Available online at www.sciencedirect.com
SCIENCE DIRECT®
Digital Signal Processing 16 (2006) 389–401
Digital Signal Processing
www.elsevier.com/locate/dsp

Detection of audio covert channels using statistical footprints of hidden messages ☆

Hamza Özer [a,c], Bülent Sankur [a,*], Nasir Memon [b], İsmail Avcıbaş [d]

[a] *Department of Electrical and Electronics Engineering, Boğaziçi University, Bebek, İstanbul, Turkey*
[b] *Department of Computer and Information Science, Polytechnic University, Brooklyn, NY, USA*
[c] *National Research Institute of Electronics and Cryptology, TUBITAK, Gebze, Turkey*
[d] *Department of Electrical and Electronics Engineering, Uludağ University, Bursa, Turkey*

Available online 22 December 2005

Abstract

We address the problem of detecting the presence of hidden messages in audio. The detector is based on the characteristics of the denoised residuals of the audio file, which may consist of a mixture of speech and music data. A set of generalized moments of the audio signal is measured in terms of objective and perceptual quality measures. The detector discriminates between cover and stego files using a selected subset of features and an SVM classifier. The proposed scheme achieves on the average 88% discrimination performance on individual steganographic algorithms and 98.5% on individual watermarking algorithms. Between 75 and 90% discrimination performance is achieved in universal tests. Correct detection performance for individual embedding algorithms is roughly 90% when the detector can encounter any one in an ensemble of different embedding algorithms.

Keywords: Steganalysis; Watermarking; Feature selection; Support vector machine

1. Introduction

Information hiding in digital audio can be used for such diverse applications as proof of ownership, authentication, integrity, secret communication, broadcast monitoring, and event annotation. There are two well-known special cases of information hiding—digital watermarking and steganography.

In digital watermarking, the embedded signal depends on a secret key as the threat model includes a malicious adversary who will try to remove or invalidate the watermark. In the watermarking context we always assume that the adversary knows that the content is watermarked and, in principle, also knows the exact technique used for watermarking. The only thing she does not know is the secret key, which, for example, can be used to disperse the watermark locations in an image. In steganography, on the other hand, the focus is secret communication. The adversary does not and should not know that there is a secret message embedded in the content.

☆ This work was supported by TUBITAK Project 102E018 and by AFRL Contract F30602-03-C-0091.
* Corresponding author.
E-mail address: bulent.sankur@boun.edu.tr (B. Sankur).

1051-2004/$ – see front matter
doi:10.1016/j.dsp.2005.12.001

In fact, the modern formulation of steganography also goes by the name of the prisoner's problem. Here Alice and Bob are in prison, and a warden, Wendy, who will punish them at the first hint of any suspicious communication, examines all communication between them. Hence, Alice and Bob must trade seemingly inconspicuous messages that actually contain hidden messages. Specifically, in the general model for steganography, we have Alice wishing to send a secret message m to Bob. In order to do so, she "embeds" m into a cover-object c, to obtain the stego-object s. The stego-object s is then sent through the public channel.

There are two versions of the problem that are usually discussed—one where the warden is passive, and only observes messages, and the other where the warden is active and modifies messages in a limited manner to guard against hidden messages. Nevertheless, in either scenario, the most important issue in steganography is that the very presence of a hidden message must be concealed. In this context, "steganalysis" refers to the body of techniques that are designed to distinguish between cover-objects and stego-objects. Recently there have been a number of studies for detection of hidden message in images [1–6], but there are relatively very few papers on audio steganalysis [3,7,8]. Wesfeld and Pfitzmann proposed a steganalysis method [3] only for LSB based steganographic methods. In another study, Westfeld [8] addressed specifically the steganalysis of MP3 stega algorithm.

The underlying basis of most steganalysis techniques is that hidden messages leave statistical fingerprints on the cover object. In other words, stego-objects, though in principle perceptually very similar to cover objects, are assumed to be statistically distinguishable. Note that this is true irrespective of the specific algorithm used for embedding. A steganalysis technique that does not make any assumptions about the steganographic algorithm that it is trying to detect, is called a universal steganalysis technique.

In this paper, we propose a universal steganalysis scheme for audio data. In other words, we develop a technique that can distinguish between cover-objects and stego-objects, differentiating between "clear" audio data that do not contain any secret message and the ones that do carry a secret message. The proposed algorithm functions without any knowledge about the embedding technique used. One can also envision functions to this scheme other than steganalysis. For example, a web crawler that is looking for watermarked content can use it as a preprocessing stage to be followed by watermark extraction and decoding operations. However, the focus of this work is steganography and steganalysis and hence we concentrate solely on universal techniques. We do use watermarking techniques in our experiments as these can be viewed as active warden steganography techniques.

Among the few steganographic algorithms in the literature, one can quote the LSB embedding applied directly to audio samples [9] or, alternatively, to its transform coefficients [10,11]. Among the plethora of audio watermarking methods, one can mention the spread-spectrum techniques in the time or in an appropriate transform domain, as well as echo hiding, frequency hopping, and phase coding [12,13]. Spread-spectrum techniques add scaled and spread version of the message into the cover object directly in the time or frequency domain, possibly with perceptual weighting in order to guarantee inaudibility. In the frequency hopping method, the spread message bits are added to spectral coefficients in some random order. In echo hiding technique, scaled and delayed version of the cover object is embedded into cover object itself, where the amount of delay codes the information. Phase coding uses insensitivity of the human ear to the phase of the signal and it modulates the phase according to the embedding message. The steganalyzer we design is expected to be operative under any of these embedding methods.

The paper is organized as follows: In Section 2, the proposed audio steganalysis method is presented and its capability discussed. The features used for audio steganalysis as well as the feature selection scheme are discussed in Section 3. The results of extensive experiments conducted on both audio and speech signals are given in Section 4. Finally, conclusions are drawn in Section 5.

