Watermark-aided pre-echo reduction in low bit-rate audio coding*

Imen Samaali[1,2], **Gaël Mahé**[1] **AND Monia Turki**[2],
(imen.samaali@mi.parisdescartes.fr) (gael.mahe@mi.parisdescartes.fr) (m.turki@enit.rnu.tn)

[1] *LIPADE, University Paris Descartes, Sorbonne Paris Cité, France*
[2] *U2S, National Engineering School of Tunis, University Tunis El Manar, Tunisia*

In the context of low bit-rate audio coding, a system for pre-echo reduction, aided by data embedding, is presented. The proposed system aims at correcting the temporal envelope of the decoded signal using side information. The restored signal is reconstructed at the receiving end from two components: the decoded signal, generated by the core decoder, and the temporal envelope of the original signal, represented by a few parameters which are transmitted as embedded data in the audio signal. The audio signals affected by the attack smoothing distortion, concerned by the pre-echo reduction system, are of percussive type such as castanet, cymbal and arabic sounds. The data describing the envelope is transparently embedded at a rate not exceeding 50 bps and with a bit error detection of approximately $1.4\ .10^{-4}$. The considered watermarking system consists in an additive time domain technique in which the spectral shaped watermark signal is generated using an auditory masking model. The performance of the proposed system is evaluated for single and multiple coding/decoding processes performed by an audio MPEG compression (AAC and MP3). Performance evaluation is done through objective perceptual measures provided by the PEMO-Q software. The experimental results exhibit an efficient reduction of the pre-echo and a significant improvement of the audio quality, especially for the MP3 codec.
keywords: Temporal envelope, pre-echo, audio coding, sound attack, audio watermarking.

0 Introduction

Perceptual audio coding at low bit-rate is used for many applications including digital radio, electronic music distribution and portable audio devices. To achieve low bit-rates audio coding while ensuring a good audio quality of the decoded signal, compression techniques have been proposed since the 90's, using a bit allocation based on psycho-acoustic models computed from the frequency representations of successive frames of the audio signal.

However, these techniques may introduce some artifacts the most important of which is known as "pre-echo". As seen in Figure 1, the pre-echo is a noise preceding a time domain transient. This problem affects especially audio transient signals. Indeed, the silence before an attack may be affected by a relatively high quantization noise, since the masking threshold is computed mainly from the portion of the frame after the attack [**?**, **?**].

Due to the properties of human auditory, a pre-echo is masked by the attack only when its duration is less than 20 ms [**?**]. Many techniques with different complexities have been proposed in the literature [**?**] in order to tackle the problem of pre-echo in transform audio coding, particulary for the case of Modified Discrete Cosine Transform (MDCT) coding. The most popular approach is to use adaptive window switching controlled by transient detection, as proposed by Edler [**?**]: typical block sizes are between N=64 and N=1024. The small blocks are only used to control pre-echo artifacts. Usually, window switching implies extra delay and complexity compared to the fixed window approach using a non-adaptive filterbank. Furthermore, short windows yield lower transform coding gains than long windows. Lastly, side information needs to be sent to the decoder to control the window switching.

In addition to the variable windowing, the author of [**?**] proposed an another solution, namely Temporal Noise Shaping (TNS), used in AAC encoders. According to this approach, the coder exercise some control on the temporal fine structure of the quantization noise even within each filterbank window.

*This work is supported by the franco tunisian CMCU project n° 08S1414 and it was part of the WaRRIS project granted by the French National Research Agency (project n° ANR-06-JCJC-0009).

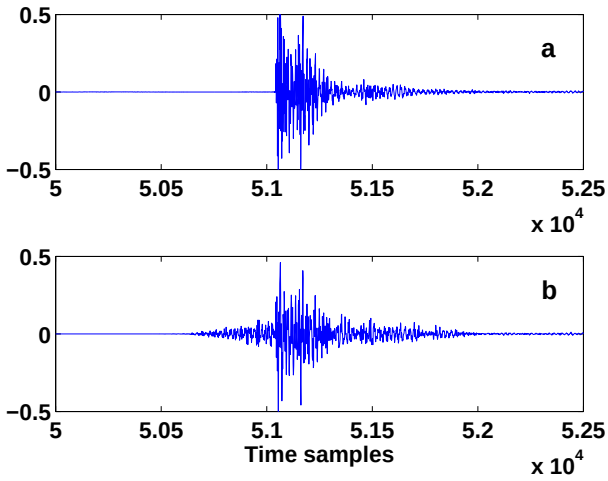

Fig. 1. Illustration of the pre-echo artefact from a castanet signal coded at 48 kbps using an MP3 coder: original signal (a), coded/decoded signal (b).

MP3 encoders can temporarily switch to a smaller window size with a higher time resolution, at the expense of frequency resolution [**?**]. In this case, the quantization error is focused in a smaller time duration and thus is less perceptible. This technique is called Temporal Masking (TM).

Although these techniques reduce significantly the pre-echo phenomenon, the problems of temporal masking and attacks softening remain harmful for some transient signals, such as castanet, cymbal or certain traditional arabic percussive sounds.

A method aiming at reducing pre-echo artifacts after decoding is proposed in this paper. The idea is to restore the attacks of the decoded signal through a correction of the temporal envelope of the frames, using a small set of parameters extracted from the original signal. In the framework presented in [**?**][**?**], the proposed system uses an auxiliary channel to transmit the side information describing the temporal envelope of the original signal. In this paper, in order to eliminate the auxiliary channel used in [**?**][**?**], we propose to use the audio signal as carrier of the useful information through auditory transparent data embedding.

This approach fits in the scope of watermark aided audio processing. Although watermarking was originally used for digital copyright protection purposes, several works exploit watermarks to enhance the performance of a particular processing of the host signal. For example, [**?**, **?**] deal with watermark aided acoustic echo cancelation (AEC); watermark aided band-width extension of speech signals is presented in [**?**, **?**]. In the latter case, high-frequency components of speech are re-synthesized after a narrow-band coding-decoding process, using information conveyed by the watermark. Following the same principle, we propose here to re-synthesize the attacks destroyed by the codec, using the description of the temporal envelope conveyed by the watermark.

This paper is structured as follows. In section 2, we present the attack restoration for audio decoded signals. This presentation includes the basic treatment before encoding and after decoding process. In section 3, the principles of the audio watermarking system used are briefly presented. Section 4 deals with the performance evaluation of the proposed technique in the case of simple and multiple successive encodings (tandem coding) using objective perceptual measures.

1 Pre-echo reduction technique

The structure of the proposed pre-echo reduction system is shown in Figure 2. The input of the system is a temporal audio signal, denoted by $x(t)$, which is fed to the envelope parameter extraction and data embedding blocks. The envelope parameters are embedded in the low-band frequency range of the audio signal by the data embedding block. The watermarked signal then undergoes a process of coding/decoding performed by an MPEG compression system. At the reception, the embedded parameters are extracted from the coded/decoded signal and used by the pre-echo reduction block to enhance the audio signal. The enhanced signal, denoted $\hat{x}(t)$, is given at time $t > 0$ by:

$$\hat{x}(t) = x(t)_{dec} \frac{\hat{e}(t)}{\hat{e}(t)_{dec}}, \quad (1)$$

where $x(t)_{dec}$ is the decoded audio signal, $\hat{e}(t)$ is an estimate of the original temporal envelope, and $\hat{e}(t)_{dec}$ is the temporal envelope of the decoded signal.

1.1 At the encoder

The processing before the encoder is illustrated by the block diagram of Figure 3. It concerns transient frames of 2048 samples. Each detected transient frame is divided in two sub-frames according to the attack position computed by the transient localization block. The temporal envelope of each sub-frame is ARMA-modeled in the frequency domain. The corresponding Line Spectrum Frequencies (LSFs) are quantized using a split vector quantization (VQ) and the selected codebook indices are transmitted to the decoder through an embedded channel based on audio watermarking (data embedding block). The main blocks of the proposed pre-echo reduction system will be described with more details in section 1.3.

1.2 At the decoder

The block diagram of the pre-echo reduction module at the decoder is depicted in Figure 4. It generates the restored audio signal from the decoded version and the embedded side information describing the original temporal envelope, $\hat{e}(t)$, of each transient sub-frame.

The temporal envelope of the decoded audio signal, $\hat{e}_{dec}(t)$, is modeled in the same manner as in the encoder process (Figure 3). Frame characterization and time attack localization are performed directly on the decoded signal. An algebraic detector [**?**][**?**] is used for the temporal localization of the attack. Indeed, this detector is quite robust to noise effects such as pre-echo phenomenon introduced by the mpeg audio compression.

