TEMPORAL ENVELOPE CORRECTION FOR ATTACK RESTORATION IN LOW BIT-RATE AUDIO CODING

Imen Samaali[1,2], *Monia Turki-Hadj Alouane*[2], *Gaël Mahé*[1]

[1] Centre de Recherche en Informatique de Paris 5 (CRIP5), Université Paris Descartes, France
email: imen.samaali@math-info.univ-paris5.fr, Gael.Mahe@math-info.univ-paris5.fr
[2]Unité Signaux et Systèmes (U2S), Ecole Nationale d'Ingénieurs de Tunis (ENIT), Tunisia
email: m.turki@enit.rnu.tn

ABSTRACT

At reduced bit rates, the audio compression affects transient parts of signals, which results in pre-echo and loss of attack character. We propose in this paper an attacks restoration method based on the correction of the temporal envelope of the decoded signal, using a small set of coefficients transmitted through an auxiliary channel. The proposed approach is evaluated for single and multiple coding-decoding, using objective perceptual measures. The experimental results for MP3 and AAC coding exhibits an efficient restoration of the attacks and a significant improvement of the audio quality.

1. INTRODUCTION

In perceptual coders at low bit-rates, variations in the masking threshold from one frame to the next leads to different bit assignments. An artefact known as pre-echo may appear in signals with transients (see Figure 1). The silence before an attack may be affected by a relatively high quantization noise, since the masking threshold is computed using the part of the frame after the attack [1][2]. However, if the pre-echo is short enough, it is post-masked by the attack. The phenomenon of pre-echo is amplified by multiple coding-decoding: the quantization noise piles up at each cycle and may become audible.

Transient signals are affected by another artefact when coding at low bit-rates. Only a small low-frequency band of the signal is fully transmitted, whereas medium frequencies are restored at the decoder without phase information and higher frequencies are simply forgotten. This smoothes attacks resulting in a reduce of the percussive quality of sounds.

In order to reduce or eliminate pre-echo and enhance the audio quality, some techniques with different complexities are developed in the literature [1]. These techniques are based on the use of an adaptive window selection algorithm to switch between long and short transform window. Long windows offer higher prediction gain, and better frequency resolution, while short windows reduce the length of pre-echo.

Another technology, the "*Temporal Noise Shaping*" (TNS) [3], used in the AAC coder, aims at resolving the problem of Temporal Masking. This approach allows the encoder to control the temporal fine structure of the quantization noise. In the same way as temporal predictive coders reshape the quantization noise spectrum, the principle of TNS, is to replace the spectral coefficients to be quantized by the coder with their frequency prediction residual (which flattens the temporal envelope). In the decoder, the inverse prediction filter is applied to the residual. As a consequence, the temporal shape of the quantization error is adapted to the temporal shape of the audio signal. For the MP3 coder, a similar correction is performed by the "*Temporal Masking*" (TM) technology [4].

This work is part of the WaRRIS project granted by the French National Research Agency (project n° ANR-06-JCJC-0009) .

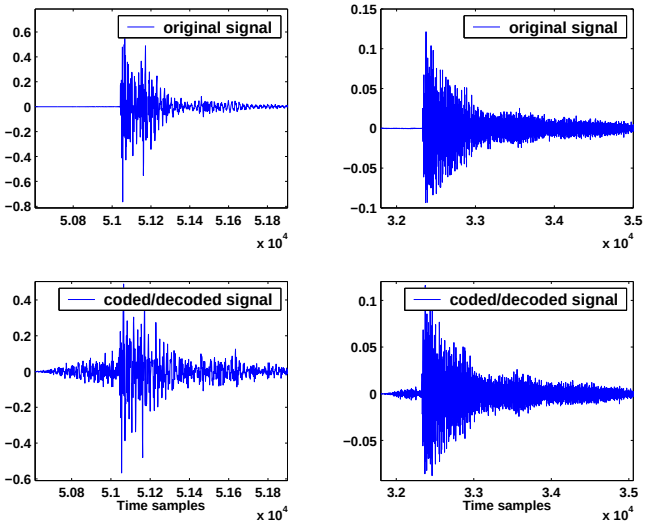

Figure 1: Original signal (top) and pre-echo (bottom) from castanet (left) and triangle (right) coded at 56 kbps using MP3 coder

Although these techniques reduce significantly the pre-echo phenomenon, the problems of temporal masking and attacks smoothing remain harmful for some transient signals, such as castanet, glockenspiel, triangle or certain types of speech signals. In this paper, we propose a novel method aiming at restoring attacks in coded-decoded signals. The idea is to perform a correction of the temporal envelope of the coded-decoded signal, using a small set of parameters (describing the temporal envelope of the original signal) transmitted over a very low bit-rate (< 1 kbps) auxiliary channel which we suppose to have (an audio watermarking channel for example).