2. Audio steganalysis method

Data hiding techniques can be modeled as an additive noise process in the time or frequency domain. More specifically, consider the cover and stego signals $x(t)$ and $y(t)$, respectively, then their difference, $z(t) = y(t) - x(t)$, is an additive noise component, and the expression for the stego-signal becomes $y(t) = x(t) + z(t)$. Notice that this is true, whether the embedding technique is cover-signal independent (as in spread-spectrum methods) or $z(t)$ is cover-signal dependent (as, for example, echo hiding). In general, a noise removal procedure applied on the stego-signal can separate the cover signal from its embedded part. While the denoised signal would correspond to the original cover object, the difference between the input and output of the denoiser, "the removed disturbance" should be an estimate of the embedded signal. We note that the denoiser will yield a residual for any input signal, whether that signal contains

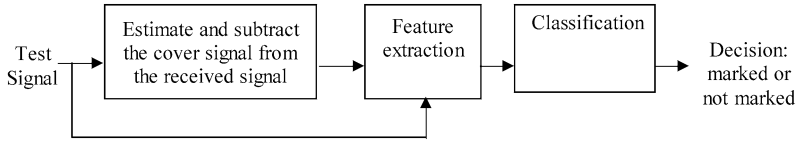

Fig. 1. Block diagram of the steganalysis method.

a hidden message or not. The idea of steganalysis lies on the conjecture that the denoising residual for cover audio signals differs statistically from that of the stego audio signals, on the basis of which the classifier is built.

The cover signal can be estimated by some denoising technique, such as wavelet shrinkage [14], independent component analysis (ICA) method, etc. [15], or, provided an appropriate probability model is available, by a decision-theoretic method such as maximum likelihood, maximum a posteriori estimate [16]. Our comparative study has shown that wavelet-based denoising, proposed by Donoho and Johnstone [14] actually works best. This wavelet-based denoising decomposes the given signal into its wavelet components, applies soft thresholding to the transform coefficients, and finally reconstructs the signal by inverse wavelet transform. We have used six-tap Daubechy filters and the maximum number of decomposition levels for wavelet transforming signal frames of 60 ms duration. The wavelet components, except for the coarsest level (low-pass components), are subjected to soft thresholding according to the formula $y' = \text{sign}(y)\text{u}[|y| - \text{threshold}]$, where $\text{sign}(\cdot)$ is the signum function and $\text{u}(\cdot)$ is the unit step function. The threshold value is calculated as a scaled version of the mean absolute difference (MAD) of the estimated noise. In other words, we use the formula $\text{threshold} = C \times \text{MAD}$, where $C = 3$ and MAD is the noise estimate given by the mean absolute deviation. It is estimated as the median of the absolute values of the processed coefficients.

An overview of the proposed steganalyzer is presented in Fig. 1. The first block estimates and removes the cover signal by a denoising algorithm, yielding an estimate of the possibly present hidden message signal. The second block extracts statistical features in order to discriminate between estimated embedded signals and spurious signals when a non-embedded signal is input to the denoiser. In the training stage, a subset of best discriminatory features is selected based on some scheme, such as analysis of variance (ANOVA) [17] and sequential forward floating search (SFFS) [18]. Finally, a two-class classifier, using the selected features, discriminates the test signal into stego- or cover signal classes. The algorithm is trained using various known data hiding algorithms and on several cover and stego-signals, then tested on unseen signals.

3. Features for steganalysis

To construct a set of features to discriminate between stego and cover signals, we resort to various speech and audio quality measures. We remark that each of these statistical speech/audio distance metrics is considered simply as a functional that converts an input signal into a measure that purportedly is sensitive to message embedding operations. The input to each functional is the denoised residual of the test signal.

While audio quality metrics have been developed in the literature for quality assessment of speech/audio signals, and to measure coding artifacts, in the steganalysis context, they are exploited solely to reveal the presence of hidden messages by measuring the statistical artifacts caused by such messages. In fact, audio distance or distortion measures can be interpreted as generalized moments of the test signal. In principle, it would be desirable to design test statistics geared just for steganalysis, perhaps by some sort of reverse engineering of the message embedding algorithm, as in [39]. However, the large variety of message-embedding techniques and the different modalities they use preclude the formulation of such measures, so that we revert to more universal distortion-based features. Among these features, we select a proper subset that achieves highest detection rate for a large variety of embedding methods and embedding strengths.

As an example, the discriminative potential of a selected subset of these features is illustrated in Fig. 2. Four of these distance metrics, namely perceptual audio quality measure (PAQM) [30], spectral phase distortion (SPD) [23], log-likelihood ratio (LLR) [24,25] and log-area ratio (LAR) [22], are computed for both a denoised stego-signal and the denoised cover-signal—the version of the signal that does not contain any embedded message. The plots in Fig. 2 have the sole purpose of illustrating the fact the measures considered do in fact differ between documents that contain hidden messages from those that do not contain. From the distance metrics, which are normalized to 1, it can be

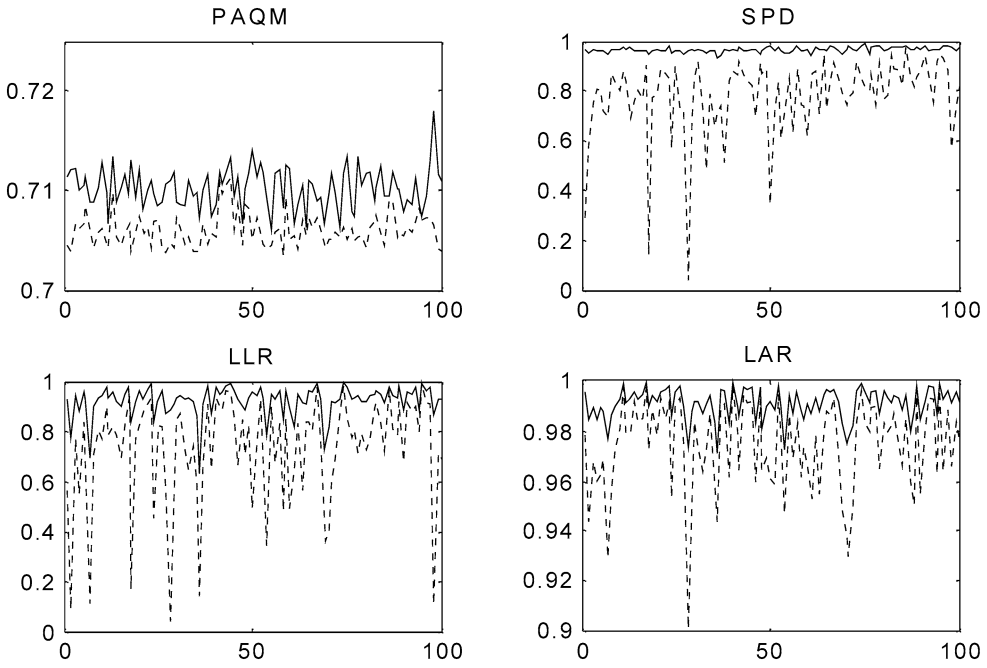

Fig. 2. Four distance metrics calculated from 100 utterances, the dotted lines are distance measures evaluated from stego-objects and the solid lines are from the cover objects. The abscissa denotes the index of utterances.

observed that the stego-signals and cover signals yield distinct scores over a string of utterances, as indexed by the abscissa.

