Information Sciences 231 (2013) 123–134

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

MP3 audio steganalysis

Mengyu Qiao [a,*], Andrew H. Sung [b,c], Qingzhong Liu [d]

[a] Department of Mathematics and Computer Science, South Dakota School of Mines and Technology, Rapid City, SD 57701, USA
[b] Department of Computer Science and Engineering, New Mexico Institute of Mining and Technology, Socorro, NM 87801, USA
[c] Institute for Complex Additive Systems Analysis, New Mexico Institute of Mining and Technology, Socorro, NM 87801, USA
[d] Department of Computer Science, Sam Houston State University, Huntsville, TX 77341, USA

ARTICLE INFO

Article history:
Received 21 April 2010
Received in revised form 1 October 2012
Accepted 15 October 2012
Available online 22 October 2012

Keywords:
Steganalysis
Steganography
Feature selection
Signal complexity
MP3
SVM

ABSTRACT

MP3, one of the most widely used digital audio formats, provides a high compression ratio with faithful quality. The widespread use enables MP3 audio files to become excellent covers to carry hidden information in audio steganography on the Internet. Our research, however, indicates that there are few steganalysis methods proposed to detect audio steganograms and that steganalysis methods for the information-hiding behavior in MP3 audio are particularly scarce. In this paper we propose a comprehensive approach to steganalysis of MP3 audio files by deriving a combination of features from quantized MDCT coefficients. We design frequency-based subband moment statistical features, accumulative Markov transition features, and accumulative neighboring joint density features on second-order derivatives. We also model the distortion by extracting the shape parameters of generalized Gaussian density from individual frames. Different feature selection algorithms are applied to improve detection accuracy. Signal complexity and modification density are introduced to provide a comprehensive evaluation. Experimental results show that our approach is successful in discriminating MP3 covers and the steganograms generated by the steganographic tool, MP3Stego, in each category of signal complexity, especially for the audio streams with high signal complexities that are generally more difficult to steganalyze.

1. Introduction

Steganography is the ancient and newly revived art and science of a type of communication that is so secretive that only the participants themselves are aware of the existence of the hidden messages. To achieve such a degree of secretiveness, the individual sending a message imperceptibly modifies an original medium, also referred to as the cover file, in order to embed encrypted messages and create a steganogram, the modified carrier, by using shared cryptographic algorithms and keys. At the receiver's end, one can extract and decrypt messages from the steganogram by using the same cryptographic knowledge. Consequently, today's use of steganography poses a serious threat to national and cyber security, since terrorists, criminals, hackers, and adversaries can use such covert communication for malicious purposes.

Audio steganography data can be hidden in several ways: (a) low-bit encoding substitutes the least significant pieces of the information in each sampling point with the binary representation of the secret message; (b) phase encoding replaces the phase of initial segment of the audio and modifies the subsequent segments to preserve the pattern of relation between

* Corresponding author.
E-mail addresses: mengyu.qiao@sdsmt.edu (M. Qiao), sung@cs.nmt.edu (A.H. Sung), qxl005@shsu.edu (Q. Liu).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.10.013

segments; (c) spread spectrum encoding embeds data across a large range of the frequency spectrum; and (d) echo data encoding injects data by varying parameters of echo and generates an imperceptible echo of the original signal.

Recognition of new steganography threats has resulted in a heightened need for the implementation of effective countermeasures and, crucial to this effort, development of steganalysis techniques with improved efficiency and reliability. The objective of steganalysis is to identify the statistical differences between cover and steganogram and to detect the existence of hidden information. Over the past few years, many steganalysis methods have been proposed for detecting information-hiding in multiple steganographic systems. Most of the previous efforts focused on detecting information-hiding in digital images. For instance, a well-known method, histogram characteristic function center of mass (HCFCOM), successfully detects the steganography of the noise-adding models [9]. Another well-known method is to construct the high-order moment statistical model in the multi-scale decomposition using a wavelet-like transform and then apply a classifier to the high-order feature set [21]. Shi et al. proposed a Markov-based approach to detect the information-hiding in JPEG images [29]. Based on the Markov approach, Liu et al. expanded the Markov features to the inter-bands of the discrete cosine transform (DCT) domains, combined the expanded features with the polynomial fitting of the histogram of the DCT coefficients. This approach successfully improved the steganalysis performance in multiple JPEG image steganographic systems [18]. For enhancing the estimation, Zhang et al. studied the Laplacian model of pixel difference distributions in least significant bit (LSB) matching steganalysis [35]. Investigations into image steganalysis by other authors are included in references [6,13,15,16,19,24].

Besides digital imagery, digital audio is another popular carrier for covert communication where the variety of audio encodings increases the difficulty of audio steganalysis. To detect the information-hiding in digital audio, Avcibas presented content-independent distortion measures as features for classifier design [3]. Ozer et al. investigated the characteristics of the denoised residuals of audio files [23]. Johnson et al. set up a statistical model by building a linear basis that captures certain statistical properties of audio signals [12]. Zeng et al. presented a new algorithm to detect echo steganography based on statistical moments of peak frequency [34]. Kraetzer and Dittmann proposed a Mel–cepstrum-based analysis to detect embedded messages [14]. Liu et al. improved the performance of audio steganalysis by combining the Mel–Cepstrum feature with a temporal derivative-based spectrum analysis [17]. Geetha et al. presented high-order statistics of Hausdorff distance as discriminative features and investigated the application of evolving decision tree for audio steganalysis [7]. Studies by other authors in audio steganalysis are presented in references [20,25–27].

Although in the past years multiple steganalysis methods were designed to detect information-hiding in uncompressed audio, the information-hiding in compressed audio, such as MPEG-1 Audio Layer 3, more commonly referred as MP3, has been barely explored due to the complexity and variety of the compression algorithms. As a result of the different characteristics between compressed and uncompressed audio, most existing methods do not work for steganalysis of audio in the compression domain, and the decompression attempt, which erases the hidden data through the dequantization step in signal reconstruction, causes these methods to fail on decompressed audio.

In this paper, we present an approach to steganalysis of MP3 audio streams by extracting frequency-based subband moment statistics as well as accumulative neighboring joint probabilities and accumulative Markov transition probabilities in the compression domain. Generalized Gaussian density (GGD) is introduced to estimate the distribution of the modified discrete cosine transform (MDCT) coefficients. We also propose moment statistics of GGD shape parameters (β), extracted from individual frames as features, and utilize the shape parameter from the whole audio clip as a measure of signal complexity. The relation between audio steganalysis performance and signal complexity is also studied experimentally. Three feature selection methods are employed to further enhance the detection accuracy. Our approach successfully detects information-hiding in MP3 audio, in each category of signal complexity and modification density, especially in audio with a high signal complexity and a low modification density.

