A HIGH-RATE DATA HIDING TECHNIQUE FOR AUDIO SIGNALS BASED ON INTMDCT QUANTIZATION

Jonathan Pinel, Laurent Girin, *

GIPSA-Lab, Grenoble Institute of Technology,
Grenoble, France

{jonathan.pinel,laurent.girin}@gipsa-lab.grenoble-inp.fr

ABSTRACT

Data hiding consists in hiding/embedding binary information within a signal in an imperceptible way. In this study we propose a high-rate data hiding technique suitable for uncompressed audio signals (PCM as used in Audio-CD and .wav format). This technique is appropriate for non-securitary applications, such as enriched-content applications, that require a large bitrate but no particular robustness to attacks. The proposed system is based on a quantization technique, the Quantization Index Modulation (QIM) applied on the Integer Modified Discrete Cosine Transform (IntMDCT) coefficients of the signal and guided by a PsychoAcoustic Model (PAM). This technique enables embedding bitrates up to 300 kbps (per channel), outperforming a previous version based on regular MDCT.

1. INTRODUCTION

Data hiding consists in imperceptibly embedding information in a media. Theoretical foundations can be found in [1], and the first papers and applications dedicated to audio signals were developed in the 90's (*e.g.* [2, 3]). The main use (and probably original use) of data hiding for audio signals is the Digital Rights Management (DRM): the embedded data are usually copyrights or information about the author or the owner of the media content (in this case data hiding is referred to as *watermarking*). For such applications, the size of the embedded data is usually small, and a crucial issue is the robustness of the watermark to malicious attacks. Therefore, researches have long focused on enhancing the security and robustness of the data hiding techniques, at the price of limited embedding bitrate.

Data hiding is now used for non-securitary applications as well (*e.g.* [4]). In this paper we focus on "enriched-content" applications where data hiding is used 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 (and somewhat opposed): here, a high embedding bitrate is generally required to provide substantial interactive features. Therefore, the technical issue is usually to maximize the embedding bitrate under the double constraint of imperceptibility and robustness. However, in contrast to security applications, robustness is here of a lesser importance because the user has no reason to impair the embedded data.

In this paper, we focus on high-rate data hiding for uncompressed audio signals (*e.g.* 44.1kHz 16-bit PCM samples, such as audio-CD, .wav, .aiff, .flac formats), with potential application to enriched-content musical processing. For example, the so-called Informed Source Separation techniques developed in [5, 6] 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 compliance with uncompressed format (16-bit time-domain PCM).

The system presented here is an improved version of the system previously presented in [7]. As in [7], it is a quantization-based embedding scheme, the quantization technique being the Quantization Index Modulation (QIM) applied on the Time-Frequency (TF) coefficients of the signal, and the computation of the embedding capacities is guided by a PsychoAcoustic Model (PAM) to ensure inaudibility. However, the TF transform used in the present study is the IntMDCT (Integer Modified Discrete Cosine Transform, [8]), which is an integer approximation of the MDCT [9] used in our previous study [7]. The interest of this transform is to provide directly the embedded signal in the PCM format. For the same reason, it was used in [10]. However, in [10] the PAM and the capacities have to be recomputed at the decoder using the *lead bits* principle. In the present study we keep the two-step embedding process of [7] that avoids such recomputation. Altogether, the combination of such two-step embedding process with the IntMDCT yields a significant gain for the embedding bitrate.

The paper is organized as follows: Section 2 is a general overview of the system while Section 3 presents in more details the core blocks of the system. Section 4 shows experiments and results, and finally conclusions and perspectives are discussed in Section 5.

2. GENERAL OVERVIEW OF THE SYSTEM

In this section we present the main principles of the data hiding system. The functional blocks will be detailed in the next section. The system consists of two main blocks (see Fig. 1): an *embedder* used to embed the data into the host signal x, and a *decoder* used to recover the data from the embedded host signal x^w; the decoder is "blind" in the sense that the original signal is assumed to be unknown from the decoding part.