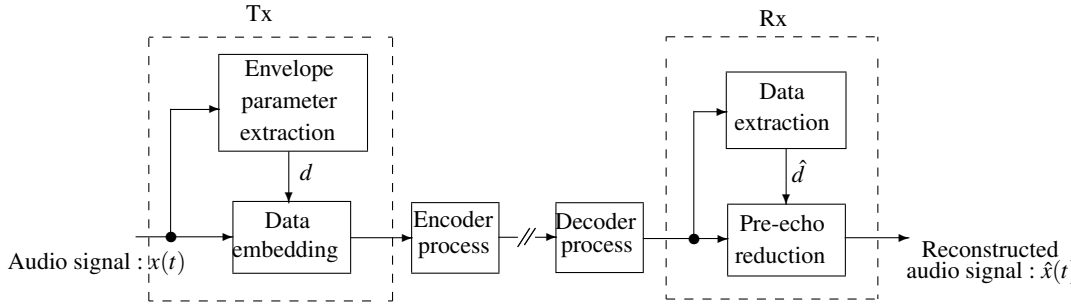

Fig. 2. Pre-echo reduction system description.

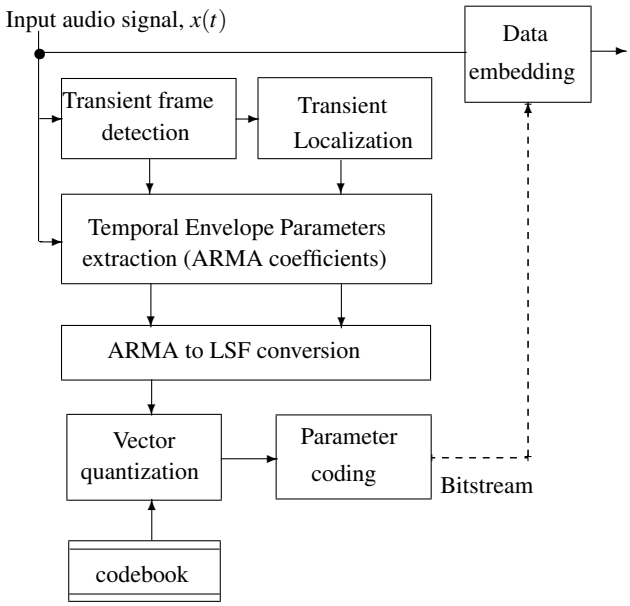

Fig. 3. Pre-echo reduction system (part Tx of Figure 2).

The restored audio signal, $\hat{x}(t)$, is then obtained according equation 1. In the following, we detail the different blocks.

1.3 Transient detector

- **Characterization of frame type:**

To detect transient frames, we adopt the technique used in standard AAC [**?**]. The technique is illustrated in Figure **??**. Each sub-frame of 1024 samples is filtered by a high pass IIR-Filter,

$$H(z) = \frac{0.7548(z-1)}{z-0.5095}. \tag{2}$$

The filtered sub-frame is divided into 8 sub-blocks of 128 samples and the energy of each sub-block i is computed as follows:

$$E_i(x) = \sum_{k=0}^{128} x_i(k)^2. \tag{3}$$

For each sub-block, an *attackR* defined by:

$$attackR(i) = \frac{E_i(x)}{mean(E_j(x)|_{j=i-8:i-1})}, \tag{4}$$

is computed.

The $attackRatio = max(attackR(i)|_{i=0,\ldots,7})$ is used as an indicator of transient presence. Indeed, the probability to have an attack is higher as the value of attackRatio is high. The attackRatio is compared to a threshold *thr* to make a final decision. Referring to [**?**] and considering that the pre-echo duration and magnitude are higher with MP3 than with AAC due to the duration of short windows, the value of *thr* is slightly modified in each case for the coded/decoded signal. Indeed, based on exhaustive empirical measures, the value of *thr* is fixed, for the coded/decoded signal, as following:

$$thr = \begin{cases} 10 & \text{for} \quad br \geq 64 \text{ kbps}, \\ 50 & \text{for} \quad br < 64 \text{ kbps}. \end{cases} \tag{5}$$

For the original signal, *thr* is set to 10.

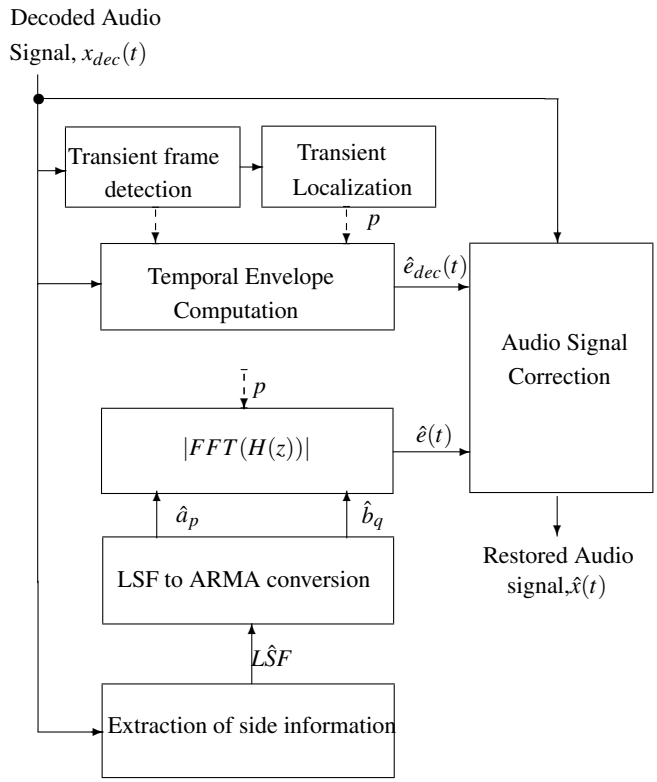

Fig. 4. Pre-echo reduction system (part Rx of Figure 2).

To illustrate the efficiency of the proposed method, Figure **??** shows three original audio frames and their corresponding *attackRatio* coefficients. It is clear that for transient frame (Frame 2), the corresponding *attackRatio* coefficient exceeds the threshold *thr* fixed to 10.

• Transient localization:

The method used to detect the transient time position is based on an algebraic detector (also called change point method) described in [**?**].

The input signal is expressed as a piecewise regular signal,

$$x(t) = \sum_{i=1}^{K} \chi_{[t_{i-1};t_i]} p_i(t - t_{i-1}) + n(t), \tag{6}$$

where

$$\begin{cases} \chi_{[t_{i-1};t_i]} : \text{the characteristic function of the interval } [t_{i-1};t_i], \\ (p_i)_{i\in[1,K]} : \text{a polynomial series} \\ n(t) : \text{additive corrupting noise.} \end{cases}$$

The accuracy of the estimator depends on the polynomial order. Let T, the time interval such as in each interval $I_\tau^T = (\tau, \tau + T)$, at most one change point occurs.

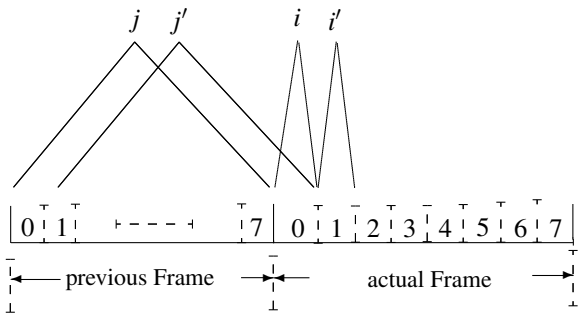

Fig. 5. Transient frame detection

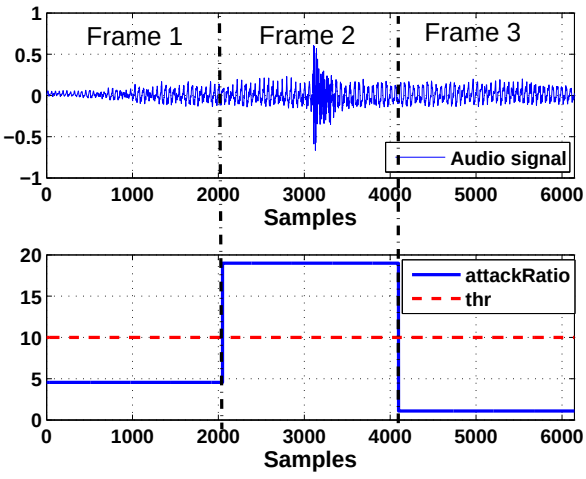

Fig. 6. The audio signal (violon+castanet) and its corresponding *attackRatio* coefficients

Let $x_\tau(t) = x(t+\tau), \quad t \in [0,T]$ for the restriction of the signal to the interval I_τ^T. We redefine the discontinuity point, say t_τ , relatively to I_τ^T with:

- $t_\tau = 0$ if $x_\tau(t)$ is smooth,
- $0 < t_\tau < T$ otherwise.

