MORPHOLOGICAL STEGANALYSIS OF AUDIO SIGNALS AND THE PRINCIPLE OF DIMINISHING MARGINAL DISTORTIONS

Oktay Altun, Gaurav Sharma, Mehmet Celik, Mark Sterling, Edward Titlebaum, Mark Bocko

Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, 14627

ABSTRACT

Steganographic methods attempt to insert data in multimedia signals in an undetectable fashion. However, these methods often disrupt the underlying signal characteristics, thereby allowing detection under careful steganalysis. Under repeated embedding, disruption of the signal characteristics is the highest for the first embedding and decreases subsequently. That is, the marginal distortions due to repeated embeddings decrease monotonically. We name this general principle as *the principle of diminishing marginal distortions (DMD)* and illustrate its validity in the audio domain using a morphological distortion metric. The principle of DMD is used to derive a steganalysis tool that detects the presence of hidden messages in uncompressed audio files. Detailed analysis and experimental results are provided for the detection of spread spectrum watermarking and stochastic modulation steganography.

1. INTRODUCTION

The goal of steganography [1, 2] is to create a covert channel within a multimedia signal—or any other digital document—to facilitate the transmission of messages, whose content and presence must be kept secret from unauthorized parties. While steganography deals with concealing information in innocuous-looking cover objects, steganalysis aims to detect the presence of these hidden messages and—if possible—to estimate their length and even extract their contents.

The game between the steganographer and the steganalysist may modeled as: *i)* the passive warden scenario, or *ii)* the active warden scenario [1]. A passive warden only monitors the channel and stops a document if there is any evidence of secret communications. On the other hand, an active warden deliberately modifies the document to prevent any secret communication. In this paper, we are primarily interested in the passive warden scenario and we restrict our attention to detection of the hidden messages.

Steganalysis problem in the passive warden scenario has been studied extensively for digital images. Authors investigated various image characteristics in order to determine the features that are the best indicators for hidden messages. These features include smoothness measures, quality metrics [3], higher-order statistics [4] and rate-distortion properties [5].

Today, audio files make up a significant portion of the multimedia traffic. This constitutes a significant opportunity for the potential users of steganographic methods. Nevertheless, interest in audio steganalysis has been relatively low, despite obvious practical implications. Although some of the methods proposed for digital images may be extended to the audio domain, it is desirable—and often necessary—to exploit specific characteristics of audio signals for improved detection.

Recently, Ozer et al. have investigated various distance measures in a statistical framework for audio steganalysis [6]. They have applied numerous distance measures to the cover signal and its denoised versions. Statistical analysis of computed distances has revealed the best discriminant metrics, which have been subsequently used to detect a variety steganography techniques.

In this paper, we describe a new approach to audio steganalysis and make two particular contributions: First, we define a specific non-linear transform for audio signals. This transform provides a morphological "distortion" measure, which is very sensitive to the changes arising from data embedding operations. Second, we show that, under repeated embedding, the marginal morphological distortions decrease monotonically with each iteration. We call this phenomenon as *the principle of diminishing marginal distortions (DMD)* and utilize it for audio steganalysis.

2. PROPOSED TECHNIQUE

Our proposed steganalysis technique is based on the effect of repeated data embedding on the morphological structure of the audio signals. In this section, we define a morphological distortion measure; observe the effect of repeated data embedding on this measure; state the principle of diminishing marginal distortions; and propose a method that utilizes this principle for audio steganalysis. A quantitative analysis of the technique is presented in the following section.

2.1. A morphological "distortion" measure for audio

Morphology is the analysis of the form or shape of objects and images. While an equivalent concept of "shape" is more difficult to define for audio signals, we use the term to define a particular class of non-linear transformations that represent the inherent structure of audio signals. These transformations are especially sensitive to minute changes caused by the data embedding procedures.

We define the first order morphology transform as the binary observation of the first difference:

$$M_1(x[n]) = \begin{cases} 1 & \text{if } x[n] - x[n-1] > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $x[n]$ is a one-dimensional monophonic audio signal.

Similarly, we may define the N^{th} order morphology transform

$$M_N(x[n]) = \begin{cases} 1 & \text{if } \Delta^N(x[n]) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

This work was supported by the Air Force Research Laboratory/IFEC under grant number F30602-02-1-0129.

where $\Delta^N(\cdot)$ is the N^{th} order difference operator.