The rest of this paper is organized as follows. In Section 2, we briefly review the encoding process of MP3 and explain the embedding algorithm of MP3Stego. Section 3 introduces the statistical model of signal complexity and the comprehensive feature design. Section 4 presents experiments and results, which are discussed in Section 5. Section 6 presents our conclusions.

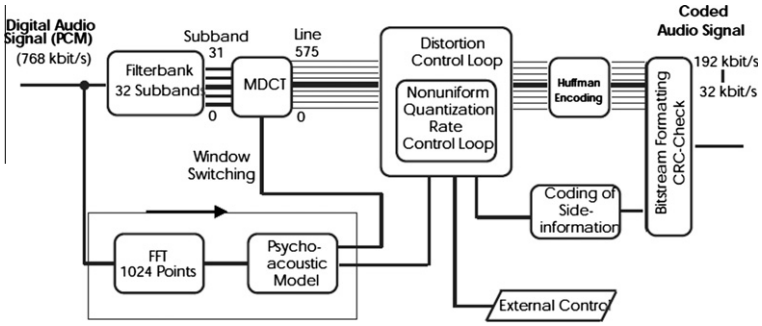

Fig. 1. MP3 audio encoding process [39].

2. Information-hiding of MP3Stego

2.1. MP3 audio compression

Pulse Code Modulation (PCM) is a standard format for storing or transmitting uncompressed digital audio. There are two parameters for PCM; sample rate (Hz) and bit-rate (kbit/s). The sample rate describes how many samples per second, while the bit-rate describes how large is the digital word that will hold the sample value. MP3 [36] is a digital audio encoding format which uses a lossy compression algorithm and significantly reduces the amount of data required to represent the audio while still providing audio quality that is, for most listeners, comparable to uncompressed audio. To ensure that the degradation of the audio quality is imperceptible to the auditory ability of most people, perceptual coding is introduced to analyze a short-term time or frequency window and to reduce less audible detail components in the signal.

Fig. 1 illustrates the entire MP3 audio encoding process, which consists of the following steps:

(1) Through a polyphase filterbank analysis, each sequence of consecutive 1152 PCM samples are filtered into 32 subbands with equal bandwidth according to the Nyquist frequency of the PCM signal. For example, if the sample frequency of the PCM signal is 44.1 kHz, the Nyquist frequency will be 22.05. Each subband will be approximately 22,050/32 = 689 Hz wide. The lowest subband will have a range of 0–689 Hz, the following subband of 689–1378 Hz, and so on.
(2) Each subband signal is further divided into 18 finer subbands by applying a modified discrete cosine transform (MDCT), and 576 subbands form a granule. To reduce artifacts caused by the edges of the time-limited signal segment, each subband signal must be windowed prior to the MDCT. If the subband signal at the present time frame shows little difference from the previous time frame, the long window type is applied to enhance the spectral resolution given by the MDCT. Alternatively, if the subband signal shows a considerable difference from the previous time frame, the short window is applied.
(3) Simultaneously, the same input PCM signal is also transformed to the frequency domain through a fast Fourier transform (FFT) in order to obtain a higher frequency resolution and information about the spectral changes over time.
(4) The frequency information from the FFT output is provided to the psychoacoustic model to identify audible parts of the audio signals. The current FFT spectra and the previous spectra are compared. If considerable differences are found, a request to adopt short windows will be sent to the MDCT block. As soon as the difference fades away, the MDCT block will be informed to change back to long windows. At the same time, the psychoacoustic model detects the dominant tonal components and masking thresholds are calculated for each critical band. Frequency components below the thresholds are masked out.
(5) Scaling and quantization are applied to 576 spectral values at a time, and carried iteratively in two nested loops: a distortion control loop (an outer loop which aims to keep the quantization noise below the masking threshold) and a rate control loop (an inner loop which determines the quantization step size).
(6) The Huffman coding, which retains MPEG-1 Layer 3 at a high quality with low bit-rates, is applied to the quantized values. All parameters generated by the encoder reside in the side information part of the frame. The frame header, side information, CRC, Huffman coded frequency lines, etc., are put together to form frames.

In filter bank analysis, PCM signals are converted into 32 subband signals through a polyphase filter $H_i[n]$:

$$H_i[n] = h[n] \times \cos\left[\frac{\pi \times (2 \times i + 1) \times (n - 16)}{64}\right] \quad i = 0, 1, 2, \ldots, 31, \quad n = 0, 1, 2, \ldots, 511 \tag{1}$$

where $h[n]$ is a low-pass filter computed by using a set of coefficients, characterized as $C[n]$ defined in [36].

$$C[k + 64j] = (-1)^j \times h(k + 64j) \quad k = 0, 1, 2, \ldots, 63, \quad j = 0, 1, 2, \ldots, 7, \quad m = 0, 1, 2, \ldots, 511 \tag{2}$$

The original PCM signal and the filtered signals are denoted by $x[n]$ and $P_i[n]$, respectively.

$$P_i[n] = \sum_{m=0}^{511} x[n - m] \times H_i[m] \quad i = 0, 1, 2, \ldots, 31 \tag{3}$$

We obtain subband signals $S_i[n]$, which are $P_i[n]$ down-sampled by 32.

$$S_i[n] = P_i[32n] = \sum_{i=1}^{n} x[32n - m] \times H_i[m] \quad i = 0, 1, 2, \ldots, 31 \tag{4}$$

To obtain MDCT coefficients, MP3 defines long and short windows to enhance the frequency resolution and temporal resolution, respectively. The product of the selected window coefficients and subband signal is denoted by Z_k. The MDCT coefficients are defined as follows:

$$X_i = \sum_{k=0}^{n-1} Z_k \cos\left[\frac{\pi}{2n}\left(2k+1+\frac{n}{2}\right)(2i+1)\right] \quad i = 0, 1, \ldots, \frac{n}{2}-1 \tag{5}$$

For the long window, n equals 36, otherwise n equals 12 for the short window.

In the psychoacoustic model II, $MNR(m)$ indicates the mask-to-noise ratio at m bits quantization, which measures the threshold of perceptible distortion. Signal-to-mask ratio at m bits quantization, $SMR(m)$, is calculated as:

$$SMR(m) = SNR - MNR(m) \tag{6}$$

Within a critical band, coding noise will not be audible as long as *MNR(m)* is negative. The MP3 compression algorithm relies on exploiting the weaknesses of human auditory perception and hiding the fact that a significant amount of information is discarded without any noticeable degradation of quality.

2.2. MP3Stego steganography

As one of the most popular audio formats on the Internet, MP3 provides a faithful reproduction of the original signal with a small amount of data. The widespread use of the format and its flexible encoding algorithm enable it to serve as a desirable carrier for covert communications. Böhme et al. [4] investigated the characteristics of MP3 encoders for potential applications in steganography or steganalysis. Although different encoders are designed to be compatible with the MP3 standard, statistical analysis also illustrates the distinctions among available MP3 encoders.

