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

1. INTRODUCTION

Data hiding consists in perceptibly embedding 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. 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 \( x^e \); 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 0). 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).
The decoding process somehow consists of the reverse operations: the embedded signal \( x^w \) is first transformed in the TF domain (Block 4) and the resulting IntMDCT coefficients are separated into high and low frequencies subvectors, similarly to the embedder. As the capacities \( C_{LF}(f) \) are known to the decoder, the information embedded in the high-frequency region is first extracted (Block 5), 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 8).

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 integral 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 \( 2 \times 2 \) Givens Rotations matrices. The permutation matrices and their inverses maps directly from integer to integer and the \( 2 \times 2 \) 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 \( SC_{C(t,f)} \) of \( 2^c(t,f) \) quantizers \( \{Q_c\}_{0 \leq c < 2^{C(t,f)} - 1} \) 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.
not the IntMDCT but the FFT. The main computations consist in
domain, however the transform used for the PAM computations is
ning inaudibility. The calculations are made in the time-frequency
power of the quantization error that can be introduced while ensur-
the PAM of the MPEG-AAC standard [12]. The output of the PAM
3.3. Psychoacoustic model
Conversely, the decoded elementary messages have to be concate-
rectly yields a PCM signal (due to the integer-to-integer mapping
At the decoder, the set of quantizers
indexed quantized value
Embedding the codeword
coefficients are integer-valued, the quantization step of each quan-
tizer indexed by 01.
where

\[ \Delta(t, f) = 2^{C(t, f)} \] (1)

Embedding the codeword \( c \) into the IntMDCT coefficient \( X(t, f) \) is simply done by quantizing \( X(t, f) \) with the quantizer \( Q_c \) in-
dexed 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 \( X^w(t, f) \):

\[ X^w(t, f) = Q_c(X(t, f)) \] (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 \( C(t, f) \) de-
coded values in low-frequency and fixed values in high-frequency.
Then, the quantizer \( Q_c \), with a level corresponding to the received
embedded coefficient \( X^w(t, f) \) is selected, and the decoded mes-
sage is the index \( c \) of the selected quantizer. As the IntMDCT di-
rectly yields a PCM signal (due to the integer-to-integer mapping
of the IntMDCT and IntMDCT), 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 concate-
nated to recover the complete message.

3.3. Psychoacoustic model
The PAM used in our system (Block \( \oplus \)) 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 ensur-
ing 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 con-
trol is applied, resulting in the FFT masking threshold. The pre-
echo control implemented is quite simple and only consists in tak-
ing the minimum of the computed masking threshold and the pre-
vious 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 thresh-
old a signal-to-mask ratio (SMR) is computed (for each frequency
bin \( f \)), and this SMR is then used to obtain the IntMDCT mask-
ing 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 embed-
ding noise (under this curve), so that it remains inaudible. The
masking threshold can also be translated by a factor of \( \alpha \) 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 pro-
posed 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 \( M(t, f) \) pro-
vided 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:

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

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

\[ C_{LF}(t, f) < \frac{1}{2} \log_2 (M(t, f)) + 1. \] (4)

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

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

where \( \lfloor \cdot \rfloor \) denotes the floor (rounding down) function. Experimen-
tally, 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 code-
words 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 \( C_{HF}(t, f) \) are fixed to 1 or 2

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 quan-
tizer indexed by 01.
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 energetic signals (like pop music or jazz-rock). For less energetic signals, for instance the scalable MPEG4-SLS format which also uses the IntMDCT.

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