Let $n(t) = 0$. In the sense of distribution theory of L. Schwartz, the N^{th} order derivative of the input signal can be written:

$$\frac{d^N}{dt^N} x_\tau(t) = [x_\tau^{(N)}(t)] + \sum_{k=1}^{N} \mu_{N-k} \delta(t - t_\tau)^{k-1}, \tag{8}$$

where:

• $[x_\tau^{(N)}]$ represents the regular part of the N^{th} order derivative of the signal.

• μ_k is the jump of the k^{th} order derivative at the point t_τ

$$\mu_k = x^{(k)}(t_{\tau^+}) - x^{(k)}(t_{\tau^-}). \tag{9}$$

If:

– $\mu_0 = \mu_1 = ... = \mu_k = 0, \quad 0 \le t_\tau \le T$, there is no transient in the given interval,

$-\exists\ k\ /\ \mu_k \neq 0,\ 0 \le t_\tau \le T$, there is a transient in the given interval at location t_τ.

The transient localization problem is now casted into estimation of the location of the signal discontinuities, t_τ. Several estimators can be derived from the equation **??**. Assuming that the polynomial order is less than N, $x_\tau^{(N)} \equiv 0$ then equation **??** becomes :

$$\frac{d^N}{dt^N} x_\tau(t) = \sum_{k=1}^{N} \mu_{N-k} \delta(t - t_\tau)^{k-1}. \tag{10}$$

The equation is solved in the operational domain and back to the time domain the estimator is deduced from.

To reduce the complexity of time resolution, the equation **??** is transferred into operational domain using the Laplace transform:

$$\begin{aligned} L\left(\frac{d^N}{dt^N} x_\tau(t)\right) &= s^N \widehat{x}_\tau(s) - \sum_{m=0}^{N-1} s^{N-m-1} \frac{d^m}{dt^m} x_\tau|_{t=0} \\ &= e^{-t_\tau s}(\mu_{N-1} + s\mu_{N-2} + ... + s^{N-1}\mu_0). \end{aligned} \tag{11}$$

Given the fact that the initial condition and the jumps of the derivative of $x_\tau(t)$ are unknown parameters, we annihilate the jumps $\mu_0, ..., \mu_{N-1}$ by applying N times derivation of the equation **??** in the operational domain. After some calculation steps, we finally obtain:

$$\sum_{k=0}^{N} \binom{N}{k} t_\tau^{N-k} (s^N \widehat{x}_\tau(s))^{(N+k)} = 0, \tag{12}$$

where $\binom{N}{k} = \frac{N!}{k!(N-k)!}$ denotes the binomial coefficient.

Recall that, by the classical rules of operational calculus, multiplication by s^ν, $\nu > 0$, corresponds to ν order derivative in the time domain. Taking derivative in the time domain is equivalent to high-pass filtering which may amplify the noise effect. In order to avoid high-pass filtering,

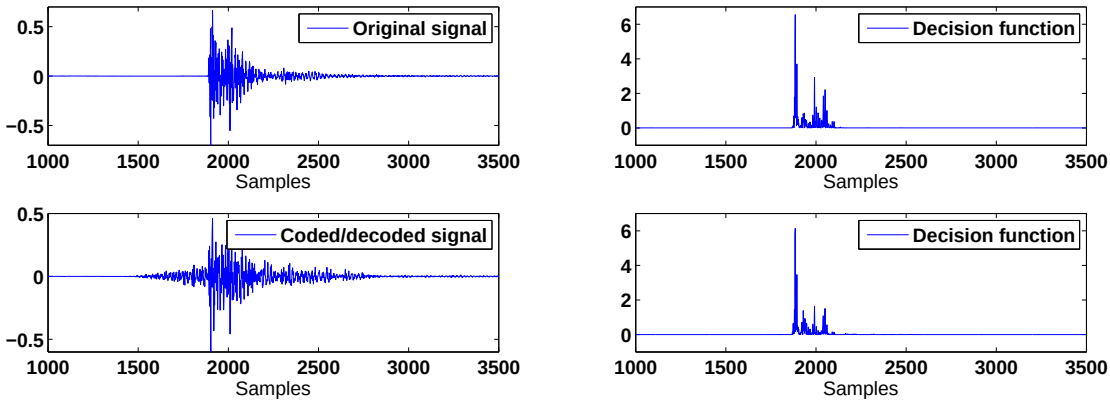

Fig. 7. Decision Function computed by 2^{nd} order Algebraic Detector for castanet signal (top) and its Mp3 coded/decoded version at 48 kbps (bottom)

the whole equation **??** is divided by s^{ν}, $\nu \geq N$, which in the time domain, will be equivalent to an integration. Thus we obtain:

$$\sum_{k=0}^{N} \binom{N}{k} t_{\tau}^{N-k} \frac{(s^N \widehat{x}_{\tau}(s))^{(N+k)}}{s^{\nu}} = 0. \tag{13}$$

Since there is no unknown parameter left, equation **??** is transformed back to the time domain. The Laplace rule that has been used is:

$$L^{-1}(s^{-\nu}\hat{u}) = \frac{1}{(\nu-1)!}\int_0^t (t-\tau)^{\nu-1} u(\tau) d\tau. \tag{14}$$

According to this rule, the transformation of equation **??** back into the time domain leads to:

$$\sum_{k=0}^{N} \binom{N}{k} t_{\tau}^{N-k} \varphi_{k+1}(t) = 0, \tag{15}$$

where

$$\begin{aligned} \varphi_{k+1}(t) &= L^{-1}\left(\frac{(s^N \widehat{x}_{\tau}(s))^{(N+k)}}{s^{\nu}}\right) \\ &= \int_0^{\infty} h_{k+1}(\tau) x(t-\tau) d\tau \end{aligned} \tag{16}$$

and

$$h_{k+1}(\tau) = \begin{cases} \left(\frac{\tau^{\nu-1}(T-\tau)^{(N+k)}}{(\nu-1)!}\right)^{(N)}, & 0 \leq \tau < T \\ 0, \quad \text{otherwise} \end{cases} \tag{17}$$

which leads to simple calculation of the unknown transient location t_{τ}.

Equation **??** is the cornerstone of the algebraic detection approach [**?**]. In fact, an interval I_{τ}^{T} is free of transient if and only if $\mu_1 = \mu_2 = ... = \mu_k = 0 \quad \forall t_{\tau}$, *i.e.* $\phi_{k+1}(t) = 0 \quad \forall k$. Thus, any of the ϕ_{k+1} could be taken as an indicator function. However, due to the noise $n(t)$ (not taken into account in the preceding equations), some values of ϕ_{k+1} may differ from zero even without transient. Hence, an indicator function is given by:

$$J(t) = \prod_{i=0}^{N} \phi_i(t). \tag{18}$$

The probability of having a discontinuity is higher as $J(t)$ is high. Thus, a discontinuity is detected if $J(t)$ is greater than a threshold $\lambda > 0$, depending on the level of the noise $n(t)$. From our experiments, we fixed the value of λ for each frame as follows:

$$\lambda = \frac{\max_{t\in[0;2047]} J(t)}{20}. \tag{19}$$

Although the mathematical framework of the algebraic detector appears complex, the implementation complexity of the detector is low. Indeed, it requires $N+1$ convolutions with the coefficients of each filter h_{k+1} computed once. Here, N is fixed to 2.

To illustrate the performance of the algebraic detector for transient detection and its robustness to the pre-echo effect, we represent in Figure **??** the decision function for both the original signal and its coded/decoded version with the presence of pre-echo. Figure **??** shows that a 2^{nd} order algebraic detector is sufficient to locate the transient. Also, as seen, the method presents a good robustness to noise. The multiple integrations performed by the algebraic detector reduce the noise, which leads to a similar decision function for both the original and the coded-decoded signals.

1.4 Temporal envelope modeling

This subsection deals with the modeling and the estimation of the temporal envelope. The proposed modeling approach is based on linear prediction in the frequency domain (FDLP) which estimates the temporal envelope of the signal, specifically, the square of its Hilbert envelope[**?**][**?**].

