Digital Signal Processing 51 (2016) 133–141

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Audio steganalysis based on reversed psychoacoustic model of human hearing

Hamzeh Ghasemzadeh [a,*], Mehdi Tajik Khass [b], Meisam Khalil Arjmandi [c]

[a] Faculty of Electrical Engineering, Islamic Azad University, Damavand Branch, Tehran, Iran
[b] Faculty of Electrical and Computer Engineering, Tabriz University, Tabriz, Iran
[c] Department of Communicative Sciences and Disorders, Michigan State University, MI, USA

ARTICLE INFO

Article history:
Available online 21 January 2016

Keywords:
Audio steganalysis
Universal steganalysis
Audio steganography
Mel frequency cepstrum coefficients
Human auditory system

ABSTRACT

During the last decade, audio information hiding has attracted lots of attention due to its ability to provide a covert communication channel. On the other hand, various audio steganalysis schemes have been developed to detect the presence of any secret messages. Basically, audio steganography methods attempt to hide their messages in areas of time or frequency domains where human auditory system (HAS) does not perceive. Considering this fact, we propose a reliable audio steganalysis system based on the reversed Mel-frequency cepstral coefficients (R-MFCC) which aims to provide a model with maximum deviation from HAS model. Genetic algorithm is deployed to optimize dimension of the R-MFCC-based features. This will both speed up feature extraction and reduce the complexity of classification. The final decision is made by a trained support vector machine (SVM) to detect suspicious audio files. The proposed method achieves detection rates of 97.8% and 94.4% in the targeted (Steghide@1.563%) and universal scenarios. These results are respectively 17.3% and 20.8% higher than previous D2-MFCC based method.

1. Introduction

Subliminal channels are types of covert channels which are used for stealth communication over innocuous-looking insecure channels. This concept was first introduced by Simmons as the prisoner's problem [37]. Two accomplices have been arrested and are kept in separate cells. They want to cook up an escape plan but they can only communicate through a vigilant warden who will deliver only innocuous messages. Steganography is among the ways to implement such subliminal channel. Steganography consists of an embedding algorithm ($\mathcal{A}_{em}$) that hides a message ($m \in \mathcal{M}$) into an innocent-looking signal called cover ($c \in \mathcal{C}$) and results in a stego signal ($s \in \mathcal{S}$). On the receiver side, another algorithm ($\mathcal{A}_{ex}$) is used to extract hidden message from the stego signal. A steganography system is called secure if the spaces of cover and stego coincide with each other:

$$\mathcal{C} = \mathcal{S} \tag{1}$$

On the other hand, steganalysis is the countermeasure of warden to detect presence of any subliminal channel. If cover signals are empirical [4] then for practical steganography systems, assumption of (1) does not hold and there would be a deviation between $\mathcal{C}$ and $\mathcal{S}$. This discrepancy can be exploited to discriminate between cover and stego signals. If $\mathcal{A}_{em}$ and statistical model of $\mathcal{C}$ are known a-priori, optimum detector can be designed using statistical decision theory, otherwise a set of suitable feature and machine learning techniques should be employed [21].

Considering different types of cover media, steganographic systems can be divided into five major categories including image, audio, video, text, and network. Reviewing literature shows that on the contrary to the image, audio steganographic systems have found less attention so far. It is noteworthy that, steganography and steganalysis like many new trends in cryptology such as multimedia encryption systems [18], multimedia secret sharing, and water marking rely heavily on signal processing techniques.

Regarding the functionality of steganalysis systems, they are either universal or targeted. In the former one, the detector does not have any prior knowledge about $\mathcal{A}_{em}$, while in the latter one the system is designed specifically for detecting signatures of a particular method. Over the past decade, different audio steganalysis systems have been proposed. Based on the nature of their features, they can be divided into two distinct types:

[*] Corresponding author.
E-mail address: hamzeh_g62@yahoo.com (H. Ghasemzadeh).

http://dx.doi.org/10.1016/j.dsp.2015.12.015

1. Methods that extract their features by comparing the input signal with a reference signal
2. Methods based on extracting features directly from the signal

Steganalysis by comparing signal with a reference:

Extracting a proper reference signal is the main issue in this category. There are different methods to generate the reference signal for this paradigm. One possible solution is applying denoising method to the input signal in order to provide an estimation of the cover signal. The first method in this area was proposed in [28]. They also used audio quality metrics (AQMs) to quantify the deviation between input signal and generated reference signal [29]. In [14], they argued that AQMs have been designed specifically to detect modifications of pure audio contents. They proposed Hausdorff distance for a better representation of dissimilarity between reference and input signal. Johnson et al. proposed another method for creating reference signal. They used a set of bases functions that were localized in both time and frequency domains to capture regularities of audio signals. These bases aimed to estimate a reference for every input signal. The deviation between reference and input signals was modeled by different moments [20]. In another work, a reference signal independent of input signal was applied [1]. Avcıbas showed that if reference is generated from input signal, then the extracted features depend on both message and content of the cover signal. This dependency on the cover signal may diminish generalization property of the system. They used a constant pair of cover-stego signal for referencing. They showed that this technique improves steganalysis results.

Steganalysis by direct inspection of input signal:

In this scenario, the features are extracted directly from input signals. First, steganalysis was integrated into an intrusion detection system. This work used ratio of ones and zeros in the LSBs to detect steghide [9]. Ru et al. used wavelet and linear prediction techniques to extract correlation between samples of input signals [34]. They employed different statistical moments calculated from the residual signal of every sub-band of wavelet tree for steganalysis. MFCC as one of the most well-known features were examined to improve steganalysis results in [24]. This work also demonstrated that removing speech relevant components of speech signal is beneficial for improving steganalysis. In [13], the first three moments of time and frequency histograms of input signal and its wavelet sub-bands were exploited for proposing a proper steganalysis scheme. Principle component analysis (PCA) was applied to reduce the features' dimension. In [23], it was argued that due to the existence of chaotic phenomena in speech signals, chaotic-based features may be employed to boost detection of audio steganalysis algorithms. It was shown that steganography noise increases chaoticity and dimension of phase space in the stego signals. Then, values of false neighbor fraction and Lyapunov spectrum were used to quantify chaotic characteristics of the analyzed signals. Markov transition probabilities were proposed in [25]. They introduced a metric for measuring complexity of different cover signals and showed that performance of their method maintained good even for complex signals. Liu et al. showed that second order derivative magnifies the differences between spectrum of cover and stego signals [26]. A steganalysis system based on auto regressive time delay neural network was proposed in [31]. A combination of different invariant moments and features was used by Bhattacharyya to model deviation between cover and stego signals [2].

