International Journal of Electronics and Electical Engineering

Volume 1 | Issue 3 Article 12

January 2013

Audio Watermarking Scheme in MDCT Domain

Santosh Kumar Singh
Electronics and Communication Engineering, Netaji Subhas Institute of Technology, Sec. 3, Dwarka, New Delhi, 110078, India, ersksingh_mtnl@yahoo.com

Jyotsna Singh
Electronics and Communication Engineering, Netaji Subhas Institute of Technology, Sec. 3, Dwarka, New Delhi, 110078, India, jsingh.nsit@gmail.com

Audio Watermarking Scheme in MDCT Domain

Santosh Kumar Singh and Jyotsna Singh

Electronics and Communication Engineering, Netaji Subhas Institute of Technology,
Sec. 3, Dwarka, New Delhi, 110078, India.
E-mails: ersksingh_mtnl@yahoo.com & jsingh.nsit@gmail.com

Abstract *- In our proposed watermark embedding system, the original audio is segmented into overlapping frames. The psychoacoustic auditory model has been utilized to calculate the global threshold in modified discrete cosine transform domain. The perceptual insignificant locations have been used to insert the appropriately scaled watermark in the transform domain. Blind detection of watermark has been performed. Simulation results indicate that the proposed watermarking system is perceptually transparent and robust against various kind of attacks such as AWGN addition and MP3 compression.*

Keywords- *digital watermarking, global threshold, modified discrete cosine transform.*

1. INTRODUCTION

To uphold commercial value of the audio, a successful digital watermarking technique should include certain characteristics such as imperceptibility, robustness and security. The main requirement for watermarking is perceptual transparency. The embedding process should not introduce any perceptible artifacts, that is, the watermark should not affect the quality of the original signal. Secondly, the watermark must be strongly resistant to unauthorized detection. It should be difficult for an unauthorized agent to forge the watermarked audio. However it is difficult to achieve robustness keeping the perceptual artifacts as low as possible. Thus, there must be a trade-off between perceptual transparency and robustness. Watermark security addresses the secrecy of the embedded information.

To determine the irrelevant components of the signal, the two 'masking' properties of HAS are utilized: Absolute Threshold of Hearing (ATH) and Auditory Masking. These two parameters, determines which portions of a signal are inaudible and indistinguishable to average human, and thus can be removed from a signal. ATH is characterized by amount of energy needed in a pure tone such that it can be detected by listener in a noiseless environment. Auditory Masking occurs whenever there is presence of a strong audio signal that makes a temporal or spectral neighborhood of weaker audio signal imperceptible [1]. Masking effects are present in time as well as in frequency domain but the frequency domain masking is more profound and hence important for audio watermarking.

In SS based watermarking schemes, watermark embedding is mostly preferred in transform domain of audio in order to obtain high robustness. Boney [2] generated the watermark by filtering a PN-sequence with a filter approximating the frequency masking characteristics of the HAS. Swanson et al. [3] presented an audio watermarking algorithm that exploits temporal and adding a perceptually shaped spread-spectrum sequence. Kirovski [4] exploited psychoacoustic auditory model to shape and embed the watermark for embedding it into Modulated Complex Lapped Transform (MCLT) coefficients of an audio signal. In [5] a new auditory masking model is proposed for DFT magnitude that can be utilized in spread spectrum audio watermarking technique.

Most of the digital audio transmission and storage are in compressed domain algorithms such as MPEG-1/2; Layer III (mp3), MPEG-2/4 AAC, Dolby AC2/AC3, and numerous experimental audio coding algorithms. Therefore, it has been desired to insert watermark in compressed audio directly. Over the last decade, modified discrete cosine transform (MDCT) has emerged as the most effective transform for audio coding due to its time domain alias cancellation property and energy compaction property [6], [7] and [8]. MDCT coefficient is widely used as it lessens the blocking artifacts that deteriorate the reconstruction of transform audio coders with overlapped transforms.

In this paper we present a technique for embedding watermark into MDCT coefficients of audio signal. The proposed technique takes advantage of the human auditory systems inability to hear watermark noise under the auditory masking effects. This masking effect occurs whenever a strong audio signal appears in the spectral neighborhood of weaker signal. The masking threshold is calculated in modified discrete cosine transform domain for watermark shaping before embedding process. This shaped watermark is then embedded additively in MDCT domain of audio signal using

spread spectrum based technique. Blind watermark detection is performed using correlation technique. The receiver only requires secret key for detection of watermark.

