectively 2 and 3 LSB for subbands 30 and 31 respectively (see Fig. 5).

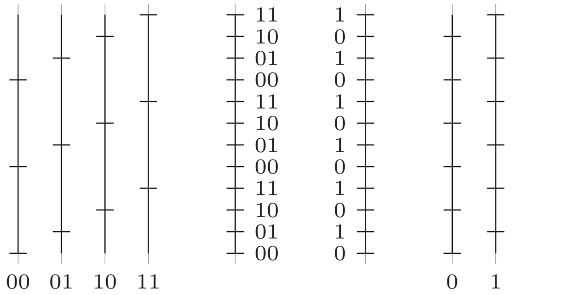


Fig. 4. Example of relation between QIM quantization with 2 bits and 1 bit. There is no need to know the number of bits used on the left to decode the last 1 bit of information. Note that in this case a Gray code must not be used for the LSB.

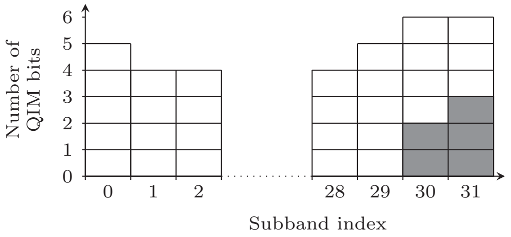


Fig. 5. Example of QIM bit allocation for the side-information (in gray). See the text for details.

The decoding of a frame is then done by:

- Decoding of the side-information in the LSB of the high frequency subbands; this provides the decoded capacities $\widehat{\mathbf{C}}_t(b)$.
- Decoding of the "useful information" using $\widehat{\mathbf{C}}_t(b)$.

3.5 Control of the Embedding Bit Rate

The "useful" embedding bit rate R is given by the average number of embedded bits per second of signal minus the bit rate of the side-information. It is obtained by summing the capacities over the TF plan, dividing the result by the signal duration D and subtracting the side-information bit rate R_{SI}:

$$R = \frac{\sum_t N_b \sum_b \widetilde{\mathbf{C}}_t(b)}{D} - R_{\mathrm{SI}}. \tag{20}$$

It is possible to control the embedding rate by translating the masking threshold of the PAM by a scaling factor α (in dB), i.e., using the following variant of Eq. (18):

$$\mathbf{C}_t^{\alpha}(f) = \left\lfloor \frac{1}{2}\log_2\left(\frac{\mathbf{M}_t(f) \cdot 10^{\frac{\alpha}{10}}}{\Delta_{\mathrm{QIM}}^2}\right) + 1 \right\rfloor. \tag{21}$$

[4] Those subbands are similar to the coding subbands used in compression: for each coding band, only one quantizer is used.

Similarly to the rate-distortion theory of source coding signal quality is expected to decrease as embedding rate increases and vice-versa. When $\alpha > 0$ dB, the masking threshold is raised. Larger values of the quantization error allows for larger capacities (and thus higher embedding rate), at the price of potentially lower quality. At the opposite, when $\alpha < 0$ dB, the masking threshold is lowered, leading to a “safety margin” for the inaudibility of the embedding process, at the price of lower embedding rate. An end-user of the proposed system can thus look for the best trade-off between rate and quality for a given application.

Let us denote by R^α the embedding rate corresponding to a translation of αdB. It can be easily shown that Eq. (21) leads to the following relationship between R^α and the basic rate $R = R^0$:[5]

$$R^\alpha \simeq R + \alpha \cdot \frac{\log_2(10)}{20}. \qquad (22)$$

This linear relation enables to easily control the embedding rate by the setting of α. Alternately, if the end-user wants to embed a given number of 32-bit codewords in the host signal $\mathbf{x}$, it is possible to translate the masking threshold “exactly” in order to reach the desired payload. This should guarantee that for a given payload, the embedding is done in the best possible way from a psychoacoustic point of view. Obviously, raising the masking threshold by too large a value in order to heavily increase the payload means that the user accepts potentially audible degradations.

3.6 Synchronization

Although we have mentioned that the proposed system is not intended to be robust to attacks, we have to mention that synchronization errors can occur and must be dealt with. We address here two special cases that are important, stand-alone and global data.

3.6.1 Stand-Alone

In this case, the message embedded in each frame is *stand-alone* and related to its host frame only. The message embedded in a given frame must be decoded without having to decode from the beginning of the musical signal. Thus the problem is to know exactly where the embedding frames within the signal are located. In the present study we propose to simply add a checksum (similarly to what is proposed in [14]) located at the same place as the transmitted $\widetilde{\mathbf{C}}_t(b)$ values. The strategy at the decoder is then the following: the side-information from the current frame is decoded and the checksum calculated. If it is different from the checksum embedded within the side-information, the frame is shifted by 1 time-domain sample, and this process is repeated until the computed checksum corresponds to the embedded one. For more robustness, several adjacent frames can be tested instead of only one. However testing many adjacent frames can hinder “real-time” decoding.

Table 1. Perceptual interpretation of ODG/SDG values.