Investigating previous audio steganalysis methods shows that:

1. As the most basic requirement, the effects of steganography should not be detectable by human perception systems. Therefore, processing a cover and its stego version with a "perfect" model of human perception system virtually should produce the same results, and they should be indistinguishable. For example in [24,26] Mel frequency cepstral coefficients (MFCC) have been used for feature extraction. MFCC is a model that mimics frequency resolution of HAS. Other characteristics of HAS that have been used in steganalysis literature include loudness, pre-masking and post-masking [1,28,29]. For instance, loudness belongs to the category of intensity sensations and it is primarily a psychological characteristic. It is known that HAS has the lowest sensitivity in the high frequencies; therefore, incorporating loudness in the feature extraction wipes out faint noises in the high frequency portions of the signal, a region that is very valuable for steganalysis. These ideas are discussed more thoroughly in the section 2 of this paper.
2. Most of the previous works have investigated only LSB steganography and its different implementations. Furthermore, to address steganography systems that resist active warden, these works have used watermarking methods [1,23,29]. We believe that reliable results for active warden are achieved if robust steganography methods are investigated. The rationale behind this claim is that undetectability is not a prerequisite for watermarking systems. Therefore, reliable detection of watermarking methods does not necessarily lead to a reliable detection of robust steganography methods.
3. Although most of previous works have claimed a universal steganalysis system but all of them (except for [29], to the best of our knowledge) have only reported results of targeted simulations.

Continuing on our seminal work [16], this paper aims to address these problems. Specifically the following contributions are made:

- A new model with the maximum deviation from HAS is proposed. Then, this model is exploited for extracting a new set of features. Finally, genetic algorithm (GA) is invoked for a near optimum feature selection.
- The reliability of the proposed steganalysis system is tested on a wide range of steganography methods including LSB, DWT, and DCT domain methods. Also, for the sake of completeness and better comparison with previous works, two watermarking methods are also considered.
- Both targeted and universal steganalysis scenarios are pursued.

The rest of this paper is organized as follows. Section 2 includes some preliminaries on the HAS and its relations with steganographic concepts. Section 3 elaborates on the proposed method. Experimental results are presented in section 4. Discussion follows in section 5 and finally conclusions are made in section 6.

2. Human auditory system

Basilar membrane within cochlea of the inner ear is the base for sensory cells of hearing. Previous studies have shown that cochlea operate as a kind of mechanical frequency analyzer [35]. Further studies have discovered that the produced effects in the inner ear are not linear or logarithmic over the whole length of the basilar. In contrast, other scales such as pitch ratio and critical-band can be plotted on linear scales along the basilar membrane. Therefore, in characterizing HAS either the critical-band scale or the pitch ratio scale are more useful than the frequency scale [12].

Pitch ratio and Mel Scale:

To measure pitch of a pure tone, one possible procedure is to present human subjects with a pure tone of frequency f_1 and ask

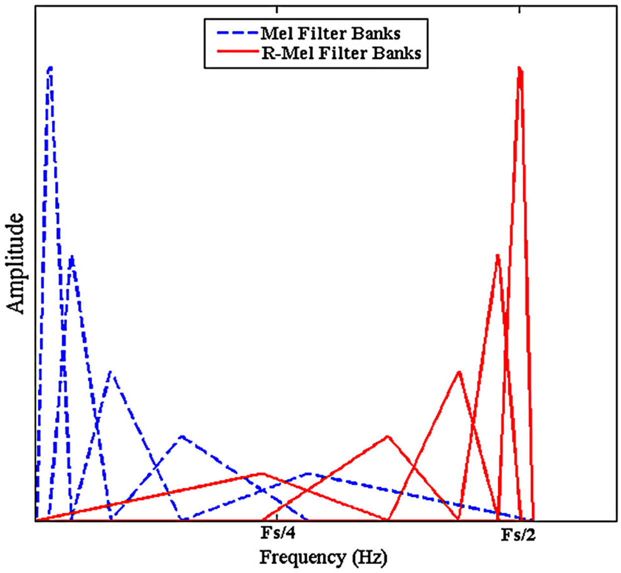


Fig. 1. R-Mel and Mel triangular filter bank.

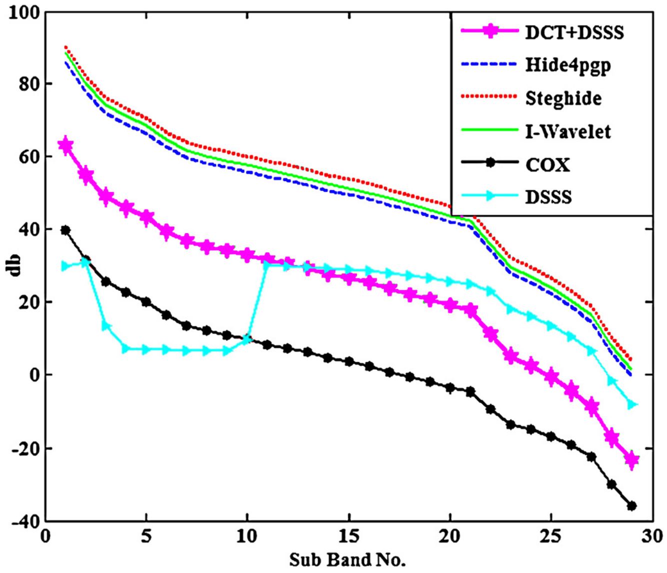


Fig. 2. SNRs of sub-band for different data hiding methods.

them to adjust a second frequency f_2 such that f_2 produces half pitch of the first tone. Subjective measurements have shown that at low frequencies, halving of the pitch sensation corresponds to the ratio of $f_1/f_2 = 2$ while in the frequencies above 1 KHz, this ratio is larger than 2 [12]. According to these observations, for each tone with an actual frequency measured in hertz, a subjective pitch is calculated on a scale called 'Mel'. Equation (2) shows the mathematical formula for converting a given frequency (f) in hertz to its corresponding Mel value.

$$Mel = 1127 \times \ln\left(1 + \frac{f}{700}\right) \tag{2}$$

Mel-frequency cepstral coefficients:

Cepstrum is the anagram of the word spectrum which reflects information about the rate of power changes in different spectrum bands. Later, this metric was tweaked to mimic characteristics of HAS. These new coefficients are commonly known as MFCC and have found numerous applications such as speaker identification [33] and speech recognition [30].

