Attack restoration in low bit-rate audio coding, using an algebraic detector for attack localization

Imen Samaali*, Monia Turki-Hadj Alouane
Unité Signaux et Systèmes (U2S)
Ecole Nationale d'Ingénieurs de Tunis, Tunisie
Email: imen.samaali@mi.parisdescartes.fr, m.turki@enit.rnu.tn

Gaël Mahé
* LIPADE
Université Paris Descartes, France
Email: Gael.Mahe@mi.parisdescartes.fr

***Abstract*—This paper deals with pre-echo reduction in low bit-rate audio compression. [1] proposed an attack restoration method based on the correction of the temporal envelop of the decoded signal. A small set of coefficients were then transmitted through a limited bit-rate auxiliary channel. However, the transmission of the transient position computed on the original audio signal was required. In this paper, we deployed a new method of attack localization based on differential algebraic, which guaranties a successful detection on the decoded audio signal. The algebraic method has also a reduced complexity compared to the index stationary detector used in [1]. The new proposed approach is evaluated for single audio coding-decoding, using objective perceptual measures. The experimental results for MP3 coding exhibits an efficient restoration of the attacks and a significant improvement of the audio quality.**

***Index Terms*—Temporal envelope, ARMA modeling, audio coding, sound attack, algebraic detector, attack restoration.**

I. Introduction

Transient waveforms, window length and psycho-acoustic bit allocation interact to produce pre-echo in low bit-rate audio coding (see Figure 1). When a transient occurs, a perceptual model allocates few bits for the quantization of the frame parameters. At the decoder, the quantization noise, supposed to be fully masked, may spread over the entire block. Therefore, this noise precede the time domain transient and then produce a potentially audible artefact known as pre-echo [2]. In addition, in a low bit-rate context, the attacks may be smoothed through coding, which reduces the percussive quality of sounds.

Many methods have been proposed to tackle the problem of echo in transform audio coding, especially for the case of modified discrete cosine transform (MDCT) coding. The most popular approach is to make the filterbank signal adaptive, using window switching controlled by transient detection [2] or close-loop decision. Usually window switching implies extra delay and complexity compared with using a non-adaptive filterbank. Another popular approach is the temporal noise shaping (TNS) [3] which allows the encoder to control the temporal fine structure of the quantization noise.

A method was proposed in [1] aiming at reducing echo artifacts after transform decoding. The principle is to restore the temporal envelope of the signal. Time envelope computation is based on linear prediction in frequency domain. Note that the restoration method requires the transmission of the coefficients of the ARMA model describing the temporal envelope and the transient position as side information. We suppose that we have an auxiliary channel to convey this side information, with a reduced bit-rate ($<$ 500 bps, for example a watermark).

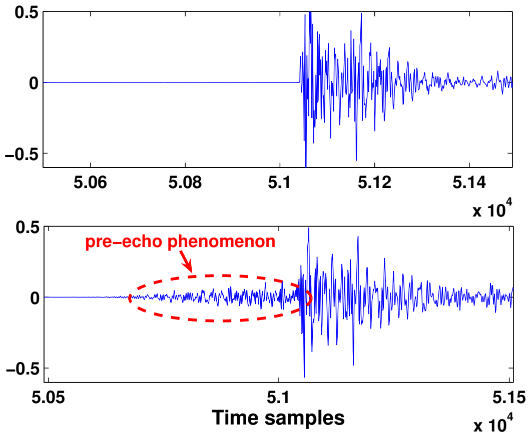

Fig. 1. illustration of the pre-echo artefact from castanet signal coded at 56 kbps using MP3 coder

In order to reduce the complexity of pre-echo reduction and to allocate all the available bit-rate to the transmission of the temporal envelope parameters, we propose a new method based on an algebraic detector, to localize transient positions on the decoded audio signal. Therefore, there is no more need to the transmit transient position information.

The remain of this paper is structured as follows: in section 2, we present the algebraic detection algorithm used to estimate transient positions. In section 3, the new approach dedicated to attack restoration is developed. Section 4 presents a performance evaluation of the proposed algorithm in the case of MP3 simple encoding.

