Robust Audio Watermarking

Based on Secure Spread Spectrum and Auditory Perception Model

PETAR HORVATIC, JIAN ZHAO, NIELS J. THORWIRTH
Fraunhofer Center for Research in Computer Graphics, Inc.
321 South Main Street
Providence, RI 02903, U.S.A.
Email: {phorvati, jzhao, nthorwir}@crcg.edu

Key words: Digital Watermarking, Auditory Perception Model, MP3, Spread Spectrum Communications

Abstract: The impetus for the work presented in this paper arose from the need to provide copyright protection for digital audio, including CD-quality music and DVD, under heavy compression rates, conventional audio processing manipulations, and transmission through noisy media. Our approach is established on modelling audio watermark as information transmitted through a probabilistic communication channel represented by digital music. Watermark information is embedded within significant portions of the audio streams making it tightly coupled with audio content and able to sustain even the most severe attacks intended to remove the hidden data. This paper presents how highly transparent, robust and secure watermark information can be embedded into digital audio streams using the model for acquiring significant audio components in conjunction with conventional spread spectrum data hiding techniques.

1. INTRODUCTION

The transition of audio technology from the analogue to the digital domain, combined with rapidly improving hardware resources for handling digital data, have enabled easy illegal copying and manipulation of audio files, especially the ones sent over the Internet. This poses a serious challenge to the security and copyright protection of IP-based audio distribution or even stored CD-quality digital music. Conventional secure audio techniques, which protect audio data using cryptographic algorithms,

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35515-3_53

S. Qing et al. (eds.), *Information Security for Global Information Infrastructures*

Table 1. Effective Watermark Features

Content-based embedding. The watermark should be embedded within the digital content itself, as opposed to the file header.
Transparency. The watermark should be transparent to users.
Robustness. The watermark should resist common signal processing attacks such as D/A-A/D, filtering, re-sampling, cropping, noise addition, as well as lossy compression and transmission of data through various mediums.
Security. The watermark should be embedded using a secret key, possessed by the user.

are impractical as they only permit the key holders to access audio data. Moreover, after decryption of the audio stream, data becomes vulnerable to piracy and unauthorised reproduction. For this reason, hiding secret information within the content of digital material can provide a meaningful way for giving authors proof of ownership, enforcing copyright protection, authenticating digital content and preventing tampering. Several key features used to describe an effective watermark are outlined in table 1.

2. RELATED WORK

Several previous watermarking methods applied to multimedia including audio have been proposed. Cox, et al. (Cox, et al. 1996) proposes embedding the watermark information in the frequency domain of select samples, based on the perceptual entropy of the image or audio material. Watermark information X is represented using an independently selected random number sequence $[x_1\ x_2\ \ldots x_n]$ within the range [0,1] with uniform distribution and reasonable but finite precision. The embedding function can be represented as:

$$v'_i = v_i*(1+\Delta_i*x_i) \quad (1)$$

where v'_i represents a sample of multimedia content with an embedded watermark, v_i represents the original sample and x_i is a single watermark sample embedded using scaling parameter Δ_i. The scaling parameter can be a constant or a variable value that is determined based on the error introduced by the watermark insertion. While the idea of embedding information using perceptually determined significant portions of multimedia content ensures watermark robustness against the majority of sophisticated attacks, the problem lies in determining a set of scaling parameters Δ_i. This can be a complicated and time-consuming task that should be based on the perceptual properties of each individual sample and should vary with each frame of music. Instead of formulating the embedding function based on Equation 1, we propose to spread a narrow band watermark over the range of significant frequencies using combination of Fourier transforms and scrambling functions. To limit the maximum

amount of watermark information added to each frequency constituent, we also propose a perceptual masking filter.

