Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Detection of quantization index modulation steganography in G.723.1 bit stream based on quantization index sequence analysis*

Song-bin LI†1,2, Huai-zhou TAO1,2, Yong-feng HUANG†‡1,2

(1Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)
(2Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China)
†E-mail: {lisb, yfhuang}@mail.tsinghua.edu.cn

Received Dec. 19, 2011; Revision accepted Mar. 29, 2012; Crosschecked July 6, 2012

Abstract: This paper presents a method to detect the quantization index modulation (QIM) steganography in G.723.1 bit stream. We show that the distribution of each quantization index (codeword) in the quantization index sequence has unbalanced and correlated characteristics. We present the designs of statistical models to extract the quantitative feature vectors of these characteristics. Combining the extracted vectors with the support vector machine, we build the classifier for detecting the QIM steganography in G.723.1 bit stream. The experiment shows that the method has far better performance than the existing blind detection method which extracts the feature vector in an uncompressed domain. The recall and precision of our method are all more than 90% even for a compressed bit stream duration as low as 3.6 s.

Key words: Steganalysis, Quantization index modulation (QIM), G.723.1, Codeword distribution characteristics
doi:10.1631/jzus.C1100374 **Document code:** A **CLC number:** TN918

1 Introduction

In recent years, voice over IP (VoIP) has become a very popular Internet streaming media communication service. The network traffic of VoIP is very large and continues to grow. This makes VoIP very suitable for steganography (Krätzer *et al.*, 2006; Mazurczyk and Szczypiorski, 2008; Tian *et al.*, 2009; Huang *et al.*, 2011a; 2011b). G.723.1 (ITU-T, 1996a; 1996b), a famous audio data compression algorithm for voice, is specially defined for VoIP applications by the International Telecommunication Union (ITU) and widely used in VoIP applications. Due to its real-time and large-scale characteristics, the G.723.1 compressed speech stream is a potentially excellent carrier for steganography and a tremendous threat to communication monitoring. Consequently, it is necessary to study the steganalysis method for the steganography that can be used in the G.723.1 stream.

Current methods of audio steganography can be divided into three main categories. The first is the least significant bit (LSB) replacement/matching method towards the pulse code modulation (PCM) format voice data (Huang and Xiao, 2008). The second hides the secret information in the transform domain. These methods firstly transform the cover's data to the transform domain, and then modify some parameters in the transform domain to embed confidential information. The transforms often used include the cepstrum transform (Li and Yu, 2000), discrete cosine transform (Wang *et al.*, 2004), discrete wavelet transform (Wu *et al.*, 2005; Tan *et al.*, 2010), etc. The third is the method based on quantization index modulation (QIM) firstly proposed by Chen and Wornell (2001). The QIM hides secret data by modifying the quantization vector and can be applied

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 60970148) and the National High-Tech R&D (863) Program of China (No. 2011AA010704)

© Zhejiang University and Springer-Verlag Berlin Heidelberg 2012

to various digital media, such as audio, image, and video. It is very suitable for use in information hiding in the encoding process for media compression, especially for information hiding in low rate speech coding.

Recently, for information hiding in an instant low bit-rate speech stream, Xiao *et al.* (2008) proposed a novel codebook partition algorithm called the complementary neighbor vertex (CNV) algorithm to optimally divide the vector codebook into two sub-codebooks required by QIM embedding. This method is referred to as CNV-QIM steganography. CNV-QIM steganography reduces the distortion of voice signal to the minimum in comparison with other codebook division methods (Lu *et al*., 2005; Wang *et al*., 2007). This makes the hiding more difficult to detect. It is one of the current most advanced information hiding methods in the low bit-rate compressed voice stream and can be applied to information hiding in the G.723.1 compressed voice stream. Therefore, in this paper, we focus on detecting the CNV-QIM steganography. For information on methods of detecting the LSB replacement/matching and transform domain steganography, the reader is referred to the methods as presented by Avcibas (2006), Liu *et al.* (2009), and Huang *et al.* (2011a; 2011b).

There has been some research into QIM information hiding methods, but these studies mainly focus on the image as a carrier. For example, Hafiz (2008) found that embedding confidential information using QIM increased the irregularity (randomness) of the quantitative image. By using the ‘approximate entropy’ to quantify the irregularity of the observed images, this method can distinguish the ‘cover’ or ‘stego’ image. Another steganalysis method proposed by Hafiz *et al.* (2008) is similar to above work; however, the former uses the kernel density estimate (KDE) to measure the local irregularity. Recently, Hafiz (2010) found that the QIM hiding introduces a very strong disturbance of local correlation into the ‘cover’ image, and the detection of QIM embedding is achieved by introducing the gamma distribution to model the disturbance. Furthermore, Wu *et al.* (2009) showed that the QIM embedding disrupts the value of image pixels and the histogram of its discrete cosine transform (DCT) coefficients, and gave a formula that depicts the relationship between changes of the histogram and length of the embedded confidential information. According to the formula, the method can estimate the embedding rate.

Obviously, all the above methods for detecting QIM steganography take advantage of the significant change of the statistical characteristics of the image caused by QIM embedding. Similarly, if we can identify the significant change characteristics of the G.723.1 speech stream caused by the QIM steganography proposed by Xiao *et al.* (2008), then we can construct its steganalysis method.

2 Changes of the quantization index sequence caused by QIM steganography

G.723.1 speech codec is based on the linear predictive coding (LPC) model, which uses an LPC filter to analyze and synthesize acoustic signals in the encoding and decoding endpoints. The LPC filter can be described as follows:

$$H(z)=1\Big/\left(1-\sum_{i=1}^{p}a_i z^{-i}\right), \tag{1}$$

where a_i is the ith order coefficient of the LPC filter. The short time stationary nature of the voice signal requires the entire signal sample be divided into frames and the LPC filter’s coefficients are then computed for each frame.