2.1. Embedding

The embedding is performed in the Time-Frequency (TF) domain. Therefore–at the embedder–the time-domain input signal x is first transformed in the time-frequency (TF) plan (Block ①). Instead of using the MDCT (as in [7]), the transform used here is its integer approximation, the IntMDCT [8]. The embedding process consists in quantizing the IntMDCT coefficients $X(t, f)$ (Block ④

* This work was supported by the French National Research Agency (ANR) in the context of the DReaM project (ANR 09 CORD 006).

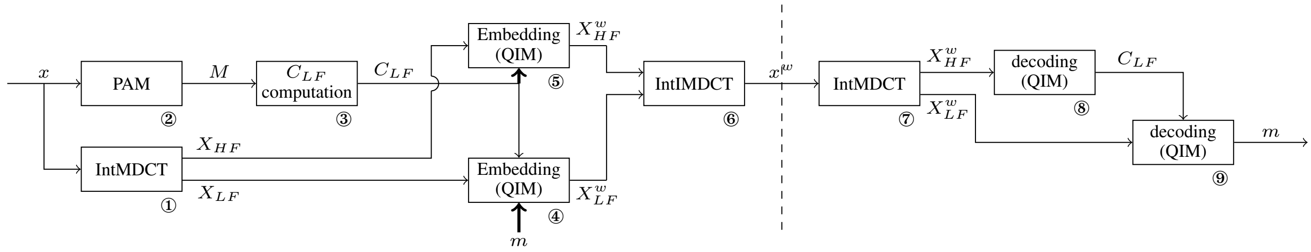

Figure 1: Block diagram of the system. Left of the dashed line is the embedder, right is the decoder. LF (resp. HF) stands for low frequency (resp. high frequency) and $.^w$ denotes vector carrying embedded data.

and ⑤) using specific sets of quantizers $\mathcal{S}_{C(t,f)}$, following the QIM technique described in [11] (see Section 3.2). Once the IntMDCT coefficients are embedded, the signal is transformed back in the time-domain using the integer inverse MDCT (IntIMDCT, Block ⑥). The resulting embedded signal has integer values. Thus there is no need to perform the time-domain 16-bit PCM quantization and take into account the effects of the resulting noise on the embedding performance, as opposed to what was done in [7] (see Section 3.1).

For each frame t and for each frequency bin f, the PAM (Block ②) provides a masking threshold $M(t, f)$ used to compute the embedding capacity $C(t, f)$ (Block ③), *i.e.* the maximum size of the binary code to be embedded in that TF bin under inaudibility constraint. The embedding capacity $C(t, f)$ determines at the same time *how much* information is embedded (at TF bin (t, f)) and *how* it is embedded and retrieved. Consequently, the set of capacity values $C(t, f)$ must be known at the decoder and they have either to be estimated from the transmitted signal at the decoder (as in [10]), or to be transmitted within the host signal x as a part of the embedded data themselves (as in [7]). We keep the line of [7] and propose the following process:

- At the embedder, after IntMDCT transform, the IntMDCT coefficients are separated into a "low-frequency" part (denoted $_{LF}$ on Fig. 1 and thereafter, accounting for 15/16 of the spectrum) and a "high-frequency" part (denoted $_{HF}$, accounting for 1/16 of the spectrum, see Section 3.4).
- Low frequencies are used to embed the "useful" side-information m that is to be transmitted within the host audio signal x. For this aim, the capacities $C_{LF}(t, f)$ are maximized under inaudibility. This is the core of the proposed method that will be described into details in Section 3.4.
- Then, high frequencies are used to embed the values of the resulting capacities $C_{LF}(t, f)$ which totally configure the data hiding process in the low-frequency region. To do this, the values of $C_{HF}(t, f)$ must be known at both the embedder and the decoder. Hence they are set to fixed values (*i.e.* independent of frame index and signal content), exploiting the fact that in the highest frequency region the human hearing system is quite inefficient. Because the high-frequency capacities do not depend on t, they are denoted $C_{HF}(f)$.