ODG/SDG	Impairment description
0.0	Imperceptible
–0.1 to –1	Perceptible, but not annoying
–1.1 to –2	Slightly annoying
–2.1 to –3	Annoying
–3.1 to –4	Very annoying

3.6.2 Global Data

In this case, the embedded message is quite large and embedded in the whole music signal. The number of decoded bits has to be the same as the number of embedded bits. This is a crucial issue in the presented system (particularly when using the classical MDCT) due to the double decoding process: if an error occurs in the decoding of the capacity values then the number of bits of the decoded message $\mathbf{m}_t$ can be wrong. To overcome this problem we add additional information to be transmitted with the capacity values: the number of 32-bit codewords embedded in the previous frames p_t and the number of 32-bit codewords embedded in the next frames n_t. The strategy at the decoder is the following: the side-information is decoded for the whole signal. Then for each frame the number of decoded bits is added with n_t and p_t. Those sums should be identical for all the frames. The frames where the sum is different are frames where an error has occurred. It is possible to know how many bits were embedded in this frame and thus the missing entries can be filled with arbitrary values (for example zeros).

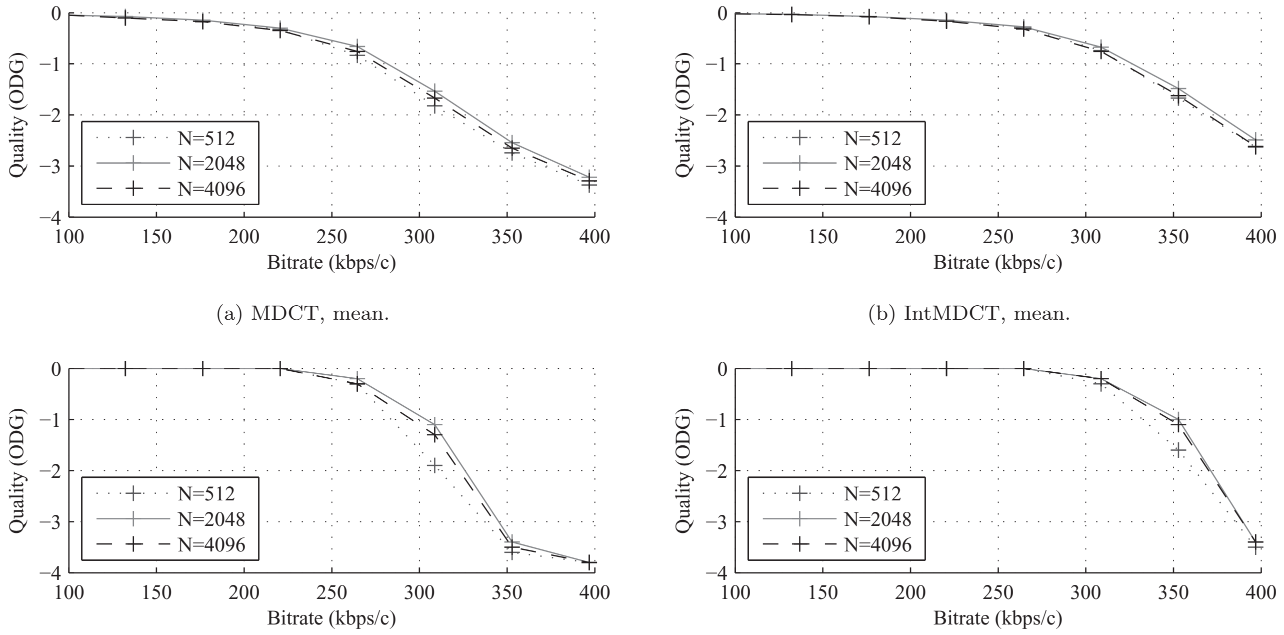
Note that in both *stand-alone* and *global data* cases, the fixed embedding location is protected by a BCH code [23].

4 EXPERIMENTS

4.1 Data and Experimental Settings

The main data set used for our experiments, *data1*, consists of 96 stereo 30-second duration excerpts (i.e., 48 minutes of stereo music) taken from commercial releases of various musical styles (pop, rock, jazz, classical, folk, reggae, latino, and rap). In Sec. 4.2 we first check the BER and the efficiency of the synchronization strategies. Then the results are presented as quality-rate curves in Secs. 4.3 and 4.4. Since there are many signals and many parameters (MDCT and IntMDCT, frame length, embedding bit rate), it was not possible to perform subjective listening tests for all the combinations. We first performed extensive objective measurements using the PEAQ algorithm [17] (the basic version was used). This algorithm compares the original and the modified signal and returns an Objective Difference Grade (ODG), which perceptual interpretation is given in Table 1. Then we conducted formal subjective listening tests on a reduced second data set, *data2* to confirm the reliability of the PEAQ measures in Sec. 4.5. This second data set consists of 8 stereo 10-second duration excerpts of the same different musical styles that were

[5] Actually, the approximation is an exact equality for α multiple of $10\log_{10}(4)$, and we have checked that the approximation is very good, since the embedding rate results from the averaging on a large number of capacity values.