The block diagram of the temporal envelope estimation is depicted in Figure **??**. The FDLP technique is implemented in two parts:

1. The Discrete Cosine Transform (DCT) is applied on frame segments of the audio signal to obtain a real valued spectral representation of the signal.
2. A linear prediction is performed on the DCT representation to obtain a parametric model of the temporal envelope. Classically, the AR model is used to estimate

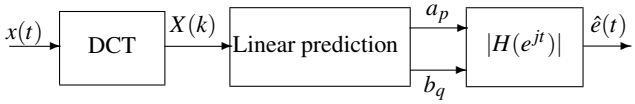

Fig. 8. Block diagram of the temporal envelope estimation.

the temporal envelope [?] [?]. In the case of our research, the ARMA model was preferred, because it ensures an accurate representation of the spectral envelope by means of a reduced number of coefficients.
The autoregressive parameters, $a_i|_{i=1,..,p}$, of the ARMA(p,q) model are estimated by minimizing the mean square prediction error defined as follows:

$$E_{pre}(k) = X(k) + \sum_{i=1}^{p} a_i X(k-i), \tag{20}$$

where $X(k) = DCT(x(t))$. The moving average parameters, $b_i|_{i=1,..,q}$, are computed by a Prony method [?].

An estimation of the temporal envelope of $x(t)$ is therefore given by:

$$\hat{e}(t) = |H(e^{jt})|, \tag{21}$$

where H is defined by the following z-transfer function:

$$H(z) = \frac{\sum_{i=0}^{q} b_i z^{-i}}{1 + \sum_{i=1}^{p} a_i z^{-i}} = \frac{H_b(z)}{H_a(z)}. \tag{22}$$

Based on exhaustive experimental results [?], an ARMA(5,3) is chosen for our application since it realizes a tradeoff between the limited bit-rate of the watermarking channel and the audio quality of the restored signal.

To illustrate the efficiency of the proposed approach, Figure ?? shows the temporal envelope based on FDLP modeling of a violon sound signal sampled at 44.1 kHz.

1.5 ARMA to LSF conversion and LSF quantization blocks

In the proposed system, only the ARMA parameters modeling the temporal envelope of the original signal must be coded and transmitted to the decoder through an audio watermarking channel. The AR and MA coefficients are characterized by a large dynamic range which requires many bits to have an accurate coding. Indeed, small quantization errors may lead to a large (temporal or spectral) distortion. The ARMA parameters are converted to LSFs (Line Spectral Frequencies) which are known for their robustness to channel noise and their limited dynamic range, since they lie in $]0;\pi[$.

To transform the Autoregressive (AR) parameters into LSFs coefficients, we use the technique described in [?].

The Moving Average (MA) parameters are transformed into LSF coefficients in the same manner as the AR parameters. However, before the conversion step, the MA filter $H_b(z)$ (equation ??) must be modified as follows:

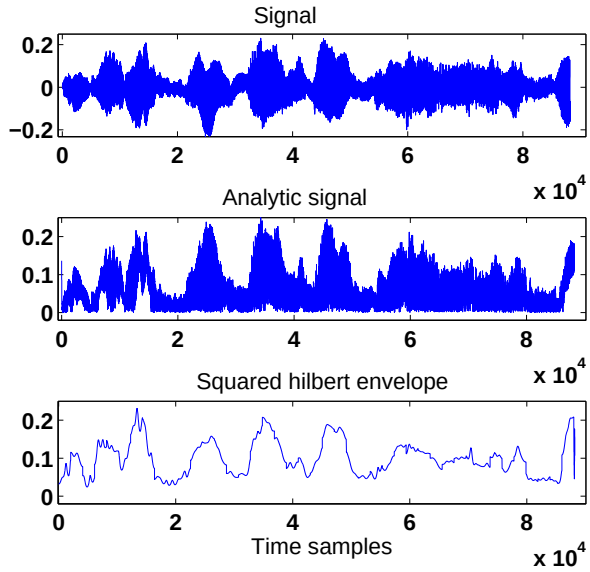

Fig. 9. Illustration of the FDLP temporal envelope modeling of a portion of a audio signal (violon, ARMA(5,3))

$$H_b(z) = 1 + \sum_{i=1}^{q} b'_i z^{-i}, \tag{23}$$

where $b'_i = \frac{b_i}{b_0}$, $i = 1,..,q$. The value of b_0 is estimated at the decoder from the decoded signal. As seen in Figure ??, the actual and estimated b_0 are very similar. As a consequence, the temporal envelopes computed respectively with the original and the estimated b_0 are very close, as illustrated in Figure ?? for a castanet signal.

As known, the computation of the LSFs requires that all roots of $H_a(z)$ (respectively $H_b(z)$) lie inside the unit circle. This property is guaranteed for $H_a(z)$. For $H_b(z)$, according to spectral factorization theory [?], such constraint can be ensured by replacing a root outside the unit circle by its reciprocal conjugate.

For each transient sub-frame, a vector grouping the 8 LSF coefficients is coded on 10 bits using a classical vector quantization technique and transmitted as embedded data in the audio signal with a bit-rate not exceeding 50 bps. The codebook C of a dimension $K \times L$ (K=1024, L=8) is obtained by training on a database of approximately 30000 vectors taken from various kinds of audio signals (percussive and harmonic). It is computed according to the Lloyd-Max algorithm [?].

2 Watermarking technique

To insert a binary message in an audio signal, several watermarking techniques are proposed in the literature [?, ?, ?] with various complexities, embedding capacities and robustness. In this paper, a generic spread-spectrum additive technique in the time-domain based on a Perceptual

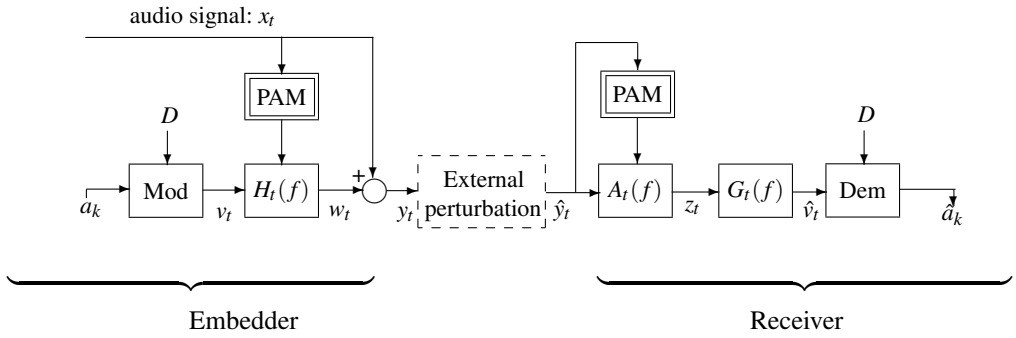

Fig. 10. The generic audio watermarking scheme in presence of channel perturbations.

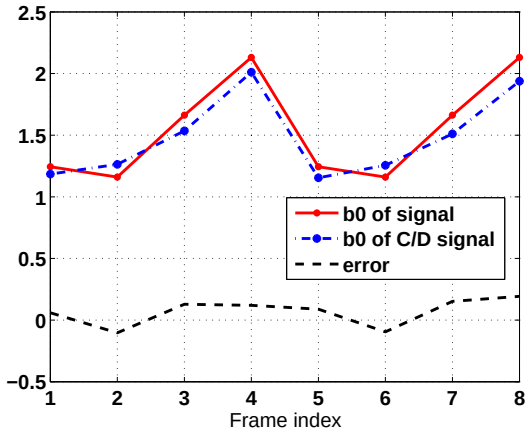

Fig. 11. b_0 of original and coded/decoded signal.

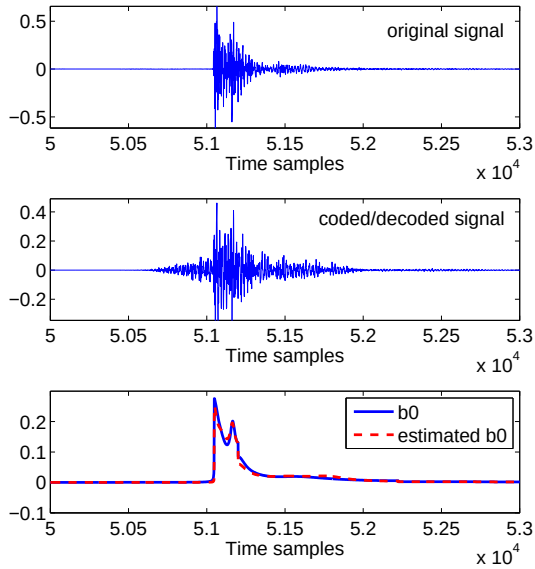

Fig. 12. Temporal envelopes (bottom) computed with the original b_0 and the one computed on the coded/decoded signal.

Auditory Model (PAM) is used [**?**], as it is robust against simple and multiple low bit-rates MPEG coding.

The considered watermarking system is depicted in figure **??**. The spectral shaped watermark signal w_t is added to the original audio x_t in the time domain. The imperceptibly watermarked audio is then given by $y_t = x_t + w_t$. The receiver is composed of two cascade filters and a demodulator. It estimates the transmitted symbol $\hat{a}_k$ (watermark) from the observed data $\hat{y}_t$.

