A Robust Watermarking Technique for Digital Audio

Sandeep R. Aedudodla and K. M. M. Prabhu

Abstract—**In this paper we have presented a robust watermarking procedure for digital audio. The procedure embeds a watermark imperceptibly in the time domain exploiting the temporal and simultaneous masking properties of the human auditory system (HAS). The spectral characterization of the audio is done using the short time Fourier transform (STFT), which provides a better description of non-stationary signals when compared to using the discrete Fourier transform. The watermark is a perceptually shaped pseudorandom sequence, which is used to modify the audio samples. The original audio is not required during the extraction of the watermark, except when the audio signal is subjected to cropping attack. The watermarking procedure exhibits resilience towards several audio distortions occurring from MPEG 1 Layer 3 coding, colored noise, cropping, temporal resampling and jittering.**

Index Terms—**Audio masking, digital watermarking, human auditory system, short time Fourier transform.**

I. Introduction

The ease of dissemination and reproduction of digital media (audio, images and video) has resulted in its widespread use. Applications such as electronic commerce and online services through the Internet are rapidly being developed. However, unrestricted duplication of digitized data has made it prone to violations of intellectual property rights (IPR). The lack of adequate copyright protection measures dissuades parties from offering their services in digital form. In this scenario, digital watermarking is being viewed as an effective tool to circumvent unauthorized reproduction of digital media. Conventional cryptographic measures are inadequate for copyright protection, as the decrypted data is once again prone to illegal duplication. Watermarking involves embedding certain copyright information into the original data such as an audio signal. The watermark should not produce any audible distortions in the data and must be able to resist distortions resulting from common signal processing operations such as compression, resampling and cropping among several others. Given the embedding algorithm and a secret key, one should be able to extract the watermark to establish true ownership of the audio.

Although most watermarking schemes in literature are devoted to images [1]-[5] and video [6], there is a great need for techniques on watermarking of digital audio, especially from the digital music industry. Certain techniques for audio watermarking have been proposed such as in [7]. A survey of multimedia watermarking techniques has been presented in [8].

The technique in [7] makes use of the discrete fourier transform (DFT) for the spectral characterization of audio signals. In this paper, we present an audio watermarking technique that makes use of simultaneous and temporal masking properties of the human auditory system. The short time fourier transform (STFT) [9] is employed for spectral characterization of the audio signal. The watermark is constructed as an independent and identically distributed (i.i.d.) Gaussian random vector. The final watermark that is embedded into the audio data is a pseudorandom sequence, shaped according to the simultaneous and temporal masking properties of the human auditory system (HAS). The use of the STFT in the proposed procedure enables a better characterization of non-stationary signals when compared to using the DFT as has been done in [7]. This enables us to design a more robust and imperceptible watermark.

The algorithm does not make use of the original audio signal during the extraction of the watermark, except when the received audio signal is cropped. This method is called blind watermarking. This makes the watermarking procedure applicable in several copyright protection scenarios. This technique is essential to real environments such as in broadcasting as well as in portable players. The watermark is perceptually unobtrusive and is robust to distortions occurring from MPEG-1 Layer 3 (MP3) coding, colored noise, temporal resampling and jittering. Section II gives a quick overview of the short time fourier transform and its significance in speech and audio processing. Section III describes human audio perception and its masking properties along with its mathematical model. The watermarking algorithm is described in Section IV and the watermark-detection procedure is described in Section V. The experiments, along with the results, are presented in Section VI and Section VII presents the conclusions.

II. Short Time Fourier Transform

Signals belonging to speech, radar, sonar and data communication are all non-stationary signals i.e., the signal properties (amplitudes, frequencies and phases) change with time. For such signals a single discrete fourier transform would not be sufficient to describe the frequency characteristics of the signal. In order to describe such signals better, the short time fourier transform, also known as the time-dependent fourier transform, is used [9]. The STFT of a signal $x[n]$ is defined as

Manuscript received September 3, 2003; revised December 10, 2004.

S. R. Aedudodla was with the Indian Institute of Technology, Guwahati, India 781005. He is now with the Department of Electrical Engineering, University of Florida, Gainesville, FL 32611 US (e-mail: sandeep@dsp.ufl.edu).

K. M. M. Prabhu, is with the Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai, India 600036 (e-mail: prabhu@ee.iitm.ernet.in).

Publisher Item Identifier S 1682-0053(05)0232

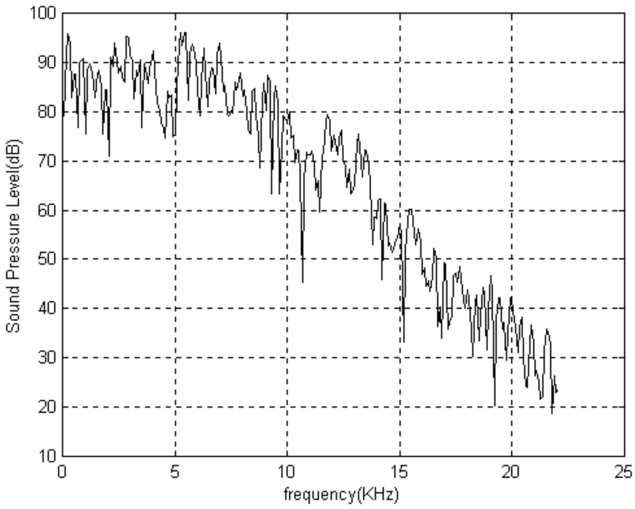


Fig. 1. Power spectrum of one of the audio blocks.

$$X[n,\lambda) = \sum_{k=-\infty}^{\infty} x[n+k]\, w[k]\, \exp[-j\lambda k], \quad (1)$$

where $w[n]$ is a window sequence. In (1) the frequency variable λ is a continuous one. The STFT tries to evaluate the way frequency content changes with time and hence better characterizes non-stationary signals.