The paper is organized as follows. In section-II, we have described the main principle of Spread Spectrum watermarking technique. Section-III presents the proposed watermarking scheme, including the embedding and detection process. Finally, in the last section-IV, we have discussed the results in terms of transparency and robustness.

II. SPREAD SPECTRUM WATERMARKING

Spread spectrum (SS) based watermarking scheme uses a concept from spread-spectrum communication. In this technique the original audio signal **x** will be called *noise* and the bit stream that forms the watermark sequence **w** will be the *data* signal. The watermark is defined as a direct SS sequence, which is a vector pseudo-randomly generated such that $\mathbf{w} \in \pm 1^N$. Each element is usually called a chip.

$$\mathbf{y} = \mathbf{x} + \alpha \mathbf{w} \qquad (1)$$

As defined in Cox et al [9] the watermarked signal **y** is obtained by embedding watermark **w** in host feature of audio signal **x.**

III. PROPOSED TECHNIQUE

A. Watermark Embedding Scheme: The scheme has been shown by figure 1,and described by following 8 steps:

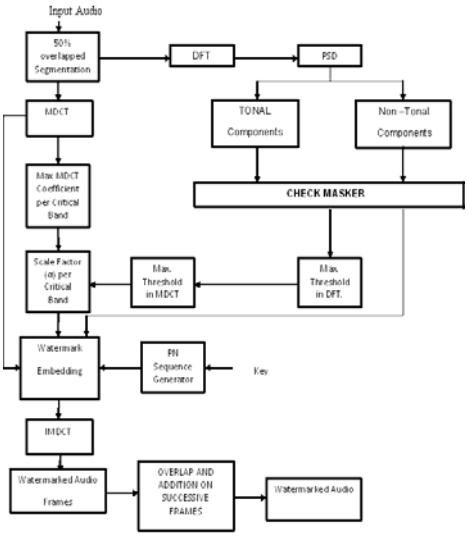

Figure 1: Watermark Embedding Blocks

Step 1 - Segmentation : Input Audio in .wav format is taken and converted to PCM stream. PCM is a method used to digitally represent a sampled analog signals. The PCM stream of input audio signal, is then segmented into 50 % overlapping frames of length 512 each, by windowing with Kaiser window of 512 sample length.

Step 2 – Peak detection: Based on fact that HAS analyzes a broad spectrum into different parts, called Critical Band. In Critical band energy of sound remains relatively constant. A more uniform measure of frequency based on critical band is *bark*. The MDCT of input audio frame of length 512 has been performed to get MDCT coefficients. The 512 sample scale is converted to frequency scale given as

$$f = \frac{N}{fs} * (1:N) \qquad (2)$$

Where, N = Frame length and f_s = sampling frequency of input audio. Further frequency scale has been converted to bark scale given as

$$z = 13arctan(0.00076 * f) + 3.5arctan\left(\frac{f}{7500}\right)^2 \qquad (3)$$

here, z is frequency in barks and f is frequency in hertz. The maximum values of MDCT coefficients are calculated in each of 25 critical bands. The MDCT Transform can be taken as an analysis synthesis filter bank where the impulse response of analysis filters as given as

$$h_k(n) = w(n)\sqrt{\frac{2}{M}} \cos\left[\frac{(2n+M+1)(2k+1)\pi}{4M}\right] \qquad (4)$$

The synthesis filters are simply obtained by a time reversal of analysis filters to satisfy the overall linear phase constant, given as

$$g_k(n) = h_k(2M-1-n) \qquad (5)$$

For audio coding application, the MDCT analysis filter is performed on a block of samples with length 2M samples and advanced only M samples after each block transform operation, i.e., with 50% overlap. The MDCT filter bank is critically sampled in a sense that only M coefficients are generated for each forward transform block. Given, an input block x(n), the transform coefficients X(k) for 0≤k<M is given as

$$X(k) = \sum_{n=0}^{2M-1} x(n) h_k(n) \qquad (6)$$

Given transform coefficients X(k), the reconstructed samples x(n), for 0≤n<M-1, will be given

$$x(n) = \sum_{k=0}^{M-1} [X(k)h_k(n) + X^P(k)h_k(n+M)] \qquad (7)$$

Where $X^P(k)$ denotes the transform coefficients of previous block. Clearly, the above 50% overlap and add operation virtually eliminates the blocking artifacts that plague the reconstruction of transform audio coders with non-overlapped transform.