3.1. Description of features

Table 1 summarizes the audio distortion measure used in the steganalysis, categorized into three groups of time-domain measures, frequency-domain measures and perceptual measures. The original signal (the cover document) is denoted as $x(i)$, $i = 1, \ldots, N$, while the distorted signal (the stego-document) as $y(i)$, $i = 1, \ldots, N$. The signal is segmented into frames of length K and the quality measures are first calculated over the K-sample segments, and then averaged over all M segments (note that $N = K \times M$). The segment sizes vary between 20 and 100 ms; hence the segments contain $K = 320$ to 1600 samples. This range was determined to yield a good solution on the basis of exhaustive search.

For a generic distortion measure D, the measure is computed by weighted averaging over the M audio segments. Thus the mth segment, $D(m)$, encompasses the samples $mK \leqslant i \leqslant (m+1)K$, $m = 0, \ldots, M-1$. Then the segmental distortion measures, $D(m)$, are averaged over the whole audio record, that is

$$D = \frac{\sum_{m=0}^{M-1} w(m) D(m)}{\sum_{m=0}^{M-1} w(m)},$$

where M is the total number of frames, and $w(m)$ is a weight associated with the mth frame. The weighting could, for example, be the energy in the reference frame. The frame durations were established experimentally to yield best classification performance individually per feature. The frame sizes for the individual measures are: BSD: 60 ms, CD: 20 ms, COSH: 100 ms, CZD: 40 ms, EMBSD: 20 ms, IS: 100 ms, LAR: 60 ms, LLR: 60 ms, MBSD: 80 ms, MNB1: 60 ms, MNB2: 60 ms, PAQM: 32 ms, PSQM: 32 ms, SNRseg: 20 ms, SPD: 40 ms, SPM: 20 ms, STFRT: 60 ms, WSS: 40 ms. In the distortion measures in Table 1, the expression for only the segmental distortion will be given, and the weighed averaging will be implicitly assumed.

More details about the metrics given in Table 1 can be found in [43] where we use the following notation for the signal x: the magnitude spectrum, $S_x(\omega)$; the phase spectrum, $\theta_x(\omega)$; the autocorrelation matrix, R_x; the linear prediction coefficients, $a_x(k)$; the autocorrelation model of autoregressive modeling (10th-order all-pole model) was

Table 1
The audio distortion measures tested for the design of the steganalyzer

Acronym	Audio distortion measures	
	Perceptual-domain measures	
BSD	Bark spectral distortion is extended by using 25 critical bands covering up to 15.5 kHz	$\mathrm{BSD} = \sum_{i=1}^{C} \big[B_x(i) - B_y(i)\big]^2$, C is the number of critical bands
MBSD	Modified Bark spectral distortion [28]	$\mathrm{MBSD} = \sum_{k=1}^{C} M(i)\big\lvert S_x(i) - S_y(i)\big\rvert$, where $M(i)$ is 1 if the difference of Bark bands i exceed a preset threshold, $\mathrm{Th}(i)$, and zero otherwise
EMBSD	Enhanced modified bark spectral distortion [29]	$\mathrm{MBSD} = \sum_{i=1}^{15} \max\big\{\big\lvert S_x(i) - S_y(i)\big\rvert - \mathrm{Th}(i), 0\big\}\big\lvert S_x(i) - S_y(i)\big\rvert$
PSQM	Perceptual speech quality measure [30]	Optimized for human auditory system for speech
PAQM	Perceptual audio quality measure [26]	Optimized for human auditory system for audio
MNB1 MNB2	Measuring normalizing block 1, measuring normalizing block 2 [34]	Based on cognition module for estimating speech distortion (computed through different time–frequency structures)
WSS	Weighted slope spectral distance [32,33]	$\mathrm{WSSD} = \sum_{k=1}^{36} w(k)\big\{\big[X(k+1) - X(k)\big] - \big[Y(k+1) - Y(k)\big]\big\}^2$
	Frequency-domain measures	
LLR	Log-likelihood ratio [24,25]	$\mathrm{LLR} = \log\left(\dfrac{a_x^T R_y a_x}{a_y^T R_y a_y}\right)$
LAR	Log-area ratio [22]	Based on PARCOR (partial correlation) coefficients
ISD	Itakura–Saito distance [40]	$\mathrm{IS} = \int_{-\pi}^{\pi} \left(\log \dfrac{Y(w)}{X(w)} + \dfrac{X(w)}{Y(w)} - 1\right) \dfrac{dw}{2\pi}$
COSH	COSH distance [41]	$\mathrm{COSH} = \int_{-\pi}^{\pi} \left[\dfrac{1}{2}\left(\dfrac{Y(w)}{X(w)} + \dfrac{X(w)}{Y(w)}\right) - 1\right] \dfrac{dw}{2\pi}$ (symmetric version of the Itakura–Saito distance)
CD	Cepstral distance [26]	$\left[\big[c_x(0) - c_y(0)\big]^2 + 2\sum_{k=1}^{L}\big[c_x(k) - c_y(k)\big]^2\right]^{1/2}$
STFRT	Short-time Fourier–Radon transform distance	$R(\rho,\theta) = \int_x \int_y S(\tau,\varpi)\delta(\tau\cos\theta + \varpi\sin\theta - \rho)\,d\tau\,d\varpi$ (some details given in Appendix A)
SPD	Spectral phase distortion [23]	$\mathrm{SPD} = \dfrac{1}{K}\sum_{w=1}^{K}\big\lvert\theta_x(w) - \theta_y(w)\big\rvert^2$
SPM	Spectral phase-magnitude distortion [23]	$\mathrm{SPM} = \dfrac{1}{K}\left(\lambda\sum_{w=1}^{K}\big\lvert\theta_x(w) - \theta_y(w)\big\rvert^2 + (1-\lambda)\sum_{w=1}^{K}\big\lvert\lvert X(w)\rvert - \lvert Y(w)\rvert\big\rvert^2\right)$, $\lambda = 0.025$
	Time-domain measures	
SNR	Signal-to-noise ratio	$\mathrm{SNR} = 10\log_{10}\dfrac{\sum_{i=1}^{N} x^2(i)}{\sum_{i=1}^{N}(x(i) - y(i))^2}$
SNRseg	Segmental signal-to-noise ratio [22]	$\mathrm{SNRseg} = \dfrac{10}{M}\sum_{m=0}^{M-1}\log_{10}\dfrac{\sum_{i=Km}^{Km+K-1} x^2(i)}{\sum_{i=Km}^{Km+K-1}(x(i) - y(i))^2}$
CZD	Czenakowski distance [23]	$C = \dfrac{1}{K}\sum_{i=0}^{K-1}\left(1 - \dfrac{2\min(x(i), y(i))}{x(i) + y(i)}\right)$