Assume that F denotes fast Fourier transform, MFCCs are calculated as:

$$S_k = \sum F\big(x(t)\big).W_k; \qquad C_m = F\big(\log(S_k)\big) \tag{3}$$

Where M is the number of filters in the Mel bank, and W_k is the triangular weighting function corresponding to the k-th filter. These filters are constructed as follows:

- In the target scale (R-Mel or Mel), linearly divide the whole spectrum into $M + 1$ parts.
- Convert stop and start points of all parts to hertz. This will lead to $M + 2$ distinct points.
- W_k is a triangle such that it starts from i-th point, reaches its peak at $i + 1$th point and returns to zero at the $i + 2$th point. Sometimes these triangles are normalized such that they have areas equal to one.

Plots of these weighting functions for both R-Mel and Mel scales are presented in Fig. 1.

Steganography and Human Auditory System

Reviewing the steganography literature shows that many works have used peak signal to noise ratio (PSNR) or signal to noise ratio (SNR) to imply security of their methods. Furthermore, Zamani et al. investigated the correlation between PSNR and the capacity of audio steganography [41]. They showed that PSNR decreases with increasing the capacity. Logically, increasing the capacity of a certain method enhances its probability of detection. Therefore, it can be inferred that lower values of PSNR lead to higher probability of detection. Based on this assumption, effect of a typical audio steganography system is investigated. Let us model the effect of steganography as an additive noise:

$$s(t) = c(t) + n(t) \tag{4}$$

We take discrete time Fourier transforms from both sides:

$$S\big(e^{jw}\big) = C\big(e^{jw}\big) + N\big(e^{jw}\big) \tag{5}$$

Then, the whole spectrum of the signal is divided into L equally spaced sub bands:

$$(i-1) \times \frac{\pi}{L} \le B_i \le i \times \frac{\pi}{L}, \quad 1 \le i \le L \tag{6}$$

We define sub-band SNR of signal as:

$$SNR_i = 10\log_{10}\left(\frac{\int_{Bi} |C(e^{jw})|^2}{\int_{Bi} |N(e^{jw})|^2}\right), \quad 1 \le i \le L \tag{7}$$

In order to investigate the effect of steganography on different sub bands, a total number of 4169 audio files were embedded with different data hiding methods. The methods included Hide4Pgp [32], Steghide [19], spread spectrum in the frequency domain [27], error-free wavelet method [36], and two watermarking methods of spread spectrum [22], and the DCT-based robust watermarking method (COX) [7]. After dividing the whole spectrum of cover and noise signals into 29 sub-bands, values of SNR_i were calculated for all files. Fig. 2 shows average values of SNR_i over all the files. (Notations are in accordance with those of Table 1.)

The main purpose of steganographic communication is to hide the mere existence of a secret message. Therefore, the most primary requirement of a steganographic system is to remain undetectable. Thus, in its most rudimentary form, it is crucial that the human perception system (ears in the case of audio steganography and eyes in the case of image) should not be able to distinguish between the stego and cover signals. In other words, effects of steganography should not be detectable by human perception systems. According to this fact, a true model of human perception system should be indifferent to steganography. Thus, it is very

Table 1
Database specifications.

	Method	Embedding domain	Capacity BPB (%)	Capacity ratio (%)	SNR (mean ± std)	Parameters	Bit error rate (%)	Ref.
Steganography method	Hide4pgp	LSB	25	100	53.7 ± 7.0	–	0	[32]
			12.5	50	65.8 ± 7.0	–	0	
			6.25	25	72.8 ± 7.0	–	0	
	Steghide		3.125	100	74.9 ± 6.4	–	0	[19]
			1.563	50	77.1 ± 6.1	–	0	
	I-Wavelet	Wavelet	25	–	35.1 ± 7.2	Haar wavelet	0	[36]
			12.5	–	58.7 ± 7.5	Haar wavelet	0	
			6.25	–	69.9 ± 8.4	Haar wavelet	0	
			3.125	–	74.6 ± 8.9	Haar wavelet	0	
	DSSS + DCT	DCT	0.0063	–	49.8 ± 7.0	$\alpha = 10$, $N = 1000$	18.76	[27]
			0.00063	–	49.9 ± 7.0	$\alpha = 10$, $N = 10\,000$	11.27	
			0.0063	–	43.8 ± 7.0	$\alpha = 20$, $N = 1000$	13.82	
			0.00063	–	43.8 ± 7.0	$\alpha = 20$, $N = 10\,000$	6.98	
Watermarking method	DSSS	Time	–	–	19.3 ± 3.7	–	–	[22]
	COX	DCT	–	–	27.7 ± 7.0	$\alpha = 0.01$	–	[7]

likely that employing features based on human perception systems leads to discarding vital information.

Investigating different audio covers shows that as frequency increases, their power spectrums decrease so they can be considered as band-limited signals. On the other hand, investigating noise of steganography indicates that it is a broadband signal. Consequently, it is expected that the value of SNR_i decreases with increasing of the frequency. Fig. 2 supports this claim.

Comparing results of Fig. 2 and characteristics of HAS reveals interesting facts. According to Fig. 1, Mel scale has its highest resolution in the lower frequencies and its lowest resolution in the higher frequencies. On the other hand, according to Fig. 2, high frequency portions of the signal tend to reveal the effects of steganography more clearly. Therefore, features based on HAS are not very suitable. It is noteworthy that steganalysis methods [1,28,29] that have incorporated other psychoacoustic characteristics of HAS (such as loudness and masking [12]) in their feature extraction routine, have produced inferior results to MFCC-based systems. We believe these inferior results are due to extracting features from a more accurate model of HAS.

Reversed Mel Scale:

Based on our previous discussions, we propose an artificial auditory model that has maximum deviation from HAS. Specifically, our suggested model employs a new scale called reversed-Mel scale (R-Mel) that has reversed frequency resolution of HAS. The new scale has its highest resolution in high frequencies and its lowest resolution in low frequencies. If F_S denotes sampling frequency of the signal, we define the R-Mel value of a given frequency f in hertz as:

$$RMel = 1127 \times \ln\left(1 + \frac{0.5 \times F_s - f}{700}\right) \tag{8}$$

Based on this new scale, a new set of triangular weighting functions is constructed. These new filters are used in equation (3) to produce reversed-Mel frequency cepstral coefficients (R-MFCC). Fig. 1 compares filter banks constructed based on Mel with R-Mel scale. Investigating filers constructed on the Mel scale shows that these filters are more concentrated in the lower frequencies. In other words because triangles in the low frequencies have smaller width, more coefficients will be extracted from low frequencies. Therefore, we say Mel scale has higher frequency resolution in the lower frequencies. On the other hand, filters constructed on the R-Mel scale have exactly the opposite characteristics. That is, the filters have finer resolutions in the higher frequencies and coarser resolutions in the lower frequencies.

