ELSEVIER

Signal Processing 66 (1998) 337–355

SIGNAL PROCESSING

Robust audio watermarking using perceptual masking[1]

Mitchell D. Swanson[a,*], Bin Zhu[a], Ahmed H. Tewfik[a], Laurence Boney[b]

[a] *Department of Electrical Engineering, Uni ersity of Minnesota, Minneapolis, MN 55455, USA*
[b] *Ecole Nationale Supérieure des Télécommunications/SIG, 46 rue Barrault, 75634 Paris Cedex 13, France*

Received 10 February 1997; received in revised form 11 November 1997

Abstract

We present a watermarking procedure to embed copyright protection into digital audio by directly modifying the audio samples. Our audio-dependent watermarking procedure directly exploits temporal and frequency perceptual masking to guarantee that the embedded watermark is inaudible and robust. The watermark is constructed by breaking each audio clip into smaller segments and adding a perceptually shaped pseudo-random sequence. The noise-like watermark is statistically undetectable to prevent unauthorized removal. Furthermore, the author representation we introduce resolves the deadlock problem. We also introduce the notion of a dual watermark: one which uses the original signal during detection and one which does not. We show that the dual watermarking approach together with the procedure that we use to derive the watermarks effectively solves the deadlock problem. We also demonstrate the robustness of that watermarking procedure to audio degradations and distortions, e.g., those that result from colored noise, MPEG coding, multiple watermarks, and temporal resampling. © 1998 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Wir stellen ein Wasserzeichen-Verfahren zur Einbettung des Urheberrechtsschutzes in digitale Audiodaten vor, wobei die Audiosignalwerte direkt modifiziert werden. Unser audioabhängiges Wasserzeichen-Verfahren nützt unmittelbar die Wahrnehmungsverdeckung in Zeit-und Frequenzbereich aus, um sicherzustellen, däs das eingebettete Wasserzeichen unhörbar und robust ist. Das Wasserzeichen wird konstruiert, indem jeder Audioabschnitt in kleinere Segmente zerteilt wird und eine wahrnehmungsgerecht geformte Pseudozufallsfolge hinzuaddiert wird. Das geräuschartige Wasserzeichen ist statistisch nicht erkennbar, um unautorisiertes Entfernen zu verhindern. Weiters löst die von uns eingeführte Autorendarstellung das Pattstellungsproblem. Wir führen auch den Begriff dualer Wasserzeichen ein: eines, das das Originalsignal während der Erkennung benutzt, und eines, das es nicht benutzt. Wir zeigen, däs der Ansatz mit dualen Wasserzeichen in Verbindung mit dem Verfahren, das wir zur Herleitung der Wasserzeichen einsetzen, das Pattstellungsproblem wirksam löst. Wir zeigen auch die Robustheit des Wasserzeichen-Verfahrens gegenüber Audiostörungen und -verzerrrungen, z.B. jenen, die von farbigem Rauschen, MPEG-Codierung, mehrfachen Wasserzeichen, und Abtastratenwandlung herrühren. © 1998 Elsevier Science B.V. All rights reserved.

Résumé

Nous présentons dans cet article une procédure de watermarking permettant d'intégrer une protection de droits d'auteur dans des données audio numériques par modification directe des échantillons audio. Cette procédure exploite

* Corresponding author.

[1] This work was supported by AFOSR under grant AF/F49620-94-1-0461. Patent pending, Media Science, Inc., 1996.

0165-1684/98/$19.00 © 1998 Elsevier Science B.V. All rights reserved.
PII S0165-1684(98)00014-0

directement les masquages perceptuels temporel et fréquentiel pour garantir que le filigrane numérique (watermark) est inaudible et robuste. Le watermark est construit en fragmentant chaque morceau audio en segments plus petits et en ajoutant une séquence pseudo-aléatoire modelée perceptuellement. Le watermark semblable à du bruit est indétectable statistiquement afin d'empêcher une suppression non autorisée de celui-ci. De plus, la représentation de l'auteur que nous introduisons résoud le problème de l'impasse. Nous introduisons également la notion de watermark dual: l'un qui utilise le signal original lors de la détection et l'autre non. Nous montrons que l'approche de watermarking dual combinée avec la procédure que nous utilisons pour dériver les watermarks résoud effectivement le problème de l'impasse. Nous mettons également en évidence la robustesse de cette procédure de watermarking vis-à-vis des dégradations et distorsions audio, telles que celles qui résultent d'un bruit coloré, d'un codage MPEG, de watermarks multiples, et de ré-échantillonnage temporel. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Copyright protection; Masking; Digital watermarking

1. Introduction

Efficient distribution, reproduction, and manipulation have led to wide proliferation of digital media, e.g., audio, video, and images. However, these efficiencies also increase the problems associated with copyright enforcement. For this reason, creators and distributors of digital data are hesitant to provide access to their intellectual property. They are actively seeking reliable solutions to the problems associated with copyright protection of multimedia data.

Digital watermarking has been proposed as a means to identify the owner or distributor of digital data. Watermarking is the process of encoding hidden copyright information in digital data by making small modifications to the data samples. Unlike encryption, watermarking does not restrict access to the data. Once encrypted data is decrypted, the media is no longer protected. A watermark is designed to reside in the host data. When the ownership of a digital work is in question, the information can be extracted to completely characterize the owner.

To function as a useful and reliable intellectual property protection mechanism, the watermark must be:

- within the host media;
- within the host media;
- to ensure security and thwart unauthorized removal;
- to manipulation and signal processing operations on the host signal, e.g., noise, compression, cropping, resizing, D/A conversions, etc.; and
- to completely characterize the copyright owner.

In particular, the watermark may not be stored in a file header, a separate bit stream, or a separate file. Such copyright mechanisms are easily removed. The watermark must be inaudible within the host audio data to maintain audio quality. The watermark must be statistically undetectable to thwart unauthorized removal by a 'pirate'. A watermark which may be localized through averaging, correlation, spectral analysis, Kalman filtering, etc., may be readily removed or altered, thereby destroying the copyright information.

The watermark must be robust to signal distortions, incidental and intentional, applied to the host data. For example, in most applications involving storage and transmission of audio, a lossy coding operation is performed on the audio to reduce bit-rates and increase efficiency. Operations which damage the host audio also damage the embedded watermark. The watermark is required to survive such distortions to identify the owner of the data. Furthermore, a resourceful pirate may use a variety of signal processing operations to attack a digital watermarking. A pirate may attempt to defeat a watermarking procedure in two ways: (1) damage the host audio to make the watermark undetectable, or (2) establish that the watermarking scheme is unreliable, i.e., it detects a watermark when none is present. The watermark should be impossible to defeat without destroying the host audio.

Finally, the watermark should be readily extracted given the watermarking procedure and the proper author signature. Without the correct signature, the watermark cannot be removed. The extracted watermark must correctly identify the owner and solve the deadlock issue (cf. Section 2) when multiple parties claim ownership.

Watermarking digital media has received a great deal of attention recently in the literature and the research community. Most watermarking schemes focus on image and video copyright protection, e.g., [1–3,7,10,14,15,18,19,22,24]. A few audio watermarking techniques have been reported. Several techniques have been proposed in [1]. Using a phase coding approach, data is embedded by modifying the phase values of Fourier transform coefficients of audio segments. Embedding data as spread spectrum noise have also been proposed. A third technique, echo coding, employs multiple decaying echoes to place a peak in the cepstrum at a known location. Another audio watermarking technique is proposed in [21], where Fourier transform coefficients over the middle frequency bands are replaced with spectral components from a signature. some commercial products are also available. The ICE system from Central Research Laboratories inserts a pair of very short tone sequences into an audio track. An audio watermarking product MusiCode is available from ARIS technologies.

Most schemes utilize the fact that digital media contain perceptually insignificant components which may be replaced or modified to embed copyright protection. However, the techniques do not exploit spatial/temporal and frequency masking. Thus, the watermark is not guaranteed inaudible. Furthermore, robustness is not maximized. The amount of modification made to each coefficient to embed the watermark are estimated and not necessarily the maximum amount possible. In this paper, we introduce a novel watermarking scheme for audio which exploits the human auditory system (HAS) to guarantee that the embedded watermark is imperceptible. As the perceptual characteristics of individual audio signals vary, the watermark adapts to and is highly dependent on the audio being watermarked. Our watermark is generated by filtering a pseudo-random sequence (author id) with a filter that approximates the frequency masking characteristics of the HAS. The resulting sequence is further shaped by the temporal masking properties of the audio. Based on pseudo-random sequences, the noise-like watermark is statistically undetectable. Furthermore, we will show in the sequel that the watermark is extremely robust to a large number of signal processing operations and is easily extracted to prove ownership.