In speech coding, the LPC filter’s coefficients of each frame are first computed and converted to line spectrum frequency (LSF) coefficients. Subsequently, the LSF coefficients are encoded using vector quantization (VQ). G.723.1 adopts split VQ and uses three split vectors to quantify the LSF coefficients. Assume that the split vectors are $\boldsymbol{f}_1$, $\boldsymbol{f}_2$, and $\boldsymbol{f}_3$, and each $\boldsymbol{f}_i$ corresponds to a codebook L_i (i=1, 2, 3) with a codeword (quantization index) space $\{c_i^1, c_i^2, ..., c_i^{|L_i|}\}$. VQ is the process of choosing the most optimal vector index c_i^k $(1\le k\le |L_i|)$ for each split vector $\boldsymbol{f}_i$ from codebook L_i to make the quantization distortion minimum according to LSF coefficients. After VQ, the LSF coefficients are represented as a complex codeword $C=(c_1^k, c_2^m, c_3^l)$, where c_1^k, c_2^m, and c_3^l are the codewords selected from codebooks L_1, L_2, and L_3, respectively.

The QIM steganography hides the secret data during the VQ process. If one codeword is selected then one secret bit can be embedded. Taking the embedded process based on L_1 as an example, the CNV-QIM steganography (Xiao *et al.*, 2008) firstly partitions L_1 into two sub-codebooks L_1^1 and L_1^2 using the CNV algorithm, where L_1^1 and L_1^2 both contain $|L_1|/2$ vector indices and satisfy

$$L_1^1 \cap L_1^2 = \varnothing, \quad L_1^1 \cup L_1^2 = L_1. \tag{2}$$

The CNV algorithm can guarantee that each codeword and its most nearest codeword in L_1 belong to different sub-codebooks. Thus, the additional signal distortion caused by QIM embedding would be minimal in comparison with other division methods. Upon the completion of partition, labels of '0' and '1' will be assigned to L_1^1 and L_1^2, respectively. When a secret bit is embedded, only the corresponding sub-codebook is used for codeword selecting. On the decoding side, the hidden bit is extracted through checking which sub-codebook the codeword belongs to.

According to the above analysis, there are three split vector sequences in the encoded speech bit stream containing N G.723.1 frames. Each split vector sequence $\boldsymbol{F}_i$ can be represented as follows:

$$\boldsymbol{F}_i = \boldsymbol{f}_{i,1}, \dots, \boldsymbol{f}_{i,k}, \dots., \boldsymbol{f}_{i,N}, \; i = 1, 2, 3, \tag{3}$$

where $\boldsymbol{f}_{i,k}$ ($i\in\{1, 2, 3\}$, $k\in[1, N]$) represents the ith split vector of frame k in the bit stream. After VQ, $\boldsymbol{F}_i$ will be converted to quantization index sequence (QIS) S_i as

$$S_i = c_{i,1}^h, \dots, c_{i,k}^u, \dots., c_{i,N}^m, \; i = 1, 2, 3, \tag{4}$$

where $c_{i,k}^u$ $(i \in \{1, 2, 3\}, k \in [1, N], u \in [1, |L_i|])$ is the quantization index of $\boldsymbol{f}_{i,k}$.

The QIM steganography (Xiao *et al.*, 2008) embeds the secret bits into the bit stream when $\boldsymbol{f}_{i,k}$ chooses the quantization index. As a result of each frame containing three split vectors, three secret bits can be hidden in each frame. Obviously, the QIM steganography will inevitably change the original quantization result, because the QIM steganography is able to convert the original quantization index $c_{i,k}^h$ of $\boldsymbol{f}_{i,k}$ into $c_{i,k}^u$ ($u \neq h$). Therefore, the original QIS S_i of $\boldsymbol{F}_i$ will produce disturbance. Fig. 1 presents an example of the QIS disturbance. In this example, we firstly encode a speech segment with a duration of 3 s according to G.723.1 and obtain the 'cover' object. Secondly, we repeat the encoding process to obtain the 'stego' object using the QIM steganography (Xiao *et al.*, 2008). We extract the QIS $S_1 = c_{1,1}^h, \dots, c_{1,50}^u, \dots, c_{1,100}^m$ from the encoded bit stream of the 'cover' and 'stego' objects. We show these two QIS in Fig. 1, and we can clearly view the difference between the original QIS and its steganography version; the QIM steganography significantly changes the quantization vector sequence. This disturbance of QIS is probable to change the distribution characteristics of the quantization index as well. Obviously, if these characteristics can be quantified then the disturbance in QIS can be measured. Taking advantage of this information, we can detect QIM steganography in G.723.1 bit stream.

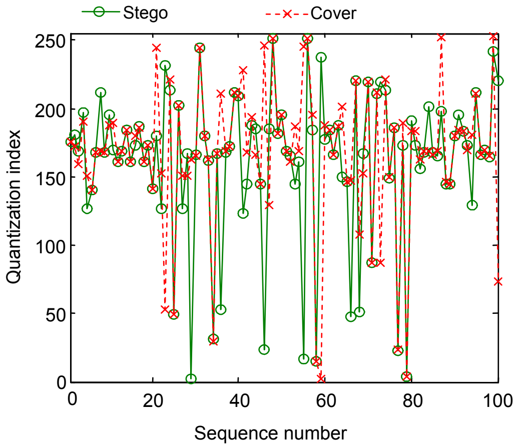


Fig. 1 Example of quantization index modulation (QIM) steganography disturbing the quantization index sequence (QIS) of the first split vector

3 Statistical models of quantization index distribution characteristics

According to acoustics of speech production, phoneme is the basic unit of human speech and is the pronunciation of one or several sequential letters (Thomas, 2002). When a person speaks, he/she continuously adjusts his/her articulators for a sequence of phonemes. For example, the pronunciation of the English word ‘shop’ is composed of sounds of the phoneme ‘sh’, ‘o’, and ‘p’ (Fig. 2). Therefore, a speech can be viewed as a sequence of phonemes, and can be divided into multiple small segments, each of which corresponds to a phoneme. We refer to this as the speech phoneme composing model (SPCM). Fig. 2 shows the principle of the SPCM and Definition 1 gives its regular description.