In order to apply the STFT to signals of indeterminate length, a sampled time-dependent fourier transform is defined as

$$X[n,k] = X[n,2\pi k/N] = \sum_{m=0}^{L-1} x[n+m]w[m]\exp[-j(2\pi/N)km], \quad (2)$$

where $0 \le k \le N-1$ and L is the length of the window. The length of the window involves a tradeoff between frequency resolution and time resolution. The short time fourier transform is an important tool in speech analysis and processing.

III. Human Auditory Perception and Masking

The human auditory system has a dynamic range from 20 to 20000 Hz. The hearing works like a non-uniform filter bank, which is characterized by 26 critical bands whose bandwidths are a function of frequency and increase with increasing frequency. The critical bands can be expressed in terms of the Bark frequency scale. The conversion into the Bark frequency scale can be approximated by the function [13]

$$Z(f) = 13\arctan(0.00076f) + 3.5\arctan((f/7500)^2), \quad (3)$$

where Z denotes the frequency in Barks and f is the frequency in Hertz.

Masking is a perceptual property of the HAS, wherein the presence of a strong audio signal makes a temporal or spectral neighborhood of weaker audio signals imperceptible. In this paper we have used the MPEG Psychoacoustic Model 1. The model involves the computation of a masking curve, from which a set of thresholds for each of the 32 equal width sub-bands is derived. These masking thresholds determine the level below which the noise will not be perceived. The

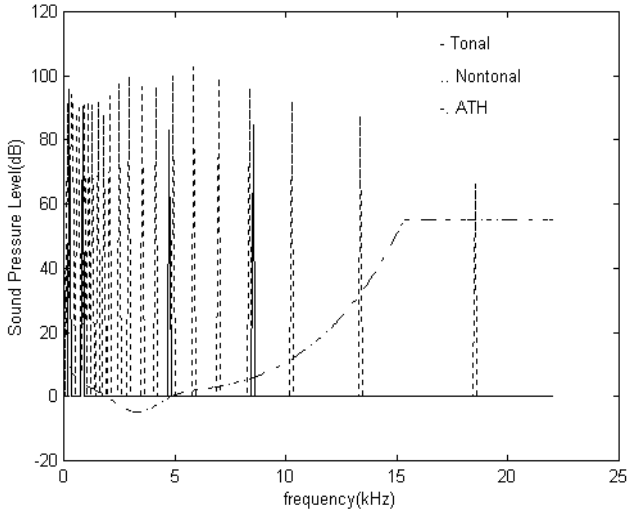


Fig. 2. Tonal and non-tonal components and the absolute threshold of hearing.

Psychoacoustic Model 1 involves the following procedure:

(i). The power spectrum of the input signal is computed using a 512-length FFT, after weighing the signal with a Hanning window of length $L = 512$, defined as

$$w[n] = \begin{cases} 0.5 - 0.5\cos[2\pi n/L], & 0 \le n \le L \\ 0, & \text{otherwise} \end{cases}, \quad (4)$$

The power spectrum is expressed in terms of the sound pressure level (SPL) [7] as in Fig. 1.

(ii). The absolute threshold of hearing (ATH) is computed according to the function [13]:

$$ATH(f) = 3.64(f/1000)^{-0.8} - 6.5\exp(-0.6(f/1000-3.3)^2) + 10^{-3}(f/1000)^4, \quad (5)$$

where $ATH(f)$ is in dB and f is the frequency in Hz.

(iii). Tonal and non-tonal components are identified, as they possess different masking properties. Determining the local maxima in the power spectrum identifies tonal components. All the nontonal components in each critical band are summed to form a single non-tonal masker. The index number of the non-tonal masker is set to the spectral line nearest to the geometric mean of the critical band. The absolute threshold and the tonal and non-tonal components are shown Fig. 2.

(iv). This step is known as decimation, where we reduce the number of maskers to be considered in determining the global masking threshold. Only the tonal and non-tonal components above the ATH are considered. Moreover, smaller tonal components within a distance of 0.5 Bark from a tonal component of greater power are removed.