(a) MDCT, mean.

(b) IntMDCT, mean.

(c) MDCT, median.

(d) IntMDCT, median.

Fig. 6. Quality-rate curves of the proposed embedding system for the MDCT (with $p_e = 10^{-4}$) (left) or the IntMDCT (right). Quality is expressed in terms of average ODG (top) or median ODG (bottom) calculated on the complete dataset *data1* (48 mn of stereo music of 8 different styles).

Table 2. Theoretical value, experimental value, and confidence intervals for the BER and SER. The confidence interval used is Wilson's confidence interval [4, 28].

Quantity	Theoretical	Estimated	CI (5%)
SER	10^{-6}	$0.96 \cdot 10^{-6}$	$[0.88, 1.04]10^{-6}$
BER	–	$1.54 \cdot 10^{-7}$	$[1.41, 1.68]10^{-7}$

deemed appropriate to test the limits of the system (e.g., strong percussive sounds).

4.2 BER and Synchronization

4.2.1 BER

In the case of MDCT, we made the following experiment to check that the experimental BER/SER corresponds to the theoretical setting of Eq. (14). Here, we set $p_e = 10^{-6}$. Assuming correct synchronization, we transmitted about $n_b = 3.2 \cdot 10^9$ bits of data, distributed among about $n_c = 5.3 \cdot 10^8$ MDCT coefficients. As can be seen in Table 2, the obtained SER experimental value $\widehat{SER}$ is very close to the theoretical one (the theoretical SER is inside the 5% confidence interval of the estimate), which confirms the relevance of the approximation that the noise on the MDCT coefficients is an AWGN. Moreover, we have $\widehat{BER} \cdot n_b/n_c \simeq \widehat{SER}$, which means that one erroneous symbol generally leads to only one erroneous bit.

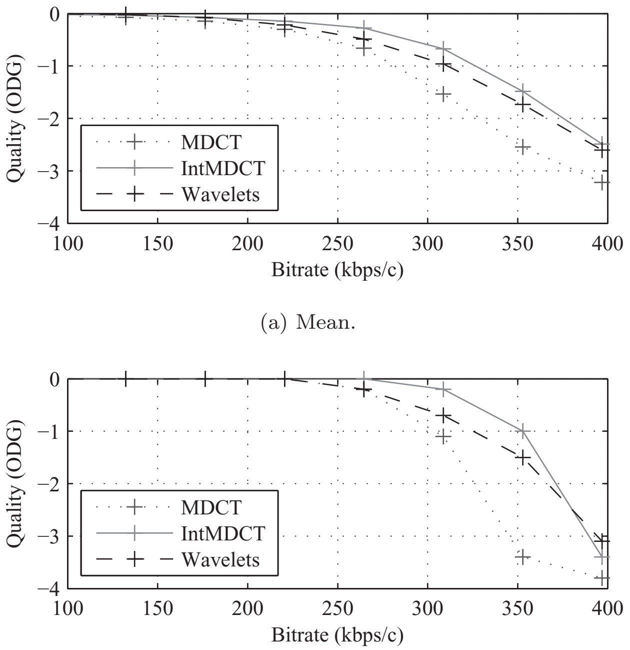
As for the IntMDCT case, as said before, because the IntMDCT is an integer-to-integer mapping there is no decoding error and thus both the theoretical and experimental BER and SER are all 0.

4.2.2 Synchronization

For both MDCT and IntMDCT, we checked the efficiency of the proposed strategy for the synchronization of embedding frames. We performed the decoding of about 80000 frames of the dataset *data1* (out of about 250000 frames) with a frame misalignment taking uniformly distributed random values within $[1, N/2 - 1]$. The checksum strategy allowed to recover frame synchronization in all cases for the IntMDCT and in all but two cases for the MDCT. Such re-synchronization errors can be due to two factors: the checksum can happen to be correct even though the frame is still not aligned; and conversely even if the frame is correctly aligned errors due to the PCM quantization can corrupt the checksum (in the MDCT case only). However, those errors happen very rarely and a multiple frame re-synchronization strategy can fix this problem (at the price of increased computational cost).

4.3 Quality-Rate Curves

In this subsection we report the results that we obtained in terms of (PEAQ) ODG, averaged on the complete dataset *data1*, for both MDCT and IntMDCT transforms, for different frame lengths N, and 8 different embedding bit rates approximately ranging from 100 to 400 kbps/c. Those bit rates were chosen to be multiples of 44.1 kbps/c to ease the comparison with the system of [10] in Sec. 4.4 and were obtained by appropriately setting the value of α in Eq. (21). The tested frame lengths were 256, 512, 1024, 2048, and 4096. The results are shown in Fig. 6, only for $N = 512$, 2048, 4096 for clarity, but the results for $N = 256$ and 1024 are consistent.



(a) Mean.

(b) Median.

