

High Payload Adaptive Audio Watermarking based on Cepstral Feature Modification

Sunita V. Dhavale

Department of Computer Science and Engineering
Defence Institute of Advanced Technology
Girinagar, Pune-411025, INDIA
sunitadhavale75@rediffmail.com

Rajendra S. Deodhar

Armament Research and Development Establishment
Pashan, Pune-411021, INDIA
rajendra.deodhar@gmail.com

Debasish Pradhan

Department of Applied Mathematics
Defence Institute of Advanced Technology
Girinagar, Pune-411025, INDIA
debasish@diat.ac.in

L. M. Patnaik

Department of Computer Science and Automation
Indian Institute of Science
Bangalore-560012, INDIA
lalit@micro.iisc.ernet.in

Received June, 2013; revised January, 2014

ABSTRACT. *In this paper, we propose two blind adaptive audio watermarking schemes based on complex cepstrum transform (CCT) domain features. In first scheme (Scheme-I), each audio segment is divided into two subsets having approximately same statistical mean value using down-sampling method. Since the human auditory system (HAS) is not much sensitive to the minute change of the wavelet high-frequency components, first level discrete wavelet transform (DWT) detail coefficients subbands of both subsets are used for embedding. Watermark is embedded by changing slightly the difference between the mean values (M_{diff}) of insignificant CCT coefficients of these subsets in order to guarantee minimal perceptual distortion. The offset value by which mean is modified is made adaptive to the local energies of the audio frames in order to increase the audio quality further. In order to enhance the payload capacity, we propose an alternate audio watermarking scheme (Scheme-II) where watermark is embedded by deciding the transition of M_{diff} value from one frame to another frame in two successive frames. In contrast to previous works, instead of embedding one bit per frame, Scheme-II can embed three bits per two frames. Thus 33.33% increase in embedding capacity is achieved. As amplitude scaling in time domain does not affect selected insignificant CCT coefficients, strong invariance towards amplitude scaling attacks is also proved theoretically. Experimental results reveal that the proposed watermarking schemes maintain high audio quality and are simultaneously robust to general attacks like MP3 compression, amplitude scaling, filtering, re-sampling, re-quantization etc.*

Keywords: Audio watermarking, Cepstral Feature, Digital Rights Management, Discrete Wavelet Transform, Complex Cepstrum Domain, Blind.

1. **Introduction.** Due to outstanding progress of digital audio technology, ease of reproducing and retransmitting digital audio has been greatly facilitated. Hence there is a need for the protection and enforcement of intellectual property rights for digital media. Digital watermarking is one of the promising ways to meet this requirement. The primary objective of digital watermarking is to hide the copyright information (e.g. owners/company name, logo etc.) into a multimedia object, without disturbing the perceivable quality of the content [1]. Watermarking of audio signals is more challenging compared to the watermarking of images or video sequences due to wider dynamic range of the human auditory system (HAS) in comparison with the human visual system (HVS). Two properties of the HAS dominantly used in audio watermarking algorithms which are frequency masking and temporal masking. According to the International Federation of the Phonographic Industry (IFPI) [2], Signal to Noise Ratio (SNR) of watermarked audio signal should be always greater than 20 dB. The embedded watermarks should not be removed or degraded using common audio processing techniques. The watermark embedding process should be faster, so that integrated watermarking functionality can be enabled in the delivery of an audio over a network. Also it should support fast watermark detection in order to authenticate audio objects, delivered over the networks. According to the IFPI [2], there should be more than 20 bits per second (bps) data payload for watermark. As these requirements are conflicting to each other, they present great challenges in designing watermarking system. Existing audio watermarking techniques are broadly categorized into time domain and transform domain techniques [3]. Time domain techniques [4] are simple to realize, but they are less robust compared to transform domain techniques [5, 6, 7, 8, 9, 10, 11, 12] like discrete wavelet transform (DWT), discrete Fourier transform (DFT), discrete cosine transform (DCT), Real/Complex Cepstrum Transform [6, 7, 8, 9, 10, 11, 12, 13] etc.

The cepstrum domain analysis is used commonly in speech analysis and recognition research. In speech recognition, the cepstral coefficients are regarded as the main features of voice. The cepstral coefficients vary less after general signal processing attacks than samples in time domain. This feature can be used to preserve the watermark information in case of attacks. In recent years, a number of efforts are made to take advantage of cepstrum domain features for audio watermarking. S.K. Lee [6] introduced a spread spectrum technique to insert a watermark into the cepstral component of the audio signal in non blind manner i.e. an original audio is required for extraction of the watermark at receiver side. As all CCT coefficients are modified in this case thus limiting the perceptual quality of watermarked audio. X. Li [7] embedded data using statistical mean manipulation (SMM) of selected cepstrum coefficients. But the embedding capacity is found lesser. Ching-Tang Hsieh [8] applied a method to combine an energy-feature basis idea in time domain to solve the synchronization problem and achieved blind audio watermarking in cepstrum domain. In this scheme, the embedding capacity mentioned is low i.e. 15bps with the fixed frame size of 2048. Tang Xianghong et. al. [9] embedded watermark into CCT components of important approximation coefficients in 3rd or 4th level wavelet transform in order to get better tradeoff between imperceptibility and robustness. But here the watermark extraction process needs the original audio. Vivekananda Bhat [10] proposed blind algorithm that embeds the watermark data into original audio signal using mean quantization of cepstrum coefficients. But in this scheme, as all cepstrum coefficients are modified, this affects SNR of watermarked audio greatly. Xiaoming Zhang [11] developed an audio watermarking algorithm using statistical mean modification of CCT coefficients of the low-frequency wavelet coefficients. Here, the time complexity of the scheme is found more due to 7th level DWT decomposition. Also embedding in several DWT subbands affects the SNR i.e. perceptual quality of watermarked audio. Chengzhong Yang et al. [12]

proposed a robust watermarking based on DCT and complex cepstrum. Here after applying DCT followed by CCT to selected frames (having absolute DCT mean value greater than predefined threshold), each CCT coefficients of frame is modified in order to quantize the mean value to embed watermark. Though robustness of scheme is increased, SNR is affected greatly. Further after attacks the absolute DCT mean value may get affected leading to false selection of the frames and thus leading to more Bit Error Rate (BER) in extracted watermark. The embedding bit rate mentioned in [12] is 2.5bps i.e. very low compared to other works. In [13], H. T. Hu et al. presented two different cepstrum-based schemes to achieve blind audio watermarking via the mean-value manipulation. Though the detection rates are high, but using the audio frame size of 2205 samples restricts the embedding capacity. Also as more than 90% of CCT coefficients in each frame are modified in order to embed watermark, this can affect the perceptual quality of an audio more. All the above mentioned schemes only focus on embedding one bit per frame based on the modification of cepstral feature i.e. mean of CCT coefficients of that frame. This approach restricts the embedding payload. Instead, we can effectively utilize the transition of this cepstral feature from one frame to another frame for embedding in order to achieve high embedding rate.