This paper is structured as follows: in section 2, we present the new approach dedicated to attack restoration for the quality enhancement of audio decoded signals. Section 3 presents a performance evaluation of the proposed algorithm in the case of simple and multiple successive encodings (codings in tandem).

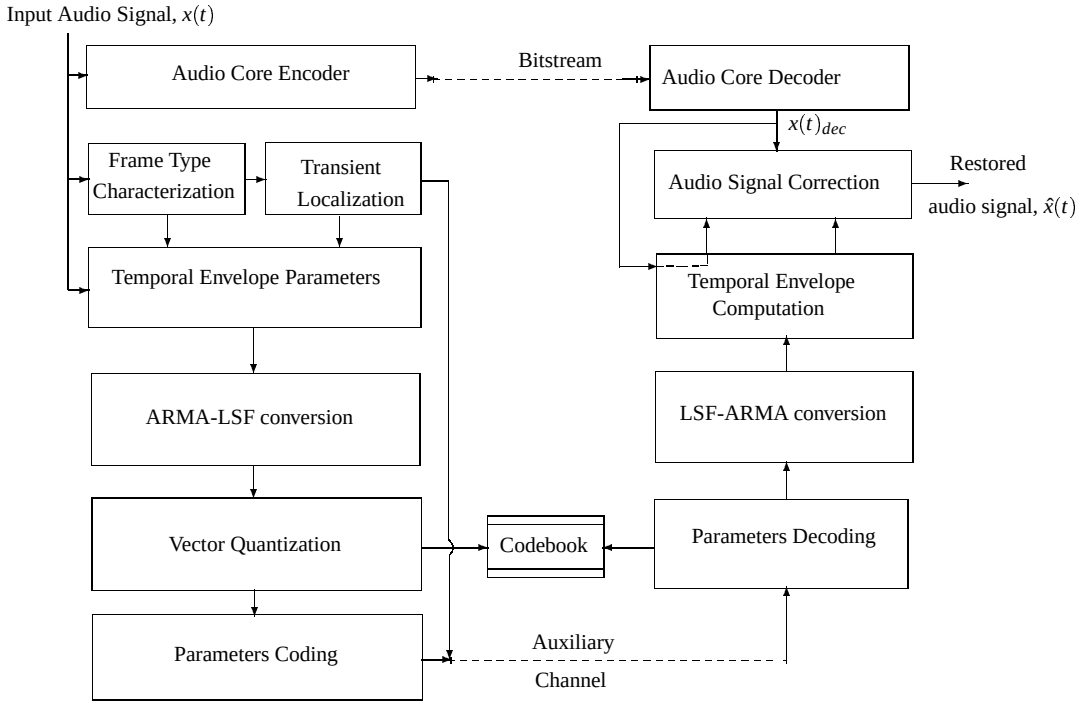

Figure 2: Block diagram of the proposed approach.

2. ENVELOPE CORRECTION

The proposed method is based on the correction of the temporal envelope of the decoded signal. The restored audio signal, $\hat{x}_t$, is given at time t 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 temporal envelope of the original signal, and $\hat{e}(t)_{dec}$ is the temporal envelope of the coded-decoded signal.

The correction approach constitutes a post-processing performed at the decoder. For this purpose, crucial parameters required for the temporal envelope estimation are extracted by the encoder and transmitted to the decoder through an auxiliary channel. Figure 2 illustrates the basic structure of the proposed approach. In addition to the standard coder/decoder blocks, the proposed system includes several components:

- Frame characterization/Transient localization: characterizes the frame type as transient or not. In particular, for the transient frames, a localization of the attack's time position is performed.
- Temporal envelope coding and vector quantization: based on linear prediction in frequency domain.
- Audio signal correction: for the restoration of the audio signal according to the relation presented in equation 1.