We illustrate the first order morphology transform and its sensitivity to data embedding noise in a simple example. A pure sine wave and its first order morphology is seen in Fig. 1(top). The first order morphology value changes on the local minima and maxima of the sinusoid. When a low power Gaussian noise signal (a spread spectrum watermark) is added to the signal, the morphology changes frequently near the local extremeties of the signal (see Fig. 1(middle)).

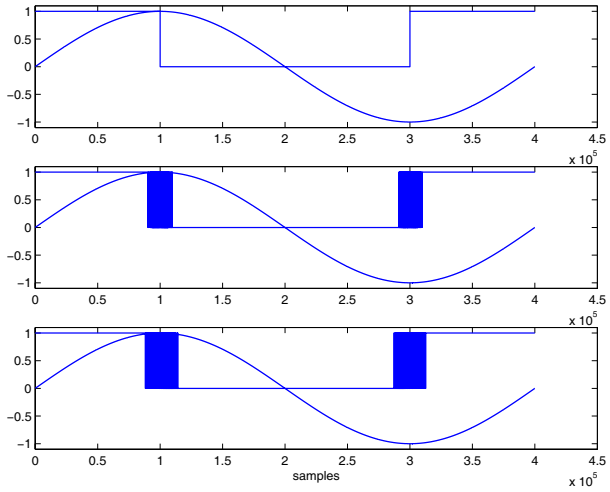

Fig. 1. *a)* A pure sine wave and its first difference binary morphological transform. *b)* Signal waveform and its transform after addition of a Gaussian watermark to the sine wave. *c)* Signal waveform and its transform after addition of a second watermark.

We define the morphological distortion between two audio signals as the Hamming distance between the N^{th} order binary morphologies of respective signals.

$$D_H(\underline{x}, \underline{y}) = D_{hamming}(M^N(\underline{x}), M^N(\underline{y})) \tag{3}$$

As seen in Fig. 1, this metric may be very sensitive to the data embedding methods.

A similar morphological distortion measure may be defined as the change in the number of transitions in the first order morphology of the signal. This metric corresponds to the change in the number of zero crossing of the signals first difference.

$$D_{ZC}(\underline{x}, \underline{y}) = |ZC(\Delta^1 \underline{x}) - ZC(\Delta^1 \underline{y})| \tag{4}$$

where $ZC(\cdot)$ is the number of zero crossings.

2.2. The principle of diminishing marginal distortions

Marginal distortion refers to the change—often decrease—in some specific measure of a signal regularity as a result of watermark embedding. *The principle of diminishing marginal distortions (DMD)* states that the marginal distortion of each additional watermark monotonically decreases under repeated embedding. That is, the impact of embedding a new watermark is less pronounced than the impact of the previous watermark. The idea is illustrated in Fig. 2.

This principle can be readily justified for distortions on the perceptual quality of a signal. Given a high-quality audio signal, introduction of noise will be extremely disturbing, i.e. introduce a large perceptual distortion[1]. Yet, the same amount of noise added to the already contaminated signal will be less disturbing, as it is masked by the first noise. The morphological distortion measure introduced in the preceding section is similar to the perceptual quality measures in the sense that the principle of DMD applies in both cases. Note that the principle does not necessarily apply to all distortion measures. A trivial example is the change in the variance of a Gaussian cover signal in response to successively added Gaussian watermarks. The marginal change (distortion) remains constant for all successive marks.

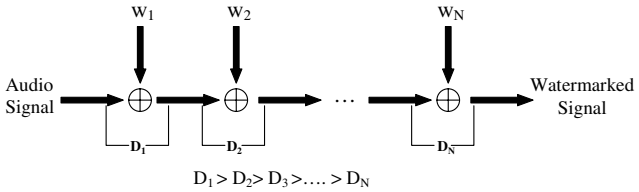

Fig. 2. The principle of diminishing marginal distortions. Successively added watermarks introduce lesser morphological distortions in each iteration.

	1^{st} wm.	2^{nd} wm.	3^{rd} wm.	4^{th} wm.
D_{ZC}	600%	40%	25%	10%
D_H	~ 5100	~ 4300	~ 4100	~ 4000

Table 1. Marginal distortion due to each additional watermark (over 10^4 samples). A Gaussian spread spectrum watermark with strength $\sigma = 5.10^{-5}$ is applied to a real audio signal.