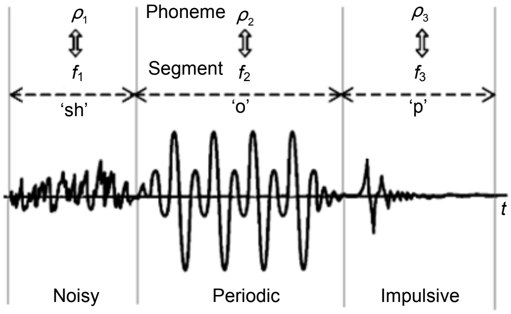


Fig. 2 Sketch map of the principle of speech phoneme composing model (SPCM) (taking ‘shop’ as an example)

Definition 1 Phoneme ρ_i is represented by triple (p_i, s_i, t_i), where p_i is the phonetic symbol, s_i is a small segment of speech which is the pronunciation of p_i, and t_i is the duration of s_i. Phoneme ρ_i is the fundamental distinctive unit of language and the phoneme set $P=\{\rho_1, \rho_2, \ldots, \rho_M\}$ of a language contains a finite number of phonemes.

Speech S with duration T can be divided into N small speech segments; all of the small segments comprise the segment set: $S=\{f_1, f_2, \ldots, f_N\}$. If segment f_k can find $\rho_i \in P$ to make $s_l = f_k$ ($k\in[1, N]$, $l\in[1, M]$), then we say f_k can be mapped to phoneme ρ_l, denoted by the mapping relation: $f_k \rightarrow \rho_l$. Supposing all $f_k \in S$ can find their mapping $\rho_l \in P$, and all the mapping relations comprise the set F, then the SPCM can be represented by the triple (P, S, F).

Based on the SPCM, a speech can be mapped to a phoneme sequence $\rho_1\rho_2\ldots\rho_N$ in theory. The durations of different phonemes are unequal; for example, vowel ‘o’ may last more than 50 ms while plosive ‘b’ may last only 10 ms. Moreover, when the speed of the pronunciation changes, the duration of phoneme will also change. Therefore, the duration t_i of phoneme ρ_i is difficult to determine in advance. However, we do not need to determine t_i accurately, because we build the SPCM only for analyzing whether or not QIM steganography exists in a G.723.1 compressed speech stream. We let t_i be the frame length of the G.723.1 approximately.

In a language, there exist some statistical characteristics of letters; for example, in English, the letter ‘e’ appears most frequently and ‘q’ is often followed by ‘u’; generally speaking, the appearance of each letter has a different probability and is correlated with its neighbor. Hence, we can infer that there are analogous statistical characteristics of the phoneme in speech. In other words, there exist unbalanced and correlated characteristics in the distribution of phoneme in human speech. According to speech process theory, the LPC filter defined by Eq. (1) quantitatively represents the state of human’s articulators in a short duration, and different phonemes correspond to different articulators’ states (Thomas, 2002). Therefore, the LPC filter can be approximately viewed as the quantitative representation of phoneme. In this way, a phoneme sequence can be represented by the complex codeword sequence $C_1C_2\ldots C_N$ and the distribution characteristics of phoneme will transfer to the LPC filter defined by C. The sequence $C_1C_2\ldots C_N$ is composed of three independent quantization index sequences:

$$C_1C_2\ldots C_N = \begin{cases} c_{1,1}^h c_{1,1}^u \ldots c_{1,N}^m, \\ c_{2,1}^h c_{2,1}^u \ldots c_{2,N}^m, \\ c_{3,1}^h c_{3,1}^u \ldots c_{3,N}^m. \end{cases} \quad (5)$$

Hence, we can infer that the quantization index in the QIS has unbalanced and correlated characteristics. Below we present the method to extract them.

3.1 Quantification of unbalanced characteristics of the quantization index

To describe the methods succinctly, we rewrite any of the three split vector sequence $\boldsymbol{F}_i = \boldsymbol{f}_{i,1}\ldots\boldsymbol{f}_{i,k}\ldots\boldsymbol{f}_{i,N}$ in G.723.1 bit stream as $\boldsymbol{F} = \boldsymbol{f}_1\ldots\boldsymbol{f}_k\ldots\boldsymbol{f}_N$, where $\boldsymbol{f}_j$ ($j\in[1, N]$) represents the jth split vector in temporal order. We suppose that L is the codebook, and the codeword belonging to L is c_i ($i\in[1, |L|]$), which is the quantization index of split vector $\boldsymbol{f}_j$. After quantization, the split vector sequence is transferred to the quantization index sequence $S=c_1\ldots c_j\ldots c_N$. The unbalanced characteristic of codeword distribution is quantified by the codeword distribution histogram (CDH) defined as follows:

$$\boldsymbol{H}=(h_1, h_2, ..., h_n), \tag{6}$$

where n is equal to $|L|$ which is the number of codewords in codebook L, and h_i ($i\in[1, n]$) represents the appearance probability of codeword c_i in S, which is defined as follows:

$$h_i=\sum_{j=1}^{N}\mathrm{Pr}_{i/j}\cdot\mathrm{Pr}_j=\frac{1}{N}\sum_{j=1}^{N}\mathrm{Pr}_{i/j}, \tag{7}$$

where Pr_j represents the probability of choosing the split vector $\boldsymbol{f}_j$ ($j\in[1, N]$) located in the jth position according to temporal order, and $\mathrm{Pr}_{i/j}$ is the conditional probability of $\boldsymbol{f}_j$ taking c_i as the quantization vector, which is defined as

$$\mathrm{Pr}_{i/j}=\begin{cases}1, & \text{if } s_j=v_i,\\ 0, & \text{else.}\end{cases} \tag{8}$$

3.2 Quantification of correlated characteristics of the quantization index

We use the first-order Markov chain for quantifying the correlated feature. According to the acoustic production model, the phoneme is the basic unit for human pronunciation. So, the process of human speaking can be viewed as a stochastic process of phoneme altering. We refer to this process as a phoneme state transition process. Therefore, a phoneme sequence $\rho_1\rho_2\ldots\rho_N$ can be viewed as a state transition sequence. According to the statistics of linguistics, the appearance of a phoneme generally relates only to its previous phoneme. In this paper, we assume that the emergence of the next phoneme is related only to the current phoneme in a phoneme sequence. This relation can be represented based on the conditional probability as follows:

$$\mathrm{Pr}\left(\frac{\rho_N}{\rho_1\rho_2\cdots\rho_{N-1}}\right)=\mathrm{Pr}\left(\frac{\rho_N}{\rho_{N-1}}\right). \tag{9}$$