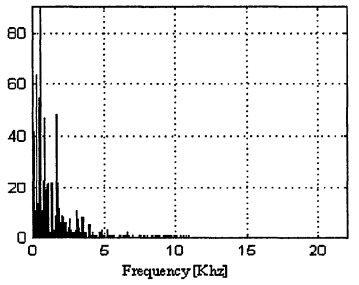

Figure 1. Frequency Spectrum for an Arbitrary Frame of Audio Produced by Perceptual Masking Model

Boney, et al. (1996) suggest embedding a watermark into significant portions of audio files by generating a PN sequence of binary numbers and filtering it using a 10^{th} order all-pole approximation of the MPEG-1, Layer-1 (MP1) psychoacoustic masking model. Prior to embedding, the filtered PN sequence is weighed in time domain to prevent pre-echoes. Pre-echoes occur when instantaneous energy present in the stream suddenly changes within 1152 consecutive samples (MPEG-1 standard) of a single audio frame. If there is a sudden temporal change of energy within this group of samples, and the watermark intended for spreading over high energy portion of a frame is actually spread over the entire frame, including the weak energy samples, the embedding process introduces audible distortions referred to as the pre-echoes. Weighing the watermark in time domainprior to embedding ensures that watermark energy is scaled according to sudden temporal changes within a single frame, preventing the audible distortions. PN sequences, used by Boney, et al. are very popular in spread spectrum communication due to their excellent auto correlation properties, resistance to interference, and noise-like characteristics. Using PN sequences to represent bits of information ensures transparency, good synchronisation properties, and resistance to attacks. While most of the embedding properties used by Boney, et al. (1996) resemble those of a transparent, robust and secure watermark model, the process of embedding the watermark into the entire audio frame is not very efficient. As it will be shown shortly, after the MPEG1/Layer-3 perceptual scaling of audio, most of the energy patterns within the frame resemble that of figure 1. Higher frequencies, which occupy the right-hand portion, carry essentially no energy, and have no redundancy. Watermark spreading should exclusively take place along the redundant portions of figure 1. Instead of embedding the watermark in to entire audio frame, our method selects a group of frame samples, representing the majority of audio energy, and uses those as candidates for watermark embedding. Candidates are selected based on power spectral analysis. Selecting fewer samples as candidates and using the perceptual masking filter, created directly from the candidates, reduces the possibility of pre-echoes to extent where satisfactory results have been achieved without temporal weighing of the watermark. Furthermore, our

watermark spectrum is shaped using a finer, more complex and perceptually transparent psychoacoustic model associated with MP3 compression.

In the following section of our study, we present the idea behind the masking phenomenon of the MP3 psychoacoustic model and relate the properties of the perceptually generated masking filter to our audio watermark transparency.

3. PSYCHOACOUSTIC MODEL

As mentioned earlier, one of the essential components used to model the audio transmission channel is the existing psychoacoustic perceptual masking model. The MPEG/Audio psychoacoustic model described well by Davis Pan (1996) uses frequency and temporal audio characteristics to effectively discard perceptually irrelevant portions of the audio frame. Most of the analysis is performed on the frequency-based frame of raw audio samples.

The model takes advantage of HAS property, which suggests that presence of a strong audio tone makes weaker tones within its spectral or temporal surrounding inaudible. On the other hand, HAS has a limited frequency resolution that is characterised by critical bands, which are a set of neighbouring regions within the human audible frequency range where HAS has uniform audibility and masking properties. The critical bands span the widths less than 100 Hz for the low frequency portions and more than 4KHz for the high-end groups. The MPEG/Audio masking model divides the spectrum of the audio frame into sub-band regions resembling those of the HAS. Figure 2 compares the HAS critical bands to those used by the MP3 psychoacoustic model. As shown, the uniform, 32 sub-band distribution does not completely resemble the actual audibility properties of the HAS. Computational efficiency requires perceptual model to have uniform sub-band regions and the model compensates well for this inconsistency as described by Davis Pan (1996). Treating each region independently and with the uniform auditory properties, the amount of available noise masking can be calculated as a function of frequency. Figure 3 illustrates a typical scenario that often occurs within the sub-band analysis window. A strong tone masks weaker spectral neighbouring components that could have noise-like or tone-like properties.