These distinctions increase the difficulty of discovering information-hiding in MP3 audios. MP3Stego [1,38] is one of the most widely used audio steganographic tools, especially for MP3 audio. Implemented by combining a novel information-hiding algorithm with an existing MP3 encoder, MP3Stego is built on the MP3 encoder and decoder from 8 Hz [37] and ISO MPEG Audio Subgroup Software Simulation Group, respectively. All payloads are encrypted using triple data encryption standard (3DES) and then embedded in frames randomly selected by using secure hash algorithm (SHA-1). With uncompressed waveform audio (WAV) as input, MP3Stego embeds data during the encoding process and generates a steganogram in MP3 format. The algorithm of MP3Stego exploits the audio degradation from lossy compression and embeds data by slightly expanding the distortion of the signal without attracting attention from the listener.

MP3Stego embeds compressed and encrypted data in an MP3 bit stream during the compression process. At the heart of layer 3 compression, two nested loops manipulate the trade-off between file size and audio quality. The hiding process occurs in the inner loop where the quantization step-size is increased to fit the available number of bits. We define the dequantized MDCT coefficients as *xr*. Round is a function that returns the closest integer value. The quantized signal, denoted by *ix*, is obtained from:

$$ix = round\left[\left(\frac{|xr|}{\sqrt[4]{2}^{stepsize}}\right)^{\frac{3}{4}} - 0.0946\right] \tag{7}$$

The part2_3_length variable in the frame header indicates the number of main data bits used for scale factors and Huffman code data in the MP3 bit stream. MP3Stego randomly modifies the part2_3_length variable to affect the loop condition of the inner loop. By increasing the step-size, the absolute values of the MDCT coefficients decrease according to the quantization level in one frame.

3. Detection methods

In MP3, each frame consists of two granules, and each granule represents 576 16-bit PCM samples in a time sequence. Through compression, each frame is first divided into 32 adjacent frequency subbands and then converted into 576 finer subbands in the MDCT domain. From our observations, we notice that the information-hiding behavior modifies, at the same time, most of the quantized MDCT coefficients with the exception of coefficients with small absolute values in one frame. This also indicates that the intra-frame distribution is preserved. On the other hand, the inter-frame pattern is altered across adjacent frames. Based on this analysis, we designed inter-frame feature sets by utilizing second-order derivative-based spectrum analysis.

3.1. Statistical model and signal complexity

In image processing, there are several statistical models [5,10,28,30,32] that illustrate the distribution of the intensity values of pixels: Markov random field models (MRFs); Gaussian mixture models (GMMs); and generalized Gaussian density models (GGD) in transform domains. Experiments show that a good probability distribution function (PDF) approximation for the marginal density of coefficients at a particular subband produced by various types of wavelet transforms may be

achieved by adaptively varying parameters of the GGD [5,22,28,33]. The GGD model contains the Gaussian and Laplacian PDFs as special cases, using $\beta = 2$ and 1, respectively.

For MP3 digital audio, the GGD model also provides a faithful approximation of the distribution of quantized MDCT coefficients which varies with compression ratio and signal complexity. Therefore, as a useful measure of signal complexity, the shape parameter of the GGD becomes another important evaluation factor, in addition to embedding strength, for MP3 steganalysis. Fig. 2 illustrates some signal samples with different values of complexity measurement β. At the same embedding strength, we surmise that the signals with low complexity are easier to be steganalyzed, but the steganalysis of the audio streams in high complexity is much harder, because the features become less discriminable in more complex signals.

For a long window, there are 576 MDCT coefficients in one frame. For short windows, three consecutive groups of 192 coefficients are combined into one frame. For an audio signal with N frames, we define the quantized MDCT coefficients as a matrix:

$$IX = \begin{pmatrix} ix_{0,0} & \cdots & ix_{0,575} \\ \vdots & \ddots & \vdots \\ ix_{N-1,0} & \cdots & ix_{N-1,575} \end{pmatrix} \tag{8}$$

In matrix *IX*, each row, denoted by *MDCT_F*, contains all quantized MDCT coefficients in one frame, and each column, denoted by *MDCT_B*, includes all quantized MDCT coefficients in one subband.

$$MDCT\_F_t = [ix_{t,0} \cdots ix_{t,i} \cdots ix_{t,575}] \tag{9}$$

$$MDCT\_B_i = [ix_{0,i} \cdots ix_{t,i} \cdots ix_{N-1,i}]^T \tag{10}$$

The GGD model of quantized MDCT coefficients of one MP3 audio is depicted in the following equation:

$$p(ix;\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|ix|/\alpha)^\beta} \tag{11}$$

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt, \quad z > 0 \tag{12}$$

where $\Gamma(z)$ is the gamma function and α models the width of the PDF peak (standard deviation), while β is inversely proportional to the decreasing rate of the peak. Sometimes α is referred to as the scale parameter, and β as the shape parameter. Generally, with the same compression ratio, the signal with complex variation has a high shape parameter of the GGD in the compression domain.

3.2. Moment statistics of GGD shape parameter

MP3Stego embeds data into MP3 audio by randomly modifying the length of the data segment in frame headers. This information-hiding behavior increases the step-size of the quantization, resulting in a slight degradation of the quality of the audio. Across spectra, the absolute values of quantized MDCT coefficients decrease in some randomly selected frames. For selected frames, the magnitudes of all MDCT coefficients are decreased simultaneously.

Based on spectrum distribution analysis, we hypothesized that information-hiding behavior alters the continuity of the distribution of adjacent frames. Therefore, we designed a moment statistical analysis method on the shape parameter of GGD on inter-frame. The GGD distribution of an individual frame is modeled as:

$$\begin{aligned} &p(ix_{t,i};\alpha_t,\beta_t) = \frac{\beta_t}{2\alpha_t\Gamma(1/\beta_t)} e^{-(|ix_{t,i}|/\alpha_t)^{\beta_t}} \\ &t = 0,1,2,\ldots,N-1; \quad i = 0,1,2,\ldots,575 \end{aligned} \tag{13}$$