The audio watermarking schemes proposed here aims to ensure minimal perceptual distortion while retaining high robustness against different attacks along with high embedding capacity. The key features of our schemes are: (1) Each audio segment is divided into two subsets (A and B) containing odd and even audio sample values using down-sampling method. Due to slow time varying feature of an audio signal, the down-sampling method can guarantee that both the audio subsets will have approximately same mean value. (2) Watermark is embedded by changing local mean value of insignificant CCT coefficients of these two subsets relative to each other. Here CCT is applied to DWT detail sub-bands of both subsets in order to retain high imperceptibility. (3) The offset value by which mean is modified is selected adaptively using the local energies of the audio frames in order to increase the audio quality. The minimum and maximum offset values decided ensure robustness against attacks. (4) Further to improve the transparency of the digital watermark, the watermark is embedded by modifying few insignificant CCT coefficients only. (5) In Scheme-I, the difference between the mean values (M_{diff}) of CCT coefficients of A and B subsets represents the embedded watermark. In Scheme-II, the transition of M_{diff} value from one frame to another frame in two successive frames is used to embed the watermark in order to increase embedding capacity without affecting imperceptibility. In contrast to previous works [6, 7, 8, 9, 10, 11, 12, 13], our algorithm can embed three bits per two frames. Thus 33.33% increase in embedding capacity is achieved. (6) In order to resist synchronization attack effectively watermark is embedded along with synchronization codes. Also Arnold encryption is applied on watermark to increase the secrecy of embedded watermark. (7) Both watermarking schemes are blind i.e. they do not require the use of the original signal for watermark detection. (8) As in our schemes, the relationship between the mean values of CCT coefficients of the subsets represent the embedded watermark bits; this makes our scheme highly invariant to various attacks.

The outline of the paper is as follows. Section 2 provides the outline of the proposed audio watermarking Scheme-I based cepstral feature. Section 3 illustrates the modified proposed audio watermarking Scheme-II based on transition of cepstral feature in consecutive frames to increase the embedding capacity. Experimental results are compared with the results of previous work in Section 4 followed by the conclusion in Section 5.

2. Proposed Scheme-I. The proposed scheme consists of watermark processing stage and audio processing stage as shown in Figure 1.

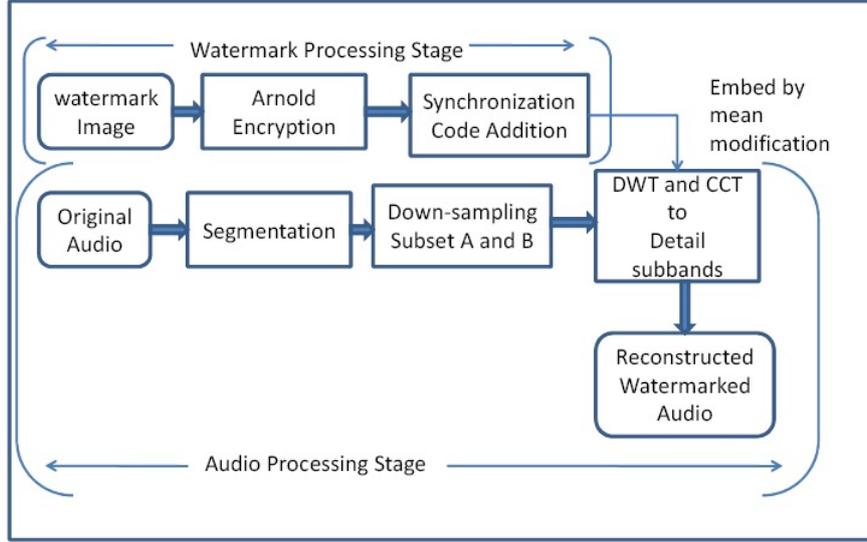


FIGURE 1. Proposed Audio Watermarking Scheme-I: based on M_{diff} cepstral feature

The detailed procedure in case of watermark embedding process is as follows,

2.1. Watermark Embedding Process. Step 1: In watermark processing stage, the binary logo image $\{B = b(i, j); b(i, j) \in \{0, 1\}, 0 < i \leq m, 0 < j \leq n\}$ is first permuted by the Arnold Transform [14] in order to enhance the security of the system.

Step 2: The resulting watermark is then converted into one dimensional bit stream and 16 bit Barker code is added as synchronization code in the bit stream to avoid false synchronization. Use of Barker code for synchronization in audio watermarking resists cropping and shifting attacks [5]. Hence in our scheme we use these codes for synchronization. Barker codes are the subsets of Pseudo-random Noise sequences. They have low correlation side lobes. A correlation side lobe is the correlation of a codeword with a time shifted version of itself. The correlation side lobe C_k for a k symbol shift of an N bit code sequence $(x(1), x(2), \dots, x(N))$ is given by (1),

$$C_k = \sum_{j=1}^{N-k} x(j)x(j+k) \quad (1)$$

where $x(j)$ is an individual code symbol taking values +1 or -1 for $j = 1, 2, \dots, N$. Low correlation side lobes make barker codes a good choice for synchronization codes. Let final watermark bit stream be $W = \{w(i)|w(i) \in (0, 1), 0 < i \leq L_w\}$ where $L_w = (m * n + 16)$ is the length of final processed watermark bit stream.

Step 3: In audio processing stage, the original host audio is first segmented into non-overlapping audio frames of size $L = 512$ samples. For convenience, we take L as an even number. If $X = (x(1), x(2), \dots, x(N))$ denotes the original audio .wav signal having size N where, $x(i) \in (-1.0, +1.0)$ are respective sample values normalized in the given range, then the audio segments are given as,

$$Y_k = (x(L(k-1) + 1), x(L(k-1) + 2), \dots, x(Lk)) \quad (2)$$

where $k = 1, 2, \dots, N_s$ and $N_s =$ total number of audio segments $= \frac{N}{L}$. The audio signal is padded with 0's if necessary. The embedding capacity also depends on this N_s , as each

frame can embed one bit of watermark.

Step 4: Each audio segment Y_k is further divided into two different subsets $Y_k^A = \{Y_k(1), Y_k(3), \dots, Y_k(L-1)\}$ and $Y_k^B = \{Y_k(2), Y_k(4), \dots, Y_k(L)\}$ using down-sampling method as shown in Figure 2. In signal processing, downsampling is the process of reducing the sampling rate of a signal. This is usually done to reduce the data rate or the size of the data. If M denote the down-sampling factor then reduce the data by picking out every M^{th} sample: $h(k) = g(Mk)$. Here $M = 2$ is chosen such that Y_k^A contains all odd samples and Y_k^B contains all even samples. Thus each subset has $\frac{L}{2}$ samples. Due to slow time varying feature of an audio signal, neighbouring sample values have very small difference. Hence, both subsets will have approximately same statistical features i.e., $Y_k(i-1) \approx Y_k(i), i = 1, 2, \dots, \frac{L}{2}$. This feature can be used effectively to embed watermark by changing the value of mean of both subsets relative to each other slightly without degrading the quality of watermarked audio.