The work presented in this paper offers several major contributions to the field, including

A - : The embedded watermark to each individual host signal. In particular, the temporal and frequency distribution of the watermark are dictated by the temporal and frequency masking characteristics of the host audio signal. As a result, the amplitude (strength) of the watermark increases and decreases with host, e.g., lower amplitude in 'quiet' regions of the audio. This guarantees that the embedded watermark is inaudible while having the maximum possible energy. Maximizing the energy of the watermark adds robustness to attacks.

A : An author is represented with a pseudo-random sequence created by a pseudo-random generator [13] and keys. One key is dependent, while the second key is dependent. The representation is able to resolve rightful ownership in the face of multiple ownership claims.

A . The watermarking scheme uses the original audio signal to detect the presence of a watermark. The procedure can handle virtually types of distortions, including cropping, temporal rescaling, etc., using a generalized likelihood ratio test. As a result, the watermarking procedure is a powerful digital copyright protection tool. We integrate this procedure with a second watermark which does require the original signal. The dual watermarks also address the deadlock problem.

In the next section, we introduce our noise-like author representation and the dual watermarking scheme. Our frequency and temporal masking models are reviewed in Section 3. Our watermarking design and detection algorithms are introduced in Sections 4 and 5. Finally, experimental results are presented in Section 6. Watermark statistics and fidelity results for four test audio signals are

presented. The robustness of our watermarking procedure is illustrated for a wide assortment of signal processing operations and distortions. We present our conclusion in Section 7.

2. Author representation, dual watermarking and the deadlock problem

Data embedding algorithms may be used to establish ownership and distribution of data. In fact, this is the application of data embedding or watermarking that has received most attention in the literature. Unfortunately, most current watermarking schemes are unable to resolve rightful ownership of digital data when multiple ownership claims are made, i.e., when a deadlock problem arises. The inability of many data embedding algorithms to deal with deadlock, first described by Craver et al. [4], is independent of how the watermark is inserted in the multimedia data or how robust it is to various types of modifications.

Today, no scheme can unambiguously determine ownership of a given multimedia signal if it does not use an original or other copy in the detection process to at least construct the watermark to be detected. A pirate can simply add his or her watermark to the watermarked data or counterfeit a watermark that correlates well or is detected in the contested signal. Current data embedding schemes used as copyright protection algorithms are unable to establish who watermarked the data first. Furthermore, none of the current data embedding schemes has been proven to be immune to counterfeiting watermarks that will correlate well with a given signal as long as the watermark is not restricted to partially depend in a non-invertible manner on the signal.

If the detection scheme can make use of the original to construct the watermark, then it may be possible to establish unambiguous ownership of the data regardless of whether the detection scheme subtracts the original from the signal under consideration prior to watermark detection or not. Specifically, [5] derives a set of sufficient conditions that watermarks and watermarking schemes must satisfy to provide unambiguous proof of ownership. For example, one can use watermarks derived from pseudo-random sequences that depend on the signal and the author. Ref. [5] establishes that this will work for watermarking procedures regardless of whether they subtract the original from the signal under consideration prior to watermark detection or not. Ref. [20] independently derived a similar result for a restricted class of watermarking techniques that rely on subtracting a signal derived from the original from the signal under consideration prior to watermark detection. The signal-dependent key also helps to thwart the 'mix-and-match' attack described in [5].

An author can construct a watermark that depends on the audio signal and the author and provides unambiguous proof of ownership as follows. The author has two random keys $_1$ and $_2$ (i.e., seeds) from which a pseudo-random sequence can be generated using a suitable pseudo-random sequence generator [16]. Popular generators include RSA, Rabin, Blum/Micali, and Blum/Blum/Shub [6]. With the two proper keys, the watermark may be extracted. Without the two keys, the data hidden in the signal is statistically undetectable and impossible to recover. Note that classical maximal length pseudo noise sequence (i.e., -sequence) generated by linear feedback shift registers are used to generate a watermark. Sequences generated by shift registers are cryptographically insecure: one can solve for the feedback pattern (i.e., the keys) given a small number of output bits .

The noise-like sequence may be used to derive the actual watermark hidden into the audio signal or control the operation of the watermarking algorithm, e.g., determine the location of samples that may be modified. The key $_1$ is dependent. The key $_2$ is dependent. The key $_1$ is the secret key assigned to (or chosen by) the author. Key $_2$ is which the author wishes to watermark. It is computed from the signal using a one-way hash function. For example, the tolerable error levels supplied by masking models (see Section 3) are hashed in [20] to a key $_2$. Any one of a number of well-known secure one-way hash functions may be used to compute $_2$, including RSA, MD4 [17], and SHA [12]. For example, the Blum/Blum/Shub pseudo-random generator uses the one way function

$= {}_n(\ ) = {}^2 \bmod$ where $=$ for primes and so that $= = 3 \bmod 4$. It can be shown that generating or from partial knowledge of is for the Blum/Blum/Shub generator.

The signal-dependent key ${}_2$ makes counterfeiting very difficult. The pirate can only provide key ${}_1$ to the arbitrator. Key ${}_2$ is automatically computed by the watermarking algorithm from the original signal. As it is computationally infeasible to invert the one-way hash function, the pirate is unable to fabricate a counterfeit original which generates a desired or predetermined watermark.

Deadlock may also be resolved using the dual watermarking scheme of [20]. That scheme employs a of watermarks. One watermarking procedure requires the original data set for watermark detection. This paper provides a detailed description of that procedure and of its robustness. The second watermarking procedure does require the original data set. A data embedding technique which satisfies the restrictions outlined in [5] can be used to insert the second watermark. The second watermark need not be highly robust to editing of the audio segment since, as we shall see below, it is meant to protect the audio clip that a pirate claims to be his . The robustness level of most of the recent watermarking techniques that do not require the original for watermark detection is quite adequate. The arbitrator would expect the original to be of a high enough quality. This limits the operations that a pirate can apply to an audio clip and still claim it to be his high-quality original sound. The watermark that requires the original audio sequence for its detection is very robust as we show in this paper.

In case of deadlock, the arbitrator simply first checks for the watermark that requires the original for watermark detection. If the pirate is clever and has used the attack suggested in [4] and outlined above, the arbitrator would be unable to resolve the deadlock with this first test. The arbitrator simply then checks for the watermark that require the original audio sequence in the audio segments that each ownership contender claims to be his . Since the original audio sequence of a pirate is derived from the watermarked copy produced by the rightful owner, it will contain the watermark of the rightful owner. On the other hand, the true original of the rightful owner will not contain the watermark of the pirate since the pirate has no access to that original and the watermark does not require subtraction of another data set for its detection.

3. Audio masking

Audio masking is the effect by which a faint but audible sound becomes inaudible in the presence of another louder audible sound, i.e., the masker [9]. The masking effect depends on the spectral and temporal characteristics of both the masked signal and the masker. Our watermarking procedure directly exploits both frequency and temporal masking characteristics to embed an inaudible and robust watermark.

3.1. Frequency masking

Frequency masking refers to masking between frequency components in the audio signal. If two signals, which occur simultaneously, are close together in frequency, the stronger masking signal may make the weaker signal inaudible. The masking threshold of a masker depends on the frequency, sound pressure level (SPL), and tone-like or noise-like characteristics of both the masker and the masked signal [13]. It is easier for a broadband noise to mask a tonal, than for a tonal signal to mask out a broadband noise. Moreover, higher-frequency signals are more easily masked.

The human ear acts as a frequency analyzer and can detect sounds with frequencies which vary from 10 to 20 000 Hz. The HAS can be modeled by a set of 26 band-pass filters with bandwidths that increase with increasing frequency. The 26 bands are known as the critical bands. The critical bands are defined around a center frequency in which the noise bandwidth is increased until there is a just noticeable difference in the tone at the center frequency. Thus, if a faint tone lies in the critical band of a louder tone, the faint tone will not be perceptible.

Frequency masking models are readily obtained from the current generation of high-quality audio

codes. In this work, we use the masking model defined in ISO-MPEG Audio Psychoacoustic Model 1, for Layer I [8]. We are currently updating our frequency masking model to the model specified by ISO-MPEG Audio Layer III. The Layer I masking method is summarized as follows for a 32 kHz sampling rate [8,11]. The MPEG model also supports sampling rates of 44.1 kHz and 48 kHz.

Step 1: *C* . Each 16 ms segment of the signal ( ), $N = 512$ samples, is weighted with a Hann window, ( ):

$$( ) = \frac{\sqrt{8/3}}{2}\left[1 - \cos\left(2\pi\frac{}{N}\right)\right]. \quad (1)$$

The power spectrum of the signal ( ) is calculated as

$$( ) = 10\log_{10}\left[\frac{1}{N}\left\|\sum_{n=0}^{N-1} ( ) ( )\exp\left(-\mathrm{j}2\pi\frac{}{N}\right)\right\|^2\right]. \quad (2)$$

The maximum is normalized to a reference sound pressure level of 96 dB. The power spectrum of a 32 kHz test signal is shown in Fig. 1.