The principle of DMD can be observed for the simple pure sinusoid example in Fig. 1. The first watermark significantly alters the first order morphology of the signal, while the second watermark has little effect on the morphology. As a result, the morphological distortion, (D_H for 10^4 samples), is 4910 for the first watermark and 2532 for the second watermark —a significant change that may be exploited for steganalysis. A similar decrease in marginal morphological distortions is observed for real audio signals as well (see Table. 1).

An intuitive insight for the for this observation may be stated as follows: Application of the first watermark breaks down the delicate morphology of the signal. As the subsequent watermarks are applied on a signal with already broken morphology, their net effect is limited. This is similar to smashing a piece of glass with an hammer. The first blow shatters the glass into many pieces, whereas the subsequent blows increase their number by an insignificant amount. A more theoretical justification of the principle is presented in the next section.

2.3. Steganalysis with DMD

The principle of DMD may be used for steganalysis purposes using the system seen in Fig. 3. Here, we test the unknown signal by introducing two test watermarks and measuring the morphological distortion D_H induced by each one. In general, these distortion

[1]Assuming the noise power is above the just noticeable distortion (JND) threshold.

values for stego signals are expected to be lower as the test watermarks are embedded after the hidden watermark. Moreover, we expect D_1 to be larger than D_2 in the light of the DMD principle. These distortion values are fed to a classifier to discriminate between original and stego signals. In our steganalysis system, we have used a single layer feed forward neural network with fifty hidden nodes.

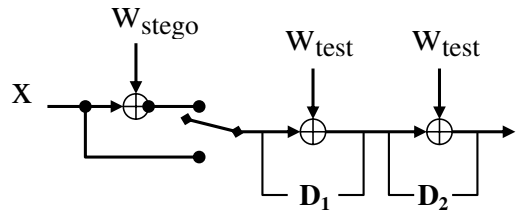

Fig. 3. The steganalysis system. Two test watermarks are used to differentiate between signals with and without stego messages. D_1 and D_2 are fed to a classifier.

3. ANALYSIS OF THE PROPOSED TECHNIQUE

Direct Sequence Spread Spectrum (DSSS) is well known technique for low SNR communications. It has been introduced in the watermarking context by Cox et al. [7]. DSSS spreads a payload bit over multiple samples/frequencies. Spreading allows for very low amplitude embedding—which is hard to detect—and high noise immunity. A DSSS watermark may be simulated by addition of a Gaussian noise sequence. This simulation also represents the stochastic modulation [8] steganography, which encodes messages as additive noise sequences. In order to analyze the technique, we need a model for the signal and the watermark processes. Audio signals are often modeled as first-order auto-regressive processes with a correlation coefficient slightly smaller than unity. We use a simplified version of this model, motivated by the limiting case when the correlation coefficient tends to 1. In this case, the first order difference of the audio signal may be considered i.i.d. distributed as $\mathcal{N}(0, \sigma_s)$, where $\mathcal{N}(\cdot)$ denotes the Gaussian distribution. The watermark is assumed to be independent of the audio signal i.i.d. and distributed as $\mathcal{N}(0, \sigma_h)$. Note that the N^{th} order difference will eventually converge to a Gaussian under the generalized central limit theorem [9], even under significantly milder assumptions.

The lemma below states that the expected Hamming distance between the "stego audio" and "stego audio plus test watermark" is less than the distance between the "unmarked audio" and "unmaked audio plus test watermark".

Lemma:

$$E[D_H(M_N(\underline{x}+\underline{w}_h), M_N(\underline{x}+\underline{w}_h+\underline{w}_t)] < E[D_H(M_N(\underline{x}), M_N(\underline{x}+\underline{w}_t)]$$

where $\underline{x}$ is the audio signal, $\underline{w}_h$ represents hidden watermark, $\underline{w}_t$ stands for test watermark.

Proof: We first derive the distribution of the N^{th} order difference for the unmarked, unmarked plus test watermark, stego, stego plus test watermark signals, respectively.

$$X_1 = \Delta^N(x[n]) \sim \mathcal{N}(0, 2^{N-1}\sigma_s^2) \quad (5)$$

$$X_2 = \Delta^N(x[n]+w_t[n]) \sim \mathcal{N}(0, 2^{N-1}\sigma_s^2 + 2^N\sigma_t^2) \quad (6)$$

$$X_3 = \Delta^N(x[n]+w_h[n]) \sim \mathcal{N}(0, 2^{N-1}\sigma_s^2 + 2^N\sigma_h^2) \quad (7)$$

$$X_4 = \Delta^N(x[n]+w_h[n]+w_t[n]) \sim \mathcal{N}(0, 2^{N-1}\sigma_s^2 + 2^N\sigma_h^2 + 2^N\sigma_t^2) \quad (8)$$