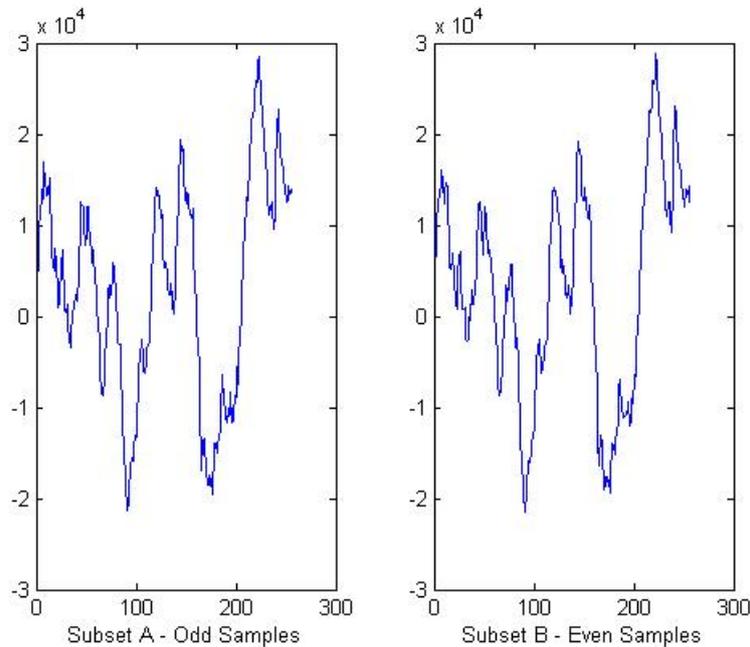


FIGURE 2. Odd and Even audio segment subsets

Step 5: Apply Single-level discrete 1-D wavelet transform to each subset. DWT which has excellent spatio-frequency localization properties is well-suited for multi-resolution analysis. It decomposes the host audio signal into several multi-resolution sub-bands. Let L^A, L^B be corresponding approximation coefficients and H^A, H^B be corresponding detail coefficients of Y_k^A and Y_k^B respectively as shown in Figure 3. Here approximation coefficients represents the low frequency components of corresponding audio signal subsets while detail coefficients represents the high frequency components of corresponding audio signal subsets. Since HAS is not much sensitive to the minute change of the wavelet high-frequency components and the coefficients of the high frequency component are smaller, so we can select detail coefficients subbands representing high frequencies i.e. H^A and H^B for embedding the watermarks.

Step 6: Apply CCT on detail coefficients subbands H^A and H^B as shown in Figure 4. Let $C_k^A = \{c_k^A(j) | j = 1, 2, \dots, \frac{L}{4}\}$ represents CCT coefficients of detail subband H^A of

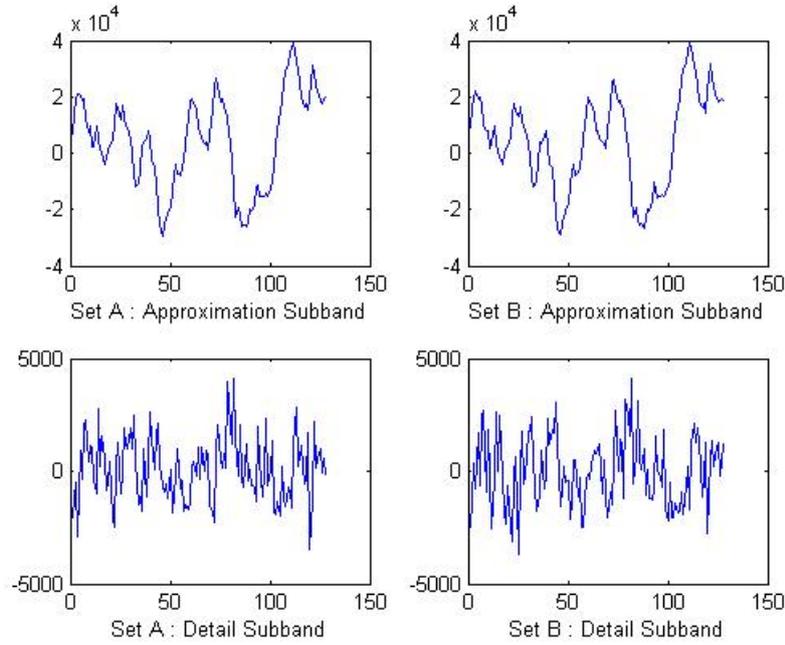


FIGURE 3. Approximation and detail coefficients of subset A and B

Y_k^A and $C_k^B = \{c_k^B(j) | j = 1, 2, \dots, \frac{L}{4}\}$ represents CCT coefficients of detail subband H^B of Y_k^B . Here, the cepstral coefficients are generally uncorrelated and experimental studies have shown that statistical mean of the cepstrum coefficients exhibits less variance than the original audio signals in the presence of common signal processing attacks. High-order cepstra are numerically quite small as shown by typical distribution of the complex cepstrum in Figure 4.

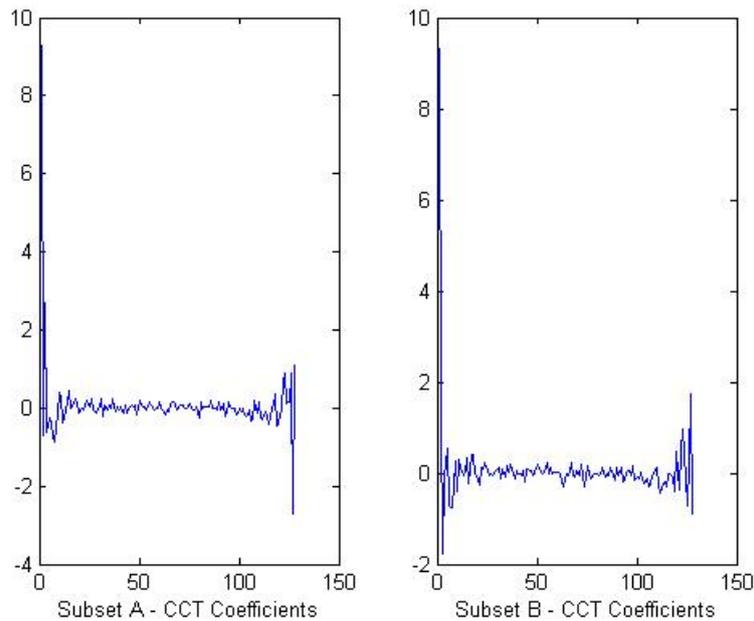


FIGURE 4. Distribution of the complex cepstrum coefficients of detail subbands of subset A and B

The complex cepstrum for a sequence $x(pT)$ is calculated by first finding the complex natural logarithm of the Fourier transform of sequence $x(pT)$ and then applying the inverse Fourier transform to the resultant. Let $F(x)_k = \sum_{p=0}^{N-1} x(pT) \exp(\frac{2\pi jpk}{N})$, $0 \leq k \leq N-1$ denote N point discrete Fourier transform (DFT) of signal $x(pT)$ sampled at every $T = \frac{1}{N}$ seconds, for an integer N , then the inverse DFT is given by $G(x)_k = \frac{1}{N} \sum_{p=0}^{N-1} F(x)_p \exp(\frac{-2\pi jpk}{N})$, $0 \leq k \leq N-1$. If \log denotes logarithm to the base e , then, cepstrum for a sequence x is calculated as,

$$C(x)_k = \text{real}(G(\log(\text{abs}(F(x)_k))))), 0 \leq k \leq N-1 \quad (3)$$

As complex cepstrum holds information about magnitude and phase of the initial spectrum, this allows the reconstruction of the signal. It can be seen that, if the detail subband size=128 then index of $c(n)$ ranges from 1 to 128 with center at 64. Here, large cepstrum coefficients are mostly perceptually significant and are not used in embedding process in order to achieve imperceptibility. The mean of center region $\{l_1, \dots, l_2\}$ containing total of $t = (l_2 - l_1 + 1)$ CCT coefficients, is very less in case of $l_1 = 55$, $l_2 = 74$, thus $t = 20$. This region contains perceptually nonsignificant components hence can be used for embedding watermark by modifying mean relative to each other. Here as only 15% of total CCT coefficients in insignificant band are affected, this retains quality of watermarked audio along with providing good robustness due to attack invariant feature of cepstrum domain. As t i.e., total number of center region insignificant CCT coefficients selected for embedding increases, robustness against 64kbps MP3 attack increases at the cost of SNR. For $t = 20$, the bit error ratio (BER) is found to be less than 1.0% with SNR above 40dB.