2.1 The embedding technique

The watermark message is a sequence of K-ary symbols chosen in a finite alphabet $A = \{a_0, ..., a_{K-1}\}$. The K-ary message is converted into signal v_t by using a dictionary, $D = \{d_0, ..., d_{K-1}\}$, containing K spread spectrum and orthogonal vectors of length N where each vector corresponds to a given symbol a_k.

The watermark w_t is obtained by filtering v_t through the shaping filter $H_t(f)$ so that its power spectrum density (PSD) approaches the masking threshold of x. This threshold and implicitly the shaping filter $H_t(f)$ are computed using a psychoacoustic model. Since the latter is computed for stationary frames, N should match the stationary duration of the audio signal. The watermarked signal is given by :

$$y_t = v_t * h_t + x_t. \tag{24}$$

2.2 The extraction technique

The technique deployed to extract the watermark signal is the one described in [**?**]. The block diagram of the technique depicted in Figure **??**, is based on a two cascade realization of the Wiener filter.

The first filter, $A_t(f)$, represents an estimate of the inverse of the shaping filter $H_t(f)$ used by the transmitter (see figure **??**). Assuming that x_t and $\widehat{y}_t$ are perceptually similar, $H_t(f)$ is approximated by $\widehat{H}_t(f)$ derived from the PAM of $\widehat{y}$. The receiver can then filter the watermarked audio signal $\hat{y}_t$ with the inverse filter,

$$A_t(f) = \frac{1}{\widehat{H}_t(f)}, \tag{25}$$

and thus reverse a great part of the spectral shaping of the watermarked signal. This method of equalization is a Zero-Forcing procedure[**?**]. The watermark is still embedded in the audio signal. Thus, a Wiener denoising filter G is used

in order to reduce the effect of the host signal [**?**]. Denoting z_t the input of $G(z)$, the output $\hat{v}_t$ is given by:

$$\hat{v}_t = \sum_{i=-\infty}^{+\infty} g_i z_{t-i}, \tag{26}$$

where the signal z_t is given by

$$z_t = \hat{y}_t * a_t \approx v_t + x_t * a_t. \tag{27}$$

Recall that a_t is the inverse Fourier Transform of $A_t(f)$.

Because v_t and $x_t * a_t$ are not correlated and by minimizing the Mean Square Error, $MSE = E[(v_t - \hat{v}_t)^2]$, the optimal non causal and infinite impulse response g_t is given by :

$$r_v(t) = \sum_{i=-\infty}^{+\infty} g_i r_z(t-i) \quad \forall\ t, \tag{28}$$

where r_v and r_z are the correlation function of respectively v_t and z_t. Note that $r_v(t)$ can be computed over the dictionary D known at the receiver.

For practical considerations, the length of the infinite impulse response g have to be truncated to finite length : $g = (g_{-L_{nc}}, ..., g_{L_c})^T$.

The demodulator bloc (DEM), presented in Figure **??** computes the correlation between $\hat{v}_t$ and the different vectors of the codebook. The selected vector is the one that maximizes the correlation. The transmitted symbol a_k is then deduced.

The approximations used in equations 22 and 24 lead to some error on A and G. As shown in [**?**], the resulting mean square deviation of A and G is about -30 to -20 dB, which has a moderate impact on the watermarking detection error: for a watermarking rate of 71 bit/s, the binary error rate is multiplied by a factor 2 (compared to exact computation of A and G) and stays in the same range, about 10^{-3}.

2.3 Selecting frames for data embedding

Recall that at reduced bit rates, the MPEG compression affects transient parts of percussive audio signals (pre-echo phenomenon in Figure 1) which can affect the embedded data. Therefore, to avoid such problem, we propose to select only non transient frames for data-embedding by using the technique described in 1.3.

2.4 Watermarking robustness

The robustness of the watermarking system against MPEG coding was studied in [**?**] in the case of MPEG1 layer 1 compression at 96 kbps and in [**?**] in the case of MPEG1 layer 3 (MP3) compression at 64 and 96 kbps. For a watermarking rate of 71 bps, the BER reported in [**?**] is $1.6.10^{-3}$ without perturbation and $2.6.10^{-3}$ when the audio signal undergoes an MP3 coding at 64 kbit/s. Thus, the watermarking system is expected to be robust to the channel perturbations considered here (MP3 and AAC encoding/decoding at bit-rates varying from 40 to 96 kbps).

- Single MP3 and AAC compression

We illustrate the robustness of the watermark extraction against MP3 and AAC coding. The embedded data, more

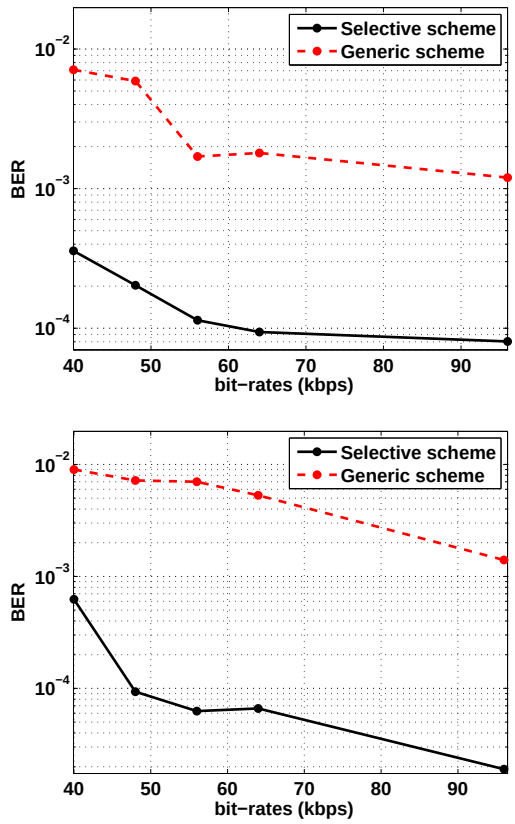

Fig. 13. Effect of bit-rates on data extraction in single encoding case: MP3 (top) and AAC (bottom)

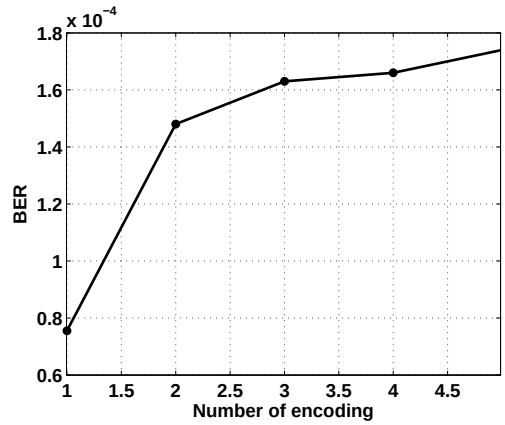

Fig. 14. Effect of number of encoding on data extraction in encoding tandem case

than 10^6 bits, was replaced by a random distribution bitstream which was modulated at 2 bits/symbol and inserted in the low band of the audio signal [0-5 kHz] with a bit rate not exceeding 50 bps. The performance in terms of BER over bit-rates varying from 40 to 96 kbps are shown in Figure **??**. The reported results correspond to the average of several Monte Carlo simulations for a percussive audio signal (castanet sampled at 44.1 kHz).

The results of Figure **??** point out that, for an audio signal with percussive character, the proposed selective scheme (only non transient frames are watermarked) is

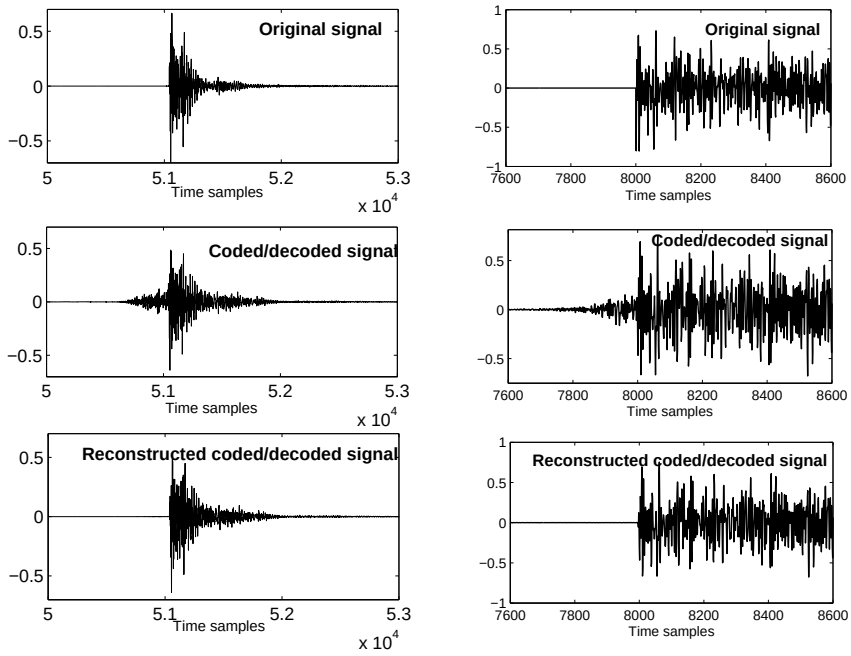

Fig. 15. Illustration of pre-echo reduction using the proposed approach for castanet (left), and cymbal (right), (MP3 coder at 64 kbps)

much more efficient than the generic scheme (all the frames are watermarked) in the presence of MPEG compression.