Fig. 7. Quality-rate curves for the proposed data hiding system with frame length 2048, for both MDCT (with $p_e = 10^{-4}$) and IntMDCT and for the reference system of [10]. Average (top) and median ODG (bottom) calculated on dataset *data1*. Bit rates are set every 44.1 kbps/c, from 88.2 kbps/c to 396.9 kbps/c.

First, it can be noted that each curve follows the same expected general trend: it is first constant at an ODG of 0 or close to 0 and then monotonically decreases. Low embedding bit rates do not impair the signal quality. Then the modifications become audible and quality drops as bit rate increases. For the MDCT, the median maximum bit rate for an ODG of 0 (no impairment) is around 220 kbps/c. The corresponding average ODG value is about –0.3. For the IntMDCT the median maximum bit rate for an ODG of 0 (no impairment) is around 265 kbps/c. The corresponding average ODG value is also about –0.3. Thus the IntMDCT seems to be systematically more efficient than the MDCT for QIM-based data embedding. This can be explained by the fact that for this experiment p_e is set to 10^{-4}. Thus, using Eq. (14) we can see that for the MDCT $\Delta_{\text{QIM}} = 2.25$ whereas $\Delta_{\text{QIM}} = 1$ for the IntMDCT Eq. (10). The fact that Δ_{QIM} in the MDCT case is about twice as large as in the IntMDCT case means that about 1 more bit can be embedded at each MDCT coefficient, thus the embedding bit rate should be greater for the IntMDCT by about 44.1 kbps/c. This can be verified in Fig. 6 (and more easily in Fig. 7). Note that to achieve $\Delta_{\text{QIM}} = 1$ for the MDCT, p_e would have to be set to around 10^{-2} which is quite a low SER.

Second, for both MDCT and IntMDCT, at a given bit rate, the quality increases as the frame length increases, up to 2048 and then decreases for 4096. The increasing trend from 256 to 2048 can be explained by two factors:

1. The frequency resolution is very important for the accuracy of the PAM, and increasing the frequency resolution is done by increasing the frame length.
2. The MDCT coefficients are split into embedding subbands of 32 coefficients. The smaller the frame length, the larger a subband (in Hz), and thus the coarser the masking curve. So when the frame length is small the accuracy of the PAM is low.

Table 3. Embedding bit rates given by the basic setting of the PAM and maximum bit rates for an ODG of 0, for the 8 excerpts of *data2* and for both MDCT and IntMDCT.

	Bit Rates (kbps)			
	MDCT		IntMDCT	
Excerpt	PAM	ODG = 0	PAM	ODG = 0
pop1	270	260	336	340
rock	356	360	422	420
rap	270	260	335	330
folk1	268	260	333	340
clas1	265	240	332	300
clas2	164	140	225	200
folk2	234	230	298	300
pop2	253	240	318	320

As for the drop in performance for 4096, this can be explained by the fact that, at a sampling frequency of 44.1 kHz and for some rapidly varying music signals, this frame length (96 ms) can be too long for a time-frequency analysis based on the local stationarity assumption. Indeed, within such a long frame, the human auditory system can sometimes separate the temporal activations of some sounds; and the PAM will apply an irrelevant frequency masking model to those sounds. The fact that the frame length of 2048 shows the best behavior is not a surprise, as it is the length commonly used for the MDCT in PAC (for example it is the basic frame length for MPEG2-AAC [15]). For the rest of the experiments, we set $N = 2048$.

Finally, it can be noted that the basic setting of the PAM (Eq. (18), or $\alpha = 0$ in Eq. (21)) corresponds quite well to the assumed limit for signals high-quality (ODG = 0). To check this we made the following complementary experiment. Each one of the 8 excerpts of dataset *data2* have been first embedded at the bit rate given by the basic setting of the PAM. We found ODG values very close to 0. We then modified the α value and used the PEAQ algorithm to find for each excerpt the maximum embedding bit rate ensuring ODG=0. The initial and modified bit rates are given Table 3. It can be noted that for the majority of the excerpts, the initial bit rate is very close to the maximum bit rate with an ODG of 0. This means that the basic setting of the PAM is appropriate to provide embedded signals without quality impairments in most cases. Furthermore, this setting is close to the limit for quality preservation.

4.4 Comparison with State-of-the-Art System

The performance of our system were compared with the performance of the system of Cvejic et al. [10], as the aim of this system was quite similar (high embedding bit rate, no particular robustness constraint). Their system works as follows:

1. The signal is split into frames of 512 samples.
2. Each frame is transformed using the Haar wavelet transform.
3. Data are embedded within the wavelet coefficients using the LSB scheme with a fixed number of bits (i.e., this number is the same for all the frames and coefficients; values in the range 2–9 are tested in the present study, corresponding to bit rates within the approximate range 100–400 kbps/c with 44.1 kbps/c spacing).
4. The signal is reverted back in the time-domain and PCM quantized.

The BER of the wavelet system is approximately 10^{-4}, therefore its performance can be compared with the ones of our MDCT system (with $p_e = 10^{-4}$), and of course with the ones of our IntMDCT system, presented in the previous section. The comparative results are given in Fig. 7.