Step 2: *I* . Tonal (sinusoidal) and non-tonal (noisy) components are identified because their masking models are different.

A tonal component is a local maximum of the spectrum (( ) > (+ 1) and ( ) ⩾ (− 1)) satisfying:

$$( ) - (+) \geqslant 7\,\mathrm{dB},$$
$$\in [-2, +2] \quad \text{if } 2 < < 63$$
$$\in [-3, -2, +2, +3] \quad \text{if } 63 \leqslant < 127$$
$$\in [-6, \ldots, -2, +2, \ldots, +6]$$
$$\text{if } 127 \leqslant \leqslant 250.$$

We add to its intensity those of the previous and following components: Other tonal components in the same frequency band are no longer considered. Non-tonal components are made of the sum of the intensities of the signal components remaining in each of the 24 critical bands between 0 and 15 500 Hz. The auditory system behaves as a bank of bandpass filters, with continuously overlapping center frequencies. These 'auditory filters' can be approximated by rectangular filters with critical

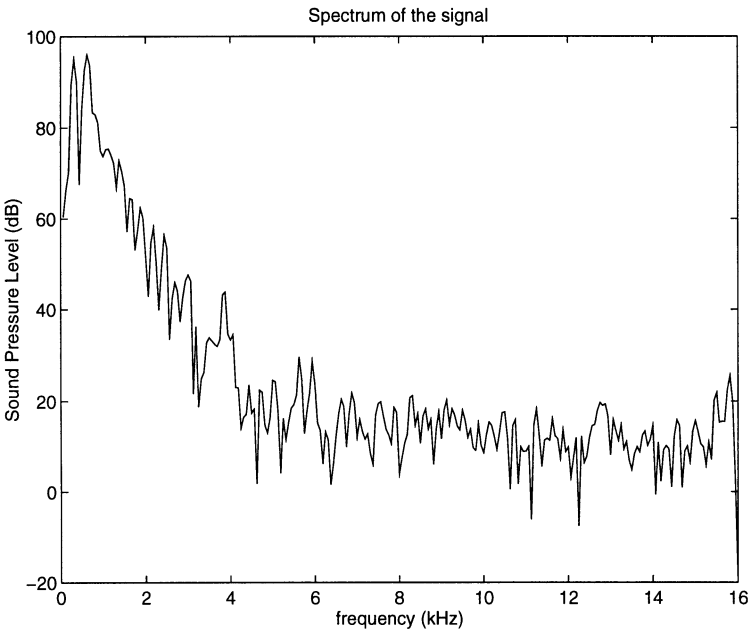

Fig. 1. Power spectrum of audio signal.

bandwidth increasing with frequency. In this model, the audible band is therefore divided into 24 non-regular critical bands. Tonal and non-tonal components of the example audio signal are shown in Fig. 2.

Step 3: *R* . Components below the absolute hearing threshold and tonal components separated by less than 0.5 Barks are removed. A plot of the removed components, along with the absolute hearing threshold is shown in Fig. 3.

Step 4: *I* .
In this step, we account for the frequency masking effects of the HAS. We need to discretize the frequency axis according to hearing sensitivity and express frequencies in Barks. Note that hearing sensitivity is higher at low frequencies. The resulting masking curves are almost linear and depend on a masking index different for tonal and non-tonal components. They are characterized by different lower and upper slopens depending on the distance between the masked and the masking component. We use $_1$ to denote the set of frequencies present in the test signal. The global masking threshold for each frequency $_2$ takes into account the absolute hearing threshold $_a$ and the masking curves P_2 of the N_t tonal components and N_n non-tonal components:

$$_m(\ _2) = 10\log_{10}\left[10^{S_a(f_2)/10} + \sum_{j=1}^{N_t} 10^{P_2(f_2,f_1,P_1)/10} + \sum_{j=1}^{N_n} 10^{P_2(f_2,f_1,P_1)/10}\right]. \quad (3)$$

The masking threshold is then the minimum of the local masking threshold and the absolute hearing threshold in each of the 32 equal width subbands of the spectrum. Any signal which falls below the masking threshold is inaudible. A plot of the original spectrum, along with the masking threshold, is shown in Fig. 4.

As a result, for each audio block of $N = 512$ samples, a masking value (i.e., threshold) for each frequency component is produced. Modifications to the audio-frequency components less than the masking threshold create no audible distortions to the audio piece.

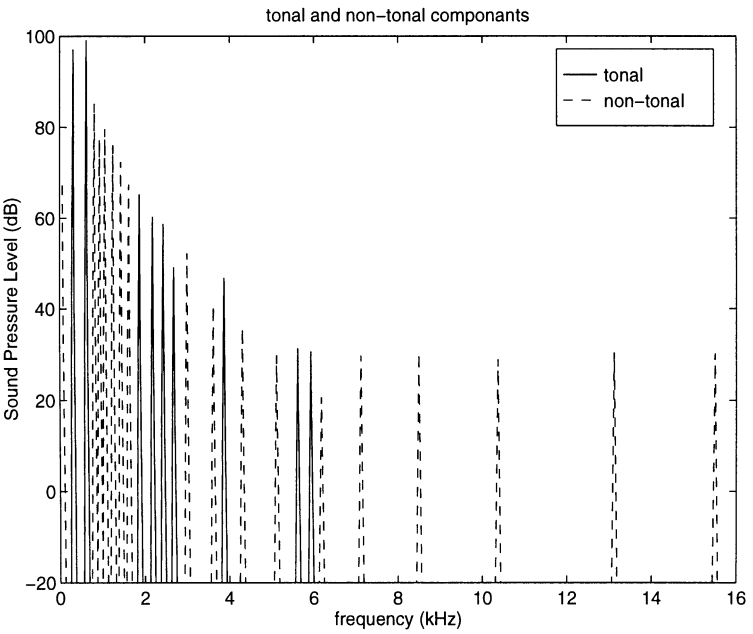

Fig. 2. Identification of tonal components.

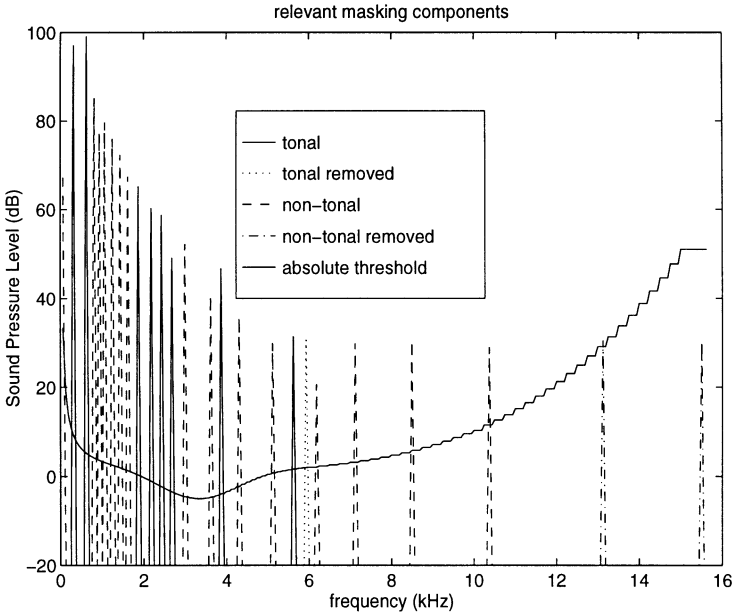

Fig. 3. Removal of masked components.

3.2. Temporal masking

Temporal masking refers to both pre- and post-masking. Pre-masking effects render weaker signals inaudible before the stronger masker is turned on, and post-masking effects render weaker signals inaudible after the stronger masker is turned off. Pre-masking occurs from 5 to 20 ms before the masker is turned on while post-masking occurs from 50 to 200 ms after the masker is turned off [13]. Note that temporal and frequency masking effects have dual localization properties. Specifically, frequency masking effects are localized in the frequency domain, while temporal masking effects are localized in the time domain.

We approximate temporal masking effects using the envelope of the host audio. The envelope is modeled as a decaying exponential. In particular, the estimated envelope ( ) of signal ( ) increases with the signal and decays as $e^{-\alpha t}$. An audio signal, along with its estimated envelope, is shown in Fig. 5.

4. Watermark design

Each audio signal is watermarked with a unique noise-like sequence shaped by the masking phenomena. The watermark consists of (1) an author representation (cf. Section 2), and (2) spectral and temporal shaping using the masking effects of the HAS.

