MP3 RESISTANT OBLIVIOUS STEGANOGRAPHY

Litao Gang, Ali N. Akansu and Mahalingam Ramkumar

New Jersey Center for Multimedia Research
ECE Dept., New Jersey Institute of Technology
University Heights, Newark, NJ 07102.
{lxg8906, ali, mxr0096}@njit.edu

ABSTRACT

Robustness to compression is a basic requirement for any data hiding scheme. In this paper, we concentrate on MP3 resistant oblivious data hiding. First the MP3 compression algorithm is briefly introduced. In the second half, we propose three effective data hiding schemes, where the message is embedded in amplitude, DFT phase domain, and noisy components, respectively. All these schemes demonstrate robustness to MP3 compression.

1. INTRODUCTION

Watermarking or data hiding is the practice of embedding some *message* into a *host* signal. It provides a potential solution to multimedia copyright protection and piracy tracking.

Nowadays popular multimedia compression schemes are ubiquitously used for Internet transmission and storage. For instance, MP3 has become the standard on the web. Steganography should therefore survive these popular compression methods.

In this paper, we focus on data hiding in audio signals. In Section 2, MP3 compression is briefly introduced. Some operations like sub-band filtering, MDCT transform and psychoacoustic analysis, and their implications on data hiding are discussed.

In Section 3, some effective data hiding schemes are investigated. In an amplitude modulation scheme, a PN sequence is embedded in the original coefficients. Psychoacoustic model can be exploited to keep the distortion under the threshold. The second scheme hides the message in the DFT phase domain, since it is believed the phase is less significant perceptually than amplitude.

To design a compression-resistant hiding scheme, it is important to make use of the "holes" in the compression [3]. In MP3 compression, frequency resolution is the same in low and high frequency bands, although it is often unnecessary at high frequency end. There exists some room to modify these high frequency coefficients without much artifacts. The third scheme, noise substitution, embeds the message in the high frequency coefficients. Our simulation demonstrates their robustness against MP3 compression.

This work is partly supported by Panasonic Technologies, NJ

2. MP3 COMPRESSION — OVERVIEW

MP3 compression is a typical audio perceptual coding scheme. It is composed of psychoacoustic analysis, sub-band filtering, MDCT transform and quantization.

2.1. Psychoacoustics

Psychoacoustic model plays a very important part in perceptual coding. Studies show that the human audible frequency range can be divided into unequal width bands (or *critical bands*). Strong component in one critical band (masker) can mask off weaker components in other bands (maskee). That is referred to as the frequency masking phenomenon. How much a masker can mask is also related to its tonality. Roughly speaking, noise-like component is a better masker than tone-like ones.

The calculation of masking curve is a quite complicated procedure. It includes Hann-windowed spectra analysis, unpredictability measure, signal energy sum-up, normalization, conversion from the threshold partitions to scale factor bands, etc. The final result is the allowed distortion ratio r in each scale factor band where one quantization step size is applied in compression.

$$r = \frac{\text{allowed distortion energy}}{\text{scale-factor band energy}}. \quad (1)$$

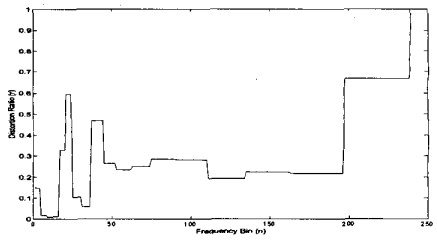


Figure 1: Scale-factor Band Distortion Ratio

The threshold ratio is used in quantization procedure to shape the quantization noise.

Another important phenomenon is temporal masking, which is solved by transform length switching (see next section for details).

2.2. Sub-band Filtering and MDCT

In MPEG-1 layer I and II, a polyphase filter bank is employed for time-frequency mapping. The filter bank is com-

0-7803-7041-4/01/$10.00 ©2001 IEEE

posed of 32 equal bandwidth filters.These filter bands however, are not equivalent to the critical bands whose bandwidth is nonuniform. To get a better frequency resolution, in MP3, each output channel is further subdivided into 18 bands via a windowed Modified Discrete Cosine Transform (MDCT).