Step 7: For k^{th} frame, let $\overline{M_{A_k}} = \frac{1}{(l_2-l_1+1)} \sum_{i=l_1}^{l_2} c_k^A(i)$ and $\overline{M_{B_k}} = \frac{1}{(l_2-l_1+1)} \sum_{i=l_1}^{l_2} c_k^B(i)$ be the mean of respective center regions of subbands C_k^A, C_k^B respectively. The cepstral feature M_{diff} representing the difference between the cepstral mean values of subset A and B is given as,

$$M_{diff_k} = \overline{M_{A_k}} - \overline{M_{B_k}} \quad (4)$$

Watermark bit is embedded according to following rules,

a) For each audio frame, only one watermark bit is embedded. The watermark bit '1' and '0' is embedded by changing the relationship between the mean values such that $\overline{M_{A_k}} > \overline{M_{B_k}}$ or $\overline{M_{A_k}} < \overline{M_{B_k}}$.

b) To embed a watermark bit='0', change the mean relative to each other to make $\overline{M_{A_k}}$ lesser than $\overline{M_{B_k}}$ i.e. $\overline{M_{diff_k}} < 0$ as follows,

$$\begin{aligned} \text{If } \overline{M_{A_k}} > \overline{M_{B_k}}, & \begin{cases} \hat{c}_k^A(i) = c_k^A(i) - \frac{|M_{diff_k}|}{2} - \delta_k \\ \hat{c}_k^B(i) = c_k^B(i) + \frac{|M_{diff_k}|}{2} + \delta_k, \end{cases} \quad l_1 \leq i \leq l_2 \\ \text{If } \overline{M_{A_k}} \leq \overline{M_{B_k}}, & \begin{cases} \hat{c}_k^A(i) = c_k^A(i) + \frac{|M_{diff_k}|}{2} - \delta_k \\ \hat{c}_k^B(i) = c_k^B(i) - \frac{|M_{diff_k}|}{2} + \delta_k, \end{cases} \quad l_1 \leq i \leq l_2 \end{aligned} \quad (5)$$

c) To embed a watermark bit='1', change the mean relative to each other to make $\overline{M_{A_k}}$ greater than $\overline{M_{B_k}}$ i.e. $\overline{M_{diff_k}} > 0$ as follows,

$$\begin{aligned}
\text{If } \overline{M_{A_k}} < \overline{M_{B_k}}, & \begin{cases} \hat{c}_k^A(i) = c_k^A(i) + \frac{|M_{diff_k}|}{2} + \delta_k \\ \hat{c}_k^B(i) = c_k^B(i) - \frac{|M_{diff_k}|}{2} - \delta_k, \end{cases} \quad l_1 \leq i \leq l_2 \\
\text{If } \overline{M_{A_k}} \geq \overline{M_{B_k}}, & \begin{cases} \hat{c}_k^A(i) = c_k^A(i) - \frac{|M_{diff_k}|}{2} + \delta_k \\ \hat{c}_k^B(i) = c_k^B(i) + \frac{|M_{diff_k}|}{2} - \delta_k, \end{cases} \quad l_1 \leq i \leq l_2
\end{aligned} \tag{6}$$

Note that, $|\hat{M}_{diff_k}| = |\hat{M}_{A_k} - \hat{M}_{B_k}| = 2 * \delta_k$ in all cases. This aids in detecting the watermark correctly at receiver.

d) Here δ_k is offset value added/deleted in order to manipulate mean of detail subband. It also represents the embedding strength. The value of δ_k can be made adaptive to the local energies of different audio frames. According to the HAS property, relatively large aberrance in the high energy audio frame is inaudible, so we may select relatively large offset value. For an audio frame of size L samples, let $E_k = \sum_{i=1}^L x_k(i)$ denote energy of the k^{th} frame, then offset value δ_k required for k^{th} frame is given as,

$$\delta_k = \max(\min(\alpha \times E_k, \delta_M), \delta_m) \tag{7}$$

where α indicates the scaling factor and here $\alpha = 1.0$ is chosen. The bound on min and max values of δ are represented by δ_m and δ_M respectively. Here $\delta_m = 0.03$ and $\delta_M = 0.06$ is chosen. Thus selection of δ_k is made automated depending upon the energy of the frame. Here, higher δ_k can be chosen for audio frames having higher energy to embed watermark bit effectively in more robust manner without sacrificing any imperceptibility. However in remaining cases, smaller δ_k has to be chosen in order to retain transparency of the embedded watermark.

Step 8: Reconstruction: Reconstruction of a watermarked audio is achieved by applying first Inverse CCT followed by the Inverse DWT to get the modified subsets $\hat{Y}_k^A = \{\hat{Y}_k(1), \hat{Y}_k(3), \dots, \hat{Y}_k(L-1)\}$ and $\hat{Y}_k^B = \{\hat{Y}_k(2), \hat{Y}_k(4), \dots, \hat{Y}_k(L)\}$. Combine these embedded odd and even samples subset to reconstruct the final embedded audio frame $\hat{Y}_k = \{\hat{Y}_k(1), \hat{Y}_k(2), \dots, \hat{Y}_k(L-1), \hat{Y}_k(L)\}$. Reconstruct watermarked audio $\hat{Y}_k(n)$ by combining all frames. The Signal to Noise Ratio (SNR) of this watermarked audio with respect to the original audio signal is calculated. The experimental results show that after embedding the 1024 bits of watermark the stego audio gives SNR value more than 35dB.

2.2. Watermark Detection Process. The extraction algorithm consists of all the audio processing steps which are carried out at the time of embedding (see Figure 1). First stego audio signal $\hat{Y}_k(n)$ is segmented into non-overlapping frames of size L samples each. Then each audio segment is divided into two different subsets $\hat{Y}_k^A = \{\hat{Y}_k(1), \hat{Y}_k(3), \dots, \hat{Y}_k(L-1)\}$ and $\hat{Y}_k^B = \{\hat{Y}_k(2), \hat{Y}_k(4), \dots, \hat{Y}_k(L)\}$ containing odd and even audio samples using down-sampling method given in Equation (3). Apply DWT to each subset to get approximation and detail coefficient subbands. Apply CCT to detail coefficient subbands of both subsets. Calculate mean \hat{M}_{A_k} and \hat{M}_{B_k} of respective center regions of detail coefficient subbands of both subsets. Figure 5 shows the histogram of M_{diff_k} before embedding, after embedding and after the MP3 attack on watermarked signal. M_{diff_k} is forced to be positive, when bit '1' is embedded and M_{diff_k} is forced to be negative, when bit '0' is embedded.