3. The proposed scheme

Our proposed method is based on taking advantages of R-MFCC coefficients discussed in the previous section. We believe that these features provide suitable discrimination between cover and stego audio files.

Analysis of the Proposed Features:

According to equations (3) and (4), the discriminating factor between the cover and stego is:

$$\begin{aligned}\mathcal{D} = {} & F\left(\log\left(\sum F(c+n).W_k\right)\right) \\ & - F\left(\log\left(\sum F(c).W_k\right)\right)\end{aligned} \tag{9}$$

Using some basic manipulation, (9) reduces to:

$$\mathcal{D} = F\left(\log\left(\frac{\sum F(c+n).W_k}{\sum F(c).W_k}\right)\right) \tag{10}$$

$$\mathcal{D} = F\left(\log\left(1 + \frac{\sum F(n).W_k}{\sum F(c).W_k}\right)\right) \tag{11}$$

Now let us investigate equation (11) for both MFCC and R-MFCC cases. According to the discussion of section 2, the most discriminative features would be extracted from high frequency regions; thus, the last weighting function of Mel and R-Mel banks are considered. Assuming $F_S = 44\,100$, and $M = 29$, the W_{29} of MFCC and R-MFCC have non-zero values in the regions of $[17\,340, 22\,050]$ Hz and $[21\,869, 22\,050]$ Hz, respectively. Apparently, W_{29} in the MFCC has larger number of non-zero components; therefore, denominator of equation (11) for MFCC feature is larger than the R-MFCC case.

$$\sum F(c).W_{29\text{-}MFCC} > \sum F(c).W_{29R\text{-}MFCC} \tag{12}$$

Furthermore, frequency components of noise are much smaller than their cover counterparts; thus, the numerator cannot compensate for this increase in the value of denominator. In other words, in the high frequency regions, the discriminating factors of (11) are larger in R-MFCC case than their MFCC counterparts.

$$\mathcal{D}_{29\text{-}RMFCC} > \mathcal{D}_{29\text{-}MFCC} \tag{13}$$

Feature Extraction:

After normalizing data to $[-1, 1]$, it was segmented into frames of 1024 samples with overlap of 512. Then, R-MFCCs were calculated for each frame. In this paper, 29 different filters were used. Features were calculated as the values of mean, standard deviation, skewness, and kurtosis of each R-MFCC coefficient over all

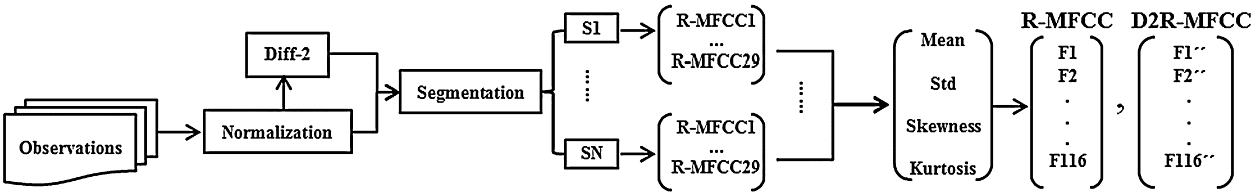


Fig. 3. Feature extraction procedure.

the frames. Previous works have shown that employing second order derivative of the signal leads to better discriminating features [25,26]. Based on this idea, the same procedure was applied on the second order derivative of the input signal. This second set of features is denoted by D2-R-Mel. Fig. 3 illustrates the feature extraction procedure.

Preprocessing:

Investigating the extracted features shows that:

1. The values of extracted features from some observations in the same class are significantly different from each other.
2. Different features tend to have different dynamic ranges.

Observations with significantly different feature values are called outliers. Outliers are either results of noisy measurements or distribution with long tails [39]. Removing the noise and outliers during training allows the learning algorithm to find more accurate classification boundaries [38]. Therefore, in the training phase, outliers were removed using the distance-based method implemented in [11]. In this method, distances between all observations from the same class were calculated. If the distance between an observation and more than 10% of other observations was larger than a threshold, it was considered as an outlier. Also, the threshold was defined as the mean of distances plus three times their standard deviation.

Another problem in classification stems from features with high values. Such features may influence cost function of the classifier more, regardless of their effectiveness in discrimination. Features were normalized to alleviate this problem. To this end, mean and variance of features over train set were calculated, and then features were normalized according to equation (14):

$$\hat{x}_{ik} = \frac{x_{ik} - m_k}{\sigma_k} \tag{14}$$

Furthermore, the values of m_k and σ_k were retained for applying normalization to test samples.

Dataset:

Our cover signals consisted of 4169 mono uncompressed audio wave files with frequency of 44 100 Hz and 16 bits resolution. The duration of each excerpt was 10 seconds and they covered wide range of music genres and languages [16]. All covers were embedded with random messages; furthermore, the message was changed for each cover. Different steganography and watermarking algorithms were used to hide message. The steganography algorithms in this study were Hide4Pgp, Steghide, spread spectrum in the frequency domain, error-free wavelet method, and watermarking methods are spread spectrum, and the DCT-based robust watermarking method (COX).

Embedding Strength:

Expressing embedding strength of steganographic methods can be accomplished through two different metrics of capacity ratio and bit-per-bit percent (BPB). Capacity ratio is the ratio of embedding rate to the maximum capacity of a particular method. Also, BPB is the ratio of message size to the size of cover. Although previous audio steganalysis studies have used capacity ratio for expressing embedding strength, BPB is much more suitable. First, steganography tries to implement a subliminal channel which its efficiency is equal to the ratio of message size to the cover size. Therefore, BPB quantifies objective of steganography more closely. Furthermore, BPB is a universal metric and can be used across different steganography methods and different bit resolutions of cover signals. Thus, BPB provides a meaningful way of comparing different steganography methods. Table 1 presents details of the employed database.

Feature Selection:

In classification tasks, there are usually some irrelevant or redundant features. In fact, there is no useful information with irrelevant features and also redundant features do not provide further information than the currently selected features. Such features increase complexity of feature extraction (as the most time-consuming part of the system) while they provide no useful information. Furthermore, high dimensional space increases the computational complexity of the classifier and it may also diminish its generalization property [40]. Due to its good performance [17], GA was invoked to choose a near-optimum subset of features. We used accuracy of the classifier as the fitness function, population size of 200 individuals, tournament selection [3], and two-point crossover [15]. For selecting k out of n features, genes were encoded as a decimal array of length k. Initially, this array was filled with k random draws from set of $[1, n]$ without replacement. To further improve the performance of GA, selection operation was followed by elitism [8] which was implemented as directly selecting 1% of the mating population from the best chromosomes. Finally, mutation with rate of 1% was implemented as replacing one of already selected features with one of the remaining ones.