Our watermarking scheme is based on a repeated application of a basic watermarking operation on smaller segments of the audio signal. A diagram of our audio watermarking technique is shown in Fig. 6. The length N audio signal is first segmented into blocks $_i(\ )$ of length 512 samples, $= 0,1,\ldots,\lfloor N/512 \rfloor - 1$, and $= 0,1,\ldots,511$. The block size of 512 samples is dictated by the frequency masking model we employ. Block sizes of 1024 have also been used. The algorithm works as follows. For each audio segment $_i(\ )$:

1. compute the power spectrum $_i(\ )$ of the audio segment $_i(\ )$ (Eq. (2));
2. compute the frequency mask $M_i(\ )$ of the power spectrum $_i(\ )$ (cf. Section 3.1);

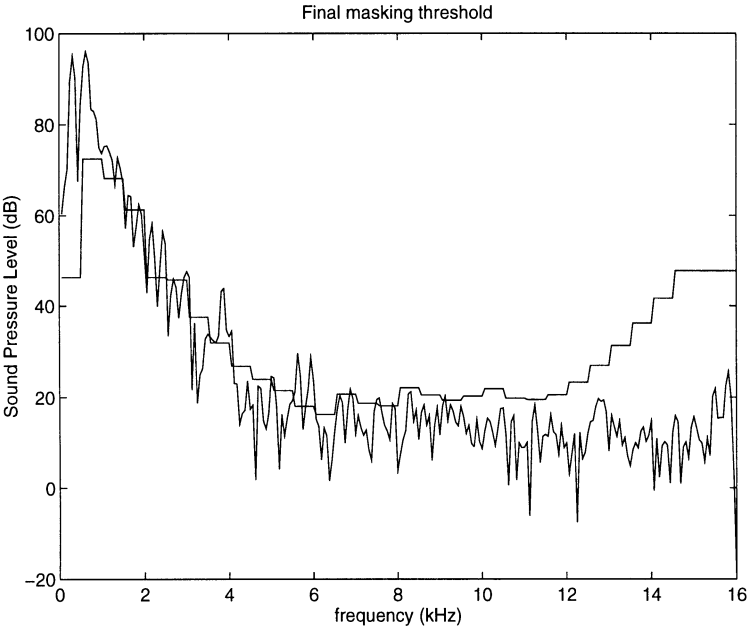

Fig. 4. Original spectrum and masking threshold.

3. use the mask $M_i(\ )$ to weight the noise-like author representation for that audio block, creating the shaped author signature $P_i(\ ) = Y_i(\ )M_i(\ )$;
4. compute the inverse FFT of the shaped noise $_i(\ ) = \mathrm{IFFT}(P_i(\ ))$;
5. compute the temporal mask $_i(\ )$ of $_i(\ )$ (cf. Section 3.2);
6. use the temporal mask $_i(\ )$ to further shape the frequency shaped noise, creating the watermark $_i(\ ) = _i(\ )\ _i(\ )$ of that audio segment;
7. create the watermarked block $'_i(\ ) = _i(\ ) + _i(\ )$.

The overall watermark for a signal is simply the concatenation of the watermark segments $_i$ for all of the length 512 audio blocks. The author signature $_i$ for block is computed in terms of the personal author key $_1$ and signal-dependent key $_2$ computed from block $_i$.

The dual localization effects of the frequency and temporal masking control the watermark in both domains. As noted earlier, frequency-domain shaping alone is not enough to guarantee that the watermark will be inaudible. Frequency-domain masking computations are based on a Fourier transform analysis. A fixed length Fourier transform does not provide good time localization for our application. In particular, a watermark computed using frequency-domain masking will spread in time over the entire analysis block. If the signal energy is concentrated in a time interval that is shorter than the analysis block length, the watermark is not masked outside of that subinterval. This leads to audible distortion, e.g., pre-echoes. The temporal mask guarantees that the 'quiet' regions are not disturbed by the watermark.

5. Watermark detection

The watermark should be extractable even if common signal processing operations are applied to the host audio. This is particularly true in the case of deliberate unauthorized attempts to remove it. For example, a pirate may attempt to add noise, filter, code, re-sample, etc., an audio piece in an attempt to destroy the watermark. As the embedded watermark is noise-like, a pirate has insufficient knowledge to directly remove the watermark. Therefore, any destruction attempts are done blindly.

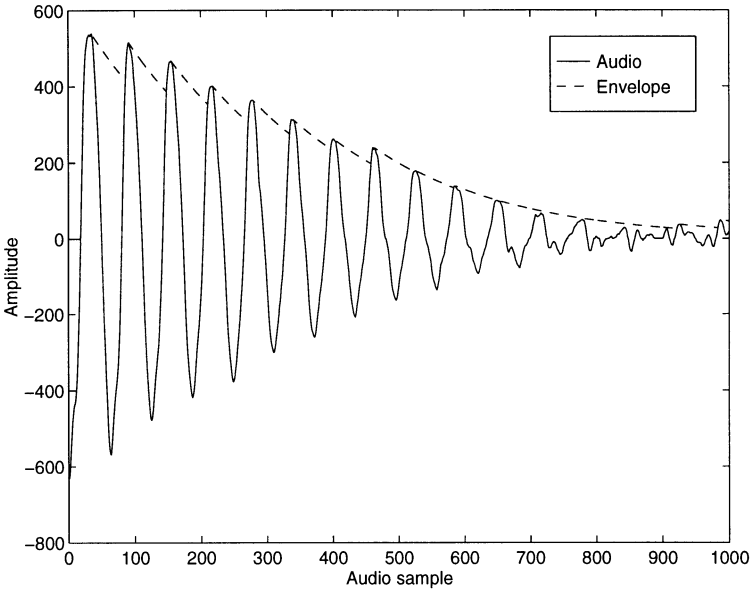

Fig. 5. Audio signal and estimated envelope.

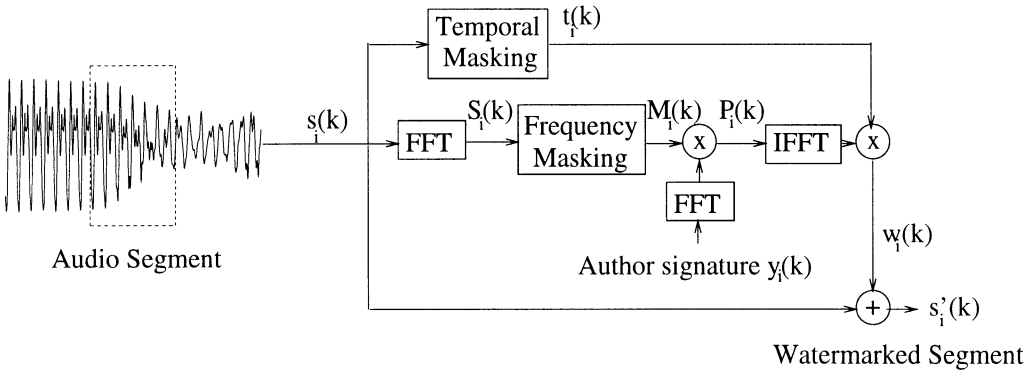

Fig. 6. Diagram of audio watermarking procedure.

Let $(), 0 \leqslant \leqslant N-1$, be N samples of recovered audio piece which may or may not have a watermark. Assume first that we know the exact location of the received signal. Without loss of generality, we will assume that $() = () + (), 0 \leqslant \leqslant N-1$, where $()$ is a disturbance that consists of noise only, or noise and a watermark. The detection scheme relies on the fact that the author or arbitrator has access to, or can compute, the original signal and the two keys $_1$ and $_2$ required to generate the pseudo-random sequence . Therefore, detection of the watermark is accomplished via hypothesis testing. Since $()$ is known, we specifically need to consider the hypothesis test

$$\begin{aligned} &\mathrm{H}_0\colon\ () = () - () = (), \\ &\quad 0 \leqslant \leqslant N-1 \text{ (No watermark)}, \\ &\mathrm{H}_1\colon\ () = () - () = '() + (), \\ &\quad 0 \leqslant \leqslant N-1 \text{ (Watermark)}, \end{aligned} \tag{4}$$

where $'(\ )$ is the potentially modified watermark, and $(\ )$ is noise. The correct hypothesis is estimated by measuring the similarity between the extracted signal $(\ )$ and original watermark $(\ )$:

$$\mathrm{Sim}(\ ,\ ) = \frac{\sum_{j=0}^{N-1} (\ )\ (\ )}{\sum_{j=0}^{N-1} (\ )\ (\ )} \tag{5}$$

and comparing with a threshold T. Note that Eq. (5) implicitly assumes that the noise $(\ )$ is white, Gaussian with a zero mean, even though this assumption may not be true. It also assumes that $(\ )$ has not been modified. These two assumptions do not hold true in most situations. However, our experiments indicate that, in practice, the detection test given in Eq. (5) is very robust (see Section 6). Our experiments also indicate that a threshold $T = 0.15$ yields a high detection performance.