Extract the embedded bit stream containing both synchronization codes and watermark according to following rules,

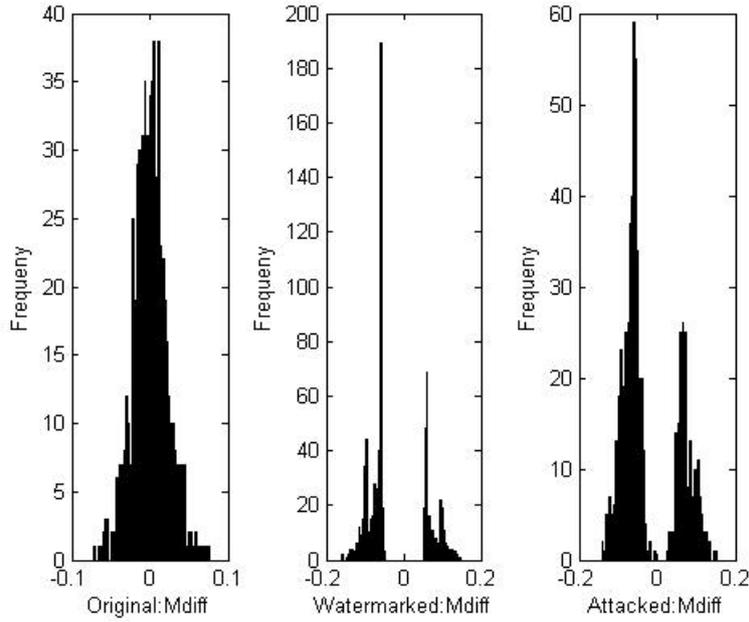


FIGURE 5. Scheme-I: Cepstral feature M_{diff} Distribution for original, watermarked and attacked audio signal

$$\hat{w} = \begin{cases} 0 & \overline{\hat{M}_{A_k}} < \overline{\hat{M}_{B_k}} \\ 1 & \overline{\hat{M}_{A_k}} > \overline{\hat{M}_{B_k}} \end{cases} \quad (8)$$

In this way, we can extract the watermark bits. Once all the bits are extracted, the watermark logo image $\{\hat{B} = \hat{b}(i, j); \hat{b}(i, j) \in \{0, 1\}, 0 < i \leq m, 0 < j \leq n\}$ can be reconstructed by first detecting the synchronization codes. The original audio is not required in the extraction process and thus the proposed algorithm is blind. The distortion caused is not perceptually audible as only few CCT coefficients are modified. Experimental results show that the offset values $\delta_m = 0.03$ and $\delta_M = 0.06$ are sufficient to provide good trade-off between robustness and imperceptibility. For the stereo audio signals, dual-channel signals are available for watermarking, while in case of a mono audio signals; only one single-channel signal is available for watermarking.

3. Proposed Scheme-II. The proposed Scheme-I discussed in section 2 is modified by simultaneously considering the cepstral features of two consecutive frames in order to achieve high embedding payload. Here all the steps upto step 6 from previous proposed scheme-I are carried out as it is. Here, only the embedding rules are modified. Let δ be a positive offset and each k^{th} embedded frame be in one of three unique states defined as below,

$$\begin{aligned} (S1) : M_{diff} &= -2 \times \delta \\ (S2) : M_{diff} &= 2 \times \delta \\ (S3) : M_{diff} &= 0 \end{aligned} \quad (9)$$

The transition of M_{diff} value from one frame to another frame in two successive frames can be used for embedding as shown in Figure 6. Here each transition reflects the change of M_{diff} value from one state to another state. S1 can be achieved by any frame using Equation (5) while S2 can be achieved using Equation (6). The offset value δ required for

achieving S1 and S2 can be made adaptive to the local energies of different audio frames using Equation (7).

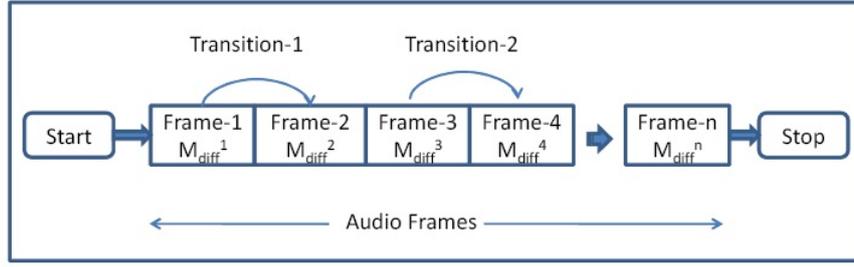


FIGURE 6. Transition of M_{diff} value

To achieve State S3, modify the mean values as,

$$\begin{aligned}
 & \text{If } \overline{M_{A_k}} < \overline{M_{B_k}}, \begin{cases} \hat{c}_k^A(i) = c_k^A(i) + \frac{|M_{diffk}|}{2} \\ \hat{c}_k^B(i) = c_k^B(i) - \frac{|M_{diffk}|}{2} \end{cases}, \quad l_1 \leq i \leq l_2 \\
 & \text{If } \overline{M_{A_k}} > \overline{M_{B_k}}, \begin{cases} \hat{c}_k^A(i) = c_k^A(i) - \frac{|M_{diffk}|}{2} \\ \hat{c}_k^B(i) = c_k^B(i) + \frac{|M_{diffk}|}{2} \end{cases}, \quad l_1 \leq i \leq l_2
 \end{aligned} \tag{10}$$

The state transitions from current frame (say k^{th} frame) to next frame (say $(k + 1)^{th}$ frame) can be effectively used to encode three bits of the watermark. Thus watermark is embedded by modifying cepstral features of consecutive frames using the following rules explained in Table 1.

TABLE 1. Watermark encoding based on M_{diff} transition

Watermark-3 bits	Transition	
	Current Frame State (k) th	Next Frame State ($k + 1$) th
[000]	S1	S1
[001]	S1	S2
[010]	S1	S3
[011]	S2	S1
[100]	S2	S2
[101]	S2	S3
[110]	S3	S1
[111]	S3	S2

Thus now instead of embedding one watermark bit per frame, we can embed 3 bits of watermark in each two successive frames. Thus 33.33% increase in embedding capacity is achieved for the increased watermarked audio quality and equal robustness as per previous scheme-I. Figure 7 shows the histogram of M_{diff} before embedding, after embedding and after the MP3 attack on watermarked signal. M_{diff} is forced to be negative for achieving state S1, M_{diff} is forced to be positive for achieving state S2, and M_{diff} is forced to be zero for achieving state S3.

From this distribution, we can see that the S3 reflects the non-embedded section or part of original audio only by default. During simulation it is found that, state S3 gets affected more due to severe attacks. Hence the transition (S3, S3) is avoided in encoding