II. Transient localization by algebraic detector

The transient localization proposed in [1] is based on a distance measurement between successive time-frequency representations of the signal, which is quite complex. Moreover, since the localization may be inaccurate in the decoded signal, two localizations are performed in the coding part: before and after coding-decoding, in order to transmit the anticipated error of localization through the side-channel.

In order to reduce the complexity of transient localization, we propose to use the change point detection method described in [4]. This latter is based on algebraic manipulations of a piecewise polynomial signal.

The input signal, $x(t)$, can be represented as a piecewise polynomial with maximum one discontinuity on the time interval $I_\tau^T = [\tau, \tau + T]$, $\tau > 0$. Let set

$$x_\tau(t) = x(\tau + t), \qquad t \in [0, T] \tag{1}$$

for the restriction of the signal in I_τ^T and 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

In the sense of the distribution theory, the N^{th} order derivative of the input signal can be written:

$$\frac{d^N x_\tau(t)}{dt^n} = x_\tau^{(N)}(t) + \sum_{k=1}^{N} \mu_{N-k}\delta(t - t_\tau)^{k-1} \tag{2}$$

where μ_k is the jump of the k^{th} order derivative at the point t_τ and $x_\tau^{(N)}$ represents the regular part of the N^{th} order derivative of the signal.

If $\mu_0 = \mu_1 = ... = \mu_k = 0$, there is no spike in the given interval. If $\mu \neq 0$, $0 \leq t_\tau \leq T$, there is a spike in given interval at location t_τ.

[4] proposes a detector $D(t)$, based on an algebraic determination of t_τ, computed through simple filtering of x. The detector D(t) relies on a decision function which must be greater than some threshold if a changing point exists in the interval. To illustrate the ability of the method to detect different change points in the same signal, the algorithm described below is implemented using only a third order derivative. The test signal is a batteries composed of 4 attacks (figure 2 (a)). The results in figures 2 (b) and (c) show how all the change-points are correctly detected: each change point position matches a corresponding attack position.

In the next experiment, we investigate the quality of algebraic detector versus the stationarity index detector [6]. Figure 3 compares error detections for both algebraic and stationarity index detector using original and mp3 decoded triangle-castanet audio signal. The error detection corresponds to the difference between the actual and estimated transient positions. The algebraic detector has high accuracy and closely approximates all the considered attacks positions.

III. PRE-ECHO REDUCTION SYSTEM

As detailled in [1], the restored audio signal, $\hat{x}_t$, is given at time $t > 0$ by:

$$\hat{x}(t) = x(t)_{dec} \frac{\hat{e}(t)}{\hat{e}(t)_{dec}}, \tag{3}$$

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

The correction constitutes a post-processing performed at the decoder. The parameters related to the temporal envelope estimation are extracted by the encoder and transmitted to the decoder through an auxiliary channel. Figures 4 and 5 illustrate the block diagrams of the basic treatments at the encoder and the decoder.

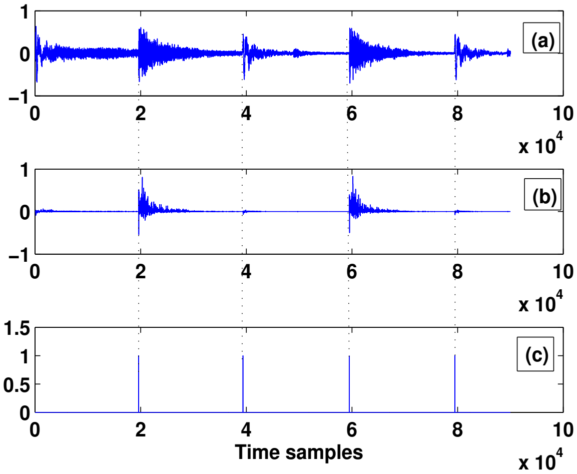

Fig. 2. Batterie signal (a), Decision function (b) and change point detector (c)

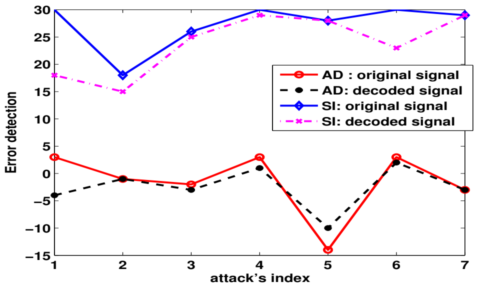

Fig. 3. Batterie signal (top) and corresponding change point detector (bottom)