2.2. Decoding

The decoding process somehow consists of the reverse operations: the embedded signal x^w is first transformed in the TF domain (Block ⑦) and the resulting IntMDCT coefficients are separated into high and low frequencies subvectors, similarly to the embedder. As the capacities $C_{HF}(f)$ are known to the decoder, the information embedded in the high-frequency region is first extracted (Block ⑧), resulting in decoded $C_{LF}(t, f)$ values. This latter information is then used to decode the "useful" information m embedded in the low-frequency region (Block ⑨).

Note that if the synchronization is not treated in this paper, at least several basic schemes are usable, like for example checksums as used in [10].

3. DETAILED PRESENTATION

3.1. Time-frequency transform

The choice of the MDCT in [7] was mainly guided by the fact that it is a TDAC (Time Domain Aliasing Cancellation) transform. It also has the perfect reconstruction property, it is critically sampled and its coefficients are real, which enables an easy use of quantization techniques. In the present study, we use the integer approximation of the MDCT, the IntMDCT, in order to get rid of the noise introduced on the MDCT coefficients by the time-domain 16-bit PCM quantization [7]. We use a frame length of 2048 to have a sufficient frequency resolution while fitting music signals dynamic.

The principle of the integer approximation is to decompose the MDCT matrix in a product of matrices that are either permutation matrices or block diagonal 2×2 Givens Rotations matrices. The permutation matrices and their inverses maps directly from integer to integer and the 2×2 Givens Rotations can be approximated using the *Lifting Scheme* (see for example [8] for a detailed explanation).

3.2. Embedding technique

The Quantization Index Modulation (QIM) is a quantization-based embedding technique introduced in [11]. The scalar version of the technique is used here[1], which means that each IntMDCT coefficient $X(t, f)$ is embedded independently from the others.

The embedding principle is the following. If $X(t, f)$ is the IntMDCT coefficient at TF bin (t, f) that has to be embedded with $C(t, f)$ bits, then a unique set $\mathcal{S}_{C(t,f)}$ of $2^{C(t,f)}$ quantizers $\{\mathcal{Q}_c\}_{0 \leq c \leq 2^{C(t,f)}-1}$ is defined with a fixed arbitrary rule. This implies that for a given value $C(t, f)$ 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

[1] Note that in this particular case the technique is similar to the *improved LSB* embedding scheme.

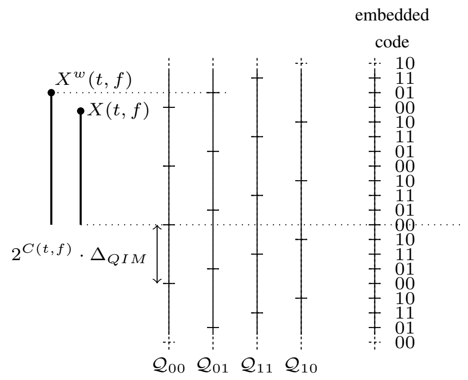

Figure 2: Example of QIM using a set $\mathcal{S}_{C(t,f)}$ of quantizers for $C(t,f) = 2$ with their respective gray code index and resulting global grid. The binary code 01 is embedded into the IntMDCT coefficient $X(t,f)$ by quantizing it to $X^w(t,f)$ using the quantizer indexed by 01.

each quantizer is indexed by a $C(t,f)$-bit codeword c. Because the quantizers are regularly intertwined and because the IntMDCT coefficients are integer-valued, the quantization step of each quantizer is given by:

$$\Delta(t,f) = 2^{C(t,f)}. \tag{1}$$

Embedding the codeword c into the IntMDCT coefficient $X(t,f)$ is simply done by quantizing $X(t,f)$ with the quantizer $\mathcal{Q}_c$ indexed by c (see Fig. 2 for an example). In other words, the IntMDCT coefficient $X(t,f)$ is replaced with its closest code-indexed quantized value $X^w(t,f)$:

$$X^w(t,f) = \mathcal{Q}_c\left(X(t,f)\right). \tag{2}$$

At the decoder, the set of quantizers $\mathcal{S}_{C(t,f)}$ is generated (and is the same as the one generated at the embedder) using the $C(t,f)$ decoded values in low-frequency and fixed values in high-frequency. Then, the quantizer $\mathcal{Q}_c$ with a level corresponding to the received embedded coefficient $X^w(t,f)$ is selected, and the decoded message is the index c of the selected quantizer. As the IntIMDCT directly yields a PCM signal (due to the integer-to-integer mapping of the IntMDCT and IntIMDCT), there is no noise introduced by the conversion (in contrast to [7] when using the MDCT).

Obviously if one wants to transmit a large binary message, this message has to be previously split and spread across the different IntMDCT coefficients according to the local capacity values, so that each 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

The PAM used in our system (Block ②) is directly inspired from the PAM of the MPEG-AAC standard [12]. The output of the PAM is a masking threshold $M(t,f)$, 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 PAM computations is not the IntMDCT but the FFT. The main computations consist in a convolution of the FFT 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, and the absolute threshold of hearing is integrated. After that, some pre-echo control is applied, resulting in the FFT masking threshold. The pre-echo control implemented is quite simple and only consists in taking the minimum of the computed masking threshold and the previous frame masking threshold multiplied by a constant $K > 1$. Taking a value close to 1 will yield a good pre-echo control but will limit the PAM efficiency (in term of embedding rate), while taking too big a value will lead to a poor pre-echo control (in this study $K = 2$). From the FFT spectrum and FFT masking threshold a signal-to-mask ratio (SMR) is computed (for each frequency bin f), and this SMR is then used to obtain the IntMDCT masking threshold $M(t,f)$ (by simply computing the ratio between the IntMDCT power spectrum coefficients and the SMR coefficients). This masking threshold $M(t,f)$ is then used to shape the embedding noise (under this curve), so that it remains inaudible. The masking threshold can also be translated by a factor of α dB so that the total payload matches exactly the size of the signal to be embedded m.

3.4. Capacities computation

The computation of the capacities $C(t,f)$ is the core of the proposed method. As the compliance to the PCM format is already ensured by the use of the IntMDCT, the problem is to optimize the embedding bitrate under inaudibility constraint. In the present study, this constraint is that the power of the embedding error in the worst case remains under the masking threshold $M(t,f)$ provided by the PAM. As the embedding is performed by uniform quantization, the embedding error in the worst case is equal to half the quantization step $\Delta(t,f)$, which is directly related to $C(t,f)$ through (1). The inaudibility constraint in a given TF bin can thus be written as:

$$\left(\frac{\Delta(t,f)}{2}\right)^2 < M(t,f). \tag{3}$$

For the low-frequency region of a given frame t, we simply combine (1) and (3) to obtain:

$$C_{LF}(t,f) < \frac{1}{2}\log_2\left(M(t,f)\right) + 1. \tag{4}$$

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

$$C_{LF}(t,f) = \left\lfloor \frac{1}{2}\log_2\left(M(t,f)\right) + 1 \right\rfloor. \tag{5}$$

where $\lfloor . \rfloor$ denotes the *floor* (rounding down) function. Experimentally, the resulting values are always lower than 15. Thus we can code those values with 4-bit codewords (from 0 to 15). However, embedding the high-frequency region with as many 4-bit codewords as there are frequency bins in the low-frequency zone is not achievable. For this reason, *embedding subbands* are defined as groups of adjacent frequency bins where the capacities $C(t,f)$ are fixed to the same value. The capacity value within each subband is given by applying (5) using the minimum value of the mask within the subband. In order to respect the inaudibility constraint in the high-frequency region, the capacities $C_{HF}(f)$ are fixed to 1 or 2