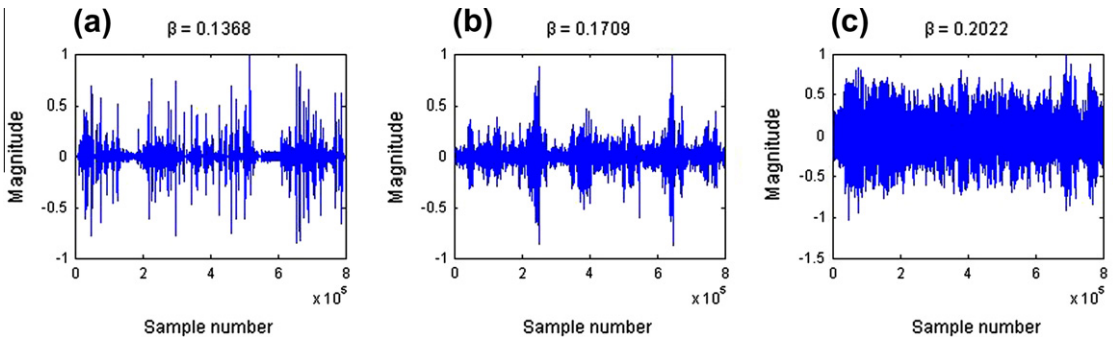

Fig. 2. MP3 audio signal samples with different values of signal complexity, β.

where t is the frame index and α_t and β_t are the scale parameter and shape parameter of the tth frame, respectively. Four moment statistical features are extracted from the spectrum of the shape parameter of the GGD. The mean value, standard deviation, skewness, and kurtosis are denoted by M_β, Δ_β, SK_β and KU_β, and calculated by the following equations:

$$M_\beta = \frac{\sum_{t=0}^{N-1}\beta_t}{N} \tag{14}$$

$$\Delta_\beta = \sqrt{\frac{1}{N}\sum_{t=0}^{N-1}(\beta_t - M_\beta)^2} \tag{15}$$

$$SK_\beta = \frac{\frac{1}{n}\sum_{t=0}^{N-1}(\beta_t - M_\beta)^3}{\left(\frac{1}{n}\sum_{t=0}^{N-1}(\beta_t - M_\beta)^2\right)^{3/2}} \tag{16}$$

$$KU_\beta = \frac{\frac{1}{n}\sum_{t=0}^{N-1}(\beta_t - M_\beta)^4}{\left(\frac{1}{n}\sum_{t=0}^{N-1}(\beta_t - M_\beta)^2\right)^{2}} - 3 \tag{17}$$

3.3. Frequency-based subband moment statistics

In image processing, second-order derivatives are widely employed for detecting isolated points, edges, etc. To the best of our knowledge, most audio steganography systems modify the bits of audio that also alter the pattern of second-order derivatives. Since MP3Stego randomly modifies the quantization step-size, the second-order derivatives of subbands also gain additional noise from information-hiding.

Let $f(x)$ $(x = 0, 1, \ldots, N-2)$ denote the MDCT coefficients of MP3 audio at a specific frequency subband. The second-order derivative is defined as follows:

$$D_f^2(x) \equiv \frac{d^2 f}{dt^2} = f(x+2) - 2 * f(x+1) + f(x) \quad x = 0 \sim N-3 \tag{18}$$

The MDCT coefficients of a stego-signal are denoted by $s(x)$, which may be modeled by adding a noise or error signal $e(x)$ into the original coefficient $f(x)$.

$$s(x) = f(x) + e(x) \tag{19}$$

The second-order derivatives of $e(x)$ and $s(x)$ are denoted by $D_e^2(x)$ and $D_s^2(x)$, respectively. We obtain:

$$D_s^2(x) = D_f^2(x) + D_e^2(x) \tag{20}$$

At this point, we present the following procedure to extract the second-order derivative-based statistics of the signals:

(1) Obtain the second-order derivatives $D_{ix}^2(t,i)$ from 576 MDCT subband signals $MDCT\_B(i)$ across all frames where $t = 0, 1, 2, \ldots, N-1$. and $i = 0, 1, 2, \ldots, 575$.
(2) Calculate statistics, including mean value, standard deviation, skewness, and kurtosis of subband signals.
(3) To reduce the number of features, the whole frequency zone is divided into Z zones or parts (Z is set to 32 in our experiments) from the lowest to the highest frequency. Then calculated the sums of the mean value, standard deviation, skewness, and kurtosis in each zone, denoted by M_Z, Δ_Z, SK_Z and KU_Z, where $Z = 0, 1, \ldots, 31$.

$$M_Z = \sum_{i=Z*18+1}^{Z*19} \frac{\sum_{t=0}^{N-3} D_{ix}^2(t,i)}{N-2} \tag{21}$$

$$\Delta_Z = \sum_{i=Z*18+1}^{Z*19} \sqrt{\frac{1}{N-2}\sum_{t=0}^{N-3}(D_{ix}^2(t,i) - M_Z)^2} \tag{22}$$

$$SK_Z = \sum_{i=Z*18+1}^{Z*19} \frac{\frac{1}{N-2}\sum_{t=0}^{N-3}(D_{ix}^2(t,i) - M_Z)^3}{\left(\frac{1}{N-2}\sum_{t=0}^{N-3}(D_{ix}^2(t,i) - M_Z)^2\right)^{3/2}} \tag{23}$$

$$KU_Z = \sum_{i=Z*18+1}^{Z*19} \frac{\frac{1}{N-2}\sum_{t=0}^{N-3}(D_{ix}^2(t,i) - M_Z)^4}{\left(\frac{1}{N-2}\sum_{t=0}^{N-3}(D_{ix}^2(t,i) - M_Z)^2\right)^{2}} - 3 \tag{24}$$

3.4. Accumulative neighboring joint density and Markov approach

The Markov process is a widely used stochastic process. In image steganalysis, Shi et al. [29] proposed a Markov-based approach to detect the information-hiding in JPEG images. Liu et al. [18] expanded the application of Markov features to the inter-blocks of the DCT domain. Although the designs of JPEG and MP3 compressions have similarities, the

information-hiding process in digital audio does not share the same pattern with image steganography. Based on our analysis in Section 2, we designed an inter-frame Markov approach (IM) and inter-frame Neighboring Joint Density (INJ) for MP3 audio steganalysis, described in the following equations, where $\delta = 1$ if its arguments are satisfied; otherwise $\delta = 0$. Similar to the references [18,29], the range of i and j is $[-4,4]$. In such a case, we have two 9×9 feature matrices with each one consisting of 81 elements or features. Fig. 3 shows the Markov transition probabilities of a cover and the steganogram in (a) and (b), the neighboring joint densities of the cover and the steganogram in (d) and (e), and the differences of the transition probabilities and the differences of the neighboring joint densities between the cover and the steganogram, in (c) and (f).

$$IM(u,v) = \sum_{i=0}^{575} \frac{\sum_{t=0}^{N-4} \delta\left(D_{ix}^2(t,i) = u, D_{ix}^2(t+1,i) = v\right)}{\sum_{t=0}^{N-4} \delta\left(D_{ix}^2(t,i) = u\right)} \tag{25}$$

$$INJ(u,v) = \frac{\sum_{i=0}^{575}\sum_{t=0}^{N-4} \delta\left(D_{ix}^2(t,i) = u, D_{ix}^2(t+1,i) = v\right)}{576 * (N-3)} \tag{26}$$

3.5. Feature selection

To achieve better performance in detection, we combined different feature sets as a comprehensive approach. However, with more features being included in the feature set, the increasing feature dimension and feature redundancy compromise the performance and the efficiency of steganalysis. Feature selection methods are designed to find an optimal feature subset by eliminating features with little discriminative information. Therefore, in a comprehensive approach, feature selection can be a useful solution to further enhance the accuracy as well as reduce the overhead.