Step 3 - Psychoacoustic Auditory Model: The simultaneous masking threshold is evaluated using following steps:

a. FFT on input audio frame **x** of 512 sample length, performed, to get 512 coefficients. After multiplication of the segmented frame with the Kaiser window, the FFT of the input block is performed

b. To derive the masking threshold, the power density spectrum of the input block is evaluated as

$$P(k) = 10\log_{10}\{(\text{abs}(\text{fft}(\mathbf{x}))^2 \text{dB} \quad (8)$$

Where, x represents audio vector, N represents frame length. k= numbers of samples in the transform domain. The maximum value of the power density spectrum X is normalized to a value of +96 dB given *as*

$$P(k) = P(k) - \max(P(k)) + 96 \text{ dB} \quad (9)$$

c. Tonal and Non Tonal components: To determine whether a frequency component is a tone requires knowing whether it has been held constant for a period of time, as well as whether it is a sharp peak in the frequency spectrum, which indicates that it is above the ambient noise of the signal. Determining whether a certain frequency is a tone (masker) can be done with the following definition: A frequency f (with FFT index k) is a tone if its power P[k] is:

1. Greater than P[k-1] and P[k+1], i.e., it is a local maxima.
2. 7 dB greater than the other frequencies in its neighborhood, where the neighborhood is dependent on f: If 0.17 Hz < f < 5.5 kHz, the neighborhood is [k-2…k+2], If 5.5 kHz =< f < 11 kHz, the neighborhood is [k-3…k+3], and If 11 kHz =< f < 20 kHz, the neighborhood is [k-6…k+6].

If a signal is not a tone, it must be noise. Thus, one can take all frequency components that are not part of a tone's neighborhood and treat them like noise. Since humans have difficulty discriminating signals within a critical band, the noise found within each of the bands can be combined to form one mask. Thus, the idea is to take all frequency components within a critical band that do not fit within tone neighborhoods, add them together, and place them at the geometric mean location within the critical band. This has been repeated for all critical bands.

The maskers which have been determined affect not only on the frequencies within a critical band, but also in surrounding bands. Studies show that the spreading of this masking has an approximate slope of +25 dB/Bark before and -10 dB/Bark after the masker. The spreading can be described as a function given by equation (10), that depends on the maskee location i, the masker location j, the power spectrum Ptm at j, and the difference between the masker and maskee locations in Barks (deltaz =z(i)-z(j)):

$$SF(i,j) = \begin{cases} 17deltaz - 0.4Ptm(j) + 11 & -3 \le deltaz < -1 \\ (0.4Ptm(j) + 6)deltaz & -1 \le deltaz < 0 \\ -17deltaz & 0 \le deltaz < 1 \\ (0.15Ptm(j) - 17)deltaz - 0.15Ptm(j)\ 1 \le deltaz < 8 & \end{cases} \quad (10)$$

There is a slight difference in the resulting mask that depends on whether the mask is a tone or noise. As a result, the masks can be modeled as

$$For\ tones: Ttm(i,j) = Ptm(j) - 0.275z(j) + SF(i,j) - 6.025\ (dB\ SPL)$$

$$For\ noise: Tnm(i,j) = Pnm(j) - 0.175z(j) + SF(i,j) - 2.025\ (dB\ SPL) \quad (11)$$

d. Check Maskers: Eliminates maskers that are either within a critical band or covered by the absolute threshold of hearing. The removal of masked sound, whose value is less then masking threshold has been removed as, shown in figure 2. After removing the irrelevant components, the location has been saved, for the purpose of inserting watermarks at these locations only.

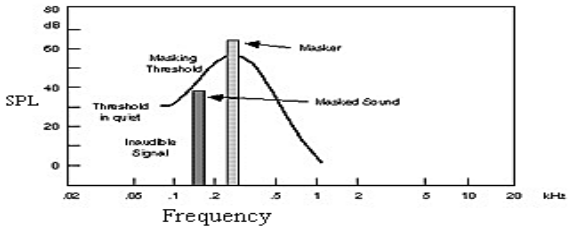

Figure 2: Masked sound has been removed and its location has been used for inserting watermark

e. Maximum global masking threshold in DFT domain has been calculated in each critical band. The de-normalisation by 96 dB, done on calculated maximum global threshold in DFT domain and then the

corresponding magnitude of Maximum Global Masking Threshold calculated given as

$$mag\ (z) = 10^{(0.1*(dB(z)-96))} \quad (12)$$

Where, z = Barks ranging from 1 to 25 ; mag is the magnitude of Maximum Global Masking Threshold; and dB is the Maximum Global Masking Threshold value in dB. Thus, overall in Step 3, the PAM of HAS has given us two important parameters for watermark embedding, firstly the locations of perceptually irrelevant components, where our intention is to insert the watermark and secondly Maximum Global Masking Threshold to further find out the scale factor to scale the inserted watermark, so that they are imperceptible by keeping them below the threshold limit.