(v). The individual masking thresholds for each of the tonal and non-tonal components are evaluated. The masking threshold at frequency z due to a masker at frequency z_m with the sound pressure level S is given by

$$T(z,z_m) = S(z_m) + \Gamma(z_m) + \mathrm{M}(z,z_m) \ \text{dB}, \quad (6)$$

where M is the masking function with different lower and upper slopes, defined as follows

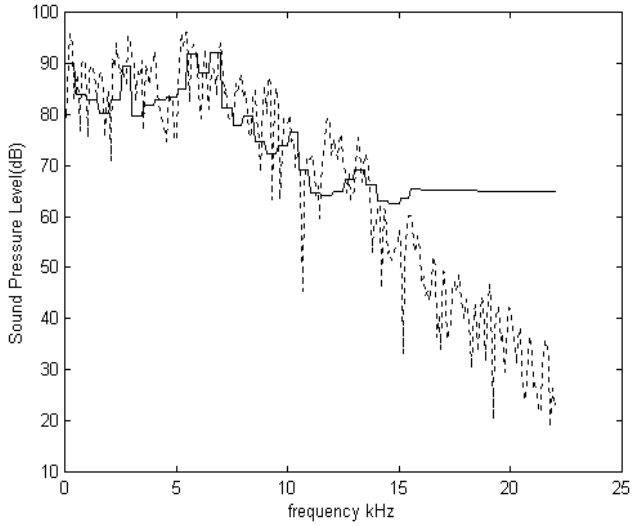


Fig. 3. Final masking threshold superimposed on the power spectrum.

$$M = \begin{cases} 17(\Delta z+1)-(0.4S(z_m)+6) \quad \text{dB}, & -3 \le \Delta z < -1 \\ (0.4S(z_m)+6)\Delta z \quad \text{dB}, & -1 \le \Delta z < 0 \\ -17\Delta z \quad \text{dB}, & 0 \le \Delta z < 1 \\ -(\Delta z-1)(17-0.15S(z_m))-17 \quad \text{dB}, & 1 \le \Delta z \le 8 \end{cases}, \quad (7)$$

where $\Delta z = z_m - z$. For the sake of reducing complexity, the masking effect of any masker is not considered outside the range $-3 \le \Delta z \le 8$. $\Gamma(z)$ is the masking index and is defined differently for tonal and non-tonal maskers. For tonal maskers,

$$\Gamma(z_m) = -1.525 - 0.275 z_m - 4.5 \quad \text{dB}, \quad (8)$$

and for non-tonal maskers,

$$\Gamma(z_m) = -1.525 - 0.175 z_m - 0.5 \quad \text{dB}. \quad (9)$$

(vi). The global masking threshold is then determined by considering all the individual masking thresholds and the ATH as follows

$$M_g(f) = 10\log_{10}\left[10^{ATH(f)/10} + \sum_j 10^{(T(z_j, Z(f))/10}\right], \quad (10)$$

The final masking threshold T_g is determined by taking the minimum value of M_g in each of the 32 equal width sub-bands of the spectrum. The final masking threshold along with the power spectrum is shown in Fig. 3.

IV. The Watermarking Algorithm

The procedure used for watermarking is an audio-adaptive one. The watermark is shaped taking into account the simultaneous and temporal masking properties of the original audio. The STFT is employed to characterize the frequency content of the audio. The window length L, used in the STFT in (2), involves a tradeoff between frequency and time resolutions. In our experiments, a Hanning window of length 512 was used. Larger window lengths could be used to obtain greater frequency resolution at the expense of a poorer time resolution and vice versa. The algorithm proceeds as follows:

(i). The watermark $y[n]$ is constructed as an independent and identically distributed (i.i.d) Gaussian random vector of zero mean and unit variance as advocated in [1].

(ii). For every sample n of the original audio signal $x[n]$, the STFT $X[n,k]$ is computed and the power spectrum is obtained.

(iii). The final masking threshold $T_g[n,k]$ is calculated as explained in Section III, from $X[n,k]$.

(iv). The STFT of the watermark $y[n]$ is calculated and $Y[n,k]$ is obtained.

(v). The mask $T_g[n,k]$ is used to weigh the watermark's frequency characteristics to obtain $P[n,k] = Y[n,k].T_g[n,k]$ and the inverse fourier transform of $P[n,k]$ is computed to give $p[n]$ of length 512 samples.

(vi). As this process is repeated for every sample number n corresponding to $x[n]$, we would obtain 512 different values of the watermark at every sample number n as $x[n]$ slides along the window in the STFT. The average of these values gives us the frequency shaped watermark at sample number n. At the end of this process, suppose the watermark signal obtained is $w'[n]$.

(vii). Next, the envelope $e[n]$, of the original audio signal $x[n]$, is computed as in [7] using a decaying exponential in order to approximate the temporal masking effects of the HAS. The choice of the decay time constant depends on the nature of the signal. A decay time constant of about 160 sec^{-1} was sufficient to guarantee imperceptibility of the watermark in silent regions for the audio signals considered in the experiments. The watermark $w'[n]$ is multiplied by $e[n]$ to get the final watermark $w[n] = w'[n].e[n]$.

(viii). The final watermark $w[n]$ is added to the original signal $x[n]$ to obtain the watermarked audio signal $x_w[n] = x[n] + w[n]$. The schematic of the algorithm is depicted in Fig. 4. The temporal mask ensures that no distortion is caused to the audio signal in the silent periods when the watermark is added to it.