2.1 Transient detector

- **Characterization of frame type:**

To detect transient frames, we use the technique described in [5]. After high-pass IIR filtering with the transfer function

$$H(z) = \frac{0.7548(z-1)}{z-0.5095}, \quad (2)$$

each frame of 1024 samples is divided into 8 sub-blocks of 128 samples and the energy of each sub-block is computed by summing up the squared samples. An attack is detected if one of these sub-block energies exceeds a sliding average of the previous energies by a constant factor *attackRatio* and is greater than a constant energy level $minAttackNrg = 10^{-3}$ [5].

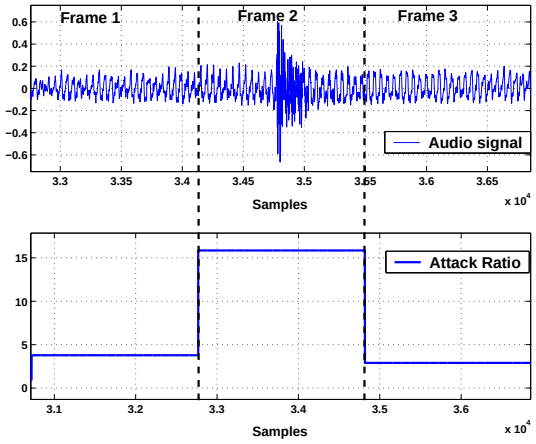

Figure 3: The audio signal (violon+castanet) and its corresponding attack ratio coefficients

Figure 3 illustrates the efficiency of the proposed method. The attack ratio coefficient corresponding to the transient frame (Frame 2) exceeds the threshold fixed to 10 as in [5].

- **Transient localization:**

The method used to detect the transient time position is based on the stationarity index which corresponds to the Kolmogorov distance measured between the time frequency representation (TFR) of the signal at different times [6]. The stationarity index is given by:

$$SI(t) = \int_{\tau=0}^{p} \int_{f=-\infty}^{+\infty} |NI_1(t;\tau,f) - NI_2(t;\tau,f)| d f d\tau. \quad (3)$$

$NI_1(t;\tau,f)$ and $NI_2(t;\tau,f)$ represent a normalization of respectively subimages $I_1(t;\tau,f)$ and $I_2(t;\tau,f)$:

$$I_1(t;\tau,f) = TFR(t-p+\tau,f); \quad (4)$$
$$I_2(t;\tau,f) = TFR(t+\tau,f). \quad (5)$$

The parameter p delimits the considered analysis duration at each instant t and allows the selectivity/sensivity control of the SIs: a high value of p leads to smoother SIs. As in [6], p is fixed to 20. As illustrated in Figure 4, the peak of SI corresponds to the attack position.

- **Attack position coding:**

Since coding directly the transient position would require a high bit-rate, we propose to code the difference between the actual attack position given by the original frame and the attack position computed from the coded/decoded frame. A 6 bits scalar quantization technique is used for this purpose.

2.2 Temporal envelope ARMA modeling

The reduced bit-rate of the auxiliary channel implies a compact representation of the transmitted temporal envelope, through a small set of coefficients for each frame. The proposed coding approach is based on linear prediction in frequency domain providing an approximation of the temporal envelope of a signal, specifically the squared Hilbert envelope [7]. The block diagrams of the temporal envelope estimation is depicted in Figure 5.

The Discrete Cosine Transform of $x(t)$ is modeled by an ARMA(p,q) model. The ARMA model was here preferred to the classically used AR model, 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

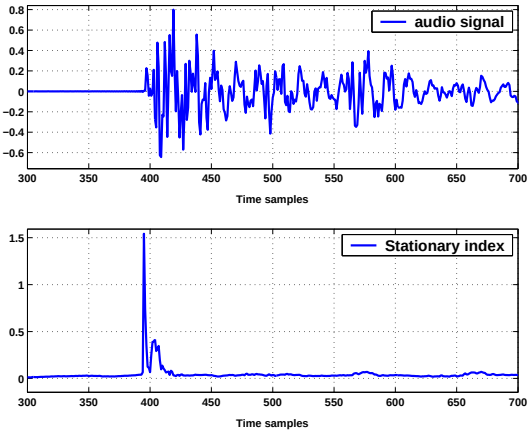

Figure 4: Transient frame (top), its corresponding stationary index (bottom)