Suppose now that we do know the location of the observed clip $(\ )$. Specifically, suppose that $(\ ) = (\ + \tau) + (\ ), 0 \leqslant \leqslant N-1$, where, as before, $(\ )$ is a disturbance that consists of noise only, or noise and a watermark, and τ is the unknown delay corresponding to the clip. Note that τ is not necessarily an integer. In this case, we need to perform a [23] to determine whether the received signal has been watermarked or not. Once more, we assume that the noise $(\ )$ is white, Gaussian with a zero mean even though this may not be true. This leads us to compare the ratio

$$\frac{\max_\tau \exp(-\sum_{n=0}^{N-1}(\ (\ ) - (\ (\ + \tau) + \ (\ + \tau)))^2}{\max_\tau \exp(-\sum_{n=0}^{N-1}(\ (\ ) - \ (\ + \tau))^2)} \tag{6}$$

with a threshold. If this ratio is higher than the threshold, we would declare the watermark to be present. Note that since τ is not necessarily an integer, computing the numerator and denominator of Eq. (6) requires that we perform interpolation or evaluate these expressions in the Fourier domain using Parseval's theorem.

A generalized likelihood ratio test is also needed if one suspects that the received signal has undergone some other types of modifications, e.g., time-scale changes.

6. Results

We illustrate the inaudible and robust nature of our watermarking scheme on four audio pieces: the beginning of the third movement of the sonata in B flat major D 960 of Schubert (Piano, duration 12.8 s), interpreted by Vladimir Ashkenazy, a castanet piece (Castanet, duration 8.2 s), a clarinet piece (Clarinet, duration 18.6 s), and a segment of 'Tom's Diner', an song by Suzanne Vega (Vega, duration 9.3 s). All of the signals are sampled at 44.1 kHz. The Castanets signal is one of the signals prone to pre-echoes. The signal Vega is significant because it contains noticeable periods of silence.

A plot of a short portion (0.5 s) of the original clarinet signal is shown in Fig. 7a. The corresponding signal with the embedded watermark is shown in Fig. 7b. The watermark is displayed in Fig. 7c. Observe that the envelope of the watermark changes over time with the signal. In particular, the magnitude increases in more powerful regions and decreases in quiet portions.

We test the robustness of the audio watermarking procedure to several degradations and distortions, including those that result from colored noise, MPEG coding, multiple watermarks, and resampling. The robustness of our water-marking approach is measured by the ability to , i.e., high probability of detection. Robustness is further based on the ability of the algorithm , i.e., low probability of false alarm. For a given distortion, the overall performance may be ascertained by the relative difference between the similarity when a watermark is present (hypothesis H_1) and the similarity when a watermark is not present (hypothesis H_0). In each robustness experiment, similarity results were obtained for both hypotheses. In particular, the degradation was applied to the audio when a watermark was present. It was also applied to the audio when a watermark was present. The similarity was computed between the original watermark and the recovered signal (which may or may not have a watermark). A large similarity indicates the presence of a watermark (H_1), while a low similarity suggests the lack of a watermark (H_0).

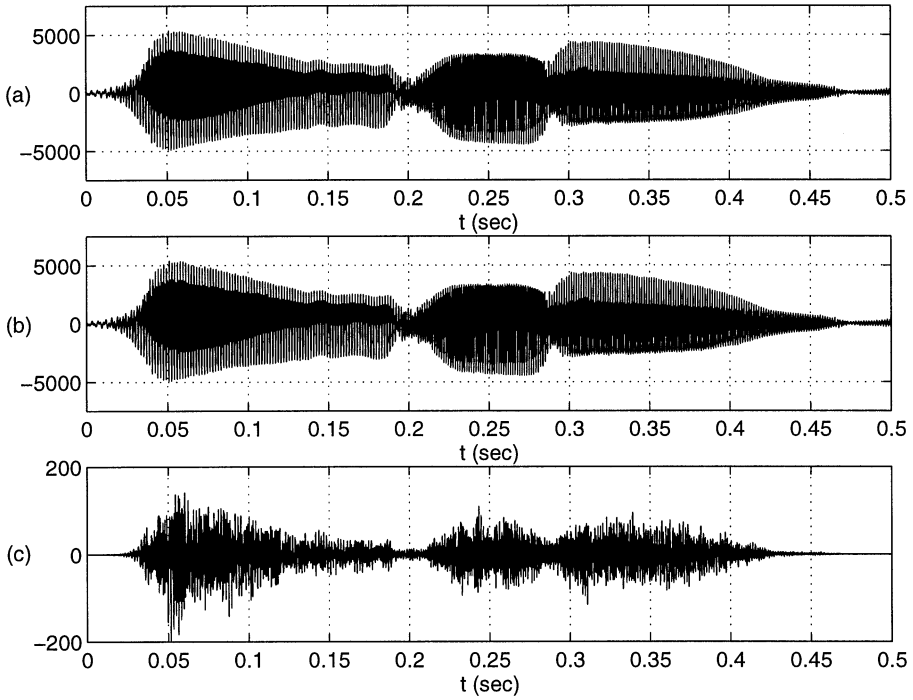

Fig. 7. A portion of the (a) original Clarinet signal, (b) watermarked Clarinet signal, and (c) corresponding watermark.

Similarity is computed on blocks of 100 consecutive 512 sample segments. Note that this corresponds to 1.16 s of audio at the 44.1 kHz sampling rate. For example, the duration of the Castanet signal is 8.2 s. A total of seven watermark detections are computed, each on 1.16 s of data. Smaller and larger blocks are easily handled.

6.1. Audio fidelity

The quality of the watermarked signals was evaluated through listening tests. In the test, the listener was presented with the original signal and the watermarked signal and reported as to whether any differences could be detected between the two signals. Eight people of varying backgrounds were involved in the listening tests. One of the listeners has the ability to perceive absolute pitch and two of the listeners have some background in music. In all four test signals, the watermark introduced no audible distortion. No pre-echoes were detected in the watermarked Castanet signal. The quiet portions of Vega were similarly unaffected. The results of the test are displayed in Table 1.

6.2. Additive colored noise

To model perceptual coding techniques and other watermarks, we corrupted the watermark with

. Noise which has the same spectral characteristics as the masking threshold provides an approximation of the worst possible additive distortion to the watermark. The additive colored noise is generated in a similar way as the watermark. Specifically, a Gaussian white noise sequence is shaped by the frequency and temporal masks. The shaped noise is then added to the audio signal. The noise level is chosen to be

Table 1
Blind testing of watermarked audio

Test audio	Original preferred to watermarked (%)
Castanets	50.33
Clarinet	49.00
Piano	49.67
Vega	48.00

barely audible. As a result, it is a good approximation of the maximum noise that we can add before strong degradations. Note that the colored noise, as constructed, is almost identical to a - interfering with the watermark we are attempting to detect. The additive colored noise test was run 1000 times for each signal, with a different noise sequence generated each time.

The similarity values obtained during testing indicate easy discrimination between the two hypotheses as shown in Fig. 8. The upper similarity curve in each plot corresponds to each of the test pieces with a watermark. The lower similarity curve correspond to each audio piece without a watermark. The error bars around each similarity value indicate the and similarity values over the 1000 runs. The -axis corresponds to block number, i.e., block number 1 consists of the first 100 audio segments. As each audio segment is of length 512 samples, this corresponds to 51 200 samples,