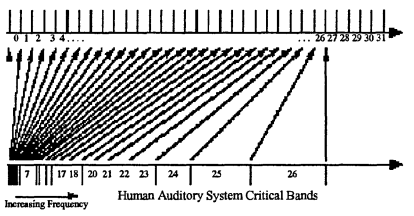

Figure 2. MPEG/Audio Frequency Bands

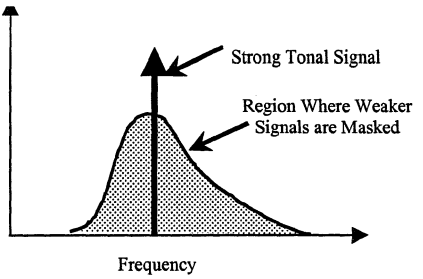

Figure 3. Masking Phenomena

The example shown here would be considered an extreme case of tonal masking. After calculating the masking thresholds, the model calculates the number of bits necessary to represent the sub-band in order to make quantization noise inaudible. Good compression models quantize the sub-bands with no more bits than necessary to keep the noise below the threshold shown in figure 3. The final result produced by the psychoacoustic model is a set of signal-to-mask ratios (SMRs) expressed in decibels (dB) and calculated as a difference between the maximum signal level and the minimum masked threshold level. SMRs are calculated for each of the 32 sub-bands used to model HAS critical bands.

Using the perceptual component extraction method in conjunction with content-based information hiding raises an important issue. Most watermarking schemes embed information as noise (Zhao & Koch 1998). In order to make the embedded watermark robust to MPEG and other perceptual lossy compression schemes, it must exceed the minimal sub-band masking threshold of the audio psychoacoustic model shown in figure 3. On contrary, the information must be embedded in such a way that it does not introduce temporal pre-echoes or similar audible distortions thus, it must be kept below the audible threshold. Our robust and transparent approach to solving this essential aspect of watermark embedding technology is outlined in section that follows.

4. WATERMARK DESIGN

The watermark embedding process is implemented by combining conventional signal processing and communication techniques, and in accordance with the MP3 encoding and decoding process. Prior to embedding, watermark information is exposed to a series of ECC (error correction coding) algorithms in order to enhance detection reliability. Existing SysCoP (Zhao & Koch 1996) ECC functions are used to encode each byte of watermark information into a larger bit stream capable of correcting an arbitrary number of errors that often occur during standard audio processing or watermark attacks. This error correction capability and the degree of reliably embedded information are part of a flexible, user-driven trade-off. As an example, consider (28,8,5) BCH (Bose, et al. 1960; Hocquenghem 1959) code used to ECC-encode an 8-byte watermark. Each byte of data is encoded into a stream of 28 bits, capable of correcting up to 5

bit-errors (18% error correction). Although for a given set of input parameters, the average bit-error contained in the stream can be held below the intended fixed value, 18% for our example, it is never the case that error distribution is uniform throughout received bit words.

When subjected to audio processing and watermark attacks, digital audio stream (modelled as a non-stationary, probabilistic transmission medium) is bound to produce burst errors, a group of several consecutive toggled bits within detected information. For a single encoded word, the total number of errors produced by a burst can easily exceed the allowed maximum, causing false decision-making during the ECC decoding process and producing single character errors in the recovered watermark. In order to enhance watermark detection performance, our algorithm spreads burst errors uniformly throughout the received bit stream, thus keeping the number of toggled bits in each received word less than the maximum allowed, a value that is dictated by error correction capability.

In the example using the (28,8,5) BCH code, each of the eight 28-bit words is combined into a single, evenly interleaved stream. In this manner, burst errors affect all encoded words equivalently, thus producing a uniform error distribution and achieving the required average of fewer than 5 bit-errors per word (18% bit-error). For cases where higher error correction capability is desired, BCH code can be modified to use longer code words, resulting in error correction performance greater than 18%. The price one pays for using higher error correction capability is a reduced capacity for information embedding. However, by keeping the trade-off among embedding features flexible, our audio watermarking technology allows the user to optimise error correction capability while retaining the maximum information desired.