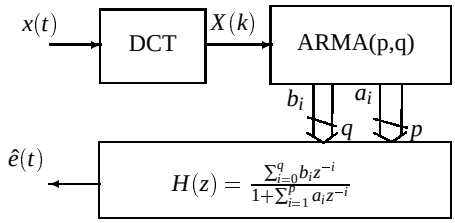

Figure 5: Block diagram of the temporal envelope estimation.

by minimizing the mean square prediction error defined as follows:

$$E_{pre}(k) = X(k) + \sum_{i=1}^{p} a_i X(k-i), \quad (6)$$

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:

$$\hat{e}(t) = |H(e^{jwt})|, \quad (7)$$

where

$$H(z) = \frac{\sum_{i=0}^{q} b_i z^{-i}}{1 + \sum_{i=1}^{p} a_i z^{-i}} = \frac{H_b(z)}{H_a(z)} \quad (8)$$

Each frame of 2048 samples is divided into two subframes, each modeled with an ARMA(2,3). For non transient frames, each sub-frames has 1024 samples. Transient frames are divided according to the transient position.

Figure 6 illustrates similarity between the estimated envelope and the original one for a castanet sound sampled at 44.1 kHz.

2.3 ARMA to LSF transformation

After the ARMA parameters are estimated, they must be coded and transmitted through the auxiliary communication channel. The autoregressive coefficients (AR) are characterized by large dynamic range and would require many bits per coefficient for accurate coding. In addition, small changes in the AR coefficients may result in instability of the synthesis filter. For these reasons, it is necessary to transform the

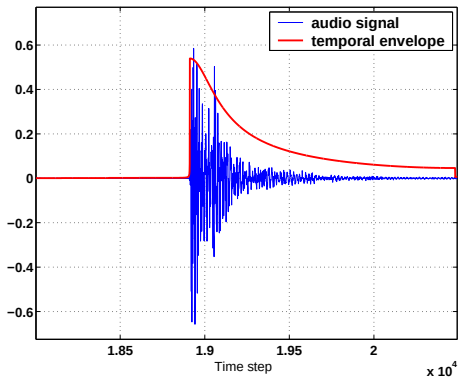

Figure 6: The castanet signal and its corresponding temporal envelope estimate

AR coefficients into an equivalent representation which ensures the stability : the Line Spectral Frequency representation (LSF) are classically used in predictive coders.

The AR filter $H_a(z)$ of order p, given by equation 8, can be represented by :

$$H_a(z) = \frac{1}{2}[P(z) + Q(z)]. \quad (9)$$

where $P(z)$ and $Q(z)$ are even and odd symmetric filters respectively. Their roots, $z_i = e^{j\theta_i}$, lie on the unit circle [8]. The $\{\theta_i,\ i = 1,..,N\}$ represent the Line Spectrum Frequency (LSF) coefficients.

To transform the Moving Average (MA) parameters into the LSF coefficients, the MA filter $H_b(z)$ (equation 8) must be modified as follows

$$H_b(z) = 1 + \sum_{i=1}^{q} b'_i z^{-i}, \quad (10)$$

where $b'_i = \frac{b_i}{b_0}$, $i = 1,..,q$. In the decoder, the value of b_0 is estimated from the decoded signal.

2.4 Vector quantization

For each sub-frame, a vector grouping 5 LSF coefficients is coded using a classical vector quantization technique. The codebook C of a dimension $K \times L$, (K=256 and L=5 here), is obtained by training on a database of approximately $100 \times K$ vectors taken from various kinds of audio signals. It can be computed by the Lloyd-Max algorithm [9].

For the proposed system, the LSF parameters are coded on 8 bits for each sub-frame. We recall that for transient frame, 6 bits are added to code the attack position. Consequently, the auxiliary channel bit-rate varies between 344 and 474 bps.

3. EXPERIMENTAL EVALUATION OF THE PROPOSED APPROACH

The experiments aim at validating our approach and comparing the restored audio signal to the original one. We used the PEMO-Q software described in [10] as a tool to measure the Objective Difference Grade (ODG) and the 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 in the interval [0,1], with 1 indicating the similarity between the reference and the test signals, whereas smaller values correspond to larger deviations between them. The perceptual measures are correlated to the Subjective Difference Grade (SDG) for audio quality.