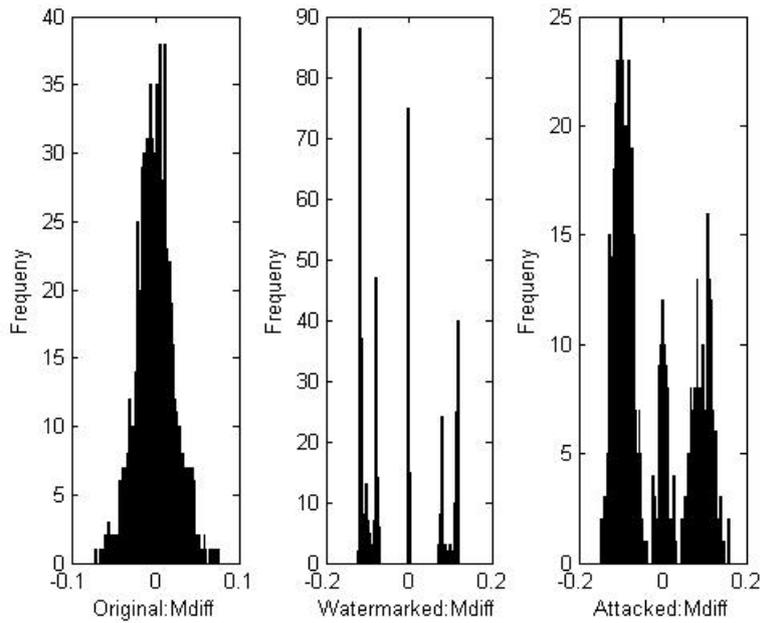


FIGURE 7. Scheme-II: Cepstral feature M_{diff} Distribution for original, watermarked and attacked audio signal

the watermark bits (see Table 1). After carrying different attacks, it is found that a threshold (τ) of 0.015 is able to detect state of k^{th} frame using Equation (11),

$$\hat{S}_k = \begin{cases} S1 & M_{diff} < -\tau \\ S2 & M_{diff} > +\tau \\ S3 & |M_{diff}| < \tau \end{cases} \quad (11)$$

During watermark extraction, first the state of each frame is decided based on above Equation (11) and using the rules in Table 1, watermark is extracted or decoded. For example, if k^{th} frame is in state S1 and $(k+1)^{th}$ frame is in the state S2 then according to Table 1, three bits of watermark containing value=[001] is extracted. Then next set of two frames i.e. $(k+2)^{th}$ frame and $(k+3)^{th}$ frames are considered to extract next 3 bits of watermark, by determining their respective states and so on.

4. Experimental results.

4.1. Experimental setup. To assess the performance of the proposed audio watermarking scheme, several experiments are carried out on different types of 250 mono audio signals of length 20 seconds each. These CD Quality audio signals are sampled at sampling rate 44.1 KHz with 16 bit resolution. These audio signals are categorized into following categories; the rock music (denoted by A1) that has very high signal energy, classical music (denoted by A2) and speech signal (denoted by A3) that has moderate signal energy (see Figure 8).

The ownership information is represented by a 32x32 binary logo image as shown in Figure 9. The logo image is first permuted using Arnold Transform and converted into a one dimensional bit stream of 1's and 0's. 16 bit synchronization code (1111100110101110) is also added in to the stream. The min and max offset values $\delta_m = 0.03$ and $\delta_M = 0.06$ are sufficient to provide good tradeoff between robustness and imperceptibility for all different types of selected audio files.

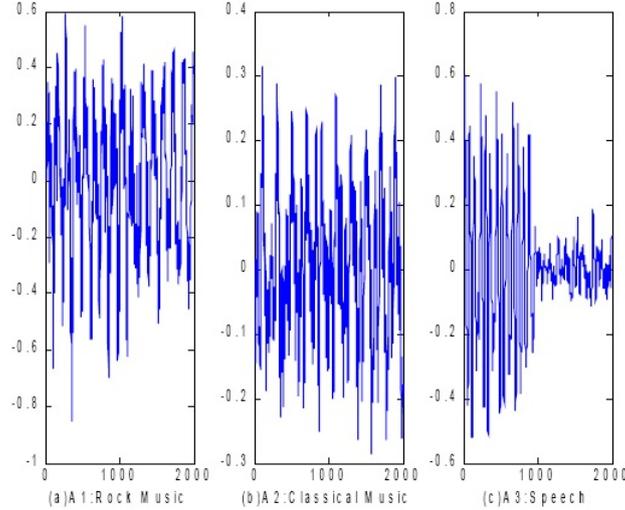


FIGURE 8. Original audio signals (a) Rock music (b) classical music and (c) speech signal



FIGURE 9. 32x32 binary logo watermark image

The data payload (D) refers to the number of bits that are embedded into the audio signal within a unit of time. It is measured in the unit of bps (bits per second). If the sampling rate of an audio signal is F (measured in Hz), the frame size is F_s (i.e. number of samples per frame) and the number of bits embedded per frame is N_b , then the data payload of the proposed scheme is given as,

$$D = \frac{F}{F_s} \times N_b \text{ (bps)} \quad (12)$$

In proposed scheme-I, for a frame containing 512 samples, the estimated data payload is 86.13 bps as $N_b = 1.0$. While in case of proposed scheme-II, for a frame containing 512 samples, the estimated data payload is 129.19 bps as $N_b = 1.5$, as two bits are embedded per two frames. It needs an audio section about 12.07 seconds in order to embed a 16 bit synchronization code along with a 32x32 binary watermark in proposed scheme-I, while it needs an audio section of 8.05 seconds only to embed same data in proposed scheme-II. Thus there is an 33.33% increase in the data payload without affecting perceptual quality of an audio.

4.2. Perceptual Quality Measures. To measure imperceptibility, we use both signal-to-noise ratio (SNR) and segmental SNR (SegSNR) as an objective measure. SNR is based on the difference between the undistorted original audio signal and the distorted watermarked audio signal. If A corresponds to the original audio signal, and \hat{A} corresponds to the watermarked audio signal then, SNR is given as,

$$SNR = 10 \log_{10} \left(\frac{\sum_n A_n^2}{\sum_n (A_n - \hat{A}_n)^2} \right) \text{ dB} \quad (13)$$

where n =total length of audio signal. The average SNRs of watermarked audio calculated are 37.18dB for A1, 37.69dB for A2 and 38.99dB for A3. The SNR is calculated only on the portion of an audio signal, where actual watermark bits are embedded.

SegSNR is defined as the average of the SNR values of short audio frames of watermarked audio signal. If N_f is total number of watermarked audio frames and r is number of samples in each audio frame then SegSNR is given as,

$$SegSNR = \frac{10}{N_f} \sum_{i=0}^{N_f-1} \log_{10} \left(\frac{\sum_{j=1}^r A_n^2(j)}{\sum_n (A_n(j) - \hat{A}_n(j))^2} \right) dB \quad (14)$$

The SegSNRs of watermarked audio calculated are 39.66dB for A1, 41.02dB for A2 and 40.52dB for A3. In our scheme as only few CCT coefficients gets modified here, this does not affect the segmental energy. Hence no rescaling of embedded audio samples is required as suggested in [13]. During simulations, the difference between the segmental energies of original and embedded audio frames is found to be almost zero.

4.3. Robustness Test. Both normalized correlation (NC) and bit error rate (BER) between the original watermark and the extracted watermark are used as an objective measure for the robustness and calculated using following equations;

$$NC = \frac{\sum_{i=0}^m \sum_{j=0}^n w(i, j) \hat{w}(i, j)}{\sqrt{\sum_{i=0}^m \sum_{j=0}^n |w(i, j)|^2} \sqrt{\sum_{i=0}^m \sum_{j=0}^n |\hat{w}(i, j)|^2}} \quad (15)$$

where $w(i, j)$ is an original watermark and $\hat{w}(i, j)$ is the extracted watermark.

$$BER = \frac{\text{Number_of_bits_in_error}}{\text{Total_number_of_bits}} \times 100 \% \quad (16)$$

The proposed algorithm gives moderate *SNR* values along with good amount of embedding capacity and lower bit error rates. The *NC* values are always above 0.9 for most of the common audio processing attacks. The center region $\{55, \dots, 74\}$ containing total $t = 20$ CCT coefficients, is found sufficient to retain quality of watermarked audio along with providing good robustness against different attacks. As t increases, robustness *NC* increases at the cost of *SNR*. For most of the audios, bit error ratio (BER) is found to be less than 1.0% with *SNR* above 40dB for $t = 20$.