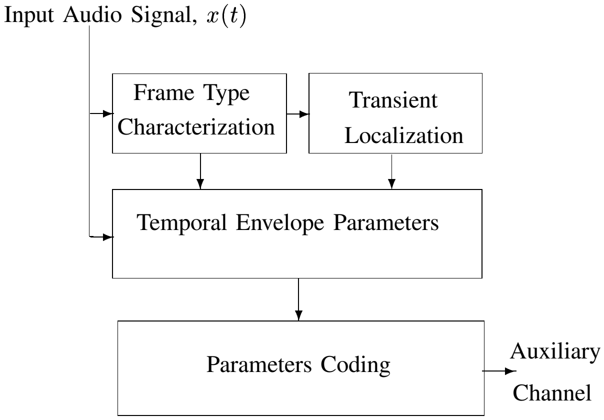

Fig. 4. Block diagram of treatment at the encoder.

The audio signal is first fed into a frame characterization in order to check if the frame is transient or not. To detect transient frames, we use the technique described in [5]. For

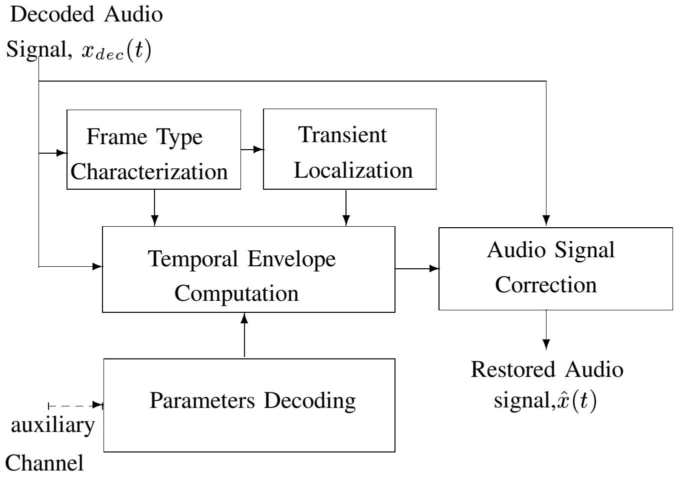

Fig. 5. Block diagram of treatment at the decoder .

the transient frames, a localization of the attack time positions is performed using the algebraic detector presented in section 2. Non transient frames are divided into two equal sub-frames. An estimate of the temporal envelope is computed using a frequency domain linear prediction model (FDLP) based on an ARMA model, which parameters are transmitted over a very low bit-rate auxiliary channel.

At the decoder, the received bitstream is decoded in order to extract the ARMA coefficients for computing the estimate of the original time envelope. In parallel, a frame type characterization and a transient localization are performed from the decoded signal. Estimates of the temporal envelopes for both original and decoded signals are computed. Finally, the restored audio signal, $\hat{x}(t)$, is obtained according to 3.

A. Temporal envelope ARMA modeling

he temporal envelope is estimated using the frequency domain linear prediction (FDLP). In fact, in the same way that TDLP (Time domain linear prediction) estimates the power spectrum, FDLPO estimates the temporal envelope of the signal, specifically the square of its Hilbert envelope [7]:

$$e(t) = \mathcal{F}^{-1}\{\int \tilde{X}(\varsigma)\tilde{X}(\varsigma - f)d\varsigma\} \tag{4}$$

i.e. the inverse Fourier transform of the autocorrelation of the single sided (positive frequency) spectrum $X\tilde{(}f)$.