Most widely used feature selection methods may be categorized into filter, wrapper, and embedded methods. Filter methods select feature subsets based on performance evaluation metrics extracted from feature set and work with no dependency on reference to machine learning algorithms. Filter methods are generally less expensive than wrapper and embedded methods. However, filter methods consider the features as independent units and ignore possible interactions among features. The combination of features does not guarantee an enhanced performance, according to the performance evaluation of individual features. Moreover, filter methods tend to select features which correspond to high evaluation scores, which might generate more redundant yet less informative feature subsets. Avcibas et al. [2] presented a universal steganalysis based on image quality metrics and utilized the one-way analysis of variance (ANOVA) for choosing good metrics. This feature selection method is a filtering approach and the final feature set may not be optimal. Wrapper methods wrap around particular machine learning algorithms that can assess the selected feature subsets by estimating classification errors and then

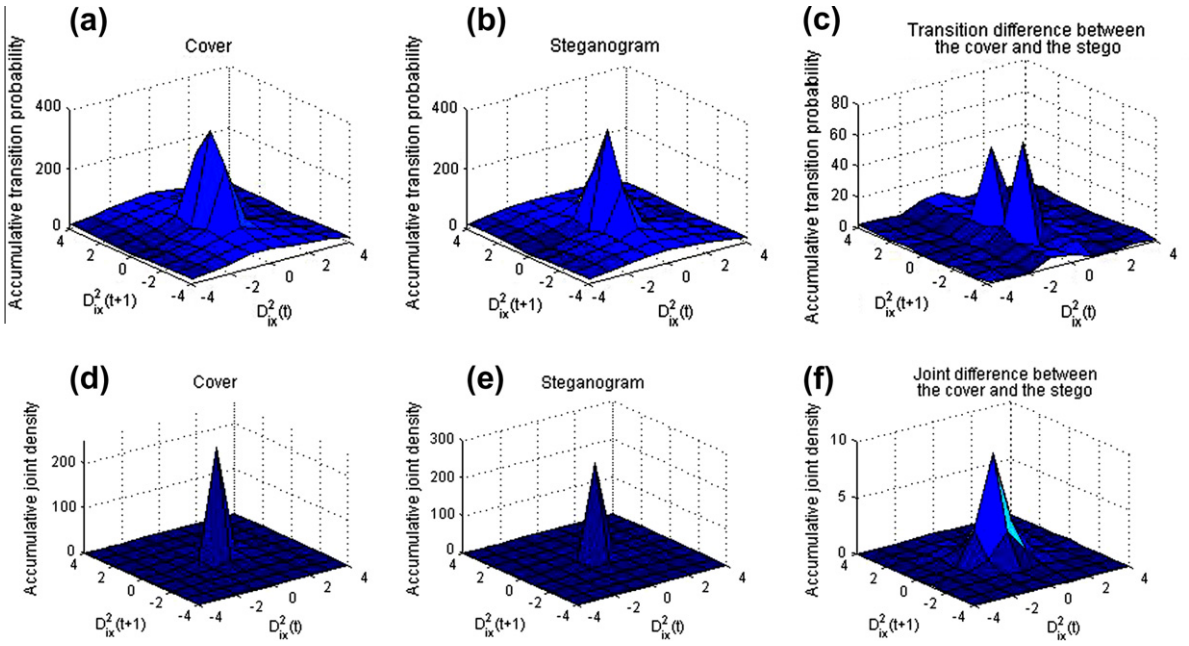

Fig. 3. The comparison of Markov transition probabilities of the second-order derivatives, shown in (a) and (b), the comparison of neighboring joint densities of the second-order derivatives, shown in (d) and (e), and the differences of the transition probability and neighboring joint density between a cover and the steganogram, shown in (c) and (f).

building the final classifiers [11]. One of the well-known methods is the Support Vector Machine–Recursive Feature Elimination (SVM–RFE), which refines the optimal feature set by using the SVM in a wrapper approach [8]. Embedded methods are built into adaptive systems while simultaneously processing feature selection with a classifier.

To deal with the issue of feature selection in MP3 audio steganalysis, we compared three feature selections: ANOVA, SVM–RFE, and a two-step approach incorporating ANOVA with SVM–RFE.

4. Experiments and results

4.1. Experimental setup

Our dataset contains 5000 mono MP3 audio clips with a bit rate of 128 kbps and a sample rate of 44.1 kHz. Each audio signal has duration of 20 s, and the file size is 313 KB. These audio files include digital speeches and songs in several languages, such as English, Chinese, Japanese, Korean, and several types of music including jazz, rock, blue, and natural sounds. The payloads include voice, video, image, text, executable codes, and random bits, with each steganogram carrying a unique payload. By embedding different amounts of data, we constructed four sets of MP3 stego-audio with approximate modification densities of 8%, 12%, 16%, and 20%, which carry payloads of 30, 60, 90, and 120 Bytes. At a modification density of 20%, the MP3Stego reaches its maximum hiding capacity. Cover MP3 audio was compressed by using the same MP3 encoder in the MP3Stego from 8 Hz [37]. In this study, we used modification density, defined as the proportion of the number of modified non-zero MDCT coefficients to the number of all non-zeros MDCT coefficients, instead of the hiding ratio to evaluate detection performance.

Four groups of features are extracted from covers and steganograms. Sixty percent of the feature sets were employed for constructing the classification model, while the other 40% of the feature sets were used for testing. For every experimental setting, we conducted the experiment 100 times, with the training and the testing sets randomly chosen every time. The classification returned results consisting of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The testing accuracy was calculated by $W^*TP/(TP + FN) + (1 - W)^*TN/(TN + FP)$, where W is a weighting factor in the range of [0,1]. Without losing generality, W was set to 0.5 in our experiments. Support vector machines (SVM) with RBF [31] kernels were used for detection.

4.2. Statistics of feature sets

We compared the significance of GGD shape statistical features calculated by (14)–(17), frequency-based subband moment statistical features, described in (21)–(24), as well as accumulative Markov transition features and neighboring joint density features, obtained from (25) and (26), respectively.

Fig. 4a and b list the F scores of the ANOVA of the features extracted from covers and steganograms produced by using the MP3Stego audio steganographic tool with 16% and 20% modification density. The *Y*-axis indicates the F score, and the *X*-axis gives the number of features.