At first, the audio quality is evaluated when a single coding is performed. For each experiment, the reference audio signals are coded by MP3 and AAC coders at different bitrates varying from 24 to 96 kbps. The considered reference audio signals for all the simulations are mono castanet and triangle, which exhibit a remarkable transient character.

As illustration of the proposed approach, the 56 kbps MP3 coded/decoded castanet and triangle signals and their corrected versions are shown in Figure 7. It can be seen that the MP3 coder introduces a pre-echo and smoothes attacks.

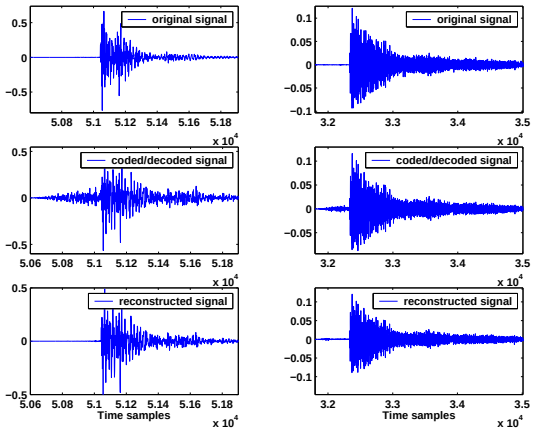

Figure 7: Attack restoration for a castanet (left) and triangle (right) signals coded by a MP3 coder at 56 kbps

In the reconstructed signal, the pre-echo is considerably reduced and the attack is restored.

Figures 8 and 9 compare the variations, over bit-rate, of the ODG and PSM_t for both coded/decoded signals and their restored versions. As illustrated in Figures 8, for MP3 coder even with TM, the proposed correction provides a significant enhancement of the PSM_t and ODG. With AAC coding (Figure 9), the improvement is slighter.

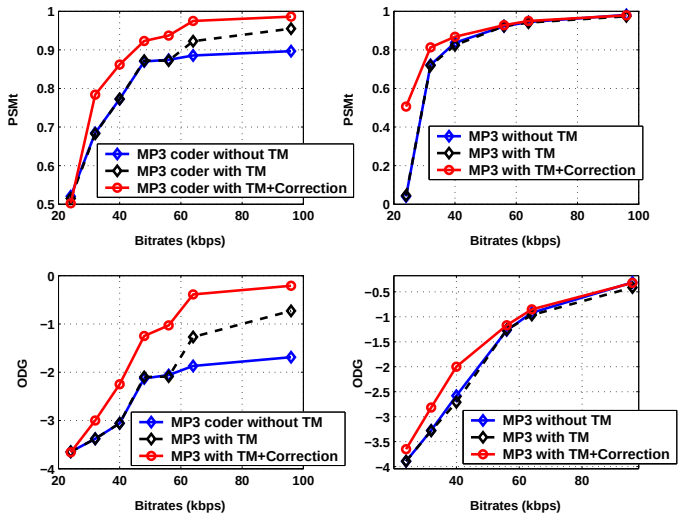

Figure 8: Mean of Perceptual Similarity Measure and Objective Difference Grade for castanet (left) and triangle (right) signals (MP3 coder)

At a second stage, the perceived quality in multi-encoding case is analyzed. Referring to the experimental results of Reiss [2] related to the stereo MP3 coder at 128 kbps, the perceptual audio quality deteriorates with the increase of encoding number. Similar results were obtained for mono MP3 coder at 64 kbps for castanet and triangle sequences.

The variations of the perceptual audio quality (PSM/ODG) with the number of encodings are shown in Figures 10 and 11. Without correction, we observe a fast deterioration. In fact, the quantization noise from each cycle piles up, so that the pre-echo, which was post-masked by the

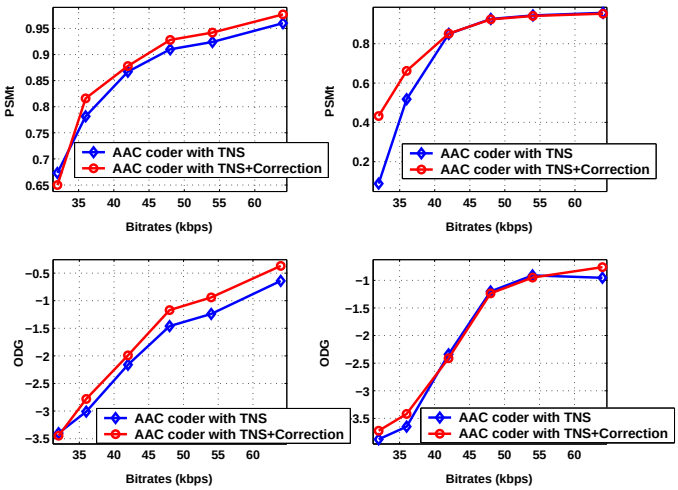

Figure 9: Mean of Perceptual Similarity Measure and Objective Difference Grade for castanet (left) and triangle (right) signals (AAC coder)