Classifier:

The process of distinguishing cover from stego samples needs a classifier to define a suitable decision boundary. This work employs support vector machine (SVM) for its superb performance [17]. SVM is basically based on Vapnik's statistical learning theory in which a maximum-margin hyper plane is created to distinguish the training vectors from different classes [6]. Furthermore, if features are not linearly separable, it is possible to map the original problem into a much higher-dimensional space and achieve better classification result. This task is accomplished by applying a suitable kernel function. In this work, SVM is applied by using the non-commercial package LIBSVM [5] with radial basis function (RBF).

4. Experimental results

Scatter plots of both MFCC and R-MFCC of the second order derivative of cover signal and steghide@1.563 BPB are shown in Fig. 4.

Comparing Figs. 4.A and 4.B shows that features based on R-Mel scale are more separated than features extracted based on Mel scale. This observation justifies our initial assumption that features extracted based on the idea of maximum deviation from HAS would lead to more discriminative features.

GA was invoked to select the best subset of features for optimum detection of steghide@1.563% BPB. The results showed that

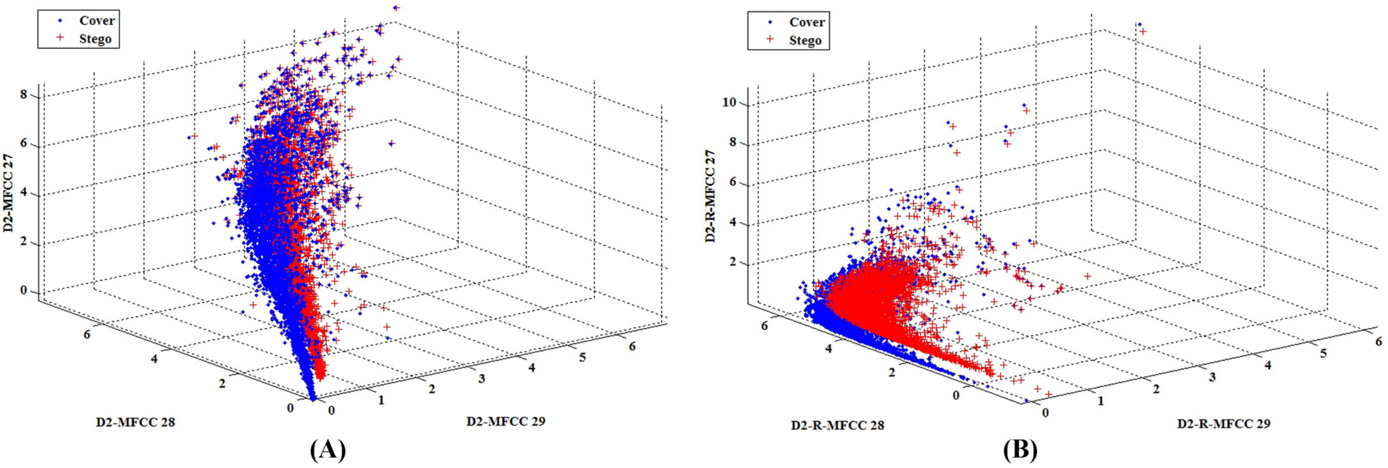


Fig. 4. Scatter plots of covers vs. stegos for (A) D2-MFCC features, and (B) D2-R-MFCC features.

Table 2
Performance of the proposed method in term of sensitivity (Se%), specificity (Sp%) and accuracy (Ac%).

Method	BPB%	MFCC [24]			D2-MFCC [25]			R-MFCC			D2-R-MFCC		
		Se	Sp	Ac	Se	Sp	Ac	Se	Sp	Ac	Se	Sp	Ac
Hide4pgp	25	91.7	83.9	87.8	99.7	94.7	97.2	99.7	99.9	99.8	**99.8**	**100**	**99.9**
	12.5	70.4	65	67.8	95.9	86.5	91.3	98.2	97.4	97.8	**99.4**	**99.7**	**99.5**
	6.25	59.4	50.5	55	89.7	79.5	84.7	93.3	90	91.7	**98.9**	**98.7**	**98.8**
Steghide	3.125	55.1	46.8	51	86.4	76.8	81.7	91.1	87.7	84.9	**98.1**	**98.6**	**98.4**
	1.563	56.6	38.9	47.8	80.5	73.8	77.2	87.1	82.6	84.9	**97.8**	**97.4**	**97.6**
I-Wavelet	25	**100**	98.9	99.4	**100**	99.2	99.6	**100**	**100**	**100**	**100**	**100**	**100**
	12.5	87.8	79.7	83.8	99.1	93.4	96.3	**99.6**	99.6	99.6	**99.6**	**99.9**	**99.8**
	6.25	64.1	60.1	62.4	94	84	89.1	96.7	95.2	96	**99.3**	**99.2**	**99.2**
	3.125	57.4	43.8	50.7	84.8	75	79.9	89	85	87	**98.5**	**97.4**	**98**
DSSS + DCT	6.3e−3	95.4	87.6	91.5	99.8	95.7	97.8	99.6	**99.9**	99.7	**99.8**	99.8	**99.8**
	6.3e−4	96.2	88.8	92.5	99.8	96.1	98	99.4	99.9	99.7	**99.9**	**100**	**99.9**
	6.3e−3	98.9	93.8	96.4	**100**	98	99	**99.8**	**100**	**99.9**	99.8	**100**	**99.9**
	6.3e−4	99.4	94.4	96.9	**100**	98	99	99.8	**100**	99.9	99.9	**100**	**100**
DSSS	–	65.4	55.7	60.7	79.1	72.3	75.8	78.3	72.8	75.6	**97.3**	**94.5**	**96**
COX	–	90.8	89.7	90.2	91.7	93	92.3	97.2	98.8	98	**98.3**	**99.8**	**99**

the best accuracy was achieved when 21 features were selected. These indexes were used for all of the simulations. The rationales behind this approach are as follows. Firstly, although the selected features may be sub-optimum for other methods, but in scenarios where embedding algorithm is not known a-prior feature selection should be independent from embedding algorithm. Secondly, among the methods considered in this paper steghide@1.563 had the highest values of SNR_i (Fig. 2); thus, it was the most challenging method to detect. Therefore, if a subset of features can detect steghide@1.563 accurately, it is more likely that they would do the same for other methods as well.