As mentioned earlier, an ECC encoded watermark is embedded in accordance with the MP3 encoding algorithm. During the process, illustrated in figure 4, significant audio portions are extracted based on the ISO psychoacoustic masking model (ISO/IEC 11172-3 1993) described earlier. Statistically determined SMRs are used to reduce the raw audio frame, shown in figure 4-a, into a far less redundant, yet perceptually firm set of quantized frequency samples (QFS). From the resulting output, a group of 64 candidates, shown in figure 4-b, is selected based on power spectral analysis. The group of 64 candidates represents the most significant of the frequency components and will be used for watermark embedding. The notion of spreading a narrow band signal (watermark) over a much larger range of frequencies (QFS) is implemented by combining the properties of the time and frequency domain transforms with a random sequence scrambling function. Selected candidates are converted into time samples via IDCT transform, shown in figure 4-c. Subsequently, their order

is scrambled according to a key-based random number sequence, as seen in figure 4-d. The time-based scrambled coefficients are then converted back into the frequency domain, producing a noise-like (zero-mean, constant variance) signal with ideal embedding properties. A single bit of watermark information is embedded into the random signal by scaling one of its samples to a larger value, as shown in figure 4-e. The scaling factor corresponds to the robustness parameter, where the value of 1 represents no information embedded and larger values represent higher watermark robustness, with degraded audio quality. Upon modifying the selected sample, the random signal is converted back into time domain, as seen in figure 4-f, in order to descramble the order of time coefficients and spread