Table 2
The discriminatory features selected, per embedding method by the SFFS method when the (SVM method is used for classification (S denotes feature selected for speech, A for audio, and S&A for both audio and speech)

Methods	SNR	SNRs	LLR	LAR	COSH	CDM	ISD	BSD	MBSD	EBSD	WSSD	PAQM	PSQM	MNB1	MNB2	CZD	SP	SPM	STFRT
DSSS		S	A	A						S&A									A
FHSS		S	A	S					A									S	A
ECHO		S&A	A	A	A		A	A	A			A					S		
DCTwHAS		S&A	S&A	A	S	S	A	S	A			A			S	S			
STEGA				A					A	S&A							S		
STOOLS				S	S&A		S			S&A		A			S		A		
StegHide		S	S							A		S&A				S	A		
Hide4PGP				S&A	S					S&A		S				A			
Watermarking		S&A	S&A	S			S&A					A		S	S		S	A	A
Steganographic				S	A		S		A	S&A							A		

used to find the filter coefficient [22]; the cepstral coefficients, $c_x(k)$; the Bark spectrum, $B_x(i)$, and the short-time Fourier spectrum, $S_x(\tau, \varpi)$. Similarly, for the stego-signal, we use the notation $S_y(\omega)$, $\theta_y(\omega)$, R_y, $a_y(k)$, $c_y(k)$, $B_y(i)$, and $S_y(\tau, \varpi)$. C is the number of critical bands.

3.2. Selection of features

For feature selection purposes we have used two approaches, which are analysis of variance (ANOVA) [17] and sequential forward floating search method (SFFS) [18,19], each coupled with two classification methods, namely linear regression (LR) and support vector machines (SVM). After having tried all the four combinations of ANOVA-LR, ANOVA-SVM, SFFS-LR, and SFFS-SVM, we found out that the features chosen with the SFFS approach coupled with the SVM classifier proved to be superior.

The SFFS method analyzes the features in ensembles and goes through cycles of elimination of redundant ones and enrollment of new ones. Pudil [18] claims that the best feature set is constructed by adding to and/or by removing from current set of features until there no more performance improvement is possible. The SFFS procedure can be described as follows:

1. Choose the best two features from the initial set of K features, which is the pair yielding the best classification result.
2. Add the most significant feature from the remaining ones, where the selection is made on the basis of the feature that contributes most to the classification result when all together are considered.
3. Determine the least significant feature from the selected set by conditionally removing features one-by-one; checking if the removal of any one improves or reduces the classification result: if it improves, remove this feature and go to step 3, else do not remove this feature and go to step 2.
4. Stop when the number of selected features equals the number of features required; otherwise go to step 2.

The chosen features are shown in Table 2. In addition, we performed SFFS tests, as also given in Table 2, for the ensemble of watermarking and steganographic techniques, in other words, when the signal could have been marked by any of the four watermarking methods or by any of the four steganographic methods considered in this work. Several observations are in order:

- Passive warden techniques necessitate fewer features as compared to active warden techniques.
- Speech signals use a smaller number of features as compared to audio segments.
- The features in most demand are LAR (log area ratio), LLR (log likelihood ratio), ISD (Itakura–Saito distance), PAQM (perceptual audio quality measure), and SPD (spectral phase distortion), as they have been selected most frequently across the embedding techniques. On the hand, the features in least demand are time-domain measures, in addition to PSQM and WSSD [21,27,31].
- The presence of some of the features can be interpreted as follows: LLR, LAR, and ISD features are also the favored features for speech recognition. PAQM feature is already the most prominent feature for speech quality

measurement in coding experiments [38]. As for the SPD spectral-phase feature, it captures waveform phase perturbations due to embedding while the others, like ISD, LAR, LLR, and PAQM are concentrating on the spectral magnitude properties.

4. Experimental results

We performed steganalysis experiments over eight different algorithms, four of which were watermarking techniques and remaining four were steganographic techniques. The watermarking techniques used were direct-sequence spread spectrum (DSSS) [11], frequency hopping with spread spectrum (FHSS) [12], frequency masking technique with DCT (DCTwHAS) [12], and echo watermarking [11]. For all of these watermarking techniques the data embedding strength is chosen just below the perceptual threshold. Notice that some watermarking techniques, such as echo hiding and frequency masking techniques (e.g., DCTwHAS watermarking), end up in significantly higher mean-square distortion as compared to the DSSS, although their subjective qualities are identical. To determine the objective distortion we use the signal-to-watermark ratio, which is defined as

$$\mathrm{SWR} = \frac{\sum x^2(n)}{\sum (x(n) - y(n))^2}.$$

Moreover we adjusted the embedding strength based on a perceptual evaluation based measure, PAQM. PAQM is known to correlate well with the mean opinion score [30], which is the most common subjective quality measure. Consequently, for embedding strengths that result in distortions just below the audible level, in other words for the PAQM value of 0.035, the resulting SWR figure are: DSSS: 38 dB, FHSS: 34 dB, DCTwHAS: 20 dB, ECHO: 18 dB.

The steganographic methods we used are Steganos Security Suite 4.13 [10], S-Tools v4.0 [9], StegHide v0.5.1 [36], and Hide4PGP v2.0 [37]. These tools were selected on the basis of being popular methods and also with readily available software. In the first three parts of the experiments, the highest allowed capacity was embedded into the cover signal. In the last experiment, the tests were done with highest allowed capacity and half of this rate, in order to assess the effect of embedding rate.