Considering search complexity of GA, the feature selection was repeated for 10 times and the numbers of required generations were calculated. The simulations showed that on average after 4.7 generations the algorithm finds the best feature set. If this number is multiplied with the number of individuals in each generation (200), average search complexity of 940 is calculated. Considering the fact that the GA was performed only once (just for steghide), this complexity is acceptable.

To measure efficacy of the proposed method, different tests were conducted. In each test, database was randomly divided into the training (70%) and the testing (30%) sets. Then, SVM was trained using the features extracted from train set. Finally, trained model was evaluated using test set. This procedure was repeated for 20 times and, the performance criteria were eventually calculated by averaging over all the iterations. Criteria used in this paper are sensitivity (SE), specificity (SP), and accuracy (ACC). These criteria are described as follows:

- True negative (TN): the number of cover samples that are classified as cover samples.
- True positive (TP): the number of stego samples that are classified as stego samples.
- False negative (FN): the number of stego samples that are classified as cover samples.
- False positive (FP): the number of cover samples that are classified as stego samples.

Sensitivity (SE) is the probability of correct detection of stego samples and is defined as:

$$SE = \frac{TP}{TP + FN} \times 100\% \tag{15}$$

Specificity (SP) is the probability of correct detection of the cover samples and is equal to:

$$SP = \frac{TN}{TN + FP} \times 100\% \tag{16}$$

Accuracy (ACC) is the probability of correct classification and is calculated as:

$$ACC = \frac{TN + TP}{TN + FP + TP + FN} \times 100\% \tag{17}$$

Targeted steganalysis scenario:

In this section, we assume that $\mathcal{A}_{em}$ is known a priori. Table 2 compares performance of the proposed features with some of pre-

Table 3
Results of targeted steganalysis N_F: Number of features N_C: No. of covers in the database.

Method	N_F	N_C	Mean Se.	Mean Sp.	Ref.
AQM	19	664	92.75	93.5	[29]
Hausdorff	25	200	88.6	72.9	[14]
MFCC	36	389	66.0	–	[24]
Chaotic	22	2554	80.3	74.2	[23]
Markov	81	12 000	Accuracy = 92.2		[25]
D2-MFCC	29	12 000	Accuracy = 85.9		[26]
D2-R-MFCC	29	4169	97.3	94.7	[16]
R-MFCC + GA	21	4169	95.3	93.9	*
D2-R-MFCC + GA	21	4169	99.1	99.0	*

* These results have been achieved in our simulations.

vious works based on HAS. Best results are presented in the bold face letters.

Results of Table 2 shows that the proposed method outperforms previous methods. Comparing accuracy of the proposed method with D2-MFCC method shows that an improvement of 20.4% has been achieved. Another important fact becomes apparent when the rate of change in the accuracy for different capacities are studied. In this fashion, for the proposed method as the capacity reduces from 25%BPB to 1.563%BPB, accuracy of the system only drops by 2.3%. On the other hand this number rises to 20% for D2-MFCC method.

To compare performance of the proposed method and previous works, average value of performance criteria over different embedding algorithms are calculated. These results with other important factors such as number of features (N_F) and number of cover files in the database (N_C) are presented in the Table 3.

Comparing results of Table 3 shows that using R-Mel instead of Mel scale improves performance of steganalysis considerably. Other points become apparent when results of D2-R-MFCC are compared with those of R-MFCC + GA and D2-R-MFCC + GA. Specifically taking second order derivative of audio signals improves sensitivity and specificity of the proposed method by 3.8% and 5.1%, respectively. Also, gains of 1.8% and 4.3% in the average values of sensitivity and specificity are achieved when higher order statistics and GA are incorporated into the proposed method.

Universal steganalysis scenario:

In this section, we assumed that $\mathcal{A}_{em}$ was not known. To simulate this scenario, all stego files were placed in a folder and then 4169 of them were selected randomly. In this fashion, stego files were uniformly selected across all data hiding algorithms. To the best of our knowledge, only in [29] result of universal scenario was reported. Therefore, we have also investigated efficacy of universal steganalysis on some of previous works. Table 4 compares results of the proposed universal system with some of previous ones.

Comparing results of D2-MFCC and D2-R-MFCC+GA shows that, the proposed method improves sensitivity and specificity of the universal scenario by 20.8% and 9.3%, respectively. Comparing sensitivity and specificity of the AQM method in the universal and targeted scenarios shows that they drop moderately and considerably in the universal case, respectively. On the other hand for the proposed method another trend is observed. That is, sensitivity and specificity of the proposed method in the universal scenario decrease and remain unchanged, respectively. These different behaviors stem directly from differences in the statistical characteristics of AQM and D2-R-MFCC features which in turn would be reflected in the support vectors of each case. That is, support vectors selected for AQM in the universal case are such that they favor more toward designating a signal as stego (therefore, higher value of sensitivity). On the other hand, support vectors selected for D2-R-MFCC are such that they favor more toward designating a signal as cover (therefore, higher value of specificity).

Table 4
Results of universal steganalysis N_F: Number of features N_C: No. of covers in the database.

Method	N_F	N_C	Mean Se.	Mean Sp.	Ref.
AQM	19	664	81.8	79.7	[29]
MFCC	29	4169	54.4	86.3	*[24]
D2-MFCC	29	4169	73.6	89.8	*[26]
R-MFCC + GA	21	4169	83.9	96.0	*
D2-R-MFCC + GA	21	4169	94.4	99.1	*

* These results have been achieved in our simulations.

Receiver Operating Characteristic (ROC):

ROC is a graphical plot that illustrates performance of the classifier, as the decision boundary is varied. An ROC plot depicts relative tradeoffs between the benefits (true positives) and the costs (false positives) [10]. So, ROC can provide a good tool for assessing classification task and selecting between different classifiers. Fig. 5 presents ROC plots of both targeted (for steghide@1.563 BPB) and universal scenarios. According to ROC of different feature sets we can infer that the proposed method is far better than its competitors. This is evident from larger value of area under the curve (AUC) for D2-R-MFCC feature set.