- Multiple MP3 compression

To investigate the robustness of watermarking extraction to multiple successive encoding/decoding process, we present in Figure **??** the performance in terms of BER over number of MP3 encoding/decoding at 64 kbps. The reported BER results are average of several Monte Carlo simulation performed on three different audio signals (castanet, violon and pop music). The considered audio watermarking technique is the selective scheme described before. As illustrated, the BER increases slightly as the number of encoding increases. Furthermore, for each encoding/decoding step, the obtained BER does not exceed $1.8\ \ 10^{(-4)}$.

3 System evaluation

The experiments aim at validating the performance of the proposed system (Figure 2) with MP3 and AAC codecs at variable bit-rates. Experimental results were performed on 6 percussive audio signals which exhibit a remarkable transient character. All the considered signals are sampled at 44.1 kHz.

The performance of the pre-echo reduction system are evaluated through the objective audio quality assessment provided by the PEMO-Q software [**?**], namely the Objective Difference Grade (ODG). The latter is a perceptual audio quality measure, which rates the difference between test and reference signals among a scale from 0 (imperceptible) to -4 (very annoying). As mentioned in [**?**], ODG is correlated with the Subjective Difference Grade (SDG) for audio quality [**?**].

As a primary evaluation of the proposed system, we present in Figure **??** original castanet and cymbal signals, their 64 kbps MP3 coded/decoded versions and their corresponding corrected versions. It can be seen that the MP3 codec introduces a remarkable pre-echo. In the corrected signal, the pre-echo is considerably reduced and the attack is restored.

3.1 Single encoding

At the first stage, we investigate the effectiveness of the pre-echo reduction according to a variable bit-rate of the audio codec. For each experiment, the reference audio signals are coded/decoded by MP3 and AAC codecs at different bit-rates varying from 40 to 96 kbps. Figure **??** shows ODG plotted versus bit-rate for six audio files in the case of MP3 codec. The application of the temporal masking option (TM) integrated in MP3 codec enhances slightly the audio quality. However, the proposed correction provides a significant enhancement of the ODGs, except for hihat signal for which the improvement is slighter. In particular, the efficiency of the proposed system to reduce pre-echo and hence enhance the audio quality is very remarkable for darbouka and tabla signals.

Figure **??** plots ODG values at different bit-rates for six AAC audio files. We note that the TNS option (Temporal Noise Shaping) integrated in the AAC codec provides a very small enhancement of the audio quality. Even with the proposed system, the quality enhancement is non relevant except for the Darbouka signal. These results can be explained by the fact that the time duration of the pre-echo generated by the AAC codec is no longer than 20 ms, which can be masked by the attack (see Figure **??**).

These interesting results were obtained with a BER $=10^{(-4)}$ for the watermark extraction. However, even

with BER=$10^{(-2)}$, the improvement of the audio quality achieved by the the proposed echo-reduction system remains significant. Indeed, as shown in Figure **??** corresponding to an MP3 castanet signal, the ODG for the corrected signal stops increasing from BER $> 10^{(-2)}$.

3.2 Multiple encoding

At the second stage, we analyze the evolution of the perceived quality in the case of successive multiple coding-decoding process. Referring to the experimental results of Vercellesi [**?**] [**?**], there is a quality loss after each coding step. Similar results were obtained here (see dotted line curves of Figure **??**). The quantization noise from each cycle piles up, so that the pre-echo, which was post-masked by the attack in the first encodings, is not masked anymore in the next encodings and becomes very annoying.

As shown in section **??**, our watermarking system is robust to multiple encoding-decoding process. Therefore, we expect to obtain a good performance of the pre-echo reduction system in the case of multiple encoding. The block diagram of the proposed system is depicted in Figure **??**. As show by Figure **??**, the watermarking step is realized only once, before the first encoding process. However, at each decoding step, the watermark extraction and envelope correction are performed.

The variations of the ODG versus the number of encodings, for the proposed pre-echo reduction system are shown in Figures **??**. The reported results correspond to six successive encoding/decoding steps using a MP3 codec at 64 kbps.

Except for Darbouka and Tabla signals, the Temporal Masking (TM) enhances the quality, but the reference temporal envelope for the n^{th} coding is the deteriorated envelope of the signal from the $(n-1)^{th}$ coding-decoding. Thus, with only TM, the curve converges quickly towards that without TM. On the contrary, for the proposed system, the reference envelope is always the original embedded one, so that the audio quality degrades slowly. In particular, for castanet and darbouka signals, the proposed system is far more efficient than with only Temporal Masking (TM) option.

The audio quality for the pre-echo reduction system when an auxiliary channel is used [**?**] remained constant over the number of encoding, especially for the case of castanet signal: the ODG stopped decreasing from encoding number=3. Indeed, due to the capacity of the auxiliary channel used in [**?**] ($\backsimeq$ 500 bps), the system proposed in [**?**] restores the whole decoded audio signal (transient and non transient frames). However, limited by the bit-rate offered by the audio watermarking channel ($\backsimeq$ 50 bps), the proposed system restores only the transient frames, which explains that the quality is not so robust to the number of encodings.

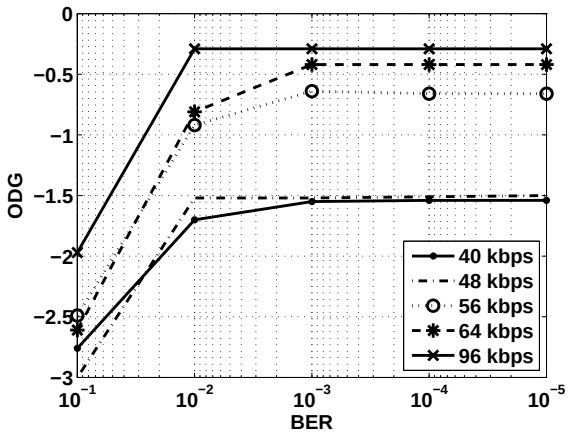

Fig. 16. Effect of BER on pre-echo reduction (MP3 castanet signal).

4 Conclusion

At low bit-rate, coding/decoding with standard AAC and MP3 coders smoothes attacks in transient signals and increases the pre-echo despite TNS/TM options. We have presented a system of pre-echo reduction dedicated for MPEG compression. The system is based on temporal envelope correction, aided by data embedding. The proposed approach uses the coded/decoded audio signal as a carrier of the side information needed for the pre-echo reduction. Therefore, the need of an auxiliary communication channel was eliminated.

The proposed method enhances significantly the audio quality as measured by Objective Difference Grade (ODG) scores, and avoids the dramatic fall of quality caused by multiple coding/decoding especially for the case of MP3

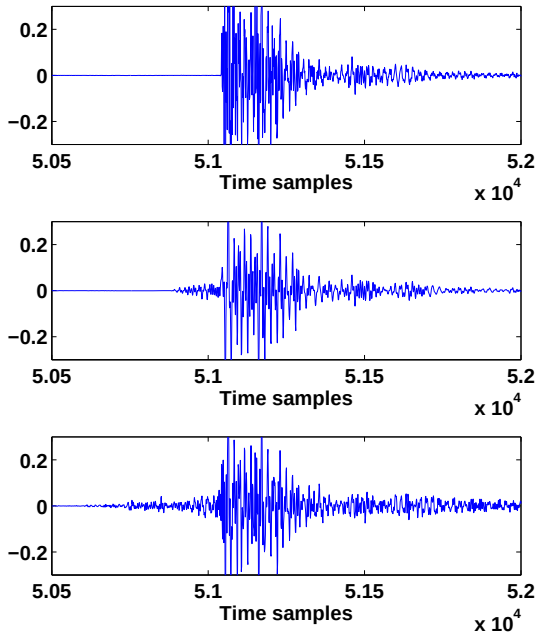

Fig. 18. Illustration of the pre-echo artefact from castanet signal coded at 40 kbps: original signal (top), AAC coded/decoded signal (middle) and MP3 coded/decoded signal (bottom).