The OSU_SVM Matlab toolbox [35] was used for SVM classifier, which employed radial basis functions as kernel type. The parameters C and gamma were optimized by exhaustive search to be 100 and 4, respectively. They were chosen to yield a 1.0% false-positive rate.

The algorithm was tested separately for three sets of data, which were speech, pure instrumental audio and music records, in addition to the ensemble of these sources. The speech segments have durations of three to four seconds, sampled at 16 kHz, and recorded in acoustically shielded medium. In the audio repertoire, three different instrumental sources and three different song records are used. The instrumental records are obtained from sound quality assessment material (SQAM) [20]. The music records are taken from the songs of famous music groups U2, and Rolling Stones. The songs are ‘One’ (a slow song), ‘Even Better than the Real Thing’ of U2, and ‘Paint It, Black’ of Rolling Stones. The audio records (songs and instrumentals) are downsampled from their 44.1 kHz version to 16 kHz to have the same sampling rate as speech. Furthermore they are separated into 5-s long segments, and half of them are used for training and the remaining half for testing. One advantage of splitting a long audio object into smaller segments is that, it enables us to pursue sequential testing and accumulation of scores (cover object versus stego object likelihoods) over the segments of the same record. In other words we can implement decision fusion over the 5-s segments. There are overall 200 speech record, 180 instrumental excerpts and 284 music excerpts. In all experiments the experimental procedure consisted of embedding messages to all available cover signals, randomly selecting half of the set of the stego and cover signals for training, leaving the other 50% for testing phase.

4.1. Design of experiments

Simulation experiments were designed and conducted with the following goals in mind:

(a) Determine the best combination of feature selection (ANOVA, SFFS) and signal classification (LR, SVM) methods.

(b) Determine the detection performance for individual embedding methods as well as in their ensembles, and find the performance differential as one moves from a non-universal (specialized for single known method) to a universal method (trained multiple methods of embedding).
(c) Determine the dependence on the cover material, that is, speech and audio as well as on the genre of audio.
(d) Determine the effect on the performance of the strength of embedding in the case of watermarking techniques and of the capacity used in the case of steganographic techniques.

4.2. The feature selection and detection methods

As mentioned in Section 3.2 we have considered the four combinations afforded by the two feature-selection and two detection methods. It has been observed that, in the overwhelming number of cases the SFFS feature selection method is superior to the ANOVA method, independently of whether linear regression or SVM classifier is used, and independently of whether speech or music material is used. Table 3 displays the results only for speech data, while quite similar results have been obtained with music. Therefore, for the classification results presented in the sequel, e.g., experiments with heterogeneous data, only SVM classification results are given.

4.3. The performance of the steganalyzer for single and multiple embedding methods

We investigate the performance differential between the cases when the steganalyzer is trained for single known method and the universal case where multiple methods of embedding are involved. The scores for the individual methods were given in Table 3. In Table 4 we give the average of the individual scores, and compare them with the detection scores of detectors trained for the pool of steganographic and watermarking methods separately and also together. As can be expected, the success rate is somewhat lower for the universal case. Here also the tests done with speech data are given. Similar performance variation occurs for other types of data.

4.3.1. Homogeneous methods (individual methods)

The experimental results for individual steganographic embedding algorithms indicate that the average success rate is 87.8%. For watermarking methods, however, the success rate is 98.5%. The steganographic methods StegHide

Table 3
The probability of misdetection (PM), and probability false alarm (PF) for individual methods, with two distinct classifier and two distinct feature selection methods

Methods	LR classifier				SVM classifier			
	ANOVA features		SFFS features		ANOVA features		SFFS features	
	PM	PF	PM	PF	PM	PF	PM	PF
DSSS	0	0	0	0	0	0	0	0
FHSS	4	0	0	0	2	0	0	0
ECHO	0	0	0	0	0	0	0	0
DCTwHAS	16	8	8	6	20	12	10	2
STEGA	10	0	8	0	2	2	2	0
STOOLS	8	12	4	6	6	6	4	6
STEGHIDE	22	32	18	26	22	30	16	22
HIDE4PGP	24	32	22	28	22	32	20	28

Table 4
Dependence of the performance of steganalyzer on the pooling of methods: comparison of the universal and the individual cases

Assembling of methods	Average of the scores of individual methods		Universal scores	
	PM	PF	PM	PF
Watermarking methods	2.5	0.5	5.0	9.0
Steganographic methods	10.5	14	12.0	15.5
Watermarking and steganographic methods	6.5	7.25	18.2	20.3

and Hide4PGP, have relatively lowest success rates (83 and 76%, respectively). Among the watermarking methods the DCTwHAS has the lowest success rate (96%), possibly due to the fact that the method uses frequency masking according to human auditory system, making it hard to track.

4.3.2. Heterogeneous active methods vs heterogeneous passive methods (semi-universal)

The ensemble of watermarking and the ensemble of steganographic methods are first pooled separately. In other words, the receiver does not know with which of the steganographic (or watermarking) methods the audio document is marked with, nor if any embedding at all has taken place. The average performance of steganographic methods is 86.3% as presented in Table 4. When the ensemble of watermarking methods is tested, the success rate is 93%.

4.3.3. Heterogeneous methods (universal)

Also the steganalyzer is designed to the detect the steganographic content, we tested with all eight steganographic and watermarking content together in order to give the scope of the detection performance. When all the watermarking and steganographic methods are tested together, the score drops down to 80.63%, a lower but still useful detection performance. In Fig. 3, the success rates are presented in a chart graphic of individual methods and of their ensembles.

4.4. The dependence of the steganalyzer on the cover material

We investigated the performance dependence of the steganalyzer on the type of document in which data hiding takes place, that is, speech and audio. Table 5 presents comparatively the steganalysis performances for different sources (speech, bass, soprano, rock, etc.).

Simulation experiments indicate that the average success rates for the speech utterances is 93.1%, for pure instrumental records it is 95.3%, and for song records is 82.7%. It can be observed that the detection performance for

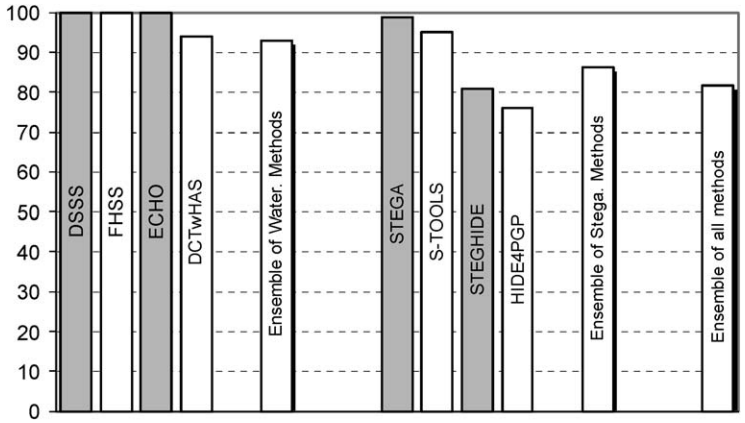

Fig. 3. Bar charts of the correct detection performance of the steganalyzer.