According to Eq. (9), we can infer that the stochastic state sequence $\rho_1\rho_2\ldots\rho_N$ is a first-order Markov chain. Therefore, phoneme sequences can be regarded as a phoneme state transition first-order Markov chain. According to our analysis of the relationship between the phoneme and LPC filter, we can infer that the split vector sequence $\boldsymbol{F}=\boldsymbol{f}_1\ldots\boldsymbol{f}_k\ldots\boldsymbol{f}_N$ is also a first-order Markov chain, and the state set is $L=\{c_1, c_2, \ldots, c_{|L|}\}$. The quantitative correlated feature of the codeword distribution can be represented by the state transition probability (STP) using the conditional probability as follows:

$$a_{ij}=\mathrm{Pr}\left(\frac{c_j}{c_i}\right),\ 1\le i,j\le M,\ \sum_{j=1}^{M}a_{ij}=1. \tag{10}$$

The conditional probability is hard to compute directly. Generally, it is often translated into calculating the joint probability as follows:

$$a_{ij}=\mathrm{Pr}\left(\frac{c_j}{c_i}\right)=\frac{\mathrm{Pr}(c_i,c_j)}{\mathrm{Pr}(c_j)},\quad c_i,c_j\in L. \tag{11}$$

Given a G.723.1 bit stream segment, an STP $\boldsymbol{A}$ with $|L|^2$ dimensions can be obtained according to Eq. (11). Obviously, STP $\boldsymbol{A}$ accurately describes the quantitative correlated characteristic of codeword distribution. However, the dimensionality of $\boldsymbol{A}$ is too large, a total of 65 536 possible transitions. Therefore, we do not directly use the matrix $\boldsymbol{A}$ as feature vectors because it is impractical.

We find that the distribution of STP is not balanced. The state transition that does not occur or rarely occurs is not very useful for reflecting the correlated feature and may be ignored. So, we choose only a subset $E=\{b_1, b_2, \ldots, b_{|L|}\}$ of $\boldsymbol{A}$ to compute the STP. Elements of the subset are determined by statistics on a large-scale speech segment dataset. Supposing the number of speech segments in the dataset is N, and a_{ij}^k denotes the transition probability a_{ij} of the kth ($1\le k\le N$) speech in the dataset, the rule of selecting each b_i ($1<i<|L|$) is as follows:

$$b_i=a_{ij},\ j=\arg\max_{j\in[1,|L|]}\left\{\sum_{k=1}^{N}a_{ij}^k\right\}, \tag{12}$$

where b_i is the transition probability from the current state to the most likely next state. We randomly select

2000 different speech segments and compute their state transition matrix. With b_i determined according to Eq. (12), the results are presented in Fig. 3 straightforwardly. After the elements of the subset are determined, the value of each b_i can be extracted from $\boldsymbol{A}$ and all the values constitute a feature vector:

$$\boldsymbol{T}=(v_1, v_2, \ldots, v_{|L|}), \tag{13}$$

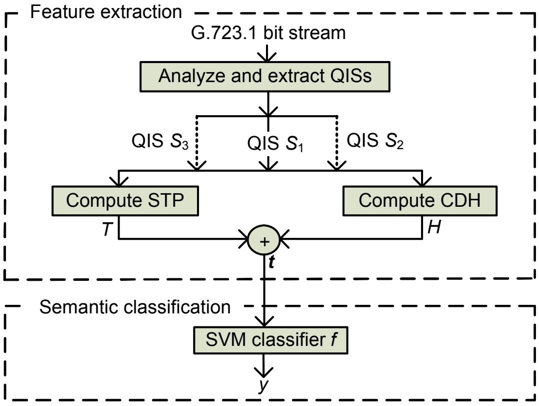
which is used to quantify the correlated characteristics of codeword distribution of the G.723.1 stream.

4 Detection methods based on supervised classification

The aim of detection is to determine whether there is CNV-QIM information hiding in a G.723.1 compressed speech stream. We assume that its duration is finite, and its data is stored in file S. If S contains hidden data, then we call it 'stego', otherwise 'cover'. The information hiding detector can be expressed as

$$y=f(\boldsymbol{t}), \quad y\in\{+1, -1\}, \tag{14}$$

where $\boldsymbol{t}$ represents the feature vector extracted from S for steganalysis and $f(\boldsymbol{t})$ is the detection process that outputs the result of steganalysis. Obviously, $f(\boldsymbol{t})$ is a two-category classifier, so the detection process is essentially a classification process: if y=+1, then S is in the 'cover' class; else, in the 'stego' class. For the classification problem, the method based on supervised classification is very effective; hence, we adopt this method as well. Fig. 4 presents the principle graph of our information hiding detector. Obviously, to run the detector, the key is to determine the feature vector and classifier.



CDH: codeword distribution histogram QIS: quantization index sequence
STP: state transition probability SVM: support vector machine

Fig. 4 Principle graph of the detector of CNV-QIM steganography in G.723.1 compressed speech bit stream