From the comparison of the F scores, we found that frequency-based subband moment statistical features outperform the other feature sets, especially those extracted from middle frequency and correspond to higher F scores. We surmised that using frequency-based subband moment statistical features would provide the best detection performance. Accumulative Markov transition features and neighboring joint density features obtain similar F scores at 16% and 20% modification density. In GGD shape statistical features, the high-order moment statistics, especially the skewness of shape parameters, are more discriminative. Although the higher F score indicates the more significant feature, the interaction and the redundancy of the feature sets also affect the classification performance. Therefore, the testing accuracy is more reliable in evaluating the performance of features.

Fig. 5a and b illustrate the comparison of SVM testing accuracies by using all samples at 16% and 20% modification density correspondingly. The detection performance of frequency-based subband moment statistical features is superior to the performance of accumulative neighboring joint density features, GGD shape statistical features, and accumulative Markov transition features. Moreover, the distribution of the testing accuracy of frequency-based subband moment statistical features shows a smaller degree of dispersion than other feature sets, which indicates a more stable classification.

4.3. Comparison of feature selection methods

To form a comprehensive approach to MP3 audio steganalysis, we combined: GGD shape statistical features; frequency-based subband moment statistical features; accumulative Markov transition features; and accumulative neighboring joint density features. However, the large feature dimension and redundancy compromise the performance and the efficiency of the detection. To further increase the classification accuracy, we employed two widely used feature selection methods; ANOVA and SVM–RFE, both introduced in Section 3. We also designed a two-step approach by combining these two methods. For the two-step approach, we picked 200 as a threshold to divide the processes of two feature selection methods. All features were ranked using ANOVA, The features with the top 200 F scores were chosen as the input for SVM–RFE. The size of a feature subset for classification continued increasing from 1 to 200 in the order of the feature rank provided by SVM–RFE.

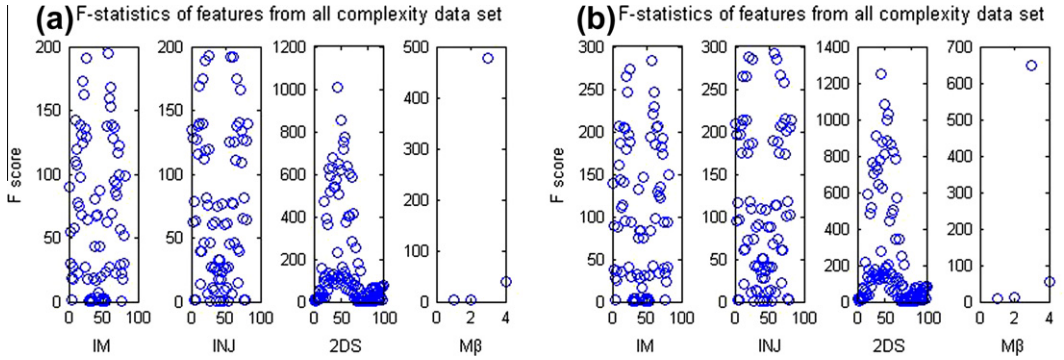

Fig. 4. One way ANOVA F scores (Y-label) of Accumulative Markov transition features (IM), Accumulative neighboring joint density (INJ), frequency-based subband moment statistical features (2DS), and GGD shape statistical features (Mβ) from whole data set including samples with all signal complexities at 16% and 20% modification density in (a) and (b) respectively.

For accurate evaluation, we divided the whole data set into low ($\beta < 0.162$), middle ($0.151 <= \beta <= 0.171$), and high ($\beta >= 0.162$) complexity zones with 50% overlap between adjacent zones using GGD shape parameter of all quantized MDCT coefficients in each sample. Furthermore, the whole dataset, including samples with all complexities, was used as another category.

Regarding the relation between detection performance and signal complexity, shown in Table 1, as the signal complexity increases, the detection performance decreases. Since the average signal complexity of the whole dataset is 0.164, the average classification accuracy of all samples is close to the accuracy of middle complexity.

Comparing the three feature selection methods, the two-step approach outperforms ANOVA and SVM–RFE in each category of signal complexity and modification density. Our study also shows that the two-step approach adopts the advantage of low standard errors and thus provides more stable detection performance.

In addition to the comparison shown in Table 1, the receiver operating characteristic (ROC) curves by using ANOVA, SVM–RFE, and the two-step approach are given in Fig. 6. The ROC curve of the two-step approach generates the largest area under the curve at 20% modification density.

5. Discussion

Frequency-based subband moment statistical features provide a more accurate and stable classification than the other feature sets. More specifically, the features corresponding to the middle part of 576 subbands are more significant than those corresponding to the high and the low parts. The reason for this phenomenon is that the high and the low parts of the subbands usually contain quantized MDCT coefficients of large and small values. The MP3Stego algorithm embeds data by modifying the quantization step-size, which affects all MDCT coefficients in the particular frame. However, the effects of different values vary greatly due to the non-uniform scale of quantization. The larger values usually refer to the more informative content of the audio signal, and they are mainly determined by the characteristic of the signal. The information-hiding behavior

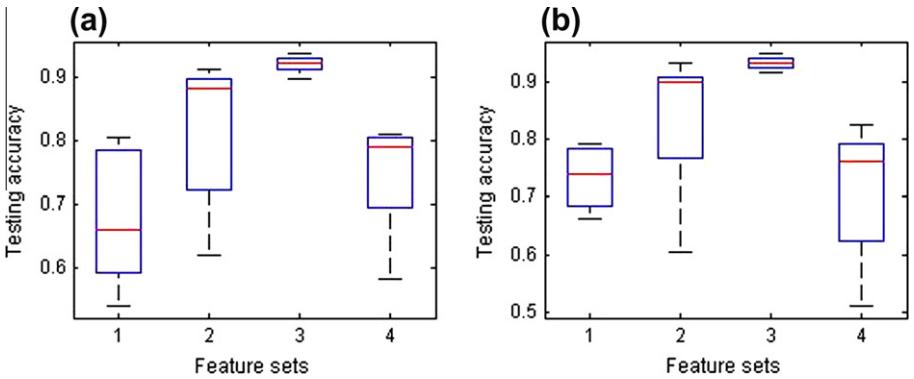

Fig. 5. SVM testing accuracies by using feature sets: Accumulative Markov transition features (1), Accumulative neighboring joint density features (2), Frequency-based subband moment statistical features (3), and GGD shape statistical features (4) from whole data set including samples with all signal complexities at 16% and 20% modification density in (a) and (b) respectively.

Table 1
Average testing accuracy values and standard errors of feature selection methods: ANOVA, SVM-RFE, and two-step approach incorporating ANOVA with SVM-RFE at the optimal feature dimension.