Table 5
Dependence of the performance of steganalyzer on audio content

Methods	Speech records		Pure instrument records		Music records	
	PM	PF	PM	PF	PM	PF
DSSS	0.0	0.0	0.0	4.4	9.8	14.1
FHSS	0.0	0.0	0.0	4.4	1.4	2.8
ECHO	0.0	0.0	13.3	4.4	16.9	20.1
DCTwHAS	10.0	2.0	6.7	6.7	29.5	16.9
STEGA	2.0	0.0	0.0	0.0	12.6	14.1
STOOLS	4.0	6.0	2.2	4.4	26.8	22.5
STEGHIDE	16.0	22.0	6.7	8.9	26.7	26.7
HIDE4PGP	20.0	28.0	6.7	6.7	16.9	19.7
Ensemble	18.2	20.3	20.2	22.3	26.2	24.9

song data, based on the 5-s segments observations, decreases somewhat. This drop could possibly be due to features selected only using solo instrumental and speech training data. However, these scores can be improved by decision fusion over consecutive segments. The above experiments also include investigation on the effects of audio genre, which is yet preliminary since few different audio sources have been employed. The indication is that the ensemble detector is most successful with speech source, followed by pure instruments and then by complex music.

4.5. Effect of the embedding strength and of the steganographic capacity

Finally we set experiments to determine the dependence of the performance upon the size of the hidden data. The steganographic methods are tested with two distinct embedding rates. In one case, 100% of the allowed capacity is used for embedding; in the other case 50% of the allowed capacity is used. Table 6b shows that the success rates do not vary significantly between 100 and 50% capacity usages in the case of Steganos, StegHide, and Hide4PGP methods. However, the success rate drops noticeably in the case of S-tools. Similar results have been reported by the method of Westfeld and Pfitzmann [3], which starts failing when less than 99.5% of the capacity is employed. The plot of average detection performance of the S-tools method versus percentage of used capacity is given in Fig. 4b. Note that in Table 6 we have taken 100 records for the tests, and therefore the actual percentages appear as integers. We have opted to give in Table 6 only results for speech, since the deterioration of detection is commensurate in the case of pure instruments and of complex music. Finally the slight performance differential between speech and pure

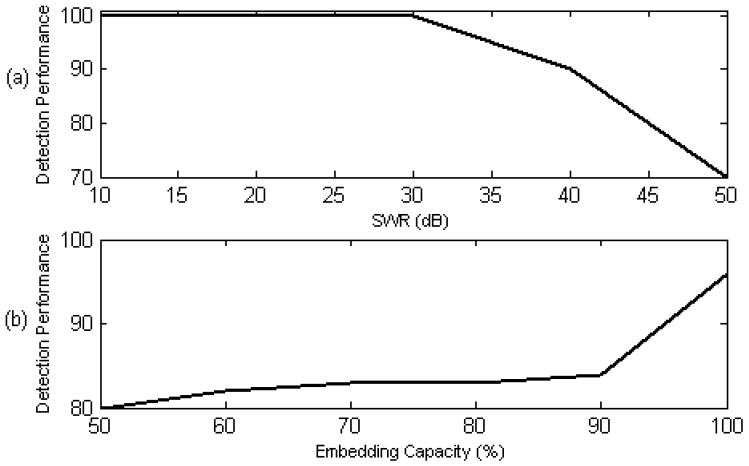

Fig. 4. (a) Dependence of steganalysis performance on the DSSS watermarking strength. (b) Dependence of steganalysis performance on the embedding capacity of the S-tools steganographic method.

Table 6
The results of experiments to determine the impact of (a) embedding strength in active methods, (b) of capacity usage in passive methods

Methods	20 dB SWR		30 dB SWR		40 dB SWR		100% of allowed capacity		50% of allowed capacity	
	PM	PF	PM	PF	PM	PF	PM	PF	PM	PF
(a) Active										
DSSS	0.0	0.0	0.0	2.0	10.0	26.0				
FHSS	0.0	0.0	0.0	0.0	4.0	8.0				
ECHO	8.0	14.0	16.0	26.0	24.0	44.0				
DCTwHAS	9.0	14.0	16.02	26.0	36.0	44.0				
(b) Passive										
STEGA							2.0	0.0	0.0	0.0
STOOLS							4.0	6.0	20.0	22.0
STEGHIDE							16.0	22.0	20.0	30.0
HIDE4PGP							20.0	28.0	24.0	30.0

instruments can be explained based on the fact that data embedding causes a larger distortion in the latter signal, leading to somewhat easier detectability.

For watermarking methods we vary the embedding strength to perform similar experiments as above. The signal to watermark ratio (SWR) is allowed to vary between 20 and 40 dB. It is known that the perceptual threshold is at about 36 dB when Gaussian noise is added. On the other hand, with the Echo and DCTwHAS watermarking methods, a much stronger watermark can be embedded and yet the distortion remains below the perception threshold. In the first method, the presence of a short delay echo is not disturbing, while in the DCTwHAS case the higher frequencies where the watermark have higher masking effect are not perceived. The results are reported in Table 6a where it is shown that the steganalyzer works well for the DSSS and FHSS methods over a large SWR interval. For the Echo and DCTwHAS methods the SWR must be around the 20 dB, which is still inaudible. The plot of average detection performance of the DSSS method versus embedding strength, measured in terms of the signal to watermark ratio, is given in Fig. 4a.