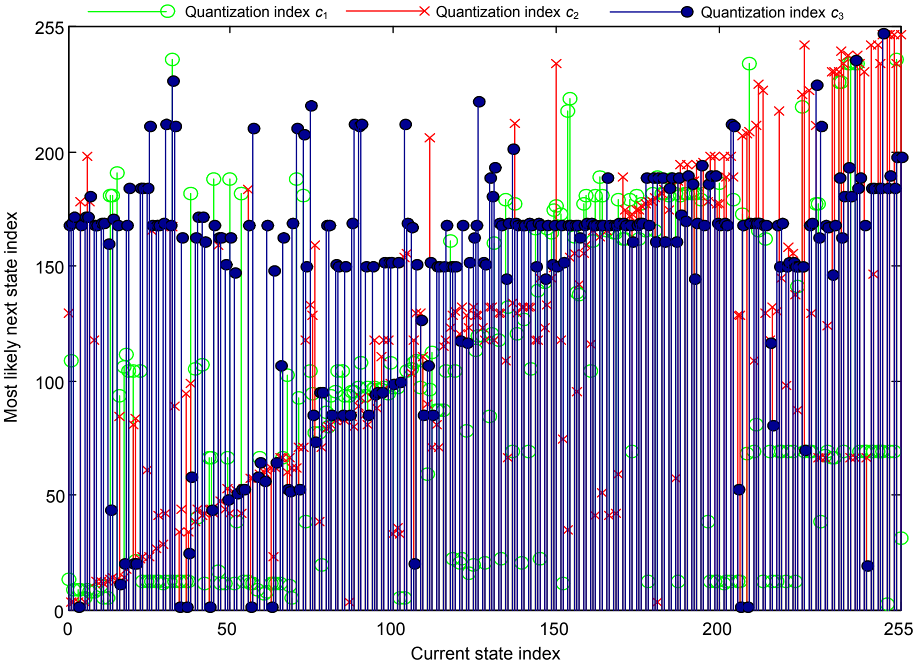


Fig. 3 Distribution of each codeword's most likely next state from the current state

In Section 3, we introduced two methods of extracting the quantitative codeword distribution characteristics. The first extracts the quantitative unbalanced characteristics of codeword distribution and obtains feature vector $\boldsymbol{H}=(h_1, h_2, \ldots, h_{|L|})$, while the second extracts the quantitative correlated characteristic and obtains feature vector $\boldsymbol{T}=(v_1, v_2, \ldots, v_{|L|})$. Obviously, either $\boldsymbol{H}$ or $\boldsymbol{T}$ expresses only part of characteristics of the codeword distribution; hence, separately using each of them as the feature vector is not sufficient for achieving a high precise detector. Therefore, we merge these two feature vectors and obtain a new hybrid feature vector (HFV): $\boldsymbol{t}=(h_1, v_1, h_2, v_2, \ldots, h_{|L|}, v_{|L|})$, which reflects all the codeword distribution characteristics. In the experiment, we will prove that the QIM steganography significantly changes the hybrid feature vector, which is a good feature for steganalysis. As regards the classifier of the supervised classification framework, we choose the support vector machine (SVM). The SVM classifier will be obtained through training. After obtaining the classifier, we use it to predict whether or not CNV-QIM steganography is in a new G.723.1 compressed bit stream.

5 Experiments and discussion

The main purpose of this section is to evaluate the performance of our steganalysis method. Firstly, we show that the QIM steganography significantly changes the unbalanced and correlated characteristics of codeword distribution. Then, we use the quantified codeword distribution characteristics as the feature vector to train the supervised classifier as the QIM steganography detector and test its classification accuracy. To the best of our knowledge, there has been no report of methods for detecting QIM steganography in a compressed speech stream. In theory, however, the blind method proposed by Liu *et al.* (2009) can also be used to detect the steganography. Thus, we compare our detection method with it.

5.1 Datasets

We performed experiments on five different large speech datasets, each of which has different types of native speakers (Table 1). The duration of each speech in these datasets is 10 s and each speech segment is stored as a PCM file. Each speech file in the datasets is encoded according to the G.723.1 standard, and its corresponding G.723.1 bit stream file that contains 333 G.723.1 frames without hidden information is obtained. We assign the category label 'cover' for these files. Each codebook used for split vector quantization in G.723.1 is optimally divided to obtain the two sub-codebooks for QIM embedding, using the CNV algorithm proposed by Xiao *et al.* (2008). Each speech segment is encoded again and the secret data is hidden using the QIM steganography. So, we obtain the G.723.1 bit stream file with hidden information. We assign the category label 'stego' for these files. Each 'cover' and its corresponding 'stego' objects belonging to each dataset constitute the samples for training and testing the classifier.

Table 1 Five speech datasets with different types of native speakers for experiments

Dataset	Number of speech segments	Native speaker type
CM	500	Chinese man
CW	532	Chinese woman
EM	818	English man
EW	824	English woman
Hybrid	2674	All above types

5.2 Changes of the codeword distribution characteristics caused by QIM steganography

To prove that the QIM steganography will make the codeword distribution characteristics change significantly, we introduce the vector variation rate (VVR) to measure the change degree of a vector. Assume $\boldsymbol{V}$ is an N-dimensional vector, and some operations can change the value in some dimensions of $\boldsymbol{V}$ into $\boldsymbol{V}^*$. The VVR is defined as follows:

$$\mathrm{VVR}=\sum_{i=1}^{N}\tau_i \Big/ \sum_{i=1}^{N}\mu_i, \tag{15}$$

where N is the dimensionality of $\boldsymbol{V}$, and μ_i and τ_i are defined as follows:

$$\mu_i=\begin{cases}1, & \text{if } a_i\neq 0,\\ 0, & \text{else},\end{cases} \tag{16}$$

$$\tau_i=\begin{cases}1, & \text{if } a_i\neq 0 \,\&\, a_i\neq b_i,\\ 0, & \text{else},\end{cases} \tag{17}$$

where a_i represents the value of the ith dimension of $\boldsymbol{V}$, and b_i is the value of $\boldsymbol{V}^*$ in the same dimension. Obviously, the larger the value of VVR, the larger the change degree of $\boldsymbol{V}$.

In analyzing a given speech segment, we firstly encode it according to the G.723.1 standard, and then compute its CDH feature vector $\boldsymbol{H}$. We apply the QIM steganography and repeat the above process to obtain the changed CDH feature vector $\boldsymbol{H}^*$. According to Eq. (15), we can compute the VVR of $\boldsymbol{H}$ caused by QIM steganography. To intuitively show the effect of QIM steganography on $\boldsymbol{H}$, we divide the range of VVR into 10 intervals, each of which is $d_i=[i\times0.1, (i+1)\times0.1]$ ($i\in\{0, 1, \ldots, 9\}$). We compute the VVR of $\boldsymbol{H}$ towards 2000 different speech segments randomly selected from the hybrid dataset and the ratio of speech segments with VVR belonging to d_i. The results are shown in Fig. 5.