Step 4 - Maximum global masking threshold in MDCT domain has been calculated in each critical band: Since the masking threshold is calculated based on the DFT of the input frame, it is not accurate to use this masking threshold for the MDCT coefficients. Instead, we consider the following relationship between the DFT and MDCT to find a more accurate masking threshold for MDCT coefficients [10], given as

$$C(k) = \sqrt{\frac{2}{M}}\ |S(k)|\cos\left[\frac{2\pi n_0(k+0.5)}{N} - \angle S(k)\right] \quad (13)$$

Where, $S(k)$ is the Fourier transform of the modulated windowed input signal, $C(k)$ is the MDCT, $n_0 = (M + 1)/2$, M and N are the number of samples in the frequency and time domain respectively. If m_{DFT} is the masking threshold corresponding to the k_{th} DFT coefficient, then in order to have the same Signal-to-Mask Ratio (SMR) at any coefficient in the DFT and MDCT domain, the relation should hold given as

$$\frac{C^2(k)}{m_{MDCT}} = \frac{|S(k)|^2}{m_{DFT}} \quad (14)$$

Step 5 - Scale Factor : Since, maximum value of MDCT Coefficient from Step 2, and Maximum value of Threshold in MDCT domain from step 4 has been found for each critical bands, thus scale factor [11] can be calculated , given as

$$\propto (z) = A * \left(\frac{\sqrt{T(Z)}}{\max(|C(Z)|)}\right) \quad (15)$$

Where, z ranging from 1 to 25. The square root of the final threshold is divided by the maximum magnitude component found in the energy of the new watermark in each critical band. Each one of these factors is scaled using the gain A, that varies from 0 to 1, and controls the overall magnitude of the watermark signal in relation with the audio signal.

Step 6 - Watermark PN sequence generator: Watermark sequence of length 'a' has been generated by a PN sequence generator using a Key of length 'b'. Key is sequence of 0 and 1. The, then watermark sequence has been changed from 0 to -1, to get w(j). Where j= 1,2,…a.

Step 7 - Watermark embedding: Watermark embedding can be done additively, given as

$$C'(k) = C(k) + \propto (z) * w(j) \quad (16)$$

such that, C'(k) = C(k) ∀ locations of dominant components, and C'(k) = $\propto (z) * w(j)$ ∀ perceptually irrelevant locations from which non dominant components has been removed in step 3. Where C'(k) is the watermarked MDCT coefficients of audio input frame, C(k) is the MDCT coefficients of audio input frame, from step 4. $\propto (z)$ is the scale factor, computed in step 5, w(j) is the watermark sequence to be embedded, computed in step 6. z ranging from 1 to 25 Barks, and j is the length of watermark sequence.

Step 8 - Watermarked audio signal in time domain obtained by taking Inverse MDCT of C'(k). Time Domain Aliasing Cancellation of overlapping signal has been done using Overlap and Add method [12], to reconstruct the watermarked audio signal. Figure 4, describes the method using three overlapping frames. Figure 5, indicates the close resemblance of the audio input samples to the watermarked audio samples.

B. Watermark Detection and Extraction Scheme

The scheme has been described by figure 3. If the correlator output is maximum, for the known Key of watermark PN sequence used at the embedder end, then it is indicated that the correct watermark has been detected.

Step 1: Lists of Keys(key1,key2…keyn) used to generate corresponding Pseudo Number (PN) sequences (w1,w2 …wn).

Step 2: Generated PN sequence in step1, has been correlated with the watermarked sequence of audio input.

Step 3: The key for which correlator has maximum value is the correct key for the desired watermark sequence that has been embedded at embedder side. Thus watermark has been extracted.