V. Detection of the Watermark

In this algorithm, the original audio signal is not required for the extraction of the watermark. As explained in Section II, such a technique is applicable in several scenarios, which involve broadcasting of data. The algorithm is able to withstand common signal processing operations such as MPEG-1 Layer 3 compression, temporal resampling, jittering, cropping and the addition of colored noise to the watermarked signal. The attacks on the watermarked signal could be unintentional or malicious. The signal in which a watermark is to be detected could possibly have undergone one or more of the above mentioned distortions and would thus be different from $x_w[n]$. Let this received signal be denoted by $x_r[n]$. The received audio signal, $x_r[n]$, is then decorrelated by applying a technique involving linear prediction coding (LPC) similar to the one used in [12]. The LPC was implemented without a pre-emphasis filter. A matched filter is then used to extract the watermark from the decorrelated audio signal using the perceptually shaped original watermark as a template. Let the extracted watermark be denoted by $w_{ex}[n]$.

In order to compare the extracted watermark with the original watermark, we need to define a measure of

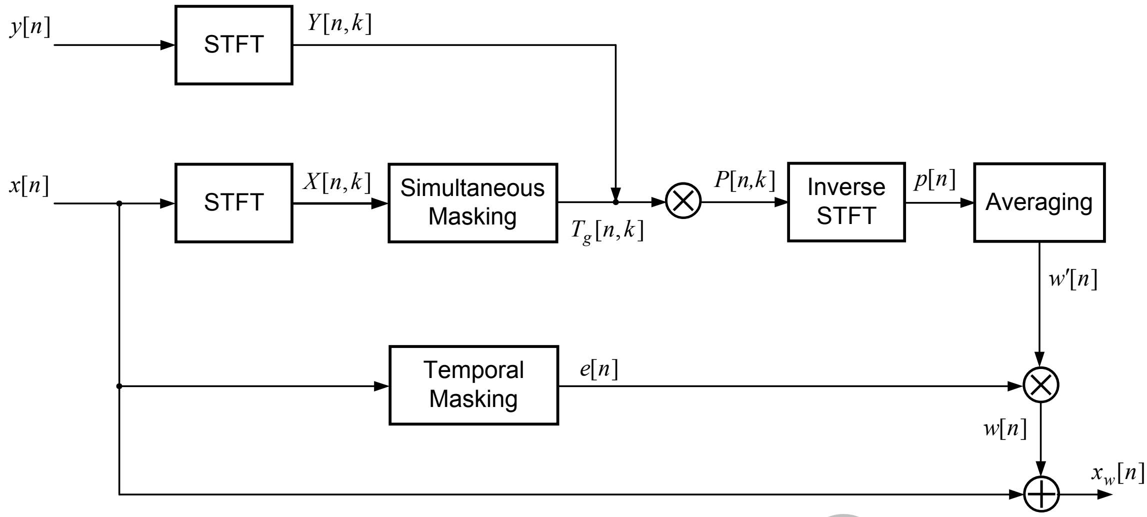


Fig. 4. Schematic of the proposed watermarking algorithm.

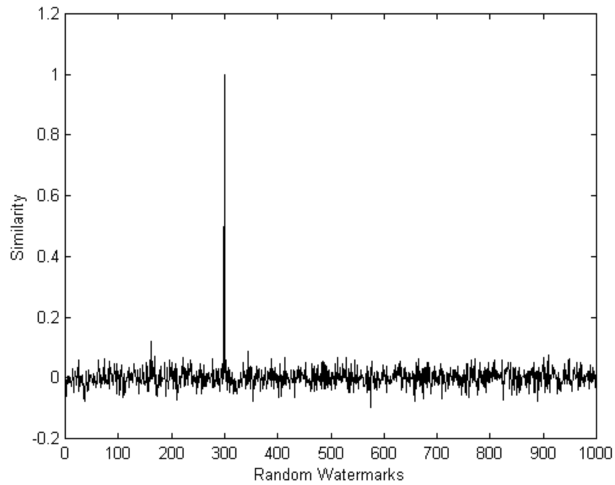


Fig. 5. Detection results of 1000 randomly generated watermarks of which only one was set to the original watermark.

similarity. In our experiments, we have used the normalized similarity measure defined by

$$Sim(w, w_{ex}) = \frac{\sum_i w_{ex}[i].w[i]}{\sum_i w[i].w[i]} \quad . \tag{11}$$

To ascertain that the received signal contains the watermark of a particular owner, *Sim* may be computed and compared with a threshold value. A watermarking algorithm should preferably give values of *Sim* that are much higher than the threshold if a watermark is present and values much lower than the threshold if no watermark is present in the received audio signal. The detection results using (11) for 1000 randomly generated watermarks are depicted in Fig. 5. Only one of the watermarks was set to the original watermark.

VI. Experiments and Results

The robustness of the watermarking algorithm is tested in the experiments, wherein the watermarked signal is subjected to distortions resulting from MPEG-1 Layer 3 (MP3) coding, colored noise, cropping and temporal resampling and jittering. The algorithm was tested on two audio signals Audio-1 and Audio-2. Audio-1 is a speech