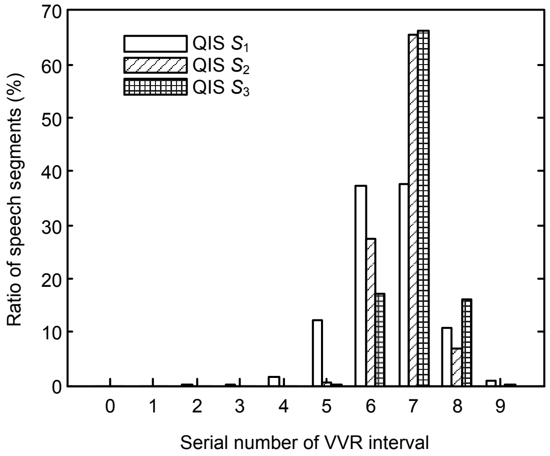


Fig. 5 Effect of quantization index modulation (QIM) steganography on the unbalanced characteristics of codeword distribution in the quantization index sequence (QIS)

We observe that VVR of $\boldsymbol{H}$ of most speech segments is greater than 0.6. This means that about 60% dimensions of the values of vector $\boldsymbol{H}$ change after QIM steganography. Using the same method, we can quantify the effect of QIM steganography on the STP feature vector $\boldsymbol{T}$. The results (Fig. 6) indicate that QIM steganography changes $\boldsymbol{T}$ in a similar way. Obviously, the extracted feature vectors are very sensitive to QIM steganography. This is very important for steganalysis.

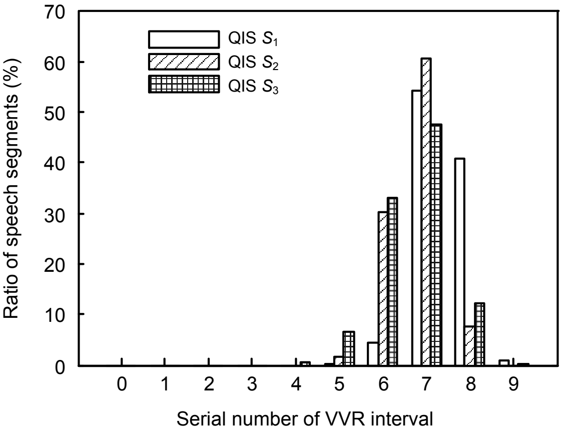


Fig. 6 Effect of quantization index modulation (QIM) steganography on the correlated characteristics of codeword distribution in the quantization index sequence (QIS)

5.3 Performance evaluation

Towards each dataset noted above, we select 75% 'cover' and its corresponding 'stego' samples as the training set and the remaining 25% samples as the testing set. We use the classifier's recall and precision to measure the performance of our detecting classifier of the QIM steganography in G.723.1 bit stream. We define the recall of classifier (RoC) and the precision of classifier (PoC) as follows:

$$\text{RoC}=\frac{Q}{Q+Q^*},\tag{18}$$

$$\text{PoC}=\frac{Q}{Q+E^*},\tag{19}$$

where Q represents the number of 'stego' samples, the category of which is correctly predicted by the classifier, and E^* and Q^* respectively represent the numbers of 'cover' and 'stego' samples, the categories of which are falsely predicted by the classifier.

We use LIBSVM (Chang and Lin, 2001), a widely used library of SVM, to train the classifiers. To make the results more comparable, all the classifiers select the radial basis function (RBF) kernel function and all related parameters of each classifier are optimally adjusted by cross-validation. Below we compare the performance of our method with that of the method proposed by Liu *et al.* (2009), called the blind detection (BD) method. The experimental results of the five datasets are presented in Table 2.

Table 2 Performance comparison of our method and the blind detection (BD) method proposed by Liu *et al.* (2009)

Dataset	Ours (%)						BD method (%)	
	QIS S_1		QIS S_2		QIS S_3			
	RoC	PoC	RoC	PoC	RoC	PoC	RoC	PoC
CM	99.2	98.4	99.2	96.9	94.0	96.7	49.6	50.0
CW	97.7	97.0	96.2	96.9	95.5	94.8	52.5	47.2
EM	100	96.2	95.6	97.0	91.2	80.9	55.4	50.9
EW	97.5	99.0	98.1	98.1	95.2	98.5	58.3	53.8
Hybrid	97.9	98.0	98.4	96.8	95.5	94.4	56.0	51.4

The proposed method extracts the feature vector for steganalysis in a compressed domain and the detection is implemented on the QIS. If we find one QIS having QIM steganography, we can know that there exists QIM steganography in the G.723.1 bit stream. So, we list the RoC and PoC of the three QISs. However, the BD method extracts the feature vector in the uncompressed domain and has only one detection result. According to Table 2, our method's performance is far better than that of the BD method: the RoC and PoC of QIS S_1 of our method are more than 96% while those of the BD method are less than 60% towards all the datasets.

Another advantage of our method is that it can adapt to different languages and native speakers. From Table 2, we can see that our method possess good performance towards different datasets. The underlying reason for this advantage is as follows: the hybrid feature vector (HFV) of codeword distribution is almost consistent for different native speakers. To prove it, we first encode all the 500 speech segments in the CM dataset according to the G.723.1 standard, and then extract QIS S_1 to compute its HFV; finally, using all 500 extracted HFVs, we compute the mean coefficient of each sub-vector of HFV as shown in Fig. 7a. Using the same process, results of other datasets are shown in Figs.7b–7d. We observe a low HFV variability coefficient distribution for different datasets. Therefore, constructing a different classifier for each native language speaker is not required. This greatly improves the adaptability of the method.

However, note that the speech sample used for testing is as long as 10 s. Our aim is to detect the CNV-QIM steganography in G.723.1 compressed stream in VoIP applications. The speech stream in VoIP application is real-time, and must be stored before information hiding detection. To make the detection fast and reduce the need for storage space, we hope that the detection can also work well when the duration of speech stream is below 10 s. Obviously, the shorter the needed duration of the speech stream, the better the detection precision. We will test the impact of duration of bit stream on the detection performance in the next experiment.