Modification density (%)	Signal complexity	Testing accuracy (mean ± std, %)		
		ANOVA	SVM-RFE	ANOVA & SVM-RFE
8	Low	81.6 ± 1.4	85.9 ± 2.8	88.6 ± 1.4
	Middle	78.1 ± 1.2	82.2 ± 2.3	84.9 ± 0.9
	High	75.6 ± 1.8	78.2 ± 1.9	81.4 ± 1.2
	All	79.4 ± 1.1	86.6 ± 1.6	86.1 ± 1.7
12	Low	88.6 ± 1.4	94.0 ± 2.2	95.1 ± 1.6
	Middle	84.9 ± 1.7	91.0 ± 1.8	91.3 ± 1.4
	High	82.6 ± 0.6	88.9 ± 2.0	90.0 ± 1.8
	All	85.0 ± 0.9	90.6 ± 1.2	90.4 ± 0.9
16	Low	90.4 ± 1.4	95.9 ± 2.8	97.0 ± 1.6
	Middle	89.3 ± 1.3	93.2 ± 2.1	94.8 ± 1.1
	High	87.7 ± 1.0	91.8 ± 2.2	92.1 ± 1.2
	All	88.5 ± 0.6	93.2 ± 0.7	93.3 ± 1.0
20	Low	94.6 ± 0.8	96.7 ± 2.6	98.6 ± 1.0
	Middle	91.1 ± 1.6	94.7 ± 2.4	95.3 ± 1.4
	High	89.9 ± 0.7	94.3 ± 2.8	93.6 ± 1.6
	All	91.0 ± 0.9	94.9 ± 0.5	95.6 ± 0.6

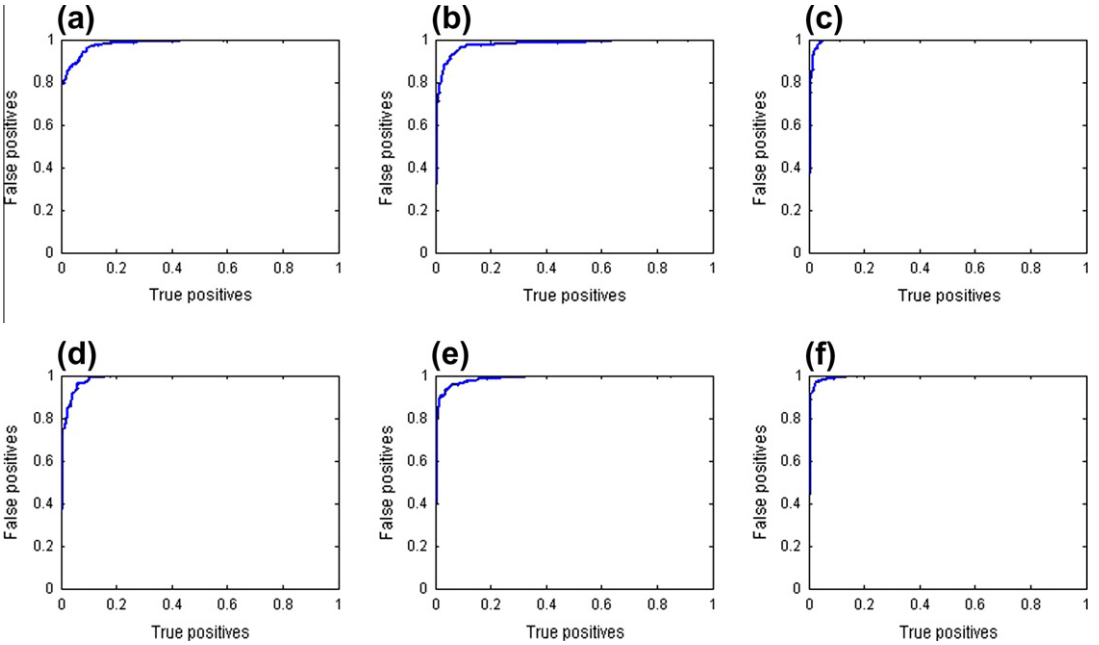

Fig. 6. ROC curves by using ANOVA (a and d), SVM-RFE (b and e), and two-step feature selection (c and f) methods, in detection of MP3 steganograms with 16% (first row) and 20% (second row) modification densities.

and the high complexity of the signal could both indicate the inconsistency of the coefficients between adjacent frames. The zero value usually does not change by modifying the quantization step-size. Therefore, the features extracted from the middle frequency are more sensitive to information-hiding.

In this study, we focus on detecting the information-hiding of MP3Stego, because MP3Stego implements a unique hiding scheme which is involved in the compression process. However, the proposed approach also has the generalization capability to steganalyze other steganographic systems, because traditional steganography introduces more distortion than MP3Stego in compressed audio.

The GGD shape parameter is introduced as an important signal complexity measure to evaluate the detection performance. With same modification density, the detection accuracy decreases as the signal complexity increases. However, the GGD shape parameter only describes the distribution of the MDCT coefficients of the entire audio and neglects the relation between MDCT coefficients in one frame or one subband. Since different complexities may have similar distributions, an

accurate measure of signal complexity with fine granularity is another important issue for MP3 audio steganalysis. Since we extract the shape parameter from quantized MDCT coefficients, the distribution of these coefficients is not only influenced by the signal complexity but also by the setting and the implementation of MP3 encoder. The MP3Stego has a tight coupling between the hiding algorithm and the MP3 encoder. Although the implementations of the MP3 encoder have to comply with ISO standards, the differences in the quantization function and distortion control will affect the performance of steganalysis. The differences in the quantization function may affect the distribution of MDCT coefficients and increase the false alarm rate with a trained model using another encoder.

6. Conclusions

In this paper, we propose a comprehensive approach to steganalysis of MP3 audio by deriving a combination of features from quantized MDCT coefficients. We extract frequency-based moment statistical features, accumulative Markov transition features, and accumulative neighboring joint density features. We also model the distortion by extracting the distribution parameters of generalized Gaussian density from individual frames in the MDCT transform domain. Different feature selection algorithms are applied to improve detection accuracy. For an accurate evaluation, signal complexity and modification density are introduced to evaluate the performance of the proposed approach. Experimental results show that our approach successfully detects the information-hiding in MP3 steganograms generated by the MP3Stego steganographic tool. The proposed approach produces reliable performance in each category of signal complexity, especially for audio signals with high signal complexity, and thus improves the state-of-the-art of audio steganalysis.

Acknowledgments

The authors gratefully appreciate the Institute for Complex Additive Systems Analysis and the Center for Graduate Studies of New Mexico Tech, as well as South Dakota School of Mines and Technology for supporting this research. This project was also supported by Award No. 2010-DN-BX-K223 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the authors and do not necessarily reflect those of the Department of Justice. SVM-RFE is the patented technology by Health Discovery Corporation.