We also investigated the MP3Stego algorithm [42]. This steganographic algorithm is different than other methods in that once decoded, the stego-message is removed from the .wav file, as compared to other schemes where the stego-message persists within the audio file. We conjectured, however, that the compression styles of the same audio file with and without a message embedded would differ. We therefore considered the compressed-and-uncompressed .wav files with the applications of MP3Stego and of 8 Hz MP3 (which is the used compression technique in MP3Stego) and extracted discriminatory features. The compression rate was 128 Kb/s. We found out that, even in this case, we were capable of detecting the presence of MP3Stego, albeit with a lower performance. The performance figures were PM: 16% and PF: 34%. Another interesting result was that there was not much of a detection performance differential between the two cover materials, that is, music or speech.

5. Conclusions

We have presented an audio steganalysis algorithm based on the generalized moments of the denoising residuals of speech and audio signals. The denoising residual is intended as an estimate of the potentially embedded signal. The generalized moments are obtained via selected speech and audio quality measures. These features are selected via the sequential forward floating search method on the basis of yielding the best detection results. Both passive-warden methods (steganography) and active-warden methods (watermarking) are investigated.

If the steganographic embedding method is known ahead, the steganalyzer yields an average success rate between 80 and 90%. With active warden techniques (watermarking) detection success is between 90 and 95%. More realistically the embedding method would not be known. If the embedding method can be guessed to be of the steganographic or watermarking variety, the respective scores become 86 and 93%. Finally, in the absence of any knowledge, that is if we are uncertain which of the eight watermarking or steganographic methods has been used, the correct detection probability becomes 80.6%. Some content dependency has been observed; in fact, the steganalyzer is more successful with speech cover material as compared to the tested music varieties. Finally, there is a critical threshold below which strength steganalysis of watermarking methods and below which capacity steganalysis of steganographic method is not possible.

We have presented the results of a universal audio steganalysis techniques, where the volume of embedded data was proportional to the file size. Another universal steganalyzer is presented in [7], where the volume of data is given in absolute terms. The results are encouraging but still need considerable improvement. Universal steganalysis techniques for example have been reported with higher performance than what we report here with audio. However, we believe that significant improvements in audio steganalysis can be achieved with feature selection targeted at artifacts caused by steganographic embedding. Our future work will focus on identifying such features. Further, methods to estimate original signal statistics need to be developed to make the steganalysis technique content independent.

Appendix A. Short-time Fourier–Radon transform measure (STFRT)

Given a short time Fourier transform (STFT) of a signal, its time projection gives us the magnitude spectrum while its frequency projection yields the magnitude of the signal itself. More generally, we can obtain the Radon transform of the STFT mass. We define the mean-square distance of Radon transforms of the STFT of two signals as a new objective audio distortion measure.

Recall that the Radon transform of a bivariate function $f(x, y)$ is defined as the projection of the mass along a line defined by its distance ρ from the origin and by its angle of inclination θ,

$$R(\rho, \theta) = \int_x \int_y f(x, y)\delta(x\cos\theta + y\sin\theta - \rho)\,dx\,dy,$$

where the delta function constrains integration only over the line. The range of θ is between 0 and π. In our experiments, projections on 90 different θ angles were used with a resolution of 2 degrees. The range of ρ is chosen 1 to the hypotenuse distance of STFT matrix. By computing the Radon transform of the STFT of the signal we get different views of evolutionary spectrum. The disturbance caused by message hiding in the signal causes changes in the STFT, which can be monitored by the Radon transform.

References

[1] I. Avcıbaş, N. Memon, B. Sankur, Steganalysis using image quality metrics, IEEE Trans. Image Process. 12 (2) (2003) 221–229.
[2] J. Fridrich, M. Goljan, R. Du, Reliable detection of LSB steganography in color and grayscale images, in: Proc. ACM Workshop on Multimedia and Security, Ottawa, CA, October 5, 2001, pp. 27–30.
[3] A. Westfeld, A. Pfitzmann, Attacks on steganographic systems, in: Information Hiding, LNCS, vol. 1768, Springer-Verlag, Heidelberg, 1999, pp. 61–66.
[4] N.F. Johnson, S. Jajodia, Steganalysis of images created using current steganography software, in: D. Aucsmith (Ed.), Information Hiding, LNCS, vol. 1525, Springer-Verlag, Berlin, 1998, pp. 32–47.
[5] J. Fridrich, Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes, in: Proc. 6th Information Hiding Workshop, Toronto, Canada, May 2004.
[6] S. Lyu, H. Farid, Steganalysis using color wavelet statistics and one-class support vector machines, in: SPIE Symposium on Electronics Imaging, San Jose, CA, 2004.
[7] M.K. Johnson, S. Lyu, H. Farid, Steganalysis of recorded speech, in: SPIE Symposium on Electronics Imaging, San Jose, CA, 2005.
[8] A. Westfeld, Detecting low embedding rates, in: F.A.P. Petitcolas (Ed.), Information Hiding. 5th International Workshop, IH 2002 Noordwijkerhout, The Netherlands, October 7–9, 2002, Springer-Verlag, Berlin, 2003, pp. 324–339.
[9] N.F. Johnson, S. Katzenbeisser, A survey of steganographic techniques, in: S. Katzenbeisser, F. Petitcolas (Eds.), Information Hiding, Artech House, Norwood, MA, 2000, pp. 43–78.
[10] Steganography software for Windows, http://members.tripod.com/steganography/stego/s-tools4.html.
[11] Steganos, http://www.steganos.com.
[12] W. Bender, D. Gruhl, N. Morimoto, A. Lu, Techniques for data hiding, IBM Syst. J. 35 (3–4) (1996) 313–336.
[13] I. Cox, J. Kilian, F.T. Leighton, T. Shamoon, Secure spread spectrum watermarking for multimedia, IEEE Trans. Image Process. 6 (12) (1997) 1673–1687.
[14] R.R. Coifman, D.L. Donoho, Translation-invariant denoising, in: A. Antoniadis, G. Oppenheim (Eds.), Wavelets and Statistics, Springer-Verlag, San Diego, 1995.
[15] A. Hyvarinen, P. Hoyar, E. Oja, Sparse code shrinkage for image denoising, in: Proc. IEEE Int. Joing. Conf. of Neural Networks, Anchorage, Alaska, pp. 859–864.
[16] S. Voloshynovsky, S. Pereira, V. Iquise, T. Pun, Attack modeling: Towards a second generation watermarking benchmark, Signal Process. 81 (2001) 1177–1214.
[17] X. Rencher, Methods of Multivariate Data Analysis, Wiley, New York, 1995.
[18] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recogn. Lett. 15 (1994) 1119–1125.
[19] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[20] Audio records, http://sound.media.mit.edu/mpeg4/audio/sqam/.
[21] B.J. McDermott, C. Scaglia, D.J. Goodman, Perceptual and objective evaluation of speech processed by adaptive differential PCM, in: IEEE ICASSP, Tulsa, April 1978, pp. 581–585.
[22] S.R. Quackenbush, T.P. Barnwell III, M.A. Clements, Objective Measures of Speech Quality, Prentice Hall, Englewood Cliffs, NJ, 1988.
[23] I. Avcıbaş, B. Sankur, K. Sayood, Statistical evaluation of image quality metrics, J. Electron. Imag. 11 (2) (2002) 206–223.
[24] F. Itakura, Minimum prediction residual principle applied to speech recognition, IEEE Trans. Acoust. Speech Signal Process. ASSP-23 (1) (1975) 67–72.
[25] B.H. Juang, On using the Itakura–Saito measure for speech coder performance evaluation, AT&T Bell Lab. Tech. J. 63 (8) (1984) 1477–1498.
[26] N. Kitawaki, H. Nagabuchi, K. Itoh, Objective quality evaluation for low-bit-rate speech coding systems, IEEE J. Select. Areas Commun. 6 (1998) 242–248.
[27] S. Wang, A. Sekey, A. Gersho, An objective measure for predicting subjective quality of speech coders, IEEE J. Select. Areas Commun. 10 (1992) 819–829.
[28] W. Yang, M. Dixon, R. Yantorno, A modified bark spectral distortion measure which uses noise masking threshold, in: IEEE Speech Coding Workshop, Pocono Manor, 1997, pp. 55–56.
[29] E. Zwicker , H. Fastl, Psychoacoustics Facts and Models, Springer-Verlag, 1990.
[30] J.G. Beerends, J.A. Stemerdink, A perceptual audio quality measure based on a psychoacoustics sound representation, J. Audio Eng. Soc. 40 (1992) 963–978.