5. Discussion

Audio media due to its remarkable redundancy and popularity can provide a suitable means to hide data. Therefore, a vast variety of schemes have been proposed to embed secret data in audio files. These methods have mainly attempted to suggest algorithms which benefits from the areas of audio file in time or frequency domain where the resulted changes from embedding were not detectable by HAS. Therefore, in order to detect the effect of steganography, it is better to employ a model that has maximum deviation from HAS. Furthermore, examining the power spectrum of steganography noise $N(e^{jw})$ and power spectrum of cover signal $C(e^{jw})$ revealed interesting facts. Steganography noise constitutes a broad-band signal with powerful high frequency components. On the other hand, cover signal is a band limited signal. That is, its power spectrum decreases with increasing of the frequency. Therefore, it is expected that the high frequency region of signals leads to more discriminating features. Result of Fig. 2 justifies this notion. On the other hand, frequency response of human ear (Mel filter bank of Fig. 1) shows that HAS has low frequency resolution at high frequencies; consequently some information will be lost. In contrast, frequency response of our proposed model matches with our goals (high resolution at high frequencies). According to Fig. 4, it is obvious that this matching has resulted in very good discriminating features. According to this figure, distributions of the proposed features are more separated than those based on Mel scale.

Comparing results of Steghide@3.125 and Hide4pgp@6.25 in the Table 2 leads to an interesting conclusion. While their accuracy of detection is the same, Steghide provide half capacity of Hide4pgp.

According to Tables 3 and 4, the proposed system has good performance in both targeted and universal scenarios. Also, results of universal systems are lower than targeted paradigm.

6. Conclusion

This paper proposed the idea of maximum deviation from human auditory system for steganalysis. Based on this idea, frequency characteristic of our proposed artificial auditory system was explained. Specifically, this artificial ear had high resolution and sensitivity in high frequencies and lower resolution and sensitivity in low frequencies. Simulation results showed that such artificial ear has potency of distinguishing between stego and cover signals effectively. Proposed method in the targeted scenario achieved accuracy of 97.6% (StegHide@1.563% BPB) and 98.8% (Hide4Pgp@6.25%

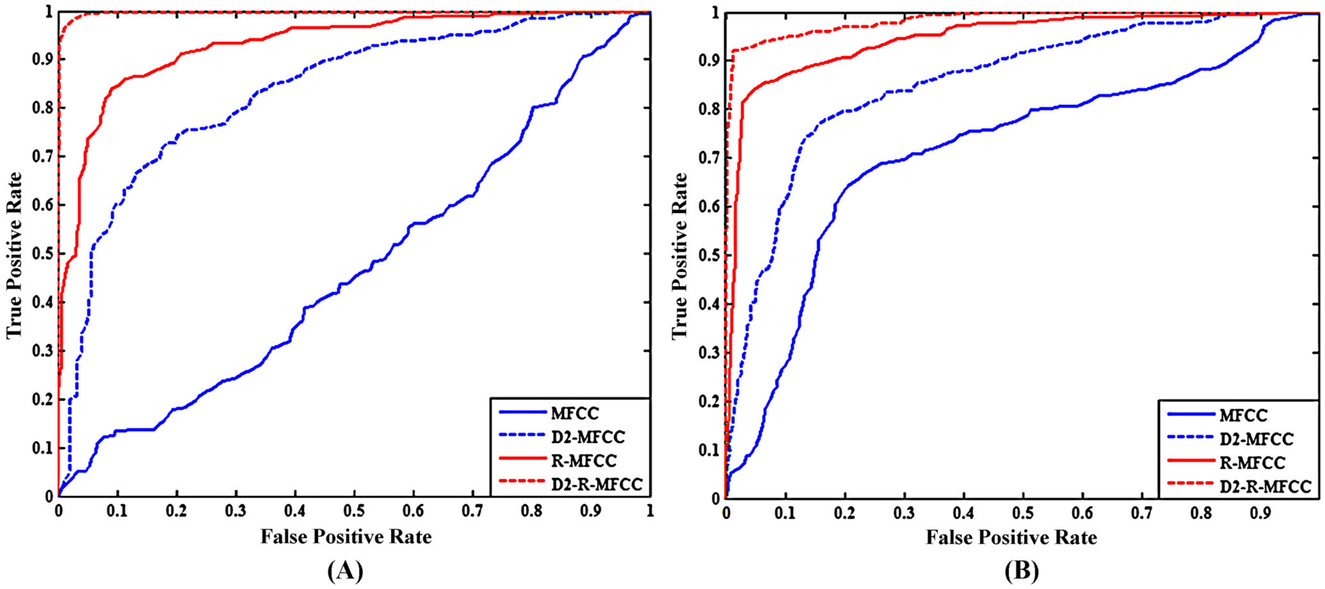


Fig. 5. ROC plots of different steganalysis system: (A) Targeted Scenario (steghide@1.563 BPB) and (B) Universal Scenario.

BPB) which were 20.4% and 14.1% higher than previous MFCC based methods. In the universal test, proposed method achieved sensitivity and specificity of 94.4% and 99.1% which were 20.8% and 9.3% higher than previously reported results.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.dsp.2015.12.015.