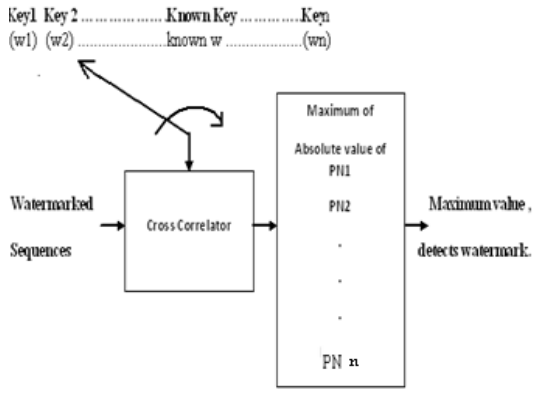

Figure 3: Watermark detection and extraction Scheme.

Figure 6, is the plot of Maximum values of Correlation of Watermarked sequences in the audio to 100 numbers of randomly generated 16 bit sequences. The correct sequence has been used at 51th place in X - axis, to compare its maximum correlation values with the other values. It has been found that the Maximum Correlation value due to the correct watermark sequence will be very much higher in comparison to other correlation values, indicating the detection of correct watermark.

IV. EXPERIMENTAL RESULTS

An experiment setup performed in MATLAB with a 16 kHz audio signal and watermarked with a 16 bit watermark sequence generated by PN sequence generator, using the proposed watermarking technique.

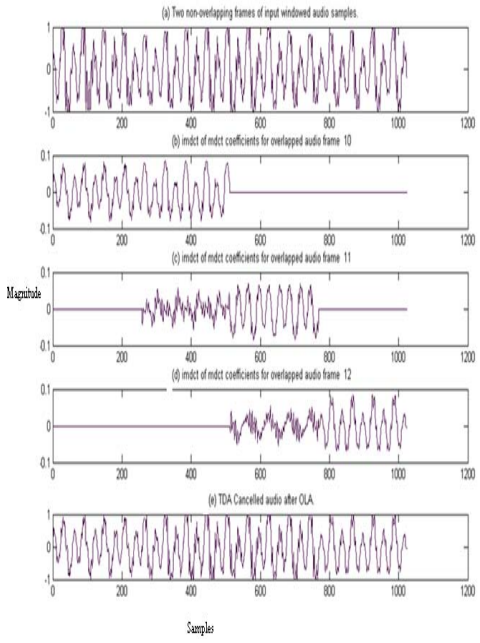

Figure 4: 4(a) Two overlapping frames of input windowed audio samples, from 1 to 1024 samples. 4(b) Inverse MDCT of MDCT Coefficients for first frame, from samples 1 to 512 samples. 4(c) Inverse MDCT of MDCT Coefficients for second 50% overlapped frame, from samples 257 to 768 samples. 4(d) Inverse MDCT of MDCT Coefficients for third 50% overlapped frame, from samples 513 to 1024 samples. 4(e) Aliased cancelled in time domain by overlap and add procedure on first, second , and third 50 % overlapped frames, of 4(b),4(c), and 4(d).The samples from 257 to 768 of 4(e) closely resembles to the samples from 257 to 768 of 4(a).

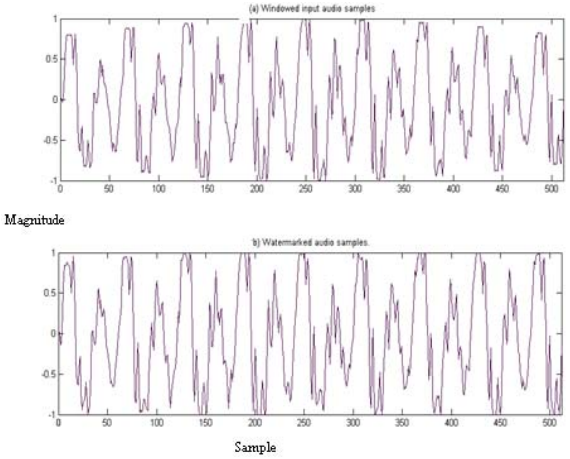

Figure 5: 5(a) Windowed input audio samples, from 2561 to 3072 of audio input corresponding to 6th non-overlapping frame, 5(b) Watermarked audio samples, from 2561 to 3072 of audio input corresponding to 11th overlapping frame

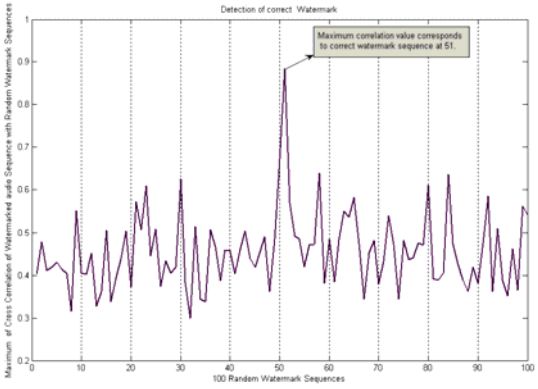

Figure 6: Detection and Extraction of watermark by correlating the watermarked audio sequence with the randomly generated 16 bit length sequences, indicated by the highest maximum value of correlation corresponding to correct watermark embedded.