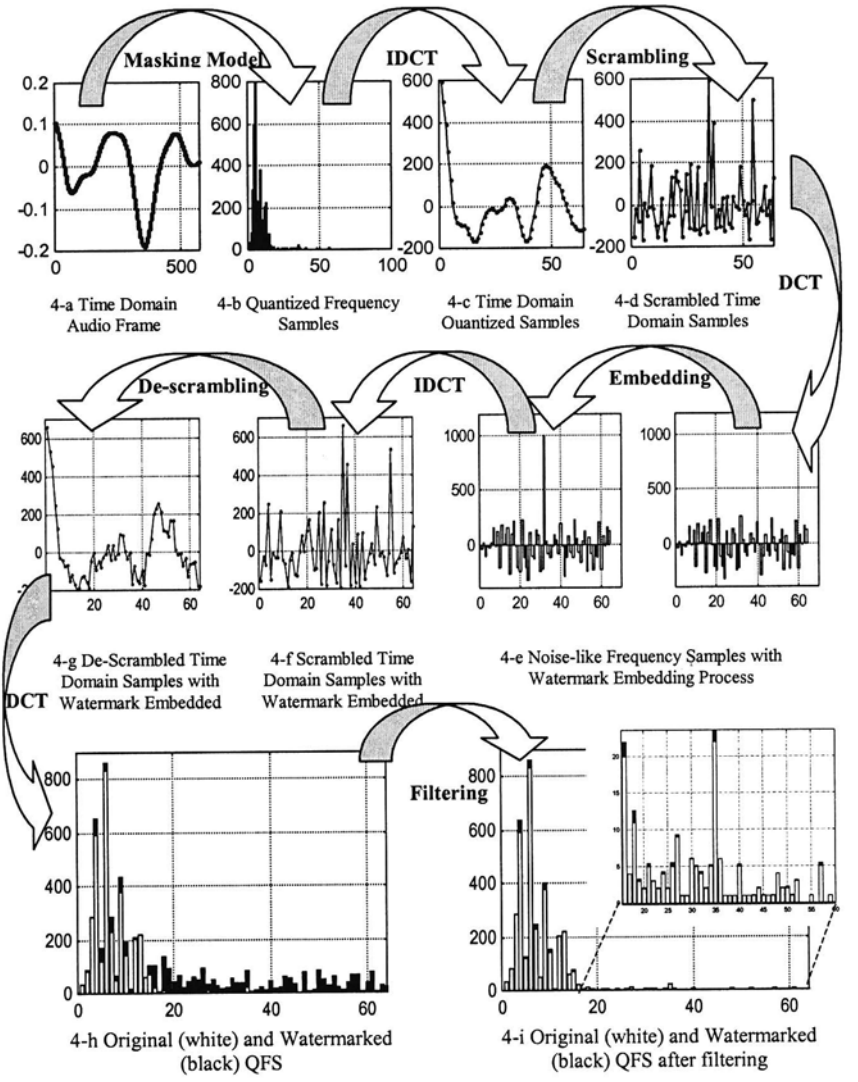

Figure 4. Watermark Embedding Process

the embedded watermark bit evenly throughout the wider band of frequencies, illustrated in figure 4-g. Finally, watermarked samples are brought back into the original QFS format using DCT, shown in the dark-coloured bar plot in figure 4-h. The difference between a newly created group of QFS (black) and the corresponding original (white) is the error resulting from watermark embedding. As shown in figure 4-h, the error has a uniform spectral distribution. The perceptual model used to extract significant components suggests that such uniform error distribution between the original and watermarked QFS will manifest itself in possible degraded audio quality since both weak and strong frequency components of the original signal are changed by the same amount. In order to prevent such consequence, watermark spectrum is filtered using the normalised version of the original QFS as a filter. The resulting plot of watermarked QFS is shown in figure 4-i. The inset chart shows the weaker magnitude samples. As shown in figure 4-i, this process shapes the error distribution according to the perceptual model and ensures that weaker frequency constituents are changed by a smaller amount, thus resulting in improved watermark transparency. Although filtering the watermark spectrum makes the hidden information weaker, users can compensate for this drawback by using a larger robustness parameter.

Watermark detection is the reverse process of spreading the watermark. First, selected group of quantized frequency candidates used for embedding the watermark is obtained. They are represented by Huffman-decoded values when watermarked audio is in MP3 form, or samples resulting from the quantizing process shown in figure 4-a, if the watermarked audio is in its raw format. Second, the quantized samples are transformed to the time domain and scrambled using the same scrambling key that was employed during watermark embedding. In this manner spread watermark energy will be coherently collected in to single frequency constituent. Finally, time samples are converted back in to frequency domain and watermark information is recovered by analysing spectral properties of produced signal. Presence of strong frequency constituent within the noise-like components indicates presence of a single watermark bit.

Along with watermark information, a key-based random number synchronisation stream is embedded as a tag-mark between consecutive watermarks. Using bit-wise correlation function, our watermark detector compares its output to a key-based synchronisation word created by the user. Once a complete synchronisation stream is obtained, watermark detection is initiated. Detected watermark bits are ECC decoded into characters. Each set of tag-marks uniquely identifies the beginning and end points of each watermark, providing reliable information recovery in case of audio stream truncation. During watermark extraction, the watermark's length is detected

first by recovering two consecutive synchronisation codes to support variable watermark length.

As the introduction briefly mentions, watermark should be embedded using a secret key, possessed by the user. Based on such key, 128 bits in length, our model uses MD5 based random number generator for producing sequence of 64 unique integers with values chosen in the range of [0,63]. These numbers are used as indices for scrambling the order of QFS as described earlier in this section. Probability of guessing the user key is contingent on breaking the 128 bit key making our watermark security as strong as the most sophisticated modern security algorithms. In order to secure the watermark against sophisticated attacks that compare consecutive audio frames in an attempt to recover the random indexing sequence, we propose using robust features unique to each audio frame for frame-based, dynamic generation of watermark key.

5. SUMMURY AND RESULTS

Table 2. Test Files

test 1	A. Vivaldi, "La Foglia", (7min, 40sec)
test 2	Sam Estes, "Romeo & Juliet-Dance Scene", (4min, 51sec)
test 3	Selected Piano Piece, (2min, 48sec)
test 4	J.S Bach, "Prelude 1", (5min, 2sec)

We selected four pieces of classical music from a popular mp3 site (http://www.mp3.com) in testing performance of our watermark technology. They were used in their CD quality form with 16-bit 44.1KHz sample format. The selected test files are shown in table 2. For all tests, watermark detection is achieved without comparison of watermarked and original audio files. Only those attacks that do not considerably affect the audio quality are taken into consideration. **Ne** is the number of watermark bytes successfully embedded before attack, and **Nd** is the number of watermark bytes present after attack. $\mathbf{P_{det}}$ is the probability with which a single watermark byte survives the attack ($\mathbf{P_{det}}$ = **Nd/Ne**). $\mathbf{P_{err}}$ is the probability that a watermark byte is incorrectly detected.