These expectations can be written in terms of the analysis window length N_A, and the probability of each element of the morphological string to change its sign:

$$E[D_H(M_N(\underline{x}), M_N(\underline{x}+\underline{w}_t)] = p_{1,2}N_A \quad (9)$$

$$E[D_H(M_N(\underline{x}+\underline{w}_h), M_N(\underline{x}+\underline{w}_h+\underline{w}_t)] = p_{3,4}N_A \quad (10)$$

These quantities may also be derived from the distributions in (5)-(8).

$$p_{i,j} = \int_{x_i,x_j \epsilon S} p_{X_i,X_j}(x_i, x_j) \cdot dx_i dx_j \quad (11)$$

where (i, j) takes the values $(1, 2)$ or $(3, 4)$.

These joint probability functions are derived by inspection:

$$p_{X_i,X_j} = \frac{1}{2\pi\sqrt{\sigma_i^2 2^N \sigma_t^2}} \exp\left(\frac{-x_i^2}{2\sigma_i^2}\right) \cdot \exp\left(\frac{-(x_j - x_i)^2}{2^{N+1}\sigma_t^2}\right) \quad (12)$$

The area where the sign change occurs is:

$$S = \{x_i, x_j \epsilon \mathbb{R} | \{x_i > 0, x_j < 0\} \text{or} \{x_i < 0, x_j > 0\}\}$$

Evaluating the integral (11) over area S provides:

$$p_{i,j} = \frac{1}{2} - \frac{\arctan\sqrt{k/2}}{\pi} \quad (13)$$

where k takes the value of $\frac{\sigma_s^2}{\sigma_t^2}$ for $p_{1,2}$ and $\frac{\sigma_s^2+\sigma_h^2}{\sigma_t^2}$ for $p_{3,4}$.

Note that $p_{i,j}$ is a strictly decreasing function of the ratio of unmarked (/stego) signal variance over the test watermark variance, or k. As $\frac{\sigma_s^2+\sigma_h^2}{\sigma_t^2} > \frac{\sigma_s^2}{\sigma_t^2}$ for any non-zero watermark variance σ_h^2, the lemma should be true. Moreover, this also proves that the first watermark addition has the highest impact, that is $p_{1,2} > p_{3,4}$.

4. EXPERIMENTAL RESULTS

We assembled a database of 200 uncompressed (16-bit PCM) audio segments, each of which is taken from a different track of various commercial music CDs. Musical styles included classical, popular, blues, country, ethnic and others. In order to simulate the DSSS [7] and stochastic modulation [8] steganographic techniques, we added a white Gaussian pseudo-random noise sequence to each segment. A neural network classifier has been trained on 150 original and 150 corresponding watermarked segments. The remaining 100 segments (50 original, 50 watermarked) are used for testing. The experiments are repeated for different watermark strengths (noise standard deviations, σ), which control the robustness, capacity and undetectability of the watermarks.

The distribution of original and stego ($\sigma = 5.10^{-4}$) segments for the Hamming morphological distortion metric (D_H, order $N = 1000$) is shown in Fig. 4. The horizontal and vertical axes are the distortion due to addition of the first and the second

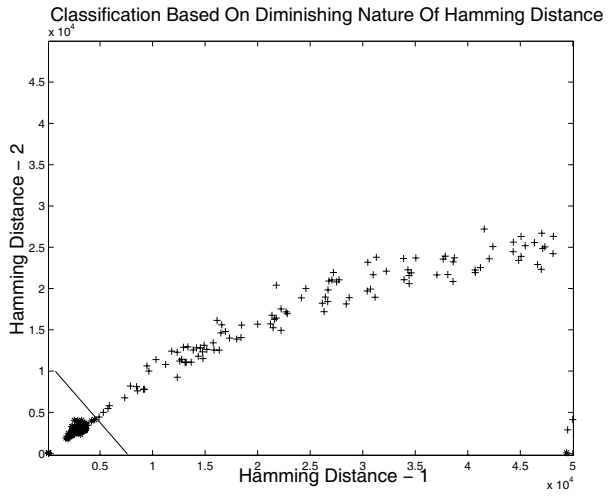

Fig. 4. Scatter plot for unmarked ("+") and stego ("*") files. Horizontal and vertical axes are the morphological Hamming distortions D_H ($N = 1000$) due to application of the first and the second test watermarks, respectively. Stego files are clustered near the origin.

test watermarks ($\sigma = 5.10^{-4}$), respectively. Note that the previously marked stego segments are clustered near the origin.