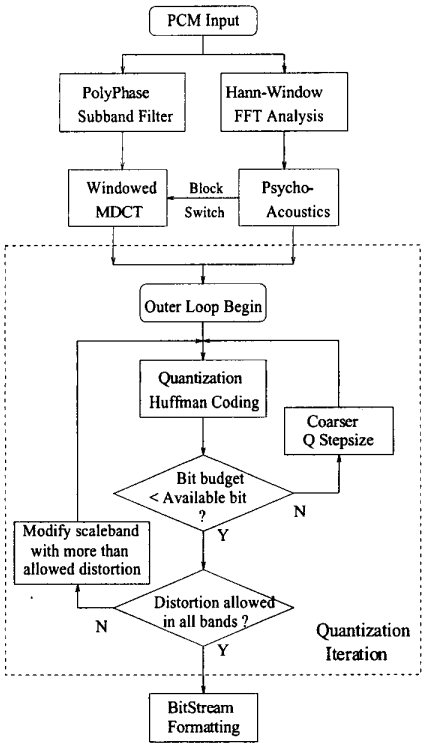


Figure 2: MP3 Encoding Flow Chart

To reduce signal redundancy, a fine frequency resolution is preferred. That favors the long transform length selection. However for the attack signals (transients), with long transform length, quantization noise tends to produce some "spreading" effect in time domain. That makes the attack signal not so "crisp". It is well known that fine time resolution and frequency resolution can not be achieved simultaneously. The transform length should be adaptive. For a stationary signal, long block is used. If an attack is present, short block should be used. The decision to switch is based on perceptual entropy [10].

2.3. Quantization and Rate Control

MP3 quantization procedure is composed of two loops — inner loop for bit rate control and outer loop for distortion control. Every MDCT coefficient in one scale factor band is multiplied by a common scale factor. Subsequently it is non-linear quantized and then Huffman encoded. If the bit budget is larger than the bit available, the scale factor is decreased until the bit consumption is acceptable. Outer loop calculates the quantization distortion in every scale factor band and compares it with the allowed distortion obtained from the psychoacoustic analysis. If it is larger than the allowed distortion, the scale factor is increased to reduce the noise audibility.

To absorb bit consumption imbalance, "bit reservoir" technique permits the current frame to "borrow" bits saved from past frames.

Fig. 2 depicts the flow chart of MP3 compression. For more detailed description on MP3, refer to [12], [4] and [7].

3. MP3 ROBUST DATA HIDING DESIGN

3.1. Amplitude Modulation

This scheme embeds a PN sequence in the original coefficients in a watermark domain. It is a direct extension of the Spread Spectrum (SS) schemes used in image and video watermarking [6], [5] and [9].

Suppose one information bit is to be embedded on a coefficient sequence $\mathbf{x}$.

The embedding procedure is

$$x_i' = \begin{cases} x_i + w_i|x_i|\alpha_i & 1 \text{ to be embedded} \\ x_i - w_i|x_i|\alpha_i & 0 \text{ to be embedded} \end{cases} \quad (2)$$

Where $\mathbf{w}$ is a random binary sequence, w_i is either -1 or +1. And α_i is the sub-band threshold ratio.

In this scheme, the same psychoacoustic model in MP3 can be used to calculate α_i. α_i in a scale factor band should be less than or equal to $\sqrt{r}$ in this scale factor band.

The oblivious correlation detector is

$$q = \sum_{i=0}^{N-1} r_i w_i. \quad (3)$$

where r_i is the received coefficient.

If $q > 0$, decision is bit value 1; Otherwise it is bit value 0.

Originally applied in escrow scenario, this scheme is not quite effective for oblivious applications. However if the sequence length N is sufficiently large, the host noise interfering term in (3) $\sum_{i=0}^{N-1} w_i x_i \approx 0$.

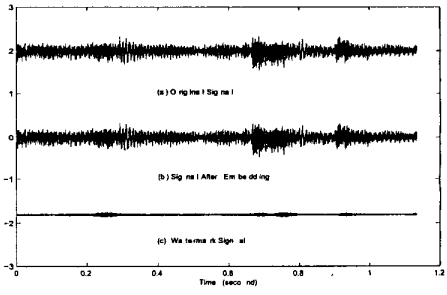


Figure 3: Amplitude Modulation

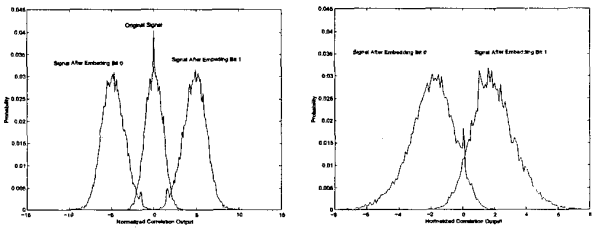
In our experiments, we use mono music clips sampled at 44.1kHz. Information bit is embedded on the MDCT coefficients from scale factor band 6 to 18, which corresponds to frequency range around 1kHz — 10kHz. To decrease the artifacts, the threshold α_i is selected smaller than $\sqrt{r}$ in the same scale factor band. In the more sensitive bands (1KHz —- 3KHz), α_i is further tuned to reduce artifacts.

The normalized detection output is

$$q' = \frac{< \mathbf{r}, \mathbf{w} >}{||\mathbf{r}||}. \quad (4)$$

One information bit is embedded every granule (576 samples) of a mono audio clip. Fig. 4 depicts the different q' distribution after embedding bit value 1 and 0. Due to the host noise interference, message extraction may not be sufficiently reliable. Some ECC code can be used or one information bit can be embedded in several granules.