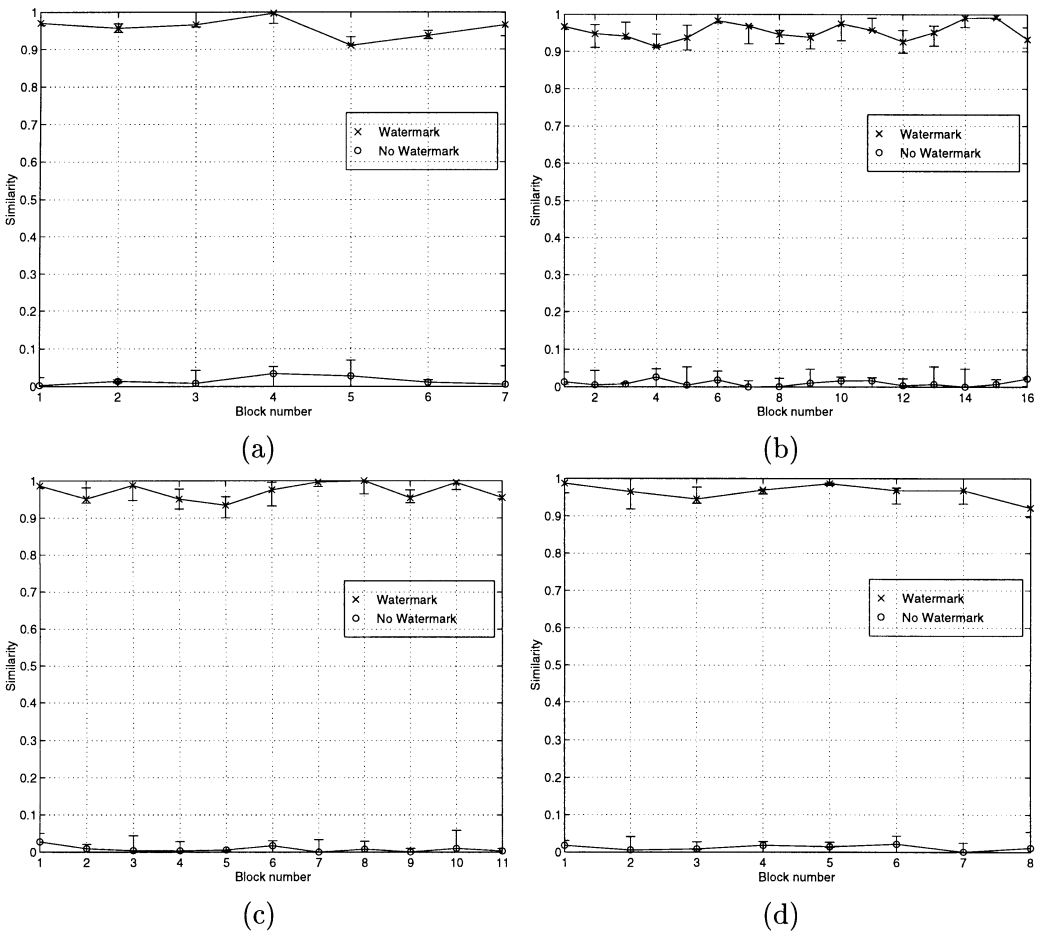

Fig. 8. Detection of watermarks is colored noise (a) Castanet, (b) Clarinet, (c) Piano, and (d) Vega. The error bars around each similarity value indicate the maximum and minimum similarity values over the 1000 runs.

i.e., 1.16 s of audio. For example, in Fig. 8a, the similarity values for block number 2 are measured over the Castanet signal from $= 1.16$ s to $= 2.32$ s. The similarity values vary over time for each test signal. This is to be expected, as power of the watermark varies temporally with the power of the host signal. Observe that the upper curve for each audio piece is widely separated from the lower curve over the entire duration of the signal. Selecting a decision threshold T anywhere in the range of approximately $0.1 \leqslant T \leqslant 0.9$ guarantees a correct hypothesis decision for the four test signals in colored noise.

6.3. Cropping and filtering

Robustness to cropping and filtering was tested. Frequently, filtering operations are performed on audio to enhance certain spectral components. Initially, five short pieces (0.1 s) were from the test signals. The cropped segments were signal (i.e., non-noise) components. We added colored noise to the cropped segments and then applied a 15-tap low-pass filter with a cutoff frequency equal to $\frac{1}{8}$ the Nyquist frequency of the signals. The test was repeated 1000 times by repeatedly generating new colored noise. During detection,

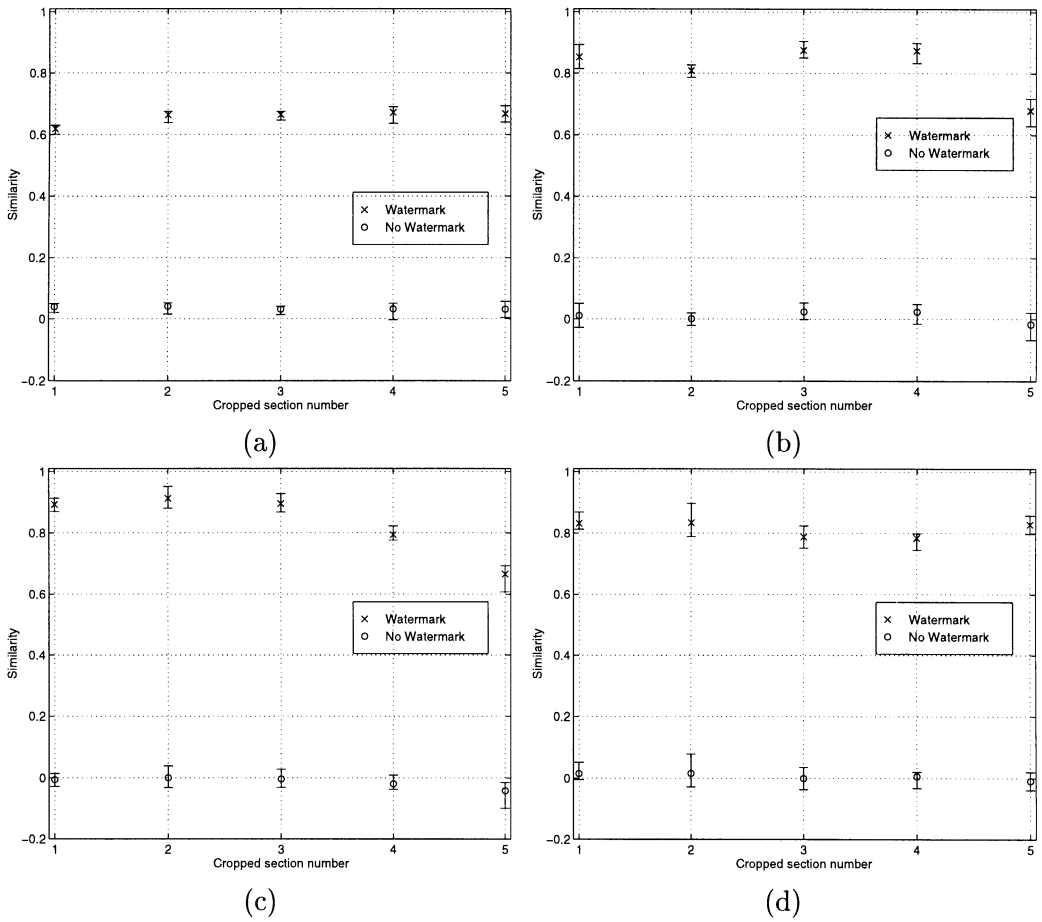

Fig. 9. Detection of watermarks after cropping and lowpass filtering (a) Castanet, (b) Clarinet, (c) Piano, and (d) Vega. The error bars around each similarity value indicate the maximum and minimum similarity values over the 1000 runs.

the GLRT described by Eq. (6) was employed to estimate the location of the crop. Detection results are presented in Fig. 9. For each test signal, a similarity with and without watermark is shown for the five cropped segments. The error bars indicate the maximum and minimum similarity over 1000 colored noise tests. The similarities of the watermarked segments are much larger than the non-watermarked segments.

6.4. MPEG coding

In many multimedia applications involving storage and transmission of digital audio, a lossy coding operation is performed to reduce bit-rates and increase efficiency. To test the robustness of our watermarking approach to coding, we added colored noise (cf. Section 6.2) to several watermarked and non-watermarked audio pieces and MPEG coded the result. The noise was almost inaudible and was generated using the technique described above. We then attempted to detect the presence of the watermark in the decoded signals.

The coding/decoding was performed using a software implementation of the ISO/MPEG-1 Audio Layer II coder with several different bit rates: 64, 96 and 128 kbits/s. The original and watermarked Castanets audio track for 1000 samples near $= 3.0$ s is shown in Fig. 10a and b. In Fig. 10d, the signal MPEG coded at 96 kbits/s is displayed. The coding error shown in Fig. 10e, which is defined as the difference between the watermarked signal Fig. 10 b and the coded signal Fig. 10 d, is on the order of 10 times greater than the watermark shown in Fig. 10c! The results of the detection tests are plotted in Fig. 11. Although the errors produced by the coders are much greater than the embedded watermarks, the plots indicate easy discrimination between the two cases. A threshold chosen in the range of 0.15–0.50 produces no detection errors.