In a general manner, the ODGs for the wavelet system are in between the ODGs of the IntMDCT and the MDCT systems. The wavelet system sticks more closely to the MDCT system for bit rates below 250 kbps/c (especially for the median ODG) and sticks more closely to the IntMDCT system for bit rates above 300 kbps/c. Except for median ODG at about 400 kbps/c, which is an irrelevant setting that corresponds to very low signal quality, the IntMDCT system outperforms the wavelet system within a range of approximately 10 to 50 kbps/c (depending on bit rate and mean/median measure). Note that the maximal difference between the IntMDCT system and the wavelet system occurs within the relevant range of bit rate (approximately 200–300 kbps/c) where the ODG obtained with the IntMDCT system is higher than −0.5.

Even if the MDCT system seems to perform less efficiently than the wavelet system, a major advantage of both the MDCT and IntMDCT systems compared to the wavelet system is the fact that the basic setting of the PAM enables for an automatic optimal setting of the embedding bitrate that ensures high quality of the embedded signals, as explained at the end of Sec. 4.3. Moreover, this quality is guaranteed for the whole signal. In contrast, there is no PAM for the control of the wavelet system, at least as proposed in [10]. Therefore there is no possibility to know beforehand how many bits can be used to embed data in the wavelet coefficients without quality impairments, hence it is very difficult to maximize the embedding bit rate. This is very problematic for long sequences of music; because the embedding setting is not adapted to the signal content, we observed that when the energy of the signal is low the embedding can be clearly audible. The proposed system (more particularly the IntMDCT system but also the MDCT system) yields better results and is easier to use when the user wants high embedding bit rates without quality impairments for long non-stationary audio sequences (which is the case for most music signals). Moreover, recall that the possibility to control the bit rate/quality trade-off through the setting of α makes our system particularly flexible.

4.5 Validation of the PEAQ Algorithm

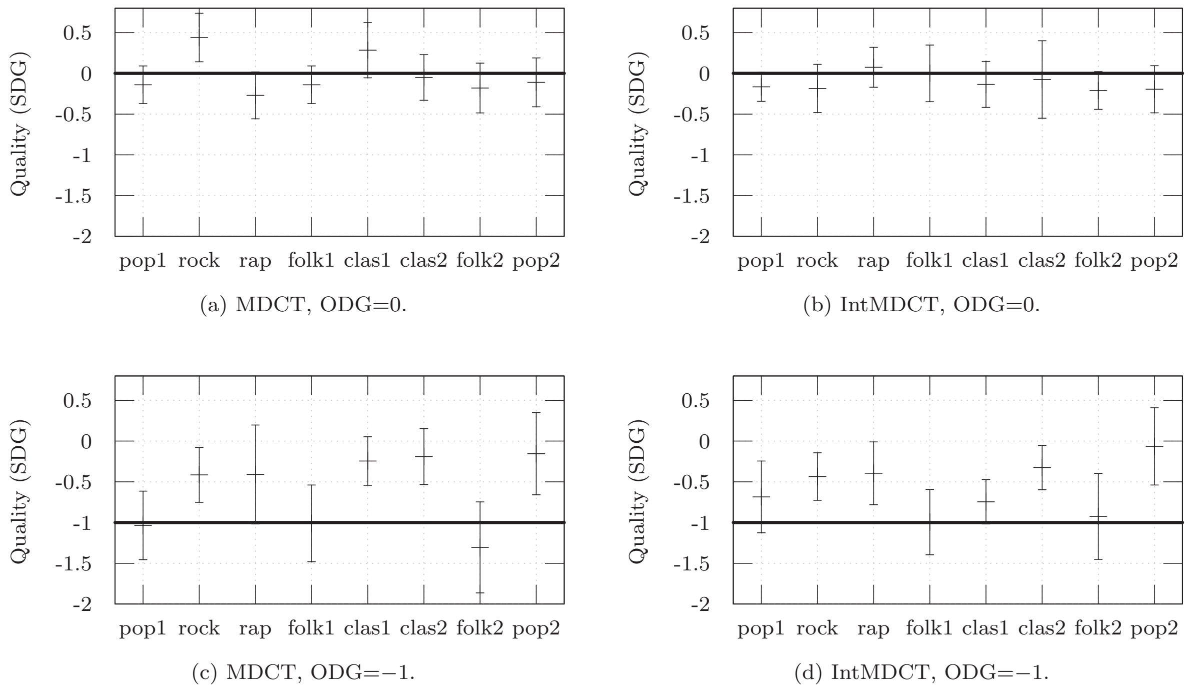
The PEAQ algorithm was not initially designed for data hiding techniques. A subjective listening test was thus performed using dataset *data2* to confirm the results reported above.