In order to assess the robustness of the proposed watermarking schemes, the watermarked audio signal is subjected to several standard audio processing attacks like,

Resampling: Here, original watermarked audio is downsampled to 22.05KHz and again upsampled back to 44.1KHz i.e. original sampling rate.

Requantization: Here, 16 bit original watermarked audio is quantized to 8 bit.

AWGN: White Gaussian noise with 15dB *SNR* per sample is added to original watermarked audio.

MPEG compression: Original watermarked audio is coded/decoded using mp3 compression algorithm at various bit rate like 224Kbps, 160Kbps, 112Kbps, 64Kbps, 56Kbps etc.

Low pass filtering: 6th order Butterworth Low pass filter is applied to original watermarked audio with different cutoff frequencies.

Amplitude Scaling: Here amplitude of all sample values of original watermarked audio is increased by 10% or 20%.

The results are summarized in Table 2. From the results, it can be seen that the proposed audio watermarking Scheme-I is more robust compared to proposed audio watermarking Scheme-II for most of the common audio processing attacks. But the *SNR* and embedding payload of Scheme-II is higher compared to scheme I. Hence depending upon the application requirement, schemes can be selected. For achieving more payload, Scheme-II can be better option while for achieving more robustness, Scheme-I can be better option.

TABLE 2. NC, BER and SNR values along with corresponding extracted watermarks for various attacks

Algorithm Attack	Proposed Scheme-I				Proposed Scheme-II			
	BER(%)	NC(%)	Watermark	SNR(dB)	BER(%)	NC(%)	Watermark	SNR(dB)
No attack	0.00	1.00		40.44	0.00	1.00		52.90
awgn-40dB	0.78	0.98		37.63	8.59	0.80		40.44
awgn-30dB	10.13	0.86		25.54	14.94	0.67		25.54
Resampling(22.05 kHz)	2.24	0.98		21.62	14.35	0.71		21.64
awgn-40dB+ Resampling-22.05kHz	1.07	0.98		37.63	15.37	0.69		21.58
Requantization(8bit)	0.00	1.00		38.60	6.05	0.86		42.51
Requantization(24bit)	0.00	1.00		40.44	0.00	1.00		52.90
LPF(35.00 kHz)	0.59	0.99		23.51	7.75	0.84		23.58
LPF(30.00 kHz)	5.01	0.88		14.59	7.81	0.82		12.32
MP3-224 kbps	0.00	1.00		31.69	0.19	0.99		32.21
MP3-160 kbps	0.00	1.00		21.16	0.19	0.99		21.19
MP3-112 kbps	0.00	1.00		17.66	1.56	0.96		17.64
MP3-64 kbps	0.09	0.98		15.37	7.03	0.84		15.38
Amplitude(10%)	0.00	1.00		20.10	0.00	1.00		20.15
Amplitude(50%)	0.00	1.00		6.16	0.00	1.00		6.16

The performance of the proposed schemes are also compared with the other cepstrum based watermarking algorithms proposed by Bhat et al. 2008 [10], Zhang et al. 2009 [11] and Hu et al. 2012 [13] and the results are summarized in Table 3. In the scheme proposed in [10], as all cepstrum coefficients are modified this affects SNR of watermarked audio greatly. In [11], besides having very low embedding capacity, using 7th level DWT with CCT increases complexity of the system. While in [13], using the audio frame size of 2205 samples restricts the embedding capacity and as more than 90% of CCT coefficients are used for embedding, this affects SNR of watermarked audio. Compared to these algorithms, our schemes give higher SNR, good payload along with the equal robustness against mp3 attacks for all three kinds of audio signals. Also our schemes exhibit simplicity in design and implementation.

TABLE 3. Algorithm Comparison

Algorithm	Domain	Synchronization	Rock		Classical		Speech		Payload (bps)
			SNR (dB)	NC	SNR (dB)	NC	SNR (dB)	NC	
Our Scheme-I	1 st Level DWT+CCT	Yes	37.18	0.99	37.69	0.99	38.99	0.99	86.13
Our Scheme-II	1 st Level DWT+CCT	Yes	39.22	0.96	39.99	0.96	40.22	0.95	129.19
[10]	Real CT	No	20.01	0.95	25.91	0.98	30.06	0.91	43.06
[11]	7 th Level DWT+CCT	No	36.39	0.98	32.26	0.98	25.71	0.99	2.5
[13]	Real CT	Yes	20.13	0.72	22.72	0.91	25.96	0.92	21.53

4.4. AWGN attack analysis. Figure 10 shows the robustness of the system against AWGN attack with different standard deviations $\sigma = 0.005$ to 0.05 and $mean = 0$ for different audios. With increase in standard deviation of AWGN noise, both NC and SNR values of speech drops more rapidly compared to music signals.

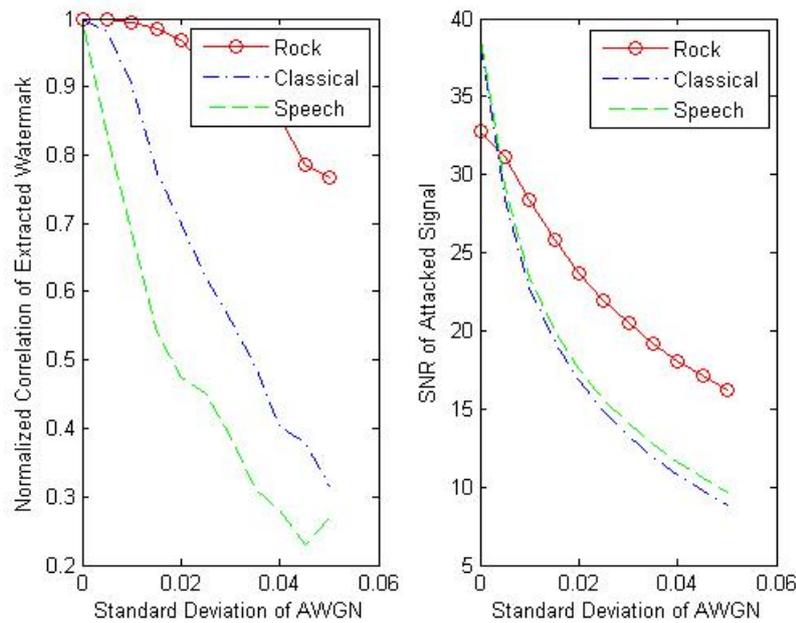


FIGURE 10. Scheme-I: Result of AWGN Attack on NC values

Watermark detection can be considered as a communication related problem requiring reliable transmission and detection of a watermark signal through a noisy channel. Thus, the watermark detection problem can be formulated as a hypothesis test where,

H_0 : Audio Signal does not contain watermark

H_1 : Audio Signal contains watermark

H_0 being the null hypothesis states that received signal is not watermarked and the alternate hypothesis H_1 states that received signal is watermarked, respectively.

The problem of hypothesis testing is to decide whether the statistic extracted from received signal supports alternate hypothesis. Due to noisy communication channels, usually it is not possible to separate all watermarked and un-watermarked audios perfectly. There is a small probability p_{FP} of accepting H_1 when H_0 is true (false positive) and a small probability p_{FN} of accepting H_0 when it is false (false negative).