5.4 Impact of the duration of bit stream on performance

With the view to evaluate the impact of duration of bit stream, we reconstruct the new datasets based on the five datasets above. We cut N ($0<N<333$) frames in front of each G.723.1 stream file in the five datasets and constitute the new corresponding CM, CW, EM, EW, and hybrid datasets. Obviously, the duration of each sample in the new dataset is $0.03N$ s. We alter the value of N and observe the variation of RoC and PoC.

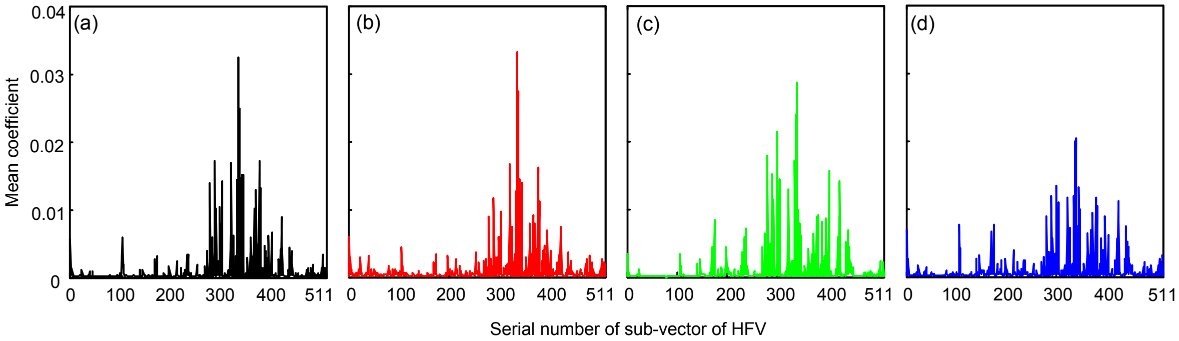


Fig. 7 Coefficients distribution of the hybrid feature vector (HFV) of different types of native speakers

(a) Chinese men; (b) Chinese women; (c) English men; (d) English women

Using the BD method (Liu *et al*., 2009), the performances at different durations are shown in Fig. 8. We observe that the RoC and PoC are all very low irrespective of the duration of the bit stream. The reason is that the BD method extracts the feature vector in the time domain but the QIM steganography has small impact on the speech signals, so the feature vector is not very effective for steganalysis.

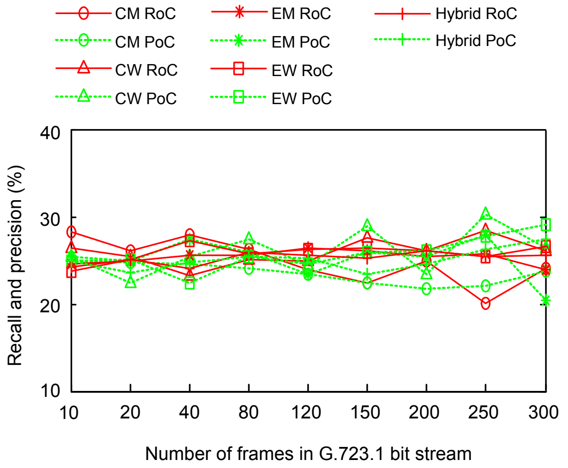


Fig. 8 RoC and PoC obtained using the BD method with different lengths of bit stream

Using our method, we experiment on the three QISs. The RoC and PoC of QIS S_i (i=1, 2, 3) at different durations are shown in Fig. 9. We observe that the RoC and PoC are also not high when N is small, but as N becomes larger RoC and PoC increase rapidly. The RoC and PoC of QIS S_2 are more than 90% on all the five datasets even though the number of frames in bit stream is only 120. This means that our method can effectively detect the QIM steganography in G.723.1 bit stream only by capturing a small segment speech stream of a monitored VoIP session, which is very important for VoIP corresponding censoring.

6 Conclusions

In this paper, we present an effective detection method for detecting QIM steganography in G.723.1 compressed speech bit stream. Our method first illustrates the significant change of the quantization vector index sequence caused by QIM steganography. We build the codeword distribution histogram and the codeword state transition model to quantify the codeword distribution characteristics. Based on these two models, we obtain the feature vectors for training the classifiers for steganalysis. The experiment shows that our method can effectively detect QIM steganography with recall and precision all more than 90%, even for a compressed bit stream duration as low as 3.6 s. This confirms that our method has far better performance than the existing blind detection method.

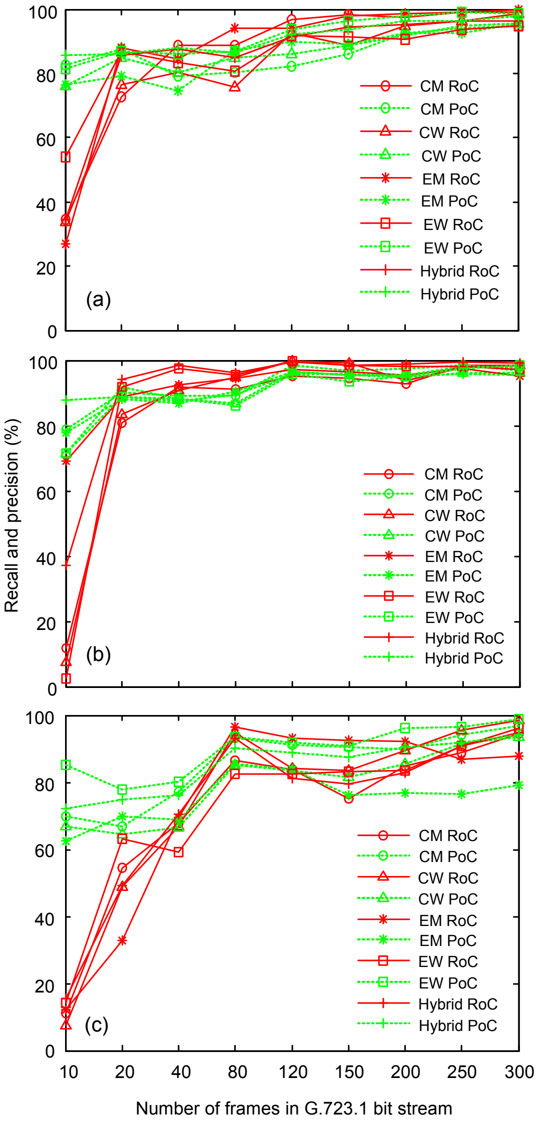


Fig. 9 RoC and PoC obtained using our method with different lengths of bit stream based on QIS S_1 (a), S_2 (b), or S_3 (c)