The detection performances of the neural network classifier trained for different watermarked strengths[2] is tabulated in Table. 2. Note that the algorithm performs very well even when strength of the embedded watermark is very low.

Watermark Strength	Detection Rate	False Alarm	Miss Rate
$\sigma = 5.10^{-2}$	100%	0%	0%
$\sigma = 5.10^{-3}$	94%	6%	0%
$\sigma = 5.10^{-4}$	88%	8%	4%
$\sigma = 5.10^{-5}$	80%	12%	8%

Table 2. Neural network classification results with D_H metric.($N = 1000$, segment length is 226 ms (10^4 samples) and watermark strength is $\sigma = 5.10^{-4}$.)

Although the analysis in the preceding section is focused on the DSSS watermarks, the principle of DMD and the proposed morphological metrics may be used for detection of other steganographic techniques. In order to support this claim, we have investigated the change in number of zero crossings of the first difference (D_{ZC}) for a number of data hiding techniques. The results are shown in Table. 3 in terms of percent changes. It is promising that the majority of methods leads to significant changes in signal morphology.

5. CONCLUSIONS AND FUTURE WORK

We have defined a morphological distortion measure and set forth the principle of diminishing marginal distortions. Based on this principle, we have developed an audio steganalysis technique using a neural network for classification. The proposed technique has been shown to be very effective against DSSS watermarking and stochastic modulation steganography techniques.

In our future work, we will investigate the use of proposed steganalysis tool for detection of other steganography techniques. Moreover, we are investigating other morphologic measures with potentially better discriminant powers. This may ultimately lead to the definition of measures which are dependent on specific steganography techniques. In that scenario, a bank of detection filters (with different morphology measures) has to be used for steganalysis.

[2]In each case, the strength of the test watermarks is matched to the strength of the embedded watermark.

	FHSS	DSSS	HAS Mask.	Phase Coding	Echo Coding
SNR (dB)	42	50	25	10	5.5
$ZC(\Delta^1 \underline{x})$ increase	13%	4%	41%	78%	360%

Table 3. Effect of various audio marking methods on the signal to noise ratio (SNR) and the morphology of the signal. In most cases, there is a significant increase in the number of zero crossings for the signal's first order difference.

6. REFERENCES

[1] G.J. Simmons, "Prisoners' problem and the subliminal channel," in *CRYPTO83-Advances in Crypto.*, 1984, pp. 51–67.

[2] Ross J. Anderson and Fabien A. P. Petitcolas, "On the limits of steganography," *IEEE Journ. of Sel. Areas in Comm.*, vol. 16, no. 4, pp. 474–481, May 1998.

[3] Ismail Avcibas, Nasir Memon, and Bulent Sankur, "Steganalysis using image quality metrics," *IEEE Trans. Info. Theory*, vol. 12, no. 2, pp. 221–229, Feb. 2003.

[4] Siwei Lyu and Hany Farid, "Detecting hidden messages using higher-order statistics and support vector machines," in *5th Int. Workshop on Info. Hiding*, 2002.

[5] M. Celik, G. Sharma, and A. M. Tekalp, "Universal image steganalysis using rate-distortion curves," in *Proc. SPIE: Sec., Stega., and Watermarking of Multimedia Cont. VI*, E. J. Delp and P. W. Wong, Eds., Jan. 2004, vol. 5306.

[6] Hamza Ozer, Ismail Avcibas, Bulent Sankur, and Nasir D. Memon, "Steganalysis of audio based on audio quality metrics," In Delp and Wong [10], pp. 55–66.

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

[8] J. Fridrich and M. Goljan, "Digital image steganography using stochastic modulation," In Delp and Wong [10], pp. 191–202.

[9] J. P. Romano and M.Wolf, "A more general central limit theorem for m-dependent randon variables with unbounded m," *Statistics And Probability Letters*, vol. 47, pp. 115–124, 2000.

[10] E. J. Delp and P. W. Wong, Eds., *Proc. SPIE: Sec. and Watermarking of Multimedia Cont. V*, vol. 5020, Jan. 2003.