The experimental protocol for the subjective listening test was the following: for each excerpt, and for both the MDCT and the IntMDCT (frame length 2048), the PEAQ algorithm was used to find the highest embedding bitrates giving ODGs of 0 and −1. The resulting 32 sound samples (8 10-second excerpts × 2 transforms × 2 target ODGs) were then evaluated by listeners according to the ITU recommendation [16], i.e., a double-blind triple stimuli test. The subjects had a training phase during which they could listen to 4 samples of different ODGs (as many times as they wanted to) to make them familiar with the effects of the data hiding system. Then they had to grade the 32 test samples within ODG/SDG scale of Table 1. Twenty subjects performed the test but only 11 were validated by t-test post-screening [16] as the differences were quite hard to detect.

The resulting Subjective Difference Grades (SDG) are given Fig. 8. For a target ODG of 0, for both the MDCT and the IntMDCT, the ODG and the SDG seem to be quite coherent. The difference between the SDG mean value and the target ODG is quite small: it is generally lower than 0.25 in absolute value. Although the corresponding medians are not shown, it can be noted that the difference between the SDG median value and the target ODG is always zero. All these results mean that when the PEAQ algorithm gives an ODG of 0, the difference is very likely to be inaudible. For a target ODG of −1, for both the MDCT and the IntMDCT, the results seem slightly less constant among the excerpts. However, except for *folk2*, the SDG values are all higher than the target ODG, which seems to indicate a “secure margin” for objective evaluation with PEAQ in our experiments, and thus strongly supports the use of this algorithm.

5 CONCLUSION AND PERSPECTIVES

The data hiding technique presented in this paper enables to embed data in PCM audio signals with adjustable embedding rate while ensuring a very good quality even for high embedding rates (up to 250–300 kbps/c depending on the musical content). The best results are obtained with the IntMDCT transform and outperform a reference system based on wavelet transform. This system can be used in “enriched-content” applications to provide additional features to a given audio media. As for perceptual audio coding, the PAM that guarantees the quality of the embedded signal is used only at the coder, and the computational cost of the decoder is very low. Therefore, this system can be used in real-time applications (for the decoding part). For example, the decoder has been integrated in the real-time



(a) MDCT, ODG=0.

(b) IntMDCT, ODG=0.

(c) MDCT, ODG=−1.

(d) IntMDCT, ODG=−1.

Fig. 8. Mean SDG with 95% confidence interval for the subjective listening test on 8 excerpts of different musical styles, for the MDCT system (left) and the IntMDCT system (right), and for target ODG $= 0$ (top) and target ODG $= -1$ (bottom). The frame length is 2048.

C/C++ implementation of the Informed Source Separation (ISS) system presented in [22]. In this application the data hiding system is used to embed in a music signal the codes that identify the predominant source signals (instruments and voices) in each bin of the TF plan, so that the source signals can be separated by a local mixture inversion process. The necessary embedding rate is here lower than 64 kbps/c, hence the inaudibility of the embedding process is guaranteed, and there is room for more voluminous information in the future improvements of the ISS system. Because the source separation is carried out in the MDCT domain, this ISS system is a good example of appropriate compliance between the proposed MDCT-based embedding system and target application.

In further works we will try to improve the proposed embedding system by improving the PAM, particularly the pre-echo phenomenon, and improving the embedding subbands distribution to gain in bit rate and quality.

6 ACKNOWLEDGMENTS

This work is supported by the French National Research Agency (ANR) as part of the DReaM project (ANR 09 CORD 006).

REFERENCES

[1] T. Bliem, G. Galdo, J. Borsum, A. Craciun, and R. Zitzmann “A Robust Audio Watermarking System for Acoustic Channels,” *J. Audio Eng. Soc.*, vol. 61, pp. 878–888 (2013 Nov.).

[2] L. Boney, T. Ahmed, and H. Khaled “Digital Watermarks for Audio Signals,” *Third IEEE Int. Conf. on Multimedia Computing and Systems*, pp. 473–480 (1996 June).

[3] K. Brandenburg and M. Bosi, “Overview of MPEG Audio: Current and Future Standards for Low Bit-Rate Audio Coding,” *J. Audio Eng. Soc.*, vol. 45, pp. 4–21 (1997 Jan./Feb.).

[4] L. D. Brown, T. T. Cai, and A. A. DasGupta “Interval Estimation for a Binomial Proportion,” *Statistical Science*, vol. 16, no. 2, pp. 101–133 (2001).

[5] 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.*, vol. 48, no. 10, pp. 1634–1637 (2000).

[6] B. Chen and G. Wornell, “Quantization Index Modulation: A Class of Provably Good Methods for Digital Watermarking and Information Embedding,” *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1423–1443 (2001).

[7] M. Costa “Writing on Dirty Paper,” *IEEE Trans. Inform. Theory*, vol. 29, no. 3, pp. 439–441 (1983).

[8] I. J. Cox, M. L. Miller, and A. L. McKellips “Watermarking as Communications with Side Information,” *Proc. IEEE*, vol. 87, no. 7, pp. 1127–1141 (1999).

[9] N. Cvejic and T. Seppänen, “Increasing the Capacity of LSB-Based Audio Steganography,” IEEE Workshop on Multimedia Signal Processing, pp. 336–338 (2002).