The main advantage of this scheme is that psychoacoustic model can be explicitly employed to control the artifact.



(a) Before 64K MP3 (b) After 64K MP3

Figure 4: Normalized Detector Output Distribution in Amplitude Modulation

3.2. Phase Modulation

It is well known human ears are more sensitive to the amplitude than the phase. A phase hiding scheme [3] is proposed to modify the initial phase while keeping the relative phase between adjacent frames unchanged.

Phase is believed to be less perceptually significant. However human beings are sensitive to phase continuity between frames. The abrupt phase change may modify the signal spectrum. Informal listening tests show small modification in DFT phase is inaudible. This property can be exploited for data hiding.

An effective oblivious scheme, Quantization Index Modulation (QIM) [2] can be applied in phase domain. Fig. 5 depicts the phase QIM signaling.

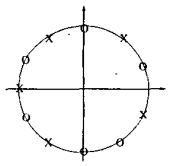


Figure 5: QIM in Phase Modulation

In our scheme, the original DFT phase value θ_i at one frequency bin is replaced by the nearest x point (to hide bit 1) or the nearest o points (to hide bit 0) on the unit circle.

To embed one bit in a phase sequence, deterministic patterns are defined to represent bit value 1 and bit value 0. For example, for a 4-coefficient sequence, we can define

Pattern A: (x o x o), to represent bit 1.
Pattern B: (o x o x), to represent bit 0.

To hide a bit, we need to modify θ_i to comply with pattern A or B.

The following suboptimal detector is used for decoding. Denote the received DFT amplitude and phase as r_i and ϕ_i respectively. Find the nearest x and o points α_i and β_i to ϕ_i and construct two sequences according to these two patterns:

$$\begin{aligned} \mathbf{u} &= (\alpha_0, \beta_1, \alpha_2, \beta_3) \\ \mathbf{v} &= (\beta_0, \alpha_1, \beta_2, \alpha_3) \end{aligned} \quad (5)$$

If $\sum r_i(\alpha_i - \phi_i)^2 < \sum r_i(\beta_i - \phi_i)^2$, decision is bit value 1; Otherwise bit value 0 is decided.

The detector is a weighted minimum distance detector in phase domain since the phase noise at smaller r_i tends to be large.

After embedding, DFT phases are fixed at x or o points. To introduce randomness, in embedding we can make $\theta_i + a_i$ be replaced by the x or o points where a_i is a random shift value.

The distortion introduced is determined by the distance d between x and o points. Smaller value of d is selected at the sensitive frequency bands while larger value of d is used at high frequency bands. In our experiments, DFT length is 512 and the DFT phases from 1KHz — 8kHz are changed. d_i varies from $\pi/12$ to $\pi/4$.

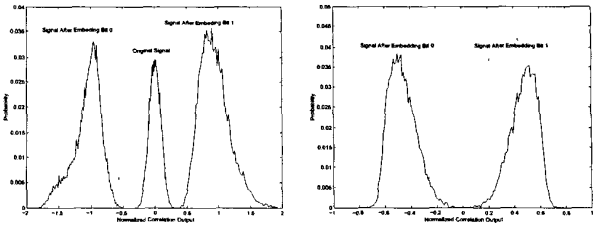
The normalized correlation output is defined

$$q' = \frac{< \mathbf{r}, ||\mathbf{u} - \mathbf{w}||^2 - ||\mathbf{v} - \mathbf{w}||^2 >}{||\mathbf{r}||}, \quad (6)$$

where $\mathbf{w}$ is the received phase sequence.

The statistical distribution of q' is shown in Fig. 6.

In this method, the audio quality can be preserved at relatively lower SNR than that in amplitude modulation.



(a) Before 64K MP3 (b) After 64K MP3

Figure 6: Normalized Output Distribution In Phase Modulation

3.3. Noise Substitution

In computer music studies, audio signal had long been regarded a combination of sine waves. X. Sierra [14] was among the first to introduce noise component in computer music. Lack of noise component makes the music "unnatural" (A good example of noise is the breathiness of a flute). Noise component is also perceptually significant.

In the advanced audio analysis model, noise component is indispensable. In the HILN model [1], signal is modeled as harmonic+individual sine+noise. While in [11], it is modeled as sine+transient+noise.

Some studies argue that for noise component, what is significant is not the fine frequency structure in noisy band, but the noise energy shape. The noise energy shape can be

described by its DCT coefficients [1] or by a source filter model (commonly used is linear predictor (LP) filter).

Goodwin *et. al* [8] proposed the Equivalent Rectangular Band (ERB) noise modeling. The author claims that the energy in ERB is more important than the noise spectral shape. Human beings do not resolve the fine frequency structure in a noisy band, only a "mixing effect" is felt.