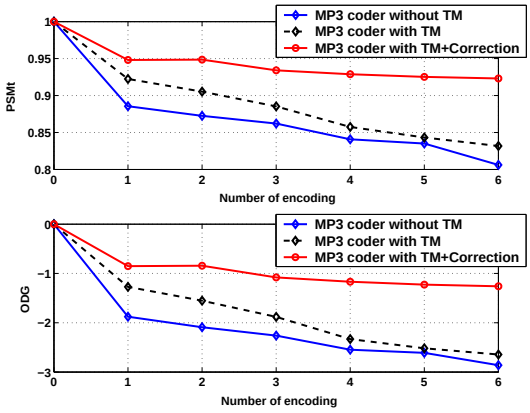

Figure 10: Effect of multiple encoding on audio quality for castanet signal (MP3 coder)

attack in the first encodings, is not masked anymore in the next encodings and becomes annoying.

For the castanet signal, the temporal masking 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 towards that without TM, whereas with our correction (with always the same envelope), the quality remains transparent.

4. CONCLUSION

Low-bit-rate coding-decoding with standard coders AAC and MP3 smoothes attacks in transients signals and increases the pre-echo despite TNS/TM. We have proposed an attack restoration method, based on 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, and avoids the dramatic fall of quality caused by multiple coding-decoding. All components of the proposed system have a reasonable complexity, except of the transient localization: other methods

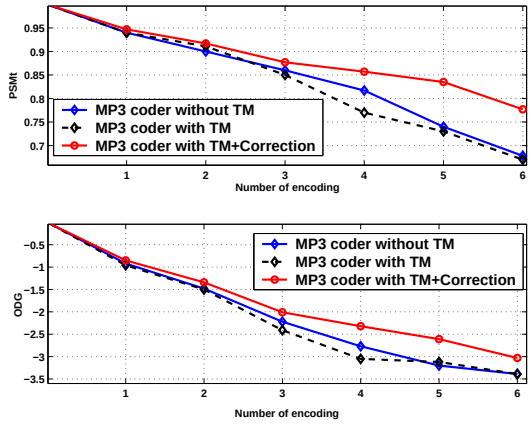

Figure 11: Effect of multiple encoding on audio quality for triangle signal (MP3 coder)

than computating stationarity indices should be considered.

Further study will aim at using the audio watermarking as an auxiliary communication channel and evaluating the influence of the watermark detection error on the proposed system.

REFERENCES

[1] Mahieux Y and Petit J.P *"High-quality audio transform coding at 64 kbps"*, IEEE Transactions on Communications, vol. 42, pp. 3010-3019, Nov. 1994.

[2] J. Reiss and M. Sandler, *"Audio Issues In MIR Evaluation"*, ISMIR, Barcelona, Spain, Oct. 2004, pp. 28-33.

[3] J. 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, Florence, Italie, Sep 1999, pp. 312-325.

[4] F. Sinaga, T. Surya Gunawan and E. Ambikairajah, *"Wavelet Packet Based Audio Coding Using Temporal Masking"*, ICICSPCM, Dec. 2003, pp. 1380-1383.

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

[6] S. Larbi and M. Jaidane, *"Audio Watermarking: A Way To Stationarize Audio Signals"*, IEEE Trans. Signal Processing, Vol. 53, pp. 816-823, Feb. 2005.

[7] M. Athineos and D. P.W.Ellis, *"Frequency-Domain linear Prediction For Temporal Features"*, ASRU'03, Nov. 2003, pp. 261-266.

[8] D. Na, S. Yeon and M. BAE, *"The High-speed LSF transformation Algorithm in CELP Vocoder"*, TENCON 2000 Proc, vol. 1, pp. 279-282, 2000.

[9] 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.

[10] R. Huber and 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, pp. 1902-1911, Nov 2006.