For the AWGN channel, $e(n)$ is Gaussian random process and statistically independent from embedded watermark $w(n)$. The normalized cross-correlation (NC) between the original watermark and the extracted watermark given in Equation (16) can be used as test statistic.

The distribution of NC under the hypothesis H_0 and H_1 is estimated using simulations. For a rock music audio, under the hypothesis H_0 , we simulated 1000 AWGN patterns with standard deviation varying from 0.0 to 0.05, and constructed the received signal in each case. Then statistic NC was evaluated for each of them. The distribution NC is shown in Figure 11 and is observed to be approximately normal with mean 0.1073 with variance of $1.1107e-04$. On the other hand, under the hypothesis H_1 , we embedded watermark into the rock audio and we subsequently applied the detection process after simulating awgn attack with different standard deviations $\sigma = 0.005$ to 0.05 and $mean = 0$. The distribution of NC in this case is shown in Figure 11 and has $mean = 0.9225$ with $variance = 0.0053$. The value of NC computed for extracted watermark can be compared with the acceptance threshold $T = 0.5$ in order to detect if received signal is watermarked or not. Both the

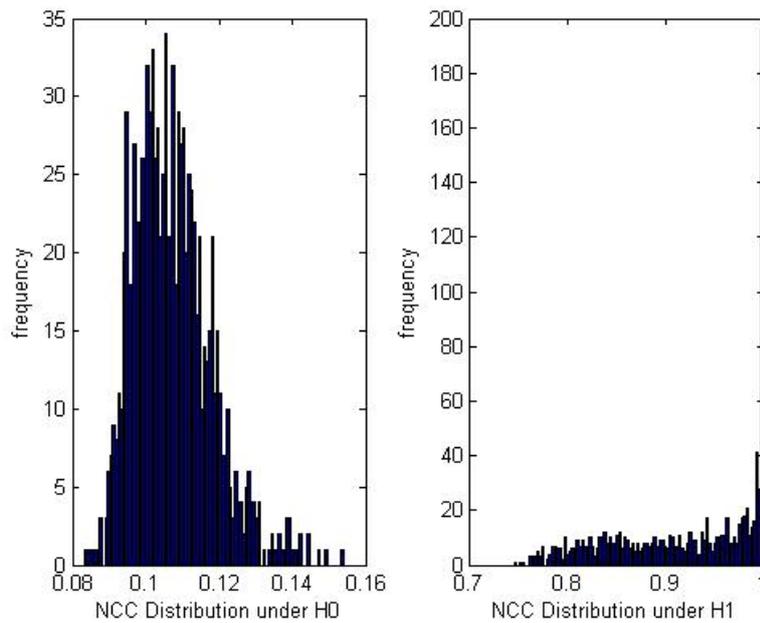


FIGURE 11. Distributions of the NC under the H0 and H1 for rock audio

probability of false positive p_{FP} and the probability of false negative p_{FN} found to be having zero value as both NC distributions are well separated.

4.5. Amplitude Scaling Attack Invariance. Referring to the definition of cepstrum mentioned in Equation (3), for any constant amplitude scaling factor λ , $C(\lambda x)_k = C(x)_k$ where $k > 0$. This means multiplication of the function $x(t)$ by a constant i.e. amplitude scaling in time domain will affect the $C(0)$ coefficient value in cepstrum domain only and not other insignificant CCT coefficients. In our case, $C(0)$ term is ignored during embedding and watermark is embedded by modifying mean of insignificant CCT coefficients only. This makes our proposed schemes invariant to amplitude scaling attacks.

5. Conclusions. In this article, we have proposed two blind adaptive audio watermarking schemes based on cepstral features. Minimal perceptual distortion is ensured using different strategies. In first scheme, few insignificant CCT coefficients of first level DWT detail subbands of downsampled subsets of each frame are modified to improve the transparency of the digital watermark. Further, CCT coefficients are modified adaptively according to the audio frame local energy to improve the performance. Simulation results show that the proposed scheme is more robust with high imperceptibility. This scheme is further modified to increase the payload capacity by embedding three bits per two frames. Thus the embedding capacity is improved by 33.33%. It is observed that in the modified scheme, the increase in embedding capacity is accompanied by improvement in imperceptibility and a slight degradation in robustness. However both the schemes perform better than the other cepstrum based audio watermarking schemes for various types of audio signals.

REFERENCES

- [1] N. Cvejic, and T. Seppänen, Audio watermarking: requirement, algorithms, and benchmarking, *Digital watermarking for digital media*, IGI Global Information Science Publishing, Pennsylvania, pp. 135-181, 2005.
- [2] IFPI(International Federation of the Phonographic Industry), <http://www.ifpi.org>

- [3] A. G. Acevedo, Audio watermarking: properties, techniques and evaluation, *Information Security and Ethics: Concepts, Methodologies, Tools, and Applications*, IGI Global, Pennsylvania, pp. 75-125, 2008.
- [4] P. Bassia, I. Pitas, and N. Nikolaidis, Robust audio watermarking in the time domain, *IEEE Trans. Multimedia*, vol. 3, no. 2, pp. 232-241, 2001.
- [5] S. Wu, J. Huang, D. Huang, and Y.Q. Shi, Self-synchronized audio watermark in DWT domain, *Proc. of the 2004 International Symposium on Circuits and Systems*, pp. 712-715, 2004.
- [6] S.K. Lee, and Y.S. Ho, Digital audio watermarking in the cepstrum domain, *IEEE Trans. Consumer Electronics*, vol. 46, no. 3, pp. 744-750, 2000.
- [7] X. Li, and H.H. Yu, Transparent and robust audio data hiding in cepstrum domain, *Proc. of IEEE International Conference on Multimedia and Expo*, pp. 397-400, 2000.
- [8] C.T. Hsieh, P.Y. Tsou, Blind cepstrum domain audio watermarking based on time energy feature, *Proc. of The 4th International Conference on Digital Signal Processing*, pp. 705-708, 2002.
- [9] X.H. Tang, Y.M. Niu, and Q.L. Li, A digital audio watermark embedding algorithm with WT and CCT, *Proc. of IEEE International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications Proceedings*, pp. 970-973, 2005.
- [10] B.K. Vivekananda, I. Sengupta, and A. Das, Audio watermarking based on mean quantization in cepstrum domain, *Proc. of The 16th International Conference on Advanced Computing and Communications*, pp. 73-77, 2008.
- [11] X.M. Zhang, Z.Y. Yu, Cepstrum-based audio watermarking algorithm against the A/D and D/A attacks, *Proc. of The 5th International Conference on Information Assurance and Security*, pp. 740-743, 2009.
- [12] C.Z. Yang, X.S. Zheng, and Y.L. Zhao, An audio watermarking based on discrete cosine transform and complex cepstrum transform, *Proc. of International Conference on Computer Application and System Modeling*, pp. 456-458, 2010.
- [13] H.T. Hu, and Wei.H. Chen, A dual cepstrum-based watermarking scheme with self-synchronization, *Signal Processing*, vol. 92, no. 4, pp. 1109-1116, 2012.
- [14] W. Ding, W.Q. Yan, D.X. Qi, Digital image scrambling technology based on arnold transformation, *Journal of Computer Aided Design and Computer Graphics*, vol. 13, no. 4, pp. 338-341, 2001.