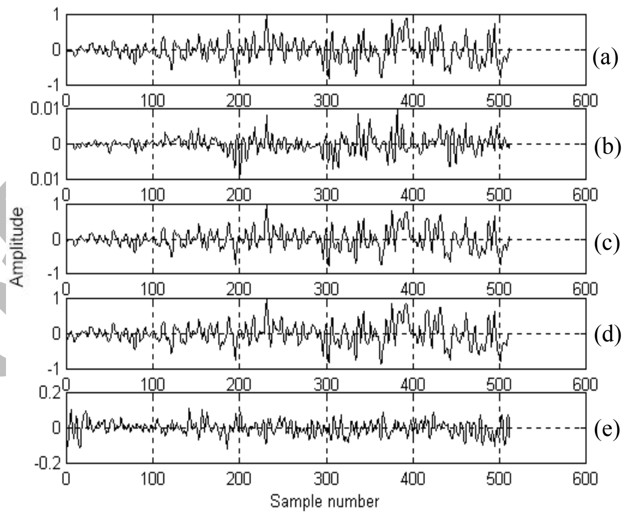
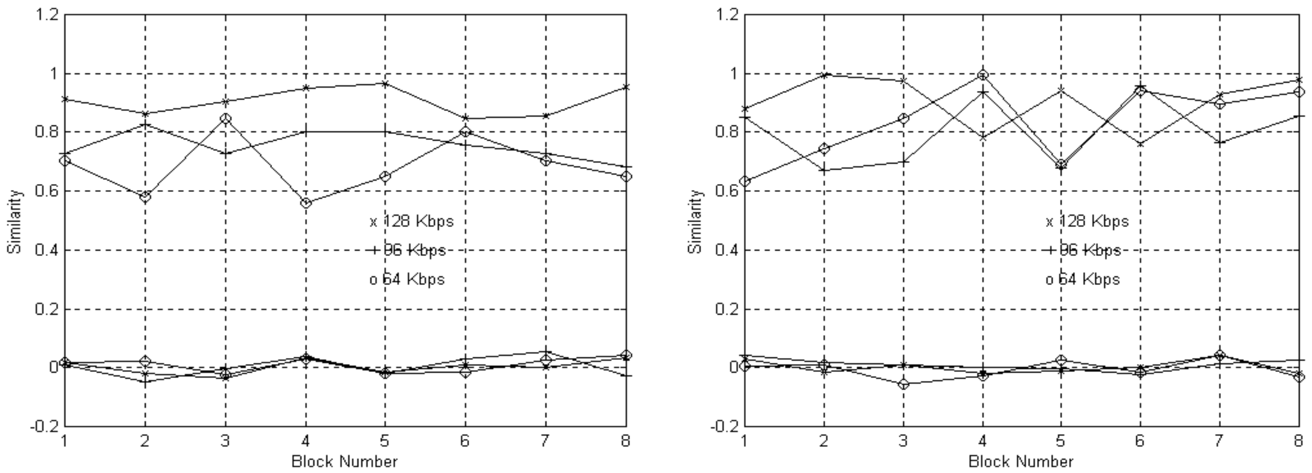


Fig. 6. One of the blocks of Audio-1 signal: (a) original, (b) watermark, (c) watermarked, (d) MP3 coded at 128kbps, and (e) coding error.

signal with a signal with sampling frequency of 44.1 kHz, and Audio-2 is a music signal with a sampling frequency of 22.05 kHz. The watermark detection experiments were carried out taking into account the length of the window and several windowing techniques. During the watermark extraction, LPC of order 30 was used for Audio-1 and LPC of order 15 was used for Audio-2. A threshold value of 0.4 was chosen during the watermark detection.

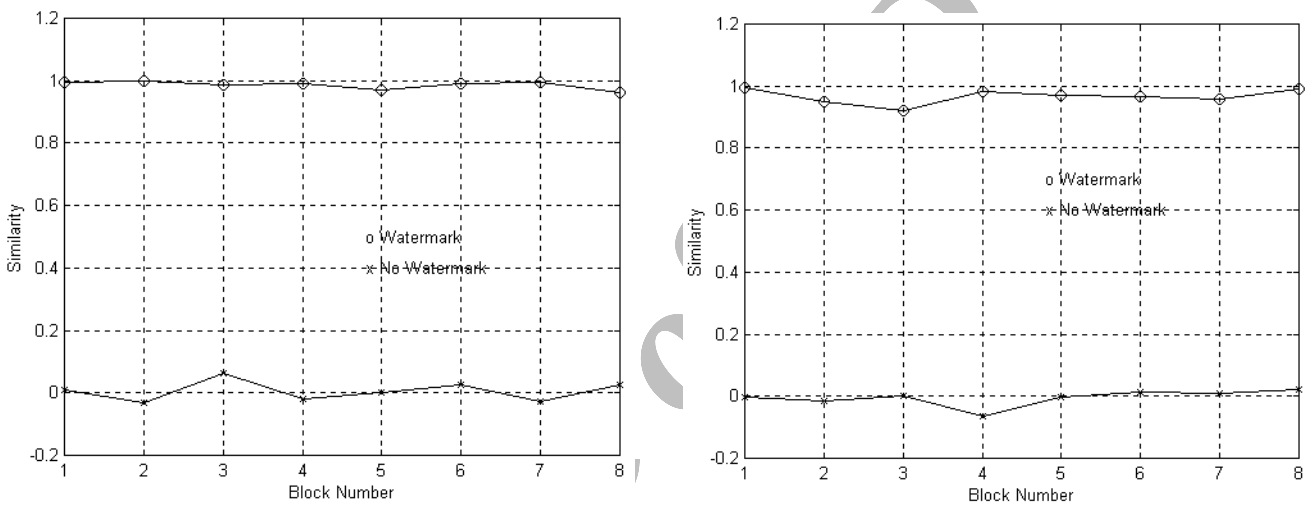
A. Perceptual Degradation and Listening Tests

In both the tested signals Audio-1 and Audio-2, the addition of the perceptually shaped watermark does not result in any perceptual distortions when compared to the original audio signals. Subjective listening tests were conducted on 8 listeners who were asked to point out any perceptual degradation in the audio samples. None of the listeners could distinguish between the watermarked audio and the original audio samples. The importance of shaping the watermark using temporal and frequency masking is highlighted in this regard. In terms of perceptual imperceptibility, this technique outperforms other time-domain methods not employing spectral and temporal masking to shape the watermark. The use of the STFT enables a better description of the spectral properties of the signal while designing the masks. When compared to the



(a) (b)

Fig. 7. Detection of watermark after MP3 coding for (a) Audio-1 and (b) Audio-2.