[10] N. Cvejic and T. Seppänen “A Wavelet Domain LSB Insertion Algorithm for High Capacity Audio Steganography,” *IEEE Digital Signal Processing Workshop*, pp. 53–55 (2002).

[11] I. Daubechies and W. Sweldens, “Factoring Wavelet Transforms into Lifting Steps,” Technical report, Bell Laboratories, Lucent Technologies (1996).

[12] W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, 1971).

[13] R. Geiger, J. Herre, J. Koller, and K. Brandenburg "IntMDCT—A Link between Perceptual and Lossless Audio Coding," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, pp. II–1813 –II–1816 (May 2002).

[14] R. Geiger, Y. Yokotani, and G. Schuller, "Audio Data Hiding with High Data Rates Based on IntMDCT," IEEE Int. Conf. on Acoustics, Speech and Signal Processing (2006).

[15] ISO/IEC JTC1/SC29/WG11 MPEG, "Information Technology—Generic Coding of Moving Pictures and Associated Audio Information—Part 7: Advanced Audio Coding (AAC)," IS13818-7(E) (2004).

[16] ITU-R, "Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems," Recommendation BS.1116-1 (1994–1997).

[17] ITU-R, "Method for Objective Measurements of Perceived Audio Quality (PEAQ)," Recommendation BS.1387-1 (2001).

[18] K. Kondo "A Data Hiding Method for Stereo Audio Signals Using Interchannel Decorrelator Polarity Inversion," *J. Audio Eng. Soc.*, vol. 59, no. 6, pp. 379–395 (2011).

[19] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard "Informed Source Separation through Spectrogram Coding and Data Embedding," *Signal Processing*, vol. 92, no. 8 (2012).

[20] S. Marchand, R. Badeau, C. Baras, L. Daudet, D. Fourer, L. Girin, S. Gorlow, A. Liutkus, J. Pinel, G. Richard, N. Sturmel, and S. Zhang "DReaM: A Novel System for Joint Source Separation and Multi-Track Coding," presented at the *133rd Convention*of the Audio Engineering Society (2012 Oct.), convention paper 8737.

[21] T. Painter and A. Spanias "Perceptual Coding of Digital Audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–515 (2000 April).

[22] M. Parvaix and L. Girin "Informed Source Separation of Underdetermined Instantaneous Stereo Mixtures Using Source Index Embedding," *IEEE Int. Conf. Acoust. and Speech, Signal Process. (ICASSP)*, Dallas, Texas (2010).

[23] W. W. Peterson and E. J. Weldon, *Error-Correcting Codes* (The MIT Press, 1972).

[24] J. P. Princen and A. B. Bradley "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 5, pp. 1153–1161 (1986).

[25] I. Samaali, G. Mahé, and M. Turki "Watermark-Aided Pre-Echo Reduction in Low Bit-Rate Audio Coding," *J. Audio Eng. Soc.*, vol. 60, pp. 431–443 (2012 June).

[26] H. Traunmüller "Analytical Expressions for the Tonotopic Sensory Scale," *J. Acoust. Soc. Am.*, vol. 88, no 4, pp. 97–100 (1990).

[27] Z. Wang "Fast Algorithms for the Discrete W Transform and for the Discrete Fourier Transform," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 32, no. 4, pp. 803–816 (1984 Aug.).

[28] E. B. Wilson "Probable Inference, the Law of Succession, and Statistical Inference," *J. Am. Stat. Assoc.*, vol. 22, no 158, pp. 209–212 (1927).

[29] E. Zwicker and U. Zwicker, *Psychoacoustics: Facts and Models* (Springer-Verlag, 1990).

APPENDIX: PCM NOISE IN THE MDCT DOMAIN

We use the same notations as defined in the main text. The following equations are valid for all frame indexes t and frequency bins $f \in \left[0, \frac{N}{2} - 1\right]$, and when relevant, for all sample indexes $n \in [0, N-1]$. Recall that MDCT and IMDCT equations are given by Eq. (1) and Eq. (2) and let us denote $c(n, f) = \cos\left(\frac{2\pi}{N} n' f'\right)$.

Let $\widehat{\mathbf{x}}_t(n)$ be the PCM-quantized version of $\mathbf{x}_t(n)$, and let $\mathbf{b}_t(n)$ be the corresponding quantization noise:

$$\widehat{\mathbf{x}}_t(n) = \mathbf{x}_t(n) + \mathbf{b}_t(n). \tag{23}$$

We assume that the noise samples $\mathbf{b}_t(n)$ are independent and that each sample follows the same uniform distribution with variance σ^2:

$$\mathbf{b}_t(n) \sim \mathcal{U}\left(-\tfrac{\Delta_{\text{PCM}}}{2}, \tfrac{\Delta_{\text{PCM}}}{2}\right), \tag{24}$$

$$\sigma^2 = \tfrac{\Delta_{\text{PCM}}^2}{12}. \tag{25}$$

Let $\widehat{\mathbf{X}}_t$ and $\mathbf{B}_t$ be the MDCT coefficient vectors of $\widehat{\mathbf{x}}_t$ and $\mathbf{b}_t$ respectively. Since the MDCT is a linear transform, we have:

$$\widehat{\mathbf{X}}_t = \mathbf{X}_t + \mathbf{B}_t, \tag{26}$$

Let us denote:

$$\mathbf{b}'_t(n, f) = \mathbf{b}_t(n)\mathbf{w}(n)c(n, f). \tag{27}$$

$\mathbf{B}_t(f)$ can be written

$$\mathbf{B}_t(f) = \frac{2}{\sqrt{N}} \sum_{n=0}^{N-1} \mathbf{b}'_t(n, f). \tag{28}$$