Table 3 shows the results of our study obtained when the watermarked audio files have been processed with several common signal-processing routines resembling those used on today's market. The first section of table 3

Table 3. Common Processing Routines

MP3 codec	N_d	N_e	P_{det}	P_{err}
test 1	167	161	.96	0
test 2	272	248	.91	0
test 3	104	104	1	0
test 4	77	70	.91	0
Re-sampling				
test 1	161	120	.75	0
test 2	248	232	.94	0
test 3	65	62	.95	0
test 4	88	88	1	0
Filtering				
test 1	168	164	.98	0
test 2	272	256	.94	0
test 3	104	104	1	0
test 4	100	94	.94	0.01

indicate high resistance of our watermark to lossy compression including MP3. This finding is in close agreement with theory suggested in this paper. Our content-based watermarking technology heavily relies on ISO/MPEG perceptual masking model, which in turn makes our watermarks immune to lossy compression. Remaining portions of table 3 indicate that re-sampling to 22 KHz or low-pass filtering to 5.5 KHz, do not effect the embedded information. The watermark contained within the spectral band of significant audio components is robust to all such re-sampling or filtering attacks that do not severely affect audio quality.

Although not shown here, it is important to note that for all signal-processing routines and attacks, the probability of false alarm is zero. This is the key feature enabling secure watermark embedding. It suggests that using any incorrect key in an attempt to recover the watermark will result in no information retrieved. On the other hand, no valid watermark will be detected from any unwatermarked audio stream.

Our method presented here extends the previous methods applied to audio and image watermarking. Our approach exploits the properties of time-frequency mapping in combination with random number sequencing and the perceptual masking model. The novelties of our approach are summarised in table 4.

Table 2. Novelties of Our Approach

The watermark is embedded in the significant portions of audio signals, which are extracted based on the psychoacoustic model..
Scrambling the order of the audio samples in time domain before embedding the watermark enhances the security and transparency of embedded information.
Optimally selected and perceptually firm set of normalised frequency samples used as a filter for shaping the watermark spectrum effectively limits the total energy embedded, thus improving watermark transparency.
Embedding is achieved in software in real-time, and any MP3 or other lossy audio codecs can be directly watermarked without conversion to raw audio.

6. REFERENCES

Zhao, J. and Koch, E. (1995). Embedding Robust Labels into Images for Copyright Protection. Proc. of the International Congress on Intellectual Property Rights for Specialized Information, Knowledge and New Technologies, Vienna, Austria, August 21-25, 1995.

Cox, I., Kilian, J., Leighton, T. and Shamoon, T. (1996). Secure Spread Spectrum Watermarking for Multimedia. Information Hiding, Lecture Notes in Computer Science, No. 1174, Springer-Verlag, 1996, pp. 39–48.

Laurence Boney, Ahmed H. Tewfik and Khaled N. Hamdy. Digital Watermarks for Audio Signals. IEEE Int. Conf. on Multimedia Computing and Systems, (Hiroshima, Japan), June 1996.

ISO/IEC 11172-3 1993. Information Technology Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s, Annex D.

Davis Pan (1996). A Tutorial on MPEG/Audio Compression. Motorola, Inc. October 7, 1996.

Zhao, J. and Koch, E. (1998). A Generic Digital Watermarking Model. In: International Journal of Computer & Graphics, Vol. 22, No. 4., July/August 1998.

R. C. Bose and D. K. Ray-Chaudhuri(1960). On a Class of Error Correcting Binary Group Codes. *Inform. Contr.*, Vol. 3, 68-79, 1960.

P. Hocquenghem. (1959). Codes correcteurs d'erreurs. *Chiffres*, Vol. 2, 147-156, 1959.