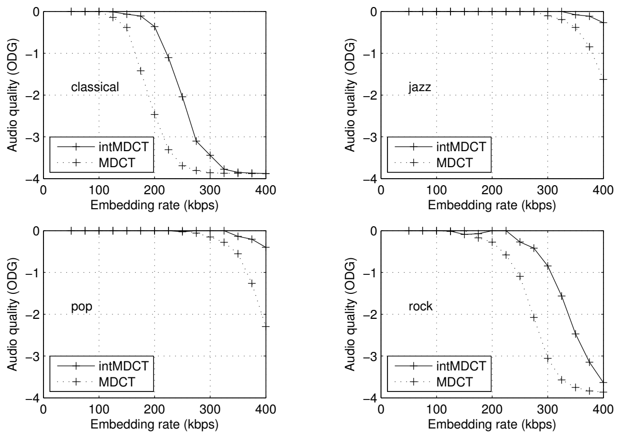

Figure 3: Audio quality as a function of embedding bitrate for a set of tracks used for the tests.

bits. In the present study the 1024 IntMDCT coefficients of each frame are split into 32 bands of 32 coefficients, and the last 2 subbands form the high-frequency zone.

4. RESULTS

The performance of the proposed data hiding system is evaluated in terms of audio quality of the embedded signal as a function of the embedding rate. The audio quality is estimated using the Perceptual Evaluation of Audio Quality (PEAQ) algorithm [13] and double-checked by informal listening tests. The PEAQ algorithm compares the embedded signal with the original signal, and provides a comparative score, called Objective Difference Grade (ODG). Grades range from 0 for inaudible effect to -4 for severe degradation. The tests were performed on twelve 10-second excerpts of 44.1-kHz 16-bit musical signals of different musical styles (classic, jazz, rock, pop…).

Fig. 3 shows some results and as we can see, the embedding bitrate is quite dependent of the audio content–as was already the case in [7]–thanks to (or due to) the PAM. When performing a comparison with the previous system [7], for the same ODG the embedding bitrate is quite higher for the new version (by about 50 kbps). It is also significantly higher than the 140 kbps announced in [10]. In particular, while maintaining inaudibility of the embedded data, bitrates up to 300 kbps can be reached for some very energetic signals (like pop music or jazz-rock). For less energetic signals (classical music) bitrates about 200 kbps are obtained. We can also see that in many cases the embedded data are inaudible with the system of the present study while it is not the case with the previous system of [7]. For most listeners, this is the case for example for an embedding rate of 375 kbps for the jazz track used in Fig.3.

5. CONCLUSION AND PERSPECTIVES

In this paper we presented a data-hiding technique for uncompressed audio signals that yields embedding bitrates of up to 300 kbps per channel for 44.1-kHz 16-bit music signals (depending on audio content). This represents more than 40% of a channel original rate and a significant gain over previous results obtained in [7] and [10]. This technique can be used for “enriched-content” applications, as for instance the informed source separation system presented in [5, 6].

As compressed signals are now widely used, in future works we plan to look at the joint compression and watermarking problem by adapting the principles presented in this paper to compressed signals, for instance the scalable MPEG4-SLS format which also uses the IntMDCT.

6. 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. In *Proc. IEEE Int. Conf. on Multimedia Computing and Systems*, Hiroshima, Japan, 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 *Proc. IEEE Int. Conf. Acoust. and Speech, Signal Proc.*, Taipei, Taiwan, 2009.

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

[7] J. Pinel, L. Girin, and C. Baras. A high-capacity watermarking technique for audio signals based on mdct-domain quantization. In *Proc. Int. Congress on Acoustics*, Sydney, Australia, 2010.

[8] R. Geiger, Y. Yokotani, and G. Schuller. Improved integer transforms for lossless audio coding. In *Proc. Asilomar Conf. Signal, Systems and Computers*, Pacific Grove, California, 2003.

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

[10] R. Geiger, Y. Yokotani, and G. Schuller. Audio data hiding with high data rates based on intMDCT. In *Proc. IEEE Int. Conf. Acoust. and Speech, Signal Proc.*, Toulouse, France, 2006.

[11] 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.

[12] 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.

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