Using a variation of the Central Limit Theorem (with Lyapunov's or Lindeberg's condition, see theorem 1 in [12, p. 548]), it can be proved that:

$$\mathbf{B}_t(f) \sim \mathcal{N}\left(0, \sigma^2_{\mathbf{B}_t(f)}\right). \tag{29}$$

with

$$\sigma^2_{\mathbf{B}_t(f)} = \frac{4}{N} \sum_{n=0}^{N-1} \sigma^2_{\mathbf{b}'_t(n,f)}. \tag{30}$$

Moreover, using Eq. (24) and Eq. (27), the variance of $\mathbf{b}'_t(n, f)$ is given by:

$$\sigma^2_{\mathbf{b}'_t(n,f)} = \frac{\Delta_{\text{PCM}}^2 \mathbf{w}^2(n) c^2(n, f)}{12}. \tag{31}$$

Then it follows from Eq. (30) and Eq. (31) that:

$$\sigma^2_{\mathbf{B}_t(f)} = \frac{\Delta_{\text{PCM}}^2}{3N} \sum_{n=0}^{N-1} \mathbf{w}^2(n) c^2(n, f) \tag{32}$$

$$= \frac{\Delta^2_{\mathrm{PCM}}}{3N}\left(\sum_{n=0}^{\frac{N}{2}-1}\mathbf{w}^2(n)c^2(n,f) + \sum_{n=\frac{N}{2}}^{N-1}\mathbf{w}^2(n)c^2(n,f)\right) \tag{33}$$

$$= \frac{\Delta^2_{\mathrm{PCM}}}{3N}\sum_{n=0}^{\frac{N}{2}-1}\mathbf{w}^2(n)(c^2(n,f) + c^2(N-1-n,f)) \qquad \text{from (3)} \tag{34}$$

$$= \frac{\Delta^2_{\mathrm{PCM}}}{3N}\sum_{n=0}^{\frac{N}{2}-1}\mathbf{w}^2(n) \tag{35}$$

$$= \frac{\Delta^2_{\mathrm{PCM}}}{3N}\frac{N}{4} \qquad \text{from (3)} \tag{36}$$

$$= \frac{\Delta^2_{\mathrm{PCM}}}{12} \tag{37}$$

$$= \sigma^2. \tag{38}$$

And finally:

$$\mathbf{B}_t(f) \sim \mathcal{N}(0,\sigma^2), \tag{39}$$

which is independent from f, t and N.

THE AUTHORS

Jonathan Pinel

Laurent Girin

Cléo Baras

Jonathan Pinel was born in Vélizy-Villacoublay, France, in 1985. He received the M.Sc. and Ph.D. degrees in signal processing from the Grenoble Institute of Technology (Grenoble-INP), Grenoble, France, respectively in 2009 and 2013. His Ph.D research was carried out at laboratory GIPSA-Lab (Grenoble Image, Speech, Signal and Control Lab), focusing on watermarking for digital audio signals and more generally dealing with digital audio signal processing. During his Ph.D. he also taught signal and image processing, control engineering, and computer science at Phelma (the Physics, Electronics and Materials department of Grenoble-INP) and ENSE3 (the Water, Energy and Environment department of Grenoble-INP).

•

Laurent Girin was born in Moutiers, France, in 1969. He received the M.Sc. and Ph.D. degrees in signal processing from the Institut National Polytechnique de Grenoble (INPG), Grenoble, France, in 1994 and 1997, respectively. In 1999, he joined the Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble (ENSERG), as an Associate Professor. He is now a Professor at Phelma (Physics, Electronics, and Materials Department of Grenoble-INP), where he lectures (baseband) signal processing, from theoretical aspects to audio applications. His research activity is carried out at GIPSA-Lab (Grenoble Laboratory of Image, Speech, Signal, and Automation). It concerns different aspects of speech and audio processing (analysis, modeling, coding, transformation, synthesis, source separation, multimodal processing).

•

Cleo Baras is Associate Professor at the Department of Image and Signal of GIPSA-Lab and at the University Institute of Technology of Joseph Fourier University in Grenoble, France. She received the engineering degree from Grenoble-INP in 2002 and the Ph.D. degree from Telecom ParisTech in 2005, after completing a thesis on audio watermarking. Her research interests include (multimedia) content protection, data hiding and communication systems. She has been involved in various French and European projects, including ARTUS, MPipe, Estampille and DReaM.