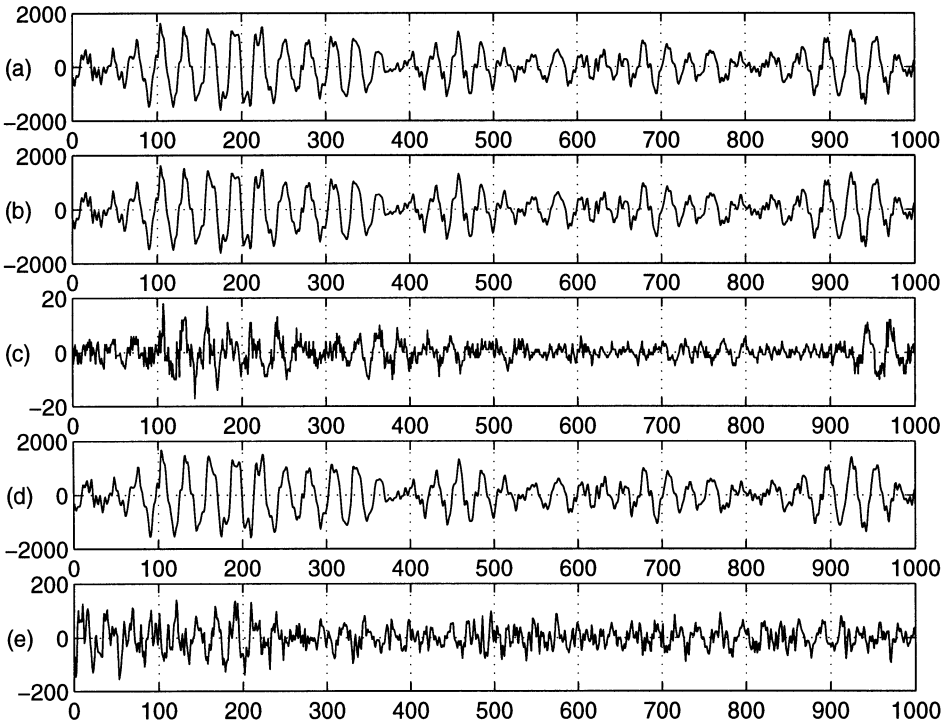

Fig. 10. Portions of Castanet signal (a) original, (b) watermarked, (c) watermark, (d) MPEG coded 96 kbits/s, and (e) coding error.

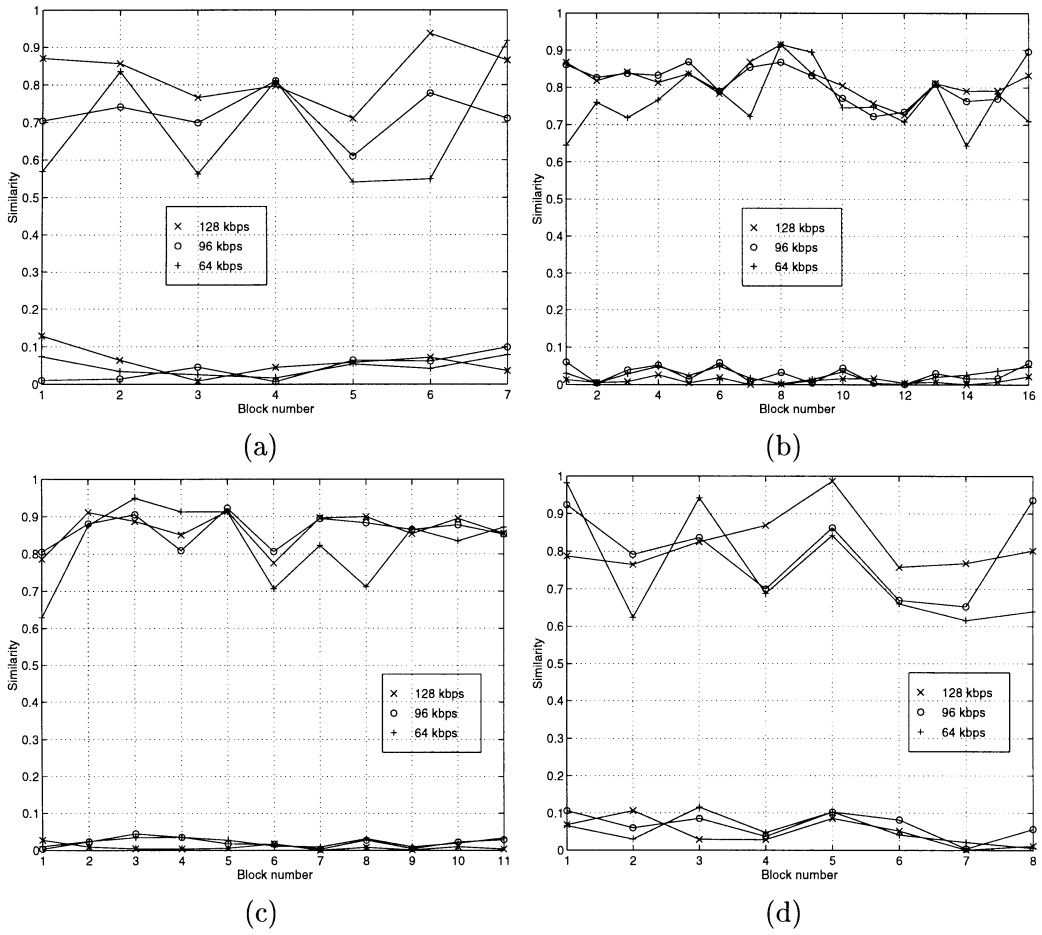

Fig. 11. Detection of watermark after MPEG coding for (a) Castanet, (b) Clarinet, (c) Piano, and (d) Vega.

6.5. Multiple watermarks

Experiments were performed to obtain results for detecting watermarks in the presence of other watermarks. In particular, the audio clips were embedded with three consecutive watermarks, and then corrupted by colored noise and MPEG coded at 128 kbits/s. As indicated in Section 6.2, where the colored noise was created using the HAS masking models, additional watermarks pose no threat to each other. The results for detecting the three watermarks are shown in Fig. 12. Again, an audio signal with a watermark is easily discriminated from an audio signal lacking a watermark.

6.6. Temporal resampling

Our experiments also indicate that the proposed watermarking scheme is robust to signal resampling. The resampled signal is obtained by oversampling by a factor 2 and then down sampling by a factor 2 by extracting the samples. The results of detection after signal resampling are shown in Fig. 13. Although a lot of damage has been

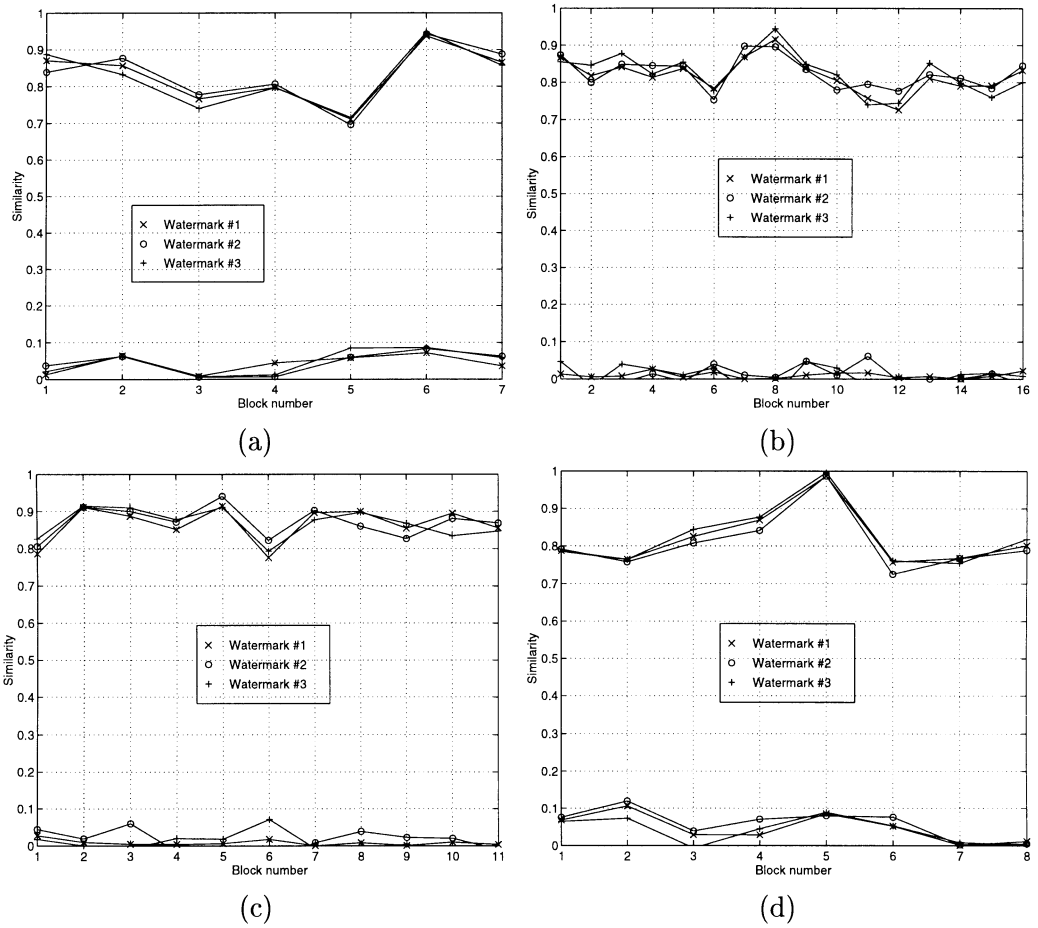

Fig. 12. Detection of three watermarks after colored noise and MPEG coding at 128 kbits/s (a) Castanet, (b) Clarinet, (c) Piano, and (d) Vega.

introduced in the host audio data, the watermarks are readily extracted.