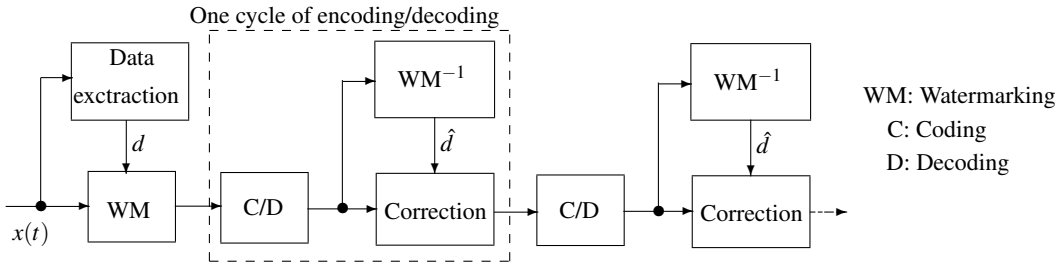

Fig. 17. Block diagram of pre-echo reduction in tandem encoding case.

codec: ODG values decrease slowly and the quality remains transparent.

To reduce the detection error of the embedded data, the bit-rate capacity used by the watermarking channel is limited to 50 bps. Therefore, the restoration of the coded/decoded signal was performed only for transient frames. Furthermore, the limited WM rate implies that the envelope parameters for restoration are embedded before and after the attack, in non-transient frames. Thus, in a context of real-time broadcast or streaming, the restored sound must be delivered with an extra delay. In further investigation work, other watermarking techniques with large embedding capacity [**?**, **?**] should be considered in order to restore all coded/decoded frames.

Acknowledgements

The authors would like to thank Sonia Larbi for her audio watermark programs and Mamadou Mboup for his Algebraic Detector programs. Also, authors thank Taoufik Majoul for many helpful discussions and ideas.

5 REFERENCES

[1] J. Reiss, M. Sandler, Audio Issues In MIR Evaluation, ISMIR, 2004.

[2] M. Erne, Pre-echo distortion, Audio Coding Artefacts Educational CD-Rom, I. Herre and J. D. Johnston, Eds.: Audio Engineering Society, 2001.

[3] B. Edler, Coding of audio signal with overlapping block transform and adative window function, Frequenz, Volume 43, pages 252-256, 1989.

[4] Jürgen HERRE, Temporal Noise Shaping, Quantization and Coding Methods in Perceptual Audio Coding: a Tutorial Introduction, AES 17^{th} International Conference on High Quality Audio Coding, September 1999.

[5] E. Ambikairajah, A. G. Wong, W. T. K, Auditory masking and MPEG-1 audio compression, Electronics and Communication Engineering Journal, Volume 9, Issue 4, pages 165-175, August 1997.

[6] I. Samaali, M.Turki and G.Mahé, Temporal envelope correction for restoration of attacks in low bit-rate audio coding, EUSIPCO 2009.

[7] I. Samaali, M.Turki and G.Mahé, Attack localization based on algebra detector for pre-echo reduction in low bit-rate audio coding, ISIVC 2010.

[8] Cléo Baras, Tatouage informé de signaux audio numériques, PhD thesis of Telecommunication, Ecole Nationale Supérieure des Télécommunications, December 2005.

[9] I. Marrakchi-Mezghani, M. Turki, S. Djaziri-Larbi, M. Jaïdane, G. Mahé, Speech processing in the watermarked domain: application in adaptive acoustique echo cancellation, EUSIPCO 2006.

[10] G. Szwoch, A. Czyzewski and A. Ciarkowski, A double-talk detector using watermarking, Journal of Audio Engineering Socity, Volume 57, pages 916-926, 2009.

[11] Bernd Geiser, Peter Jax, and Peter Vary, Artificial Bandwidth Extension of Speech Supported by Watermark-Transmitted Side Information, INTERSPEECH 2005.

[12] A. Sagi and D. Malah, Bandwidth extension of telephone speech aided by data embedding, EURASIP Journal on Applied Signal Processing, Volume 2007 Issue 1, 1 January 2007.

[13] Yong Xiang, Dezhong Peng, Natgunanathan, I. Wanlei Zhou, Effective Pseudonoise Sequence and Decoding Function for Imperceptibility and Robustness Enhancement in Time-Spread Echo-Based Audio Watermarking, IEEE Transactions on Multimedia, Vol. 14, No. 6, February 2011.

[14] G. Vercellesi, A. Vitali and M. Zerbini, MP3 audio quality for single and multiple encoding, ICME, 2007.

[15] A. Vitali, G. Vercellesi, and M. Zerbini, A multi-level approach to direct process MP3 codes in compression domain, ST Journal of research, volume 3, number 2. 2006, Multimedia Stream Technologies.

[16] W. Guibene, A. Hayar, and M. Turki, Distribution discontinuities detection using algebraic technique for spectrum sensing in cognitive radio, CROWNCOM, September 2010.

[17] M.Mboup, A Volterra filter for neuronal spike detection, inria, 2012.

[18] S. Larbi, M. Jaïdane and N. Moreau, A new Wiener filtering based detection scheme for time domain perceptual audio watermarking, ICASSP 2004.

[19] E. Wolff, C. Baras, and C. Siclet Toward robustness of audio watermarking systems to acoustic channels, EUSIPCO, 2010.

[20] F. C. C. B. Diniz and S. L. Netto, A package tool for general-purpose signal denoising, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, USA, pp. 573-576, March 2005.

[21] 3GPP TS 26.403, Advanced Audio Coding (AAC) part, September 2004.

[22] S. Hu and S. M. Wu, Prony estimation of AR parameters of an ARMA time series, Mechanical Systems and Signal Processing (1989) 3(2), 207-211.
[23] S. Larbi, M. Jaidane, Audio Watermarking: A Way To Stationarize Audio Signals, IEEE Trans. Signal Processing, Vol. 53 (2), pp. 816–823, 2005.
[24] M. Athineos, D. P.W.Ellis, Frequency-Domain linear Prediction For Temporal Features, ASRU 2003.
[25] M. Athineos and D. P.W.Ellis, Autoregressive modeling of temporal envelopes, IEEE transaction on signal processing, vol. 55, No. 11, November 2007.
[26] N. Moreau, Techniques de compression des signaux, Dunod, December 1997.
[27] S. Hayes, Satistical digital signal processing and modeling, Mars 1996.
[28] V.S. Jayanthi, K.S. Marothi, T.M. Ishaq, M. Abbas and A. Shanmugam, Performance Analysis of Vector Quantizer using Modified Generalized Lloyd Algorithm, IJISE, vol.1, pp. 11-15, Janury 2007.
[29] R. Huber, B.Kollmeier, PEMO-Q–A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception, IEEE Transactions on audio, speech and language processing, vol. 14, No. 6, November 2006.
[30] Subjective assessment of sound quality,ITU, Geneva, Germany, 1990, ITU-R recommendation BS. 562-3.

THE AUTHORS

Imen Samaali

Gaël Mahé

Monia Turki

Imen Samaali received the engineering degree in telecommunications and the M.Sc. in communication systems from National Engineering School of Tunis (ENIT),Tunisia, in 2006. She is a member of the research unit Signals and systems (U2S) at ENIT in Tunisia and the Laboratory of Informatics of Paris Descartes University (LIPADE) in France and currently a Ph.D. candidate in telecommunications at ENIT conducting research on the new uses of watermarking to improve low bit-rate audio communication systems.

●

Gaël Mahé received the engineering degree in telecommunications and the M.Sc. in signal and communications from Télécom Bretagne, France, in 1998. From 1999 to 2002 he has worked at Orange Labs in Lannion, France, where he received the PhD in signal and telecommunications from the University of Rennes 1 in 2002. Since 2003, he has been with the Laboratory of Informatics of Paris Descartes University (LIPADE), where he is currently assistant professor. His research deals mainly with new uses of watermarking in audio processing.

●

Monia Turki was born in Mahdia , Tunisia . She received the Principal Eng. degree in 1989, M . Eng. degree in 1991, and the Ph. D. degree in 1997 from the Department of Electrical Engineering, National School of Engineers of Tunis (ENIT), Tunisia. From 1991 to 1997, she was a teaching assistant in the Department of Electrical Engineering, ENIT. From 1997 to 2007, she was been an assistant professor at ENIT. She joined the department of Information Technology and Communications of ENIT since 2001. Since 2007 she has been an associate professor and since 2009 she joined the Tunisia Polytechnic School. Her present research concerns signal processing applied to telecommunications and audio processing applied to denoising and coding.