References

Avcibas, I., 2006. Audio steganalysis with content-independent distortion measure. *IEEE Signal Process. Lett.*, **13**(2):92-95. [doi:10.1109/LSP.2005.862152]

Chang, C.C., Lin, C.J., 2001. LIBSVM: a Library for Support Vector Machines. Available from http://www.csie.ntu.edu.tw/~cjlin/libsvm [Accessed on May 9, 2011].

Chen, B., Wornell, G.W., 2001. Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Inf. Theory*, **47**(4):1423-1443. [doi:10.1109/18.923725]

Hafiz, M., 2008. Steganalysis of QIM Steganography Using Irregularity Measure. Proc. 10th ACM Workshop on Multimedia and Security, p.149-158. [doi:10.1145/1411328.1411355]

Hafiz, M., 2010. Statistical Modeling of Footprints of QIM Steganography. IEEE Int. Conf. on Multi-media and Expo, p.1487-1492. [doi:10.1109/ICME.2010.5582954]

Hafiz, M., Subbalakshmi, K.P., Chandramouli, R., 2008. Nonparametric steganalysis of QIM data hiding using approximate entropy. *SPIE*, **6819**:681914. [doi:10.1117/12.767313]

Huang, Y., Xiao, B., 2008. Implementation of Covert Communication Based on Steganography. Proc. 4th Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing, p.1512-1515. [doi:10.1109/IIH-MSP.2008.174]

Huang, Y., Tang, S., Zhang, Y., 2011a. Detection of covert voice-over Internet protocol communications using sliding window-based steganalysis. *IET Commun.*, **5**(7):929-936. [doi:10.1049/iet-com.2010.0348]

Huang, Y., Tang, S., Bao, C., Yip, Y.J., 2011b. Steganalysis of compressed speech to detect covert voice over Internet protocol channels. *IET Inf. Secur.*, **5**(1):26-32. [doi:10.1049/iet-ifs.2010.0032]

ITU-T, 1996a. ITU-T Recommendation G.723.1: Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s. ITU-T, Geneva.

ITU-T, 1996b. ITU-T Recommendation G.723.1-Annex A: Silence Compression Scheme for G.723.1. ITU-T, Geneva.

Krätzer, C., Dittmann, J., Vogel, T., Hillert, R., 2006. Design and Evaluation of Steganography for Voice-Over-IP. Proc. IEEE Int. Symp. on Circuits and Systems, p.2397-2400. [doi:10.1109/ISCAS.2006.1693105]

Li, X., Yu, H.H., 2000. Transparent and Robust Audio Data Hiding in Cepstrum Domain. IEEE Int. Conf. on Multimedia and Expo, p.397-400. [doi:10.1109/ICME.2000.869624]

Liu, Q., Sung, A.H., Qiao, M., 2009. Temporal derivative-based spectrum and mel-cepstrum audio steganalysis. *IEEE Trans. Inf. Forens. Secur.*, **4**(3):359-368. [doi:10.1109/TIFS.2009.2024718]

Lu, Z.M., Yan, B., Sun, S.H., 2005. Watermarking combined with CELP speech coding for authentication. *IEICE Trans. Inf. Syst.*, **E88-D**(2):330-334. [doi:10.1093/ietisy/E88-D.2.330]

Mazurczyk, W., Szczypiorski, K., 2008. Steganography of VoIP Streams. Proc. 3rd Int. Symp. on Information Security, p.1001-1018. [doi:10.1007/978-3-540-88873-4-6]

Tan, L., Wu, B., Liu, Z., Zhou, M., 2010. An audio information hiding algorithm with high-capacity which based on chaotic and wavelet transform. *Acta Electron. Sin.*, **38**(8):1812-1824 (in Chinese).

Thomas, Q.F., 2002. Discrete-Time Speech Signal Processing: Principles and Practice. Prentice Hall PTR, NJ, USA, p.45-60.

Tian, H., Zhou, K., Jiang, H., Huang, Y., Liu, J., Feng, D., 2009. An Adaptive Steganography Scheme for Voice Over IP. Proc. IEEE Int. Symp. on Circuits and Systems, p.2921-2925. [doi:10.1109/ISCAS.2009.5118414]

Wang, C.T., Chen, T.S., Chao, W.H., 2004. A New Audio Watermarking Based on Modified Discrete Cosine Transform of MPEG/Audio Layer III. Proc. IEEE Int. Conf. on Networking, Sensing and Control, p.265-277. [doi:10.1109/ICNSC.2004.1297081]

Wang, F.H., Jain, L.C., Pan, J.S., 2007. VQ-based watermarking scheme with genetic codebook partition. *J. Network Comput. Appl.*, **30**(1):4-23. [doi:10.1016/j.jnca.2005.08.002]

Wu, Q., Li, W., Yu, X.Y., 2009. Revisit Steganalysis on QIM-Based Data Hiding. Proc. 5th Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing, p.929-932. [doi:10.1109/IIH-MSP.2009.316]

Wu, S., Huang, J., Huang, D., 2005. Efficiently self-synchronized audio watermarking for assured audio data transmission. *IEEE Trans. Broadcast.*, **51**(1):69-76. [doi:10.1109/TBC.2004.838265]

Xiao, B., Huang, Y., Tang, S., 2008. An Approach to Information Hiding in Low Bit-Rate Speech Stream. Proc. IEEE Global Communications Conf., p.1940-1944. [doi:10.1109/GLOCOM.2008.ECP.375]