The block diagrams of the temporal envelope estimate is depicted in Figure 6. To get an approximation of the Hilbert envelope, first the Discrete Cosine Transform (DCT) is applied to a given audio segment. Next, a linear prediction is applied to the DCT transformed signal in order to get an ARMA model. An estimation of the temporal envelope of $x(t)$ is therefore given by:

$$\hat{e}(t) = |H(e^{jwt})|, \tag{5}$$

where

$$H(z) = \frac{\sum_{i=0}^{q} b_i z^{-i}}{1 + \sum_{i=1}^{p} a_i z^{-i}} = \frac{H_b(z)}{H_a(z)} \tag{6}$$

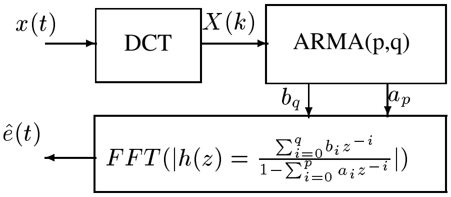
$x(t)$
DCT
$X(k)$
ARMA(p,q)
b_q
a_p
$\hat{e}(t)$
$FFT(|h(z) = \frac{\sum_{i=0}^{q} b_i z^{-i}}{1-\sum_{i=0}^{p} a_i z^{-i}}|)$

Fig. 6. Block diagram of the temporal envelope estimation.

where $(a_i)_{1\leq i\leq p}$ and $(b_i)_{0\leq i\leq q}$ are the ARMA coefficients.

The selection of the FDLP model order is guided by the temporal structure of signal in the same way as the TDLP model order is dictated by the formant structure. To illustrate the importance of model order, figure 7 shows a violon segment at 44.1 kHz sampling rate and its corresponding time envelope estimates obtained by ARMA(2,3) and ARMA(7,3) models. As expected, with an ARMA(2,3), only a smooth version of the time envelope is given, however, in the case of ARMA(7,3), the envelope almost fits the pitch pulses.

An evaluation of the FDLP model order based on objective measurements of the audio quality will be presented in section IV.

B. Parameter coding

After the ARMA parameters are estimated, they must be coded and transmitted through the auxiliary communication channel. The autoregressive coefficients (ARMA) are characterized by large dynamic range and would require many bits per coefficient for accurate coding. For this reasons, it is necessary to transform the ARMA coefficients into reflection coefficient (RC).

For each sub-frame, a vector grouping the RC coefficients is coded using a classical vector quantization technique. The codebook C is obtained by training on a database taken from various kinds of audio signals. It can be computed by the Lloyd-Max algorithm [8]. The size of the codebook and the corresponding bit-rate will be discussed in section IV.

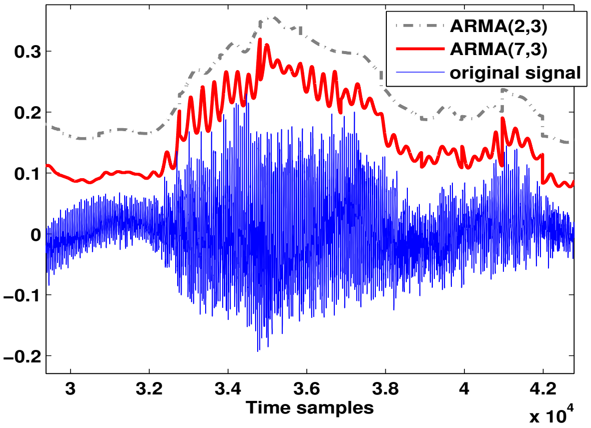

Fig. 7. Original violon segment and temporal envelopes using ARMA(2,3) and ARMA(7,3) model respectively.

IV. Experimental evaluation of the proposed approach

The experiments aim at validating our approach and comparing the restored audio signal to the original one. We use the PEMO-Q software described in [9] as an objective measurement to compare the score of Objective Difference Grade (ODG) and instantaneous Perceptual Similarity Measure (PSM$_t$). The ODG 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). The values of PSM$_t$ vary at the interval [0,1], with 1 indicating the best similarity between the reference and the test signals. These perceptual measures are correlated to the Subjective Difference Grade (SDG) for audio quality.

As a primary evaluation of the proposed approach, the coded/decoded castanet signal at 56 kbps and the corrected version are shown in Figure 8. It can be seen that the MP3 coder introduces a pre-echo and smoothes attacks. In the reconstructed signal, the pre-echo is considerably reduced and the attack is restored.