References

[1] R. Anderson, F. Petitcolas, On the limits of steganography, IEEE Journal of Selected Areas in Communications 16 (4) (1998) 474–481.
[2] I. Avcibas, N. Memon, B. Sankur, Steganalysis using image quality metrics, IEEE Transactions on Image Processing 12 (2) (2003) 221–229.
[3] I. Avcibas, Audio steganalysis with content-independent distortion measures, IEEE Signal Processing Letters 13 (2) (2006) 92–95.
[4] R. Böhme, A. Westfeld, Statistical characterisation of MP3 encoders for steganalysis, in: Proceedings of the Workshop on Multimedia and Security, 2004, pp. 25–34.
[5] M.N. Do, M. Vetterli, Wavelet-based texture retrieval using generalized Gaussian density and Kullback–Leibler distance, IEEE Transactions on Image Processing 11 (2002) 146–158.
[6] J. Fridrich, J. Kodovský, V. Holub, M. Goljan, Steganalysis of content-adaptive steganography in spatial domain, Information Hiding, Lecture Notes in Computer Science (2011).
[7] S. Geetha, N. Ishwarya, N. Kamaraj, Evolving decision tree rule based system for audio stego anomalies detection based on Hausdorff distance statistics, Information Sciences 180 (13) (2010) 2540–2559.
[8] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (1–3) (2002) 389–422.
[9] J. Harmsen, W. Pearlman, Steganalysis of additive noise modelable information hiding, Proceedings of SPIE Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents 5020 (2003) 131–142.
[10] J. Huang, D. Mumford, Statistics of natural images and models, Proceedings of Computer Vision and Pattern Recognition 1 (1999).
[11] I. Inza, B. Sierra, R. Blanco, P. Larranaga, Gene selection by sequential search wrapper approaches in microarray cancer class prediction, Journal of Intelligent and Fuzzy Systems 12 (1) (2002) 25–33.
[12] M. Johnson, S. Lyu, H. Farid, Steganalysis of recorded speech, Proceedings of SPIE 5681 (2005) 664–672.
[13] M. Kharrazi, T.H. Sencar, N. Memon, Improving steganalysis by fusion techniques: a case study with image steganography, LNCS Transactions on Data Hiding and Multimedia Security I 4300 (2006) 123–137.
[14] C. Kraetzer, J. Dittmann, Pros and cons of Mel–cepstrum based audio steganalysis using SVM classification, Lecture Notes in Computer Science 4567 (2008) 359–377.
[15] Q. Liu, A. Sung, Z. Chen, J. Xu, Feature mining and pattern classification for steganalysis of LSB matching steganography in grayscale images, Pattern Recognition 41 (1) (2008) 56–66.
[16] Q. Liu, A. Sung, B. Ribeiro, M. Wei, Z. Chen, J. Xu, Image complexity and feature mining for steganalysis of least significant bit matching steganography, Information Sciences 178 (1) (2008) 21–36.
[17] Q. Liu, A. Sung, M. Qiao, Temporal derivative-based spectrum and Mel–cepstrum audio steganalysis, IEEE Transactions on Information Forensics and Security 4 (3) (2009) 359–368.
[18] Q. Liu, A. Sung, M. Qiao, Z. Chen, B. Ribeiro, An improved approach to steganalysis of JPEG images, Information Sciences 180 (9) (2010) 1643–1655.
[19] Q. Liu, A. Sung, M. Qiao, Neighboring joint density based JPEG steganalysis, ACM Transactions on Intelligent Systems and Technology 2 (2) (2011), http://dx.doi.org/10.1145/1899412.1899420.
[20] Q. Liu, A. Sung, M. Qiao, Derivative based audio steganalysis, ACM Transactions on Multimedia Computing, Communications and Applications 7 (3) (2011), http://dx.doi.org/10.1145/2000486.2000492.
[21] S. Lyu, H. Farid, Steganalysis using higher-order image statistics, IEEE Transactions on Information Forensics and Security 1 (1) (2006) 111–119.
[22] P. Moulin, J. Liu, Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors, IEEE Transactions on Information Theory 45 (1999) 909–919.
[23] H. Ozer, B. Sankur, N. Memon, I. Avcibas, Detection of audio covert channels using statistical footprints of hidden messages, Digital Signal Processing 16 (4) (2006) 389–401.

[24] T. Pevny, J. Fridrich, Multi-class detector of current steganographic methods for JPEG format, IEEE Transactions on Information Forensics and Security 3 (2) (2008) 247–258.
[25] M. Qiao, A. Sung, Q. Liu, Steganalysis of MP3Stego, in: Proceedings of 22nd International Joint Conference on Neural Networks, 2009, pp. 2566–2571.
[26] M. Qiao, A. Sung, Q. Liu, Predicting embedding strength in audio steganography, in: Proceedings of 9th IEEE International Conference on Cognitive Informatics, pp. 925–930, 2010.
[27] M. Qiao, A. Sung, Q. Liu, B. Ribeiro, Locating information-hiding in MP3 audio, in: Proceedings of 3rd International Conference on Agents and Artificial Intelligence, vol. 1, 2011, pp. 504–507
[28] K. Sharifi, A. Leon-Garcia, Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video, IEEE Transactions on Circuits and Systems for Video Technology 5 (1995) 52–56.
[29] Y. Shi, C. Chen, W. Chen, A Markov process based approach to effective attacking JPEG steganography, Lecture Notes in Computer Sciences 437 (2007) 249–264.
[30] A. Srivastava, A. Lee, E.P. Simoncelli, S. Zhu, On advances in statistical modeling of natural images, Journal of Mathematical Imaging and Vision 18 (1) (2003) 17–33.
[31] V. Vapnik, Statistical Learning Theory, John Wiley, 2008.
[32] G. Winkler, Image Analysis, Random Fields and Dynamic Monte Carlo Methods, Springer, 1996.
[33] G. Wouwer, P. Scheunders, D. Dyck, Statistical texture characterization from discrete wavelet representations, IEEE Transactions on Image Processing 8 (1999) 592–598.
[34] W. Zeng, H. Ai, R. Hu, An algorithm of echo steganalysis based on power cepstrum and pattern classification, in: Proceedings of International Conference on Information and Automation, 2008, pp. 1667–1670.
[35] T. Zhang, W. Li, Y. Zhang, E. Zheng, X. Ping, Steganalysis of LSB matching based on statistical modeling of pixel difference distributions, Information Sciences 180 (23) (2010) 4685–4694.
[36] ISO/IEC JTC1/SC29/WG11, IS11172-3, Information Technology–Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s, Part 3: Audio, 1992.
[37] http://www.8hz.com/.
[38] http://www.petitcolas.net/fabien/steganography/.
[39] http://www.telossystems.com/techtalk/hosted/Brandenburg_mp3_aac.pdf/.