(a) (b)

Fig. 8. Detection of watermark after colored noise attack in (a) Audio-1 and (b) Audio-2.

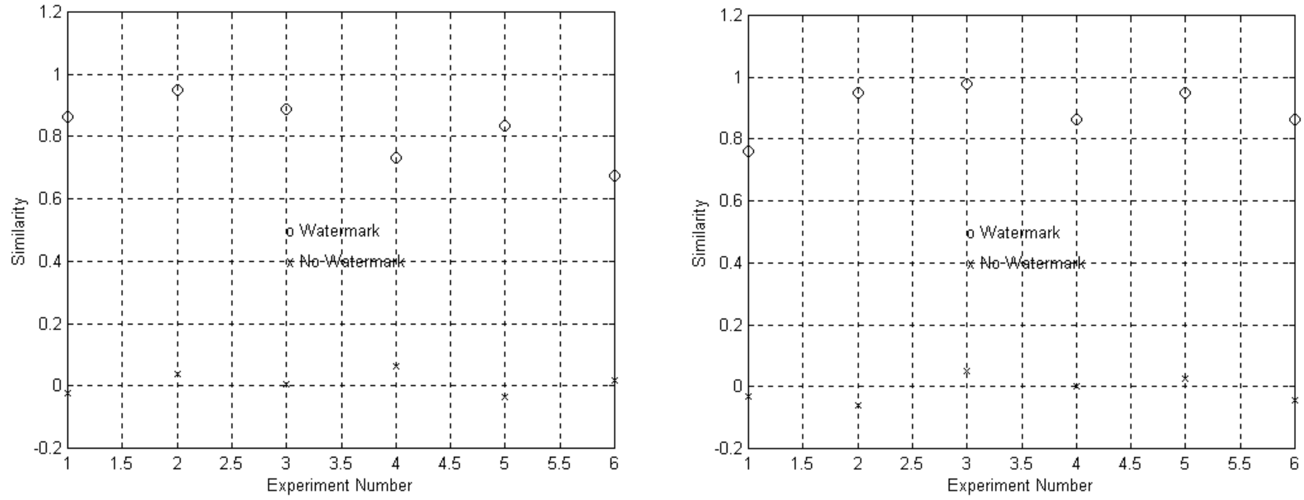
methods using the discrete fourier transform [7], our technique performs better as it can embed a watermark of similar energy but with lesser amount of perceptual degradation. Hence, for a particular tolerable amount of degradation, we can embed a watermark of greater energy, which increases the robustness of the watermark.

B. MPEG-1 Layer 3 (MP3) Compression

The MP3 is a widely used compression standard on the Internet. It achieves very high compression ratios and is thus useful in efficient real-time transmission of audio and enables the reduction of memory required to store the audio. This is a lossy coding method, which results in a coding error defined as the difference between the uncoded and coded signals. The watermarked signals of Audio-1 and Audio-2 are subjected to MP3 coding at bit rates of 64, 96 and 128 kb/s. The MP3 coder used was a software implementation of the standard. Fig. 6 compares the coding error magnitude with the magnitude of the watermark, the former being almost 20 times greater than the latter. The watermark detection test results are shown in Fig. 7. We give the *Sim* values for consecutive 512-length blocks of the audio signal using the Hanning window. The watermark detection results for when different types of windows (Hanning, Hamming, Blackman, Bartlett and Kaiser) used with varying window-lengths (128-1024) have been found to be quite similar.