7. Conclusion

We presented a watermarking procedure to embed copyright protection into digital audio by directly modifying the audio samples. The watermarking technique directly exploits the masking phenomena of the human auditory system to guarantee that the embedded watermark is imperceptible. The owner of the digital audio piece is represented by a pseudo-random sequence defined in terms of two secret keys. One key is the owner's personal identification. The other key is calculated directly from the original audio piece. The signal dependent watermarking procedure shapes the noise-like author representation according to the temporal and frequency masking effects of the host signal. The embedded watermark is inaudible and statistically undetectable. We also introduce the notion of a dual watermark. We show that the dual watermarking approach together with the procedure that we use to derive the watermarks effectively solves the deadlock problem.

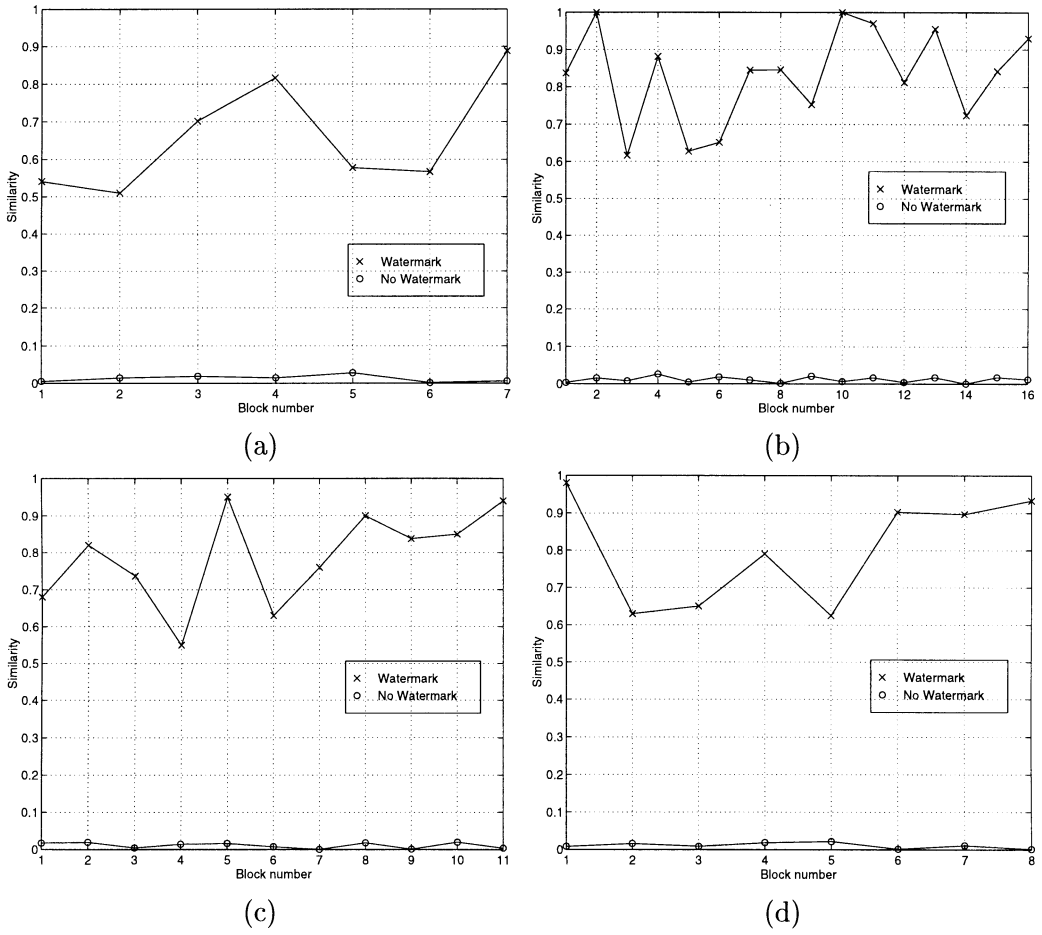

Fig. 13. Detection of watermark after resampling (a) Castanet, (b) Clarinet, (c) Piano, and (d) Vega.

Several tests have shown the robustness of the watermarking procedure to several audio degradations, including colored noise, MPEG coding, multiple watermarks, and temporal resampling. The watermark was readily detected in the experiments on short duration (1.16 s) segments of the audio signals.

References

[1] W. Bender, D. Gruhl, N. Morimoto, Techniques for data hiding, Tech. Rep., MIT Media Lab, 1994.

[2] O. Bruyndonckx, J.-J. Quisquater, B. Macq, Spatial method for copyright labeling of digital images, Proc. 1995 IEEE Nonlinear Signal Processing Workshop, Thessaloniki, Greece, 1995, pp. 456 459.

[3] I. Cox, J. Kilian, T. Leighton, T. Shamoon, Secure spread spectrum watermarking for multimedia, Tech. Rep. 95-10, NEC Research Institute, 1995.

[4] S. Craver, N. Memon, B.-L. Yeo, M. Yeung, Can invisible watermarks resolve rightful ownerships? IBM Research Technical Report RC 20509, IBM Cyber Journal, July 1996.

[5] S. Craver, N. Memon, B.-L. Yeo, M. Yeung, Resolving rightful ownerships with invisible watermarking techniques: limitations, attacks, and implications, IBM Research Technical Report RC 20755, IBM Cyber Journal, March 1997.

[6] S. Goldwasser, M. Bellare, Lecture notes on cryptography, preprint, July 1996.

[7] F. Hartung, B. Girod, Digital watermarking of raw and compressed video, Proc. SPIE Dig. Comp. Tech. and

Systems for Video Comm., Vol. 2952, October 1996, pp. 205–213.
[8] ISO/CEI, Codage de l'image animee et du son associe pour les supports de stockage numerique jusqu'a environ 1,5 mbit/s, Tech. Rep. 11172, ISO/CEI, 1993.
[9] J. Johnston, K. Brandenburg, Wideband coding-perceptual considerations for speech and music, in: S. Furui, M. Sondhi (Eds.), Advances in Speech Signal Processing, Dekker, New York, 1992.
[10] K. Matsui, K. Tanaka, Video steganography: how to secretly embed a signature in a picture, IMA Intellectual Property Project Proc., Vol. 1, 1994, pp. 187–206.
[11] N. Moreau, Techniques de Compression des Signaux, Masson, Paris, 1995.
[12] National Institute of Standards and Technology (NIST), Secure Hash Standard, NIST FIPS Pub. 180-1, April 1995.
[13] P. Noll, Wideband speech and audio coding, IEEE Commun. Vol. 31 (11) (November 1993) 34–44.
[14] I. Pitas, A method for signature casting on digital images, Proc. 1996 Int. Conf. on Image Proc., Vol. III, Lausanne, Switzerland, 1996, pp. 215–218.
[15] I. Pitas, T. Kaskalis, Applying signatures on digital images, Proc. 1995 IEEE Nonlinear Signal Processing Workshop, Thessaloniki, Greece, 1995, pp. 460–463.
[16] R. Rivest, Cryptography, in: J. van Leeuwen (Ed.), Handbook of Theoretical Computer Science, Vol. 1, Ch. 13, MIT Press, Cambridge, MA, 1990, pp. 717–755.
[17] R. Rivest, The MD4 message digest algorithm, Advances in Cryptology, CRYPTO 92, Springer, Berlin, 1991, pp. 303–311.
[18] J.J.K.O. Ruanaidh, W.J. Dowling, F.M. Boland, Phase watermarking of digital images, Proc. 1996 Int. Conf. on Image Proc., Vol. III, Lausanne, Switzerland, 1996, pp. 239–242.
[19] M.D. Swanson, D. Zhu, A.H. Tewfik, Transparent robust image watermarking, Proc. 1996 Int. Conf. on Image Proc., Vol. III, Lausanne, Switzerland, 1996, pp. 211–214.
[20] M.D. Swanson, B. Zhu, A. Tewfik, Multiresolution video watermarking using perceptual models and scene segmentation, to appear IEEE J. Selected Areas Commun. June 1998.
[21] J.F. Tilki, A.A. Beex, Encoding a hidden digital signature onto an audio signal using psychoacoustic masking, Proc. 1996 7th Int. Conf. on Sig. Proc. Apps. and Tech., Boston, MA, 1996, pp. 476–480.
[22] R.G. van Schyndel, A.Z. Tirkel, C.F. Osborne, A digital watermark, Proc. 1994 IEEE Int. Conf. on Image Proc., Vol. II, Austin, TX, 1994, pp. 86–90.
[23] H.L. Van Trees, Detection, Estimation, and Modulation Theory, Vol. 1, Wiley, New York, 1968.
[24] R. Wolfgang, E. Delp, A watermark for digital images, Proc. 1996 Int. Conf. on Image Proc., Vol. III, Lausanne, Switzerland, 1996, 219–222.