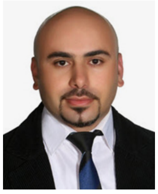
References

[1] I. Avcıbas, Audio steganalysis with content-independent distortion measures, IEEE Signal Process. Lett. 13 (2006).
[2] S. Bhattacharyya, G. Sanyal, Feature Based Audio Steganalysis (FAS), Int. J. Comput. Netw. Inf. Secur. 4 (2012).
[3] T. Blickle, L. Thiele, A comparison of selection schemes used in genetic algorithms, TIK-report, 1995.
[4] R. Böhme, Advanced Statistical Steganalysis, Springer, 2010.
[5] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27.
[6] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[7] I.J. Cox, J. Kilian, F.T. Leighton, T. Shamoon, Secure spread spectrum watermarking for multimedia, IEEE Trans. Image Process. 6 (1997) 1673–1687.
[8] K.A. De Jong, Analysis of the behavior of a class of genetic adaptive systems, 1975.
[9] J. Dittmann, D. Hesse, Network based intrusion detection to detect steganographic communication channels: on the example of audio data, in: IEEE 6th Workshop on Multimedia Signal Processing, IEEE, 2004, pp. 343–346.
[10] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, 2012.
[11] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. De Ridder, D. Tax, S. Verzakov, PRTools4: A Matlab Toolbox for Pattern Recognition, Delft University of Technology, Netherlands, 2007.
[12] H. Fastl, E. Zwicker, Psychoacoustics: Facts and Models, Springer, Berlin, 2001.
[13] J.-W. Fu, Y.-C. Qi, J.-S. Yuan, Wavelet domain audio steganalysis based on statistical moments and PCA, in: International Conference on Wavelet Analysis and Pattern Recognition, vol. 4, ICWAPR'07, IEEE, 2007, pp. 1619–1623.
[14] S. Geetha, N. Ishwarya, N. Kamaraj, Audio steganalysis with Hausdorff distance higher order statistics using a rule based decision tree paradigm, Expert Syst. Appl. 37 (2010) 7469–7482.
[15] H. Ghasemzadeh, A metaheuristic approach for solving jigsaw puzzles, in: Iranian Conference on Intelligent Systems, ICIS, IEEE, 2014, pp. 1–6.
[16] H. Ghasemzadeh, M. Khalil Arjmandi, Reversed-Mel cepstrum based audio steganalysis, in: 4th International Conference on Computer and Knowledge Engineering, ICCKE2014, IEEE, 2014.
[17] H. Ghasemzadeh, M.T. Khass, M.K. Arjmandi, M. Pooyan, Detection of vocal disorders based on phase space parameters and Lyapunov spectrum, Biomed. Signal Process. Control 22 (2015) 135–145.
[18] H. Ghasemzadeh, H. Mehrara, M.T. Khas, Cipher-text only attack on hopping window time domain scramblers, in: 4th International Conference on Computer and Knowledge Engineering, ICCKE, IEEE, 2014, pp. 194–199.
[19] S. Hetzl, P. Mutzel, A graph-theoretic approach to steganography, in: Communications and Multimedia Security, Springer, 2005, pp. 119–128.
[20] M.K. Johnson, S. Lyu, H. Farid, Steganalysis of recorded speech, in: Electronic Imaging 2005, International Society for Optics and Photonics, 2005, pp. 664–672.
[21] A.D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, T. Pevný, Moving steganography and steganalysis from the laboratory into the real world, in: Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security, ACM, 2013, pp. 45–58.
[22] D. Kirovski, H.S. Malvar, Spread-spectrum watermarking of audio signals, IEEE Trans. Signal Process. 51 (2003) 1020–1033.
[23] O.H. Koçal, E. Yürüklü, I. Avcibas, Chaotic-Type Features for Speech Steganalysis, IEEE Trans. Inf. Forensics Secur. 3 (2008) 651–661.
[24] C. Kraetzer, J. Dittmann, Mel-cepstrum-based steganalysis for VoIP steganography, in: Electronic Imaging 2007, International Society for Optics and Photonics, 2007, pp. 650505–650512.
[25] Q. Liu, A.H. Sung, M. Qiao, Novel stream mining for audio steganalysis, in: Proceedings of the 17th ACM International Conference on Multimedia, ACM, 2009, pp. 95–104.
[26] Q. Liu, A.H. Sung, M. Qiao, Temporal derivative-based spectrum and Mel-cepstrum audio steganalysis, IEEE Trans. Inf. Forensics Secur. 4 (2009) 359–368.
[27] R.M. Nugraha, Implementation of direct sequence spread spectrum steganography on audio data, in: International Conference on Electrical Engineering and Informatics, ICEEI, IEEE, 2011, pp. 1–6.
[28] H. Ozer, I. Avcibas, B. Sankur, N.D. Memon, Steganalysis of audio based on audio quality metrics, in: Electronic Imaging 2003, International Society for Optics and Photonics, 2003, pp. 55–66.
[29] H. Özer, B. Sankur, N. Memon, İ. Avcıbaş, Detection of audio covert channels using statistical footprints of hidden messages, Digit. Signal Process. 16 (2006) 389–401.
[30] L.R. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, vol. 14, PTR Prentice Hall, Englewood Cliffs, 1993.
[31] S. Rekik, S.-A. Selouani, D. Guerchi, H. Hamam, An autoregressive time delay neural network for speech steganalysis, in: 11th International Conference on Information Science, Signal Processing and their Applications, ISSPA, IEEE, 2012, pp. 54–58.
[32] H. Repp, Hide4PGP. Available at: www.heinz-repp.onlinehome.de/Hide4PGP.htm, 1996.
[33] D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. Audio Speech Lang. Process. 3 (1995) 72–83.
[34] X.-M. Ru, H.-J. Zhang, X. Huang, Steganalysis of audio: attacking the steghide, in: Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol. 7, IEEE, 2005, pp. 3937–3942.
[35] J. Schnupp, I. Nelken, A. King, Auditory Neuroscience: Making Sense of Sound, MIT Press, 2011.
[36] S. Shirali-Shahreza, M. Manzuri-Shalmani, High capacity error free wavelet Domain Speech Steganography, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008, IEEE, 2008, pp. 1729–1732.

[37] G.J. Simmons, The prisoners' problem and the subliminal channel, in: Advances in Cryptology, Springer, 1984, pp. 51–67.
[38] M.R. Smith, T. Martinez, Improving classification accuracy by identifying and removing instances that should be misclassified, in: The 2011 International Joint Conference on Neural Networks, IJCNN, IEEE, 2011, pp. 2690–2697.
[39] S. Theodoridis, K. Koutroumbas, fourth edition, Academic Press, 2009.
[40] S. Theodoridis, A. Pikrakis, K. Koutroumbas, D. Cavouras, Introduction to Pattern Recognition: A Matlab Approach, Academic Press, 2010.
[41] M. Zamani, A.B.A. Manaf, S.M. Abdullah, S.S. Chaeikar, Correlation between PSNR and bit per sample rate in audio steganography, in: 11th International Conference on Signal Processing, Saint Malo, Mont Saint-Michel, France, April, 2012, pp. 2–4.

Hamzeh Ghasemzadeh was born in Tehran in 1984. He received his B.S. degree in Electrical Engineering from Ferdowsi University of Mashhad in 2007. He received his M.S. degree in Communications Engineering from Malek-e-Ashtar University of Technology in 2011. His primary research interests are Multimedia Security, Steganography, Computer Forensic, Pattern Recognition, and Random Signal Processing. He is currently a lecturer at Electrical Engineering Department of Islamic Azad University of Damavand.

Mehdi Tajik Khass (born 1985) has received a B.S. degree and an M.S. degree in Telecommunication Engineering from Tabriz University, Iran. His research interests are in signal and image processing, cryptography, human voice processing, and bioengineering especially in the field of human auditory system.



Meisam Khalil Arjmandi received his B.Sc. in Electrical and Electronic Engineering from the Islamic Azad University (IAU) South Tehran Branch, Iran in 2006. He earned his master's degree in Biomedical Engineering at Shahed University, Iran in March 2010. He is currently a Ph.D. student in Communication Sciences and Disorders at Michigan State University. His research area of interests is mainly focused on speech/music perception and production, cognitive neuroscience of speech and language, automatic voice disorders assessment, and digital signal processing.