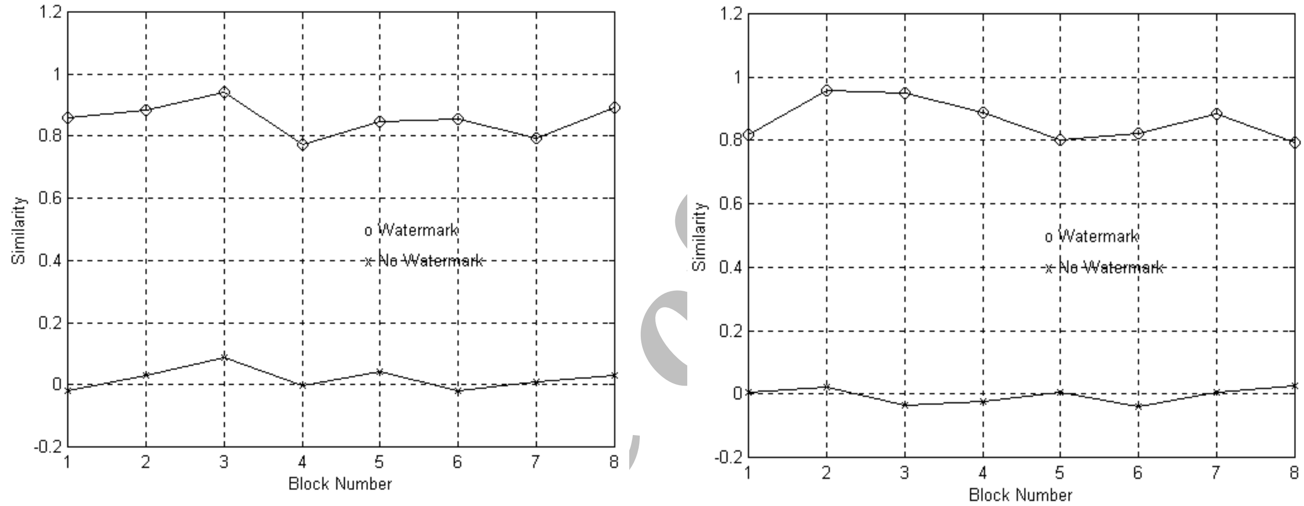
C. Colored Noise Addition

Gaussian noise, which is shaped in exactly the same manner as the watermark, when added to the watermarked signal, can be considered to be the worst possible case of attack by noise. This attack can be considered to be an attempt by the attacker to embed his own watermark into the watermarked audio and claim that it is his/her own. However, using our original audio sample, we can easily prove that the attacker's watermark is not present in our original audio signal. Since the attacker would only have access to the watermarked audio, he/she would try to modify the watermarked signal to construct his/her 'original'. Since any such attempt would only result in distortion being caused to the audio, there would be a certain limit to the extent of processing the attacker would employ. We would be able to detect our watermark in the audio sample, which the attacker would claim to be his/her 'original', thus proving the piracy of the attacker. Our method easily detects the actual watermark in the signal corrupted by colored noise. Colored noise obtained as said above, was added to the watermarked signal and the detection test was performed. Colored noise was generated 1000 times and each time the detection test was performed. The average values for each of the consecutive 512-length



(a) (b)

Fig. 9. Detection of watermark after cropping in (a) Audio-1 and (b) Audio-2.



(a) (b)

Fig. 10. Detection of watermark after temporal resampling in (a) Audio-1 and (b) Audio-2.

blocks using the Hanning window in the STFT domain are shown in Fig. 8, for both Audio-1 and Audio-2.

D. Cropping

More than half of the watermarked signal was cropped at random places in the watermarked signal. Colored noise was added to the remaining segments. During detection, the cropped portions were replaced by portions from the original audio, and the similarity between the original audio and the recovered audio was computed. This test was repeated 6 times and each time 1000 different generations of colored noise were used and the average value of similarity was computed. The detection results are shown in Fig. 9. This is the only attack when the original audio has been used in the watermark detection process.

E. Temporal Resampling

The watermarking algorithm exhibits resilience towards temporal resampling. Specifically, the watermarked signal was first downsampled by a factor of 2 and then using interpolation, was upsampled by a factor of 2. The detection results on the resampled audio are shown in Fig. 10. As shown in Fig. 11, temporal resampling results in significant distortions to the audio.

F. Jittering Attack

The robustness of the algorithm to the jittering attack [10], [11] was tested during the course of the experiments. Fig. 12 shows the performance of the technique when the watermarked audio is subjected to various degrees of the jittering attack, during which samples from the watermarked audio were randomly modified. The technique performs reasonably well even when the watermarked audio is subjected to serious jittering attack (up to 10%).

VII. Conclusion

A watermarking algorithm for digital audio in the time domain has been presented. The watermark is constructed as an i.i.d Gaussian random vector. The algorithm employs the short time fourier transform for better spectral characterization of the audio signal during the process of estimating the frequency masking characteristics of the human auditory system. The frequency mask along with a temporal mask is used to shape the watermark. The algorithm improves the imperceptibility and the robustness of the watermark as demonstrated by its resilience to MP3 coding, colored noise addition, cropping, temporal resampling and jittering attacks.

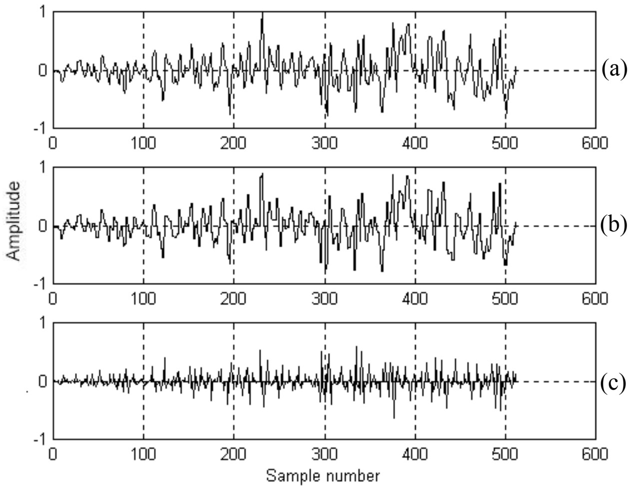


Fig. 11. Effect of temporal resampling (a) original, (b) downsampled by a factor of 2 and then upsampled through interpolation, and (c) error.