Given that the bit-rate offered by the auxiliary channel is limited to 20 bits per frame (434 bps), we study the influence of the ARMA order on the audio quality. Figure 9 compares the variations, over bit-rates, of the ODG and PSM$_t$ for both coded/decoded signal and its restored version. For comparison, we present the system evaluation using two ARMA model orders with two different bit-rate coding: ARMA(2,3) coded at 16 bits per frame (347 bps) as used in [1] and ARMA(5,3) coded at 20 bits per frame (434 bps). Remain that in addition to the ARMA(2,3) parameter coefficients, 4 bits are allowed to code the transient position in [1]. As illustrated in Figure 9, the proposed correction using ARMA(5,3) provides a significant enhancement of the PSMt and ODG. With only ARMA(2,3) the improvement is slighter.

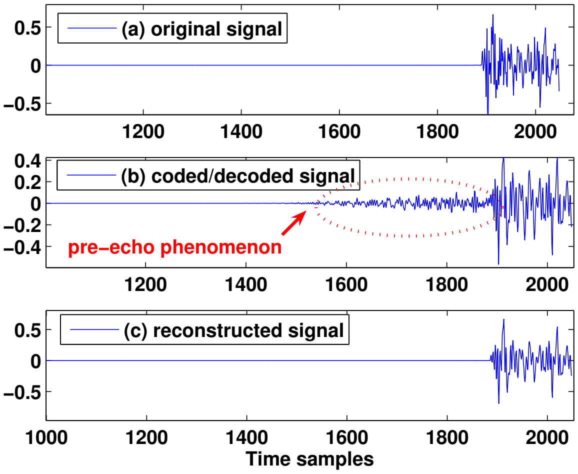

Fig. 8. Attack restoration for a castanet signal coded by a MP3 coder at 56 kbps

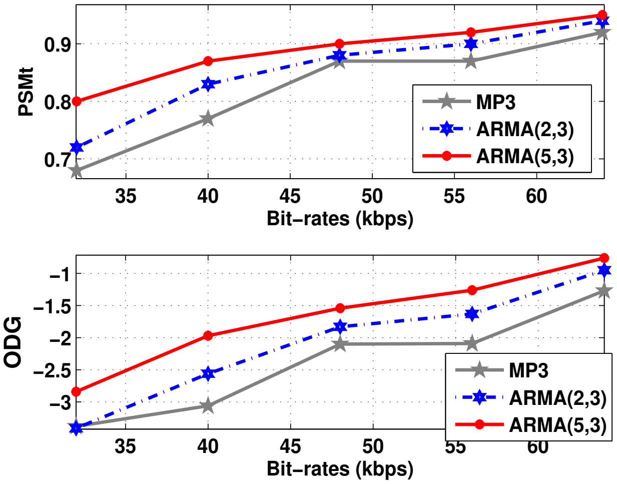

Fig. 9. Mean of Perceptual Similarity Measure and Objective Difference Grade for castanet signals using MP3 coder, with (ARMA) and without (MP3) correction.

V. Conclusion

Low-bit-rate coding-decoding with standard MP3 coder smoothes attacks in transients signals and increases the pre-echo. We have proposed an attack restoration method, based on accurate transient localization and temporal envelope correction, using a small set of information transmitted through an auxiliary channel. Our method enhances significantly the audio quality as measured by ODG and PSM$_t$.

References

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

[2] K. Iwai Kyle, *Pre-echo Detection and Reduction*, Master of science, Massachusetts Institute of Technologie, May 1994.

[3] Jrgen HERRE *Temporal Noise Shaping, Quantization and Coding Methods in Perceptual Audio Coding: a Tutorial Introduction* AES 17^th Internationanl Conference on High Quality Audio Coding.

[4] M.Mboup, C.Join and M.Fliess, *A delay estimation approach to change-point detection*, ICASSP 2008.

[5] 3GPP TS 26.403, *Advanced Audio Coding (AAC) part* September 2004.

[6] S. Larbi, M. Jaidane, *Audio Watermarking: A Way To Stationarize Audio Signals*, IEEE Trans. Signal Processing, Vol. 53 (2), pp. 816–823, 2005.

[7] M. Athineos, D. P.W.Ellis, *Frequency-Domain linear Prediction For Temporal Features*, ASRU 2003.

[8] 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, Jan 2007.

[9] 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.