[31] J.G. Beerends, J.A. Stemerdink, A perceptual speech quality measure based on a psychoacoustic sound representation, J. Audio Eng. Soc. 42 (1994) 115–123.
[32] D.H. Klatt, A digital filter bank for spectral matching, in: Proc. 1976 IEEE ICASSP, April 1976, pp. 573–576.
[33] D.H. Klatt, Prediction of perceived phonetic distance from critical-band spectra: A first step, in: Proc. 1982 IEEE ICASSP, Paris, May 1982, pp. 1278–1281.
[34] S. Voran, Objective estimation of perceived speech quality. I. Development of the measuring normalizing block technique, IEEE Trans. Speech Audio Process. 7 (4) (1999) 371–382.
[35] http://eewww.eng.ohio-state.edu/~maj/osu_svm/.
[36] http://steghide.sourceforge.net/.
[37] http://www.heinz-repp.onlinehome.de/Hide4PGP.htm.
[38] J.D. Gordy, L.T. Bruton, Performance evaluation of digital audio watermarking algorithms, in: IEEE International Midwest Symposium on Circuits and Systems, Michigan, 2000.
[39] R. Böhme, A. Westfeld, Statistical characterisation of MP3 encoders for steganalysis, in: ACM Multimedia and Security Workshop, Magdeburg, Germany, 2004.
[40] F. Itakura, S. Saito, Analysis synthesis telephony based on the maximum likelihood method, in: Proc. 6th Int. Congr. Acoust., Tokyo, Japan, 1968, pp. C-17–C-20.
[41] A.H. Gray Jr., J.D. Markel, Distance measures for speech processing, IEEE Trans. Acoust. Speech Signal Process. ASSP-24 (1976) 380–391.
[42] http://www.petitcolas.net/fabien/steganography/mp3stego.
[43] H. Ozer, İ. Avcıbaş, B. Sankur, N. Memon, Steganalysis of audio based on audio quality metrics, in: SPIE Electronic Imaging Conf. on Security and Watermarking of Multimedia Contents, vol. V, Santa Clara, January 20–24, 2003, pp. 55–66.

Hamza Özer was born in Erzincan, Turkey, on August 10, 1973. He received the B.Sc., M.Sc., and Ph.D. degrees, all in electrical and electronics engineering, from Middle East Technical University, Başkent University, and Boğaziçi University, in 1996, 1998, and 2005, respectively. From 1996 to 1999 he was with the Department of Electrical and Electronics Engineering at the Başkent University as a research assistant. Since June 1999 he is a senior researcher at the National Research Institute of Electronics and Cryptology (UEKAE). His research interests include signal processing and applications, data hiding, audio watermarking, robust audio hashing, software defined radio, time-frequency signal analysis, speech processing, image processing, development of test, and measurement plan and setup.

Bülent Sankur has received his B.Sc. degree in electrical engineering at Robert College, Istanbul, Turkey, and completed his M.Sc. and Ph.D. degrees at Rensselaer Polytechnic Institute, New York, USA. He has been teaching at Boğaziçi (Bosphorus) University in the Department of Electric and Electronic Enginering. His research interests include digital signal processing, image and video compression, biometry, cognition, and multimedia systems. Dr. Sankur has held visiting positions at University of Ottawa, Technical University of Delft, and Ecole Nationale Supérieure des Télécommications, Paris. He was the chairman of ICT '96: International Conference on Telecommunications and EUSIPCO '05: The European Conference on Signal Processing, as well as as technical chairman of ICASSP '00.

Nasir Memon is a Professor in the Computer Science Department, Polytechnic University, New York, USA. His research interests include data compression, computer and network security and multimedia communication, computing, and security. He has published more than 200 articles in journals and conference proceedings. He was a Visiting Faculty at Hewlett–Packard Research Labs during the academic year 1997–1998. He has won several awards including the NSF CAREER Award and the Jacobs Excellence in Education Award. He was an Associate Editor for the IEEE Transactions on Image Processing from 1999 to 2002. He is currently an Associate Editor for the IEEE Transactions on Information Security and Forensics, ACM Multimedia Systems Journal, and the Journal of Electronic Imaging.

İsmail Avcıbaş received the B.Sc. and M.Sc. degrees in electronics engineering from Uludağ University, Turkey, in 1992 and 1994, and the Ph.D. degree in electrical and electronics engineering from Boğaziçi University, Turkey, in 2001. He is currently with the Electronics Engineering Department of Uludağ University as a lecturer. His research interests include image processing, data compression, information hiding, and multimedia communications.