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their detailed constructive comments that immensely helped to improve the presentation of this paper.

REFERENCES

[1] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. on Image Processing*, vol. 6, no. 12, pp. 1673-1687, Dec. 1997.

[2] C. -S. Lu, S. K. Huang, C. J. Sze, and H. M. Liao, "Cocktail watermarking for digital image protection," *IEEE Trans. on Multimedia*, vol. 2, no. 4, pp. 209-224, Dec. 2000.

[3] S. D. Lin and C. F. Chen, "A robust DCT-based watermarking for copyright protection," *IEEE Trans. on Cons. Elec*., vol. 46, no. 3, pp. 415-421, Aug. 2000.

[4] A. Nikolaidis and I. Pitas, "Robust watermarking of facial images based on salient geometric pattern matching," *IEEE Trans. on Multimedia*, vol. 2, no. 3, pp. 172-184, Sep. 2000.

[5] J. R. Hernandez, M. Amado, and F. P. Gonzalez, "DCT-domain watermarking techniques for still images: detector performance analysis and a new structure," *IEEE Trans. on Image Processing*, vol. 9, no. 1, pp. 55-68, Jan 2000.

[6] L. De Strycker, *et al.,* "Implementation of a real-time digital watermarking process for broadcast monitoring on a Trimedia VLIW processor," *IEE Proc. Vision, Image and Signal Processing*, vol. 147, no. 4, pp. 371-376, Aug. 2000.

[7] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Processing*, vol. 66, no. 3, pp. 337-355, May 1998.

[8] F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proc. of the IEEE*, vol. 87, no. 7, pp. 1079-1106, Jul. 1999.

[9] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, NJ, 1989.

[10] X. Li and H. H. Yu, "Transparent and robust audio data hiding in subband domain," in *Proc. Int. Conf. on Information Technology: Coding and Computing*, pp. 397-400, New York, 2000.

[11] X. Li and H. H. Yu, "Transparent and robust audio data hiding in cepstrum domain," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, pp. 397-400, 2000.

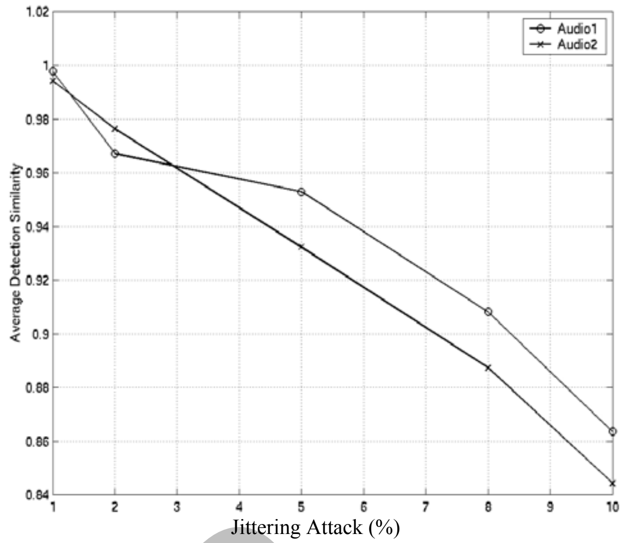


Fig. 12. Average detection similarity over 20 intervals after varying degrees of jittering attack.

[12] J. Seok, J. Hong and J. Kim, "A novel watermarking algorithm for copyright protection of digital audio," *ETRI Journal*, vol. 24, no. 3, pp. 181-189, Jun. 2002.

[13] —, *W.A.V.S. Compression: Background on the Psychoacoustic* Model, available online at http://is.rice.edu/~welsh/elec431/psychoAcoustic.html

Sandeep Aedudodla received the B.Tech. degree in electronics and communication engineering from the Indian Institute of Technology, Guwahati, in 2002 and the M.S. degree in electrical and computer engineering from the University of Florida, Gainesville, in 2004. He is currently working towards the Ph.D. degree at the University of Florida.

His research interests include wireless communication system design and signal processing for wireless communications.

K. M. M. Prabhu obtained the B.Sc. (Eng.) degree in Electronics and Communication Engineering from the Government College of Engineering, Trivandrum, the M.Sc. (Eng.) degree in Applied Electronics Engineering from M.I.T., Madras and the Ph.D. degree in Digital Signal Processing from the Indian Institute of Technology (I.I.T.) Madras, Chennai, India, in 1971, 1973 and 1981, respectively. He has been on the faculty of Department of Electrical Engineering, IIT Madras, where he is currently a full Professor.

While at IIT Madras, he has been a Principal Investigator of many projects funded by the Defense Research and Development Organization. He has published over 110 research papers in the International Journals and the International Conference Proceedings. Under his supervision, 17 M.S. (by research) and Ph.D. students have completed their degrees. He has conducted 10 short-term courses in the area of DSP for the benefit of Engineers and Scientists from various organizations. He spent his sabbatical, from August 1999 to August 2000, with the Signal Processing Systems Group, Eindhoven University of Technology, The Netherlands. He has been a reviewer for a number of IEEE, IEE and other International Journals. His current research interests include DSP Algorithms and their Implementations, Digital filter structures, time-frequency signal analysis and wavelet-based signal processing. Prof. Prabhu is a Senior Member of the IEEE.