The **objective difference grade** (ODG) is calculated by perceptual evaluation of the audio quality (PEAQ) algorithm [13] specified in ITU BS.1387-1.It corresponds to the subjective difference grade used in human-based audio tests. The ODG ranges from 0 to −4 and is defined such that 0 value, means imperceptible (the Best grading), and -4 value as most annoying (the worst grading). Equal length samples of input audio (Reference signal) and output watermarked audio (Test Signal) has been taken and .wav files generated with equal sampling rate, 48kHz. Then using PEAQ Algorithm for sample length of 512 samples and sampling rate of 48kHz, objective difference grade (ODG) comes equal to -0.34.

Watermark to Noise Ratio (WNR) is the ratio of watermark signal energy to noise signal energy . The average value of normalized correlation obtained is 0.95 for WNR 15dB.

V. CONCLUSION

This paper has provided a audio Watermark embedding and detection schemes, utilizing MDCT coefficients , relationship between DFT and MDCT and Psychoacoustic Auditory Model of HAS, to achieve highly transparent and robust watermarking as indicated by results in Section IV. It will be much helpful to watermark in compressed domain, which is nowadays more popular, in MDCT domain, without going into any other domain. Further it is suggested that there is much scope for future work in this paper by enhancing the watermark capacity, and research the other locations where watermark can be inserted without degrading the PEAQ and maintaining the robustness, security and transparency.

REFERENCES

[1] K *Brandenburg, " Perceptual coding of high quality digital audio", in Applications of Digital Signal Processing to Audio and Acoustics, M. Kahrs and K. Brandenburg, Eds. Boston, MA: Kluwer Academic, 1998.*

[2] L. Boney, A. H. Tewfik, and K. H. Hamdy, "Digital watermarks for audio signals," in Proc. Int. Conf. Multimedia Computing, Systems, Trieste, Italy, Sep. 1996, pp. 473–480.

[3] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual coding," Signal Process. (Special Issue Copyright Protection Access Control Multimedia Services), vol. 66, no. 3, pp. 337–355, May 1998.

[4] D. Kirovski and H. S. Malvar, "Spread-spectrum watermarking of audio signals," IEEE Transactions on Signal Processing, vol. 51, no. 4, pp. 1020–1033, Apr. 2003

[5] Jyotsna Singh, Parul Garg, and Alok Nath De, "Audio Watermarking Using Spectral Modifications" World Academy of Science, Engineering and Technology 59 2009

[6] A. W. Johnson, and A. B. Bradley, "Adaptive transform coding incorporating time domain aliasing cancellation," Speech Comm. Vol. 6, pp. 299-308, 1987.

[7] Henrique S. Malvar, Signal Processing with Lapped Transforms, Artech House, Norwood MA, 1992.

[8] Rongshan Yu, and Xiao Lin, "A Statistics study of the MDCT coefficient distribution for audio", IEEE International Conference on Multimedia and Expo (ICME), 2004.

[9] I. J. Cox, J. Kilian, F. T. Leigton, and T. Shamoon,"Secure Spread Spectrum watermarking for Multimedia," IEEE Transactions on Image Processing, vol. 6, pp. 1673-1687, 1996.

[10] Hossein Najafzadeh-Azghandi and Peter Kabal "improving perceptual coding of narrowband audio signals at low rates" Electrical & Computer Engineering McGill University, Montreal, Canada

[11] Ricardo a. Garcia, "digital watermarking of audio signals using a Psychoacoustic auditory model and spread spectrum Theory" School of Music Engineering Technology, University of Miami Coral Gables, FL 33146, USA.

[12] Ye Wang, AES Member, and Miikka Vilermo, "Modified Discrete Cosine Transform – Its Implications for Audio Coding and Error Concealment", Nokia Research Center,FI-3372I Tamere, Finland.

[13] T. Thiede and Al., "PEAQ The ITU Standard for Objective Measurement of Perceived Audio Quality," J.Audio Eng.Soc., Vol. 48, No 1/2, 2000., pp. 3-29.

◈◈◈