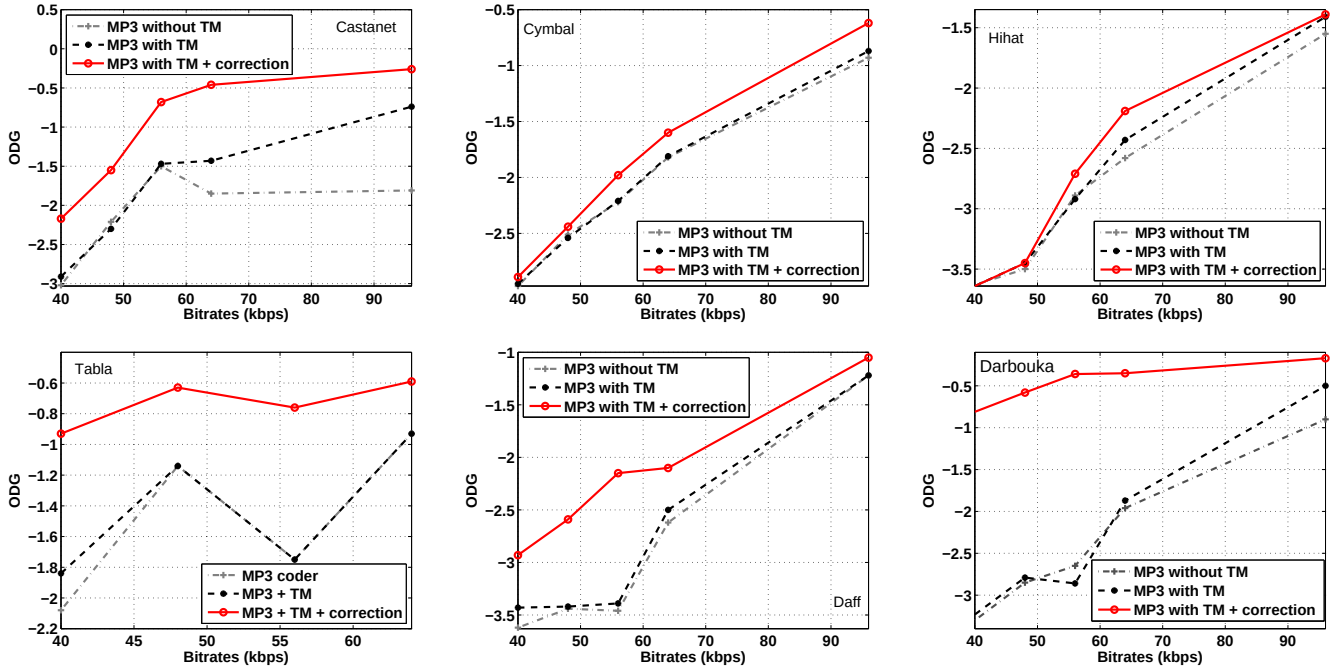

Fig. 19. Evaluation of the perceived quality for six single MP3 encoded/decoded sounds

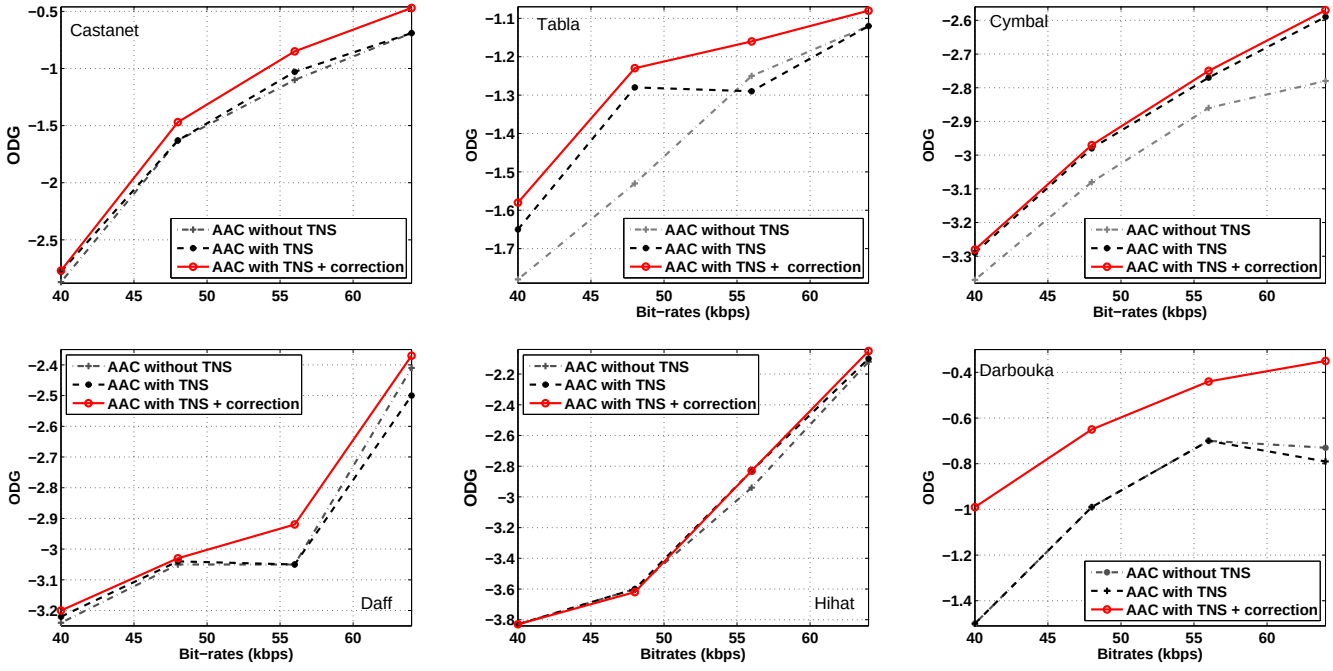

Fig. 20. Evaluation of the perceived quality for six single AAC encoded/decoded sounds.

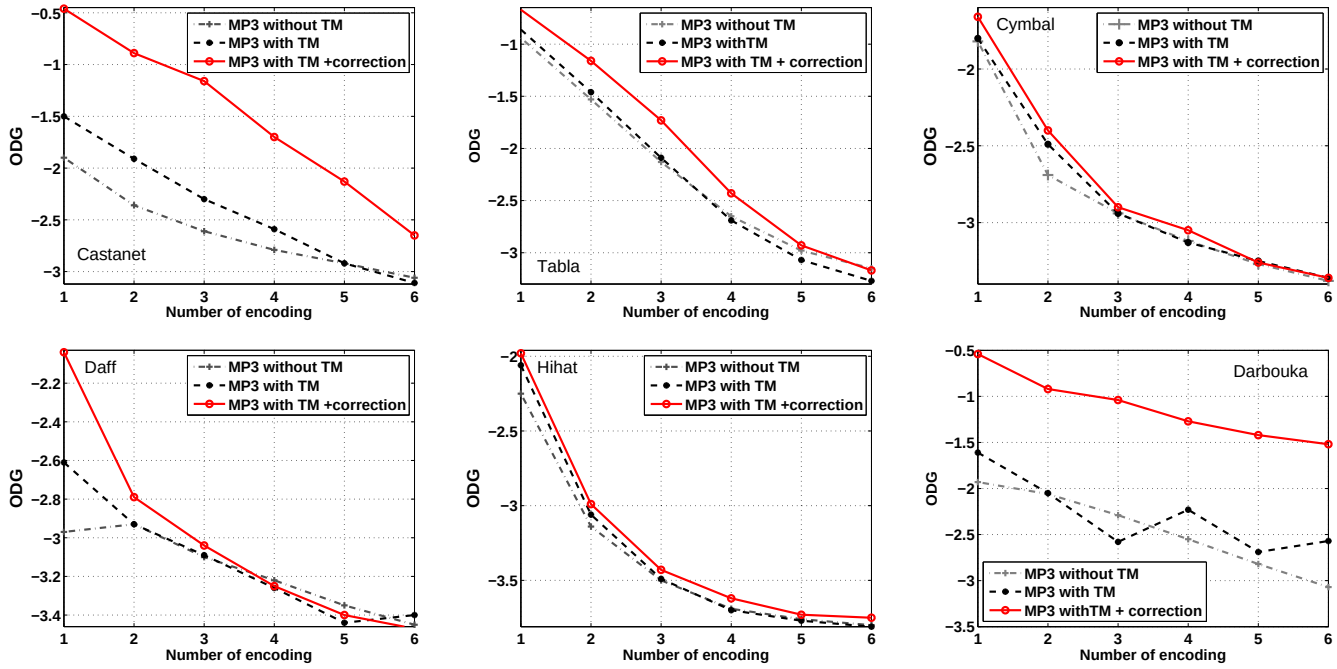

Fig. 21. Evaluation of perceived quality for six tandem MP3 encoding.