Recently Levine *et. al* [11] proposed a similar approach, *bark band noise modeling*. In noise coding, at noisy bark bands, only the noise energy gain is coded and transmitted. The reconstructed noise spectrum is flat over the frequency range of each bark band. The author claims higher quality compared with DCT spectral envelope and LPC-smoothed representations.

In MP3, the fine spectral structure is encoded, including noisy bands. We thus can take advantage of it and embed message in these noisy bands. That is *noise substitution*.

In this method, the sign of MDCT coefficient x_i in noisy bands can be changed by a random pattern. Hiding procedure is

$$x_i' = \begin{cases} p_i|x_i|, & \text{to hide bit value 1} \\ -p_i|x_i|, & \text{to hide bit value 0} \end{cases} \quad (7)$$

where $\mathbf{p}$ is a binary random sequence and p_i is either -1 or $+1$.

Given a received sequence $\mathbf{r}$, correlation detector is

$$q = \sum r_i p_i. \quad (8)$$

When $q > 0$, the extracted bit value is 1; Otherwise bit value 0 is decided.

To accurately distinguish the noisy bands from non-noisy ones is not an easy job. Not all high frequency coefficients are noisy, some may be the high frequency components of a transient signal. It is reported that over 80% of the high frequency coefficients are "non-edged". In [13], several algorithms are proposed to make a distinction between noisy and non-noisy bands. For simplicity, we suppose frequency bands over 5kHz is noisy. Informal listening test shows it is a reasonable assumption.

Although the scheme does not survive the Low-pass filtering, it is robust to MP3 compression. MP3 may quantize a small value coefficient to zero, but it never inverts its sign. This sign conservative property promises its robustness against the compression.

In our experimental studies, the methods have achieved around $20 \sim 60$ bits/second hiding capacity.

4. CONCLUSIONS

In this paper, we briefly introduced the MP3 compression and proposed some data hiding schemes. The spread spectrum modulation is extended in the audio case. Its advantage is psychoacoustic model can be used to control artifacts, although it is not quite effective in host noise suppression. Two new methods are discussed after. Phase modulation embeds the message in phase domain. It is effective in oblivious scenarios and can preserve quality at lower SNR than amplitude modulation. Taking advantage of MP3 compression, we can also embed the message in high frequency coefficients. That corresponds to modification of noisy components. Simulations demonstrate the robustness to MP3 compression.

5. REFERENCES

[1] ISO/IEC FDIS 14496-3 Sec 2. "Information technology–Coding of audio-visual objects, Part 3:audio, Section 2: Parametric Audio Coding". 1999.

[2] B.Chen and G.W. Wornell. "Dither Modulation: A new approach to digital watermarking and information embedding". *Proc. of SPIE: Security and Watermarking of Multimedia Contents*, 3657:344–353, Jan 1999.

[3] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. "Technique for data hiding". *IBM System Journal*, 35(3-4):313–336, 1996.

[4] Karlheinz Brandenburg and Gerhard Stoll. "ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quantity Digital Audio". *J. Audio Eng. Soc.*, 42(10):780–792, Oct. 1994.

[5] I. Cox and M.L.Miller. "A review of watermarking and the importance of perceptual modeling". *Proceeding of Electronic Imaging*, Feburary 1997.

[6] I. J. Cox, Joe Kilian, Tom Leighton, and Talal Shamoon. "A Secure, Robust Watermark for Multimedia". *Workshop on Information Hiding*, May 1996.

[7] E. Eberlein, H. Popp, B. Grill, and J. Herre. "Layer III A Flexible Coding Standard". *Audio Eng. Soc. preprint 3493, 94th Convention, Berlin, Germany*, March 1993.

[8] M. Goodwin. "Adaptive Signal Models: Theory, Algorithms, and Audio Applications". *Ph.D. thesis, University of California, Berkley*, 1997.

[9] F. Hartung and B. Girod. "Watermarking of Uncompressed and Compressed Video". *Signal Processing*, 66(3):283–301, May 1998.

[10] J.D.Johnston. "Transform coding of audio signals using perceptual noise criteria". *IEEE Journal Sel.Areas Comm*, 6:314–323, Feb. 1988.

[11] S. Levine. "Audio Representations for Data Compression and Compressed Domain Processing". *Ph.D. thesis, Stanford University*, 1998.

[12] Davis Pan. "A Tutorial on MPEG/Audio Compression". *IEEE Multimedia Journal*, 1995.

[13] D. Schulz. "Improving Audio Codecs by Noise Substitution". *J. Audio Eng. Soc.*, 44(7/8):593–598, Jul./Aug. 1996.

[14] X. Sierra and J. Smith. "Musical sound modeling with sinusoids plus noise". http://www.iua.upf.es/ sms/.