基于统计模型及SVM的低速率语音编码QIM隐写检测

李松斌[1),3)] 黄永峰[1)] 卢记仓[2)]

[1)](清华大学电子工程系 北京 100084)

[2)](解放军信息工程大学网络工程系 郑州 450002)

[3)](中国科学院声学研究所南海研究站 海口 570105)

摘 要 QIM(Quantization Index Modulation,量化索引调制)隐写在标量或矢量量化时嵌入机密信息,可在语音压缩编码过程中进行高隐蔽性的信息隐藏,文中试图对该种隐写进行检测.文中发现该种隐写将导致压缩语音流中的音素分布特性发生改变,提出了音素向量空间模型和音素状态转移模型对音素分布特性进行了量化表示.基于所得量化特征并结合SVM(Support Vector Machine,支持向量机)构建了隐写检测器.针对典型的低速率语音编码标准G.729以及G.723.1的实验表明,文中方法性能远优于现有检测方法,实现了对QIM隐写的快速准确检测.

关键词 QIM隐写;隐写检测;低速率语音编码器;音素分布特性

中图法分类号 TP309 **DOI号** 10.3724/SP.J.1016.2013.01168

Detection of QIM Steganography in Low Bit-Rate Speech Codec Based on Statistical Models and SVM

LI Song-Bin[1),3)] HUANG Yong-Feng[1)] LU Ji-Cang[2)]

[1)](*Department of Electronic Engineering, Tsinghua University, Beijing* 100084)

[2)](*Zhengzhou Information Science and Technology Institute, Zhengzhou* 450002)

[3)](*Haikou Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences, Haikou* 570105)

Abstract Quantization Index Modulation (QIM) steganography, which embeds the secret information during the Vector Quantization, can hide information in low bit-rate speech codec with high imperceptibility. This paper tries to detect this type of steganography. For this purpose, starting from the speech generation and compress coding theory, this paper firstly analyzes the possible significant feature degradation through the QIM steganography in compressed audio stream deeply. And it finds that the QIM steganography will disturb the phoneme sequence in the stream, and inevitably make the imbalance and correlation characteristics of phoneme distribution in the sequence change. According to this discovery, this paper adopts the phoneme distribution characteristics as the key for the detection of the QIM steganography. In order to get the quantitative features of phoneme distribution characteristics, this paper designs the Phoneme Vector Space Model and the Phoneme State Transition Model to quantify the imbalance and correlation characteristics respectively. By combining the quantitative vector features with supervised learning classifier, this paper builds a high performance detector towards the QIM steganography in low bit-rate speech codec. The experiments show that, for the two typical low bit-rate speech codec: G.729 and G.723.1, the proposed method has an excellent performance compared to existing method.

Keywords QIM steganography; steganalysis; low bit-rate speech codec; phoneme distribution characteristics

收稿日期:2011-07-12;最终修改稿收到日期:2011-10-17.本课题得到国家自然科学基金(60970148,61271392)资助.**李松斌**,男,1981年生,博士,副研究员,主要研究方向为多媒体信息处理、信息隐藏等.E-mail:lisongbin_work@126.com.**黄永峰**,男,1967年生,博士,教授,主要研究领域为多媒体网络安全及下一代因特网等.**卢记仓**,男,1985年生,博士研究生,主要研究方向为隐写及隐写分析.

1 引　言

VoIP(Voice over IP)是非常流行的流媒体通信服务，在全球范围内得到了广泛应用，彻底变革了语音通信市场格局. 由 VoIP 带来的语音数据流具有量大且实时瞬态等特征，非常适合作为信息隐藏载体，这使 VoIP 很可能被用于在 IP 网络中进行隐蔽通信[1]. 当前在语音中进行信息隐藏的方法可大致分为以下几类：(1) 针对脉冲编码调制语音数据的最低有效位替换或匹配方法[2]；(2) 变换域方法，该方法先将载体数据变换到变换域，然后通过在变换域修改一些参数实现机密信息的嵌入，常用的变换包括倒谱变换[3]、离散余弦变换[4]、离散小波变换[5]等；(3) 基于量化索引调制(Quantization Index Modulation，QIM)的方法，该方法由 Chen 等人[6]提出，适用于包含矢量量化的数字音频、图像和视频编码，可用于在压缩编码过程中进行信息隐藏；(4) 一些针对特定压缩语音标准的信息隐藏方法，例如，最近文献[7]提出了一种在 G.723.1 码流的静音帧中嵌入机密信息的方法.

QIM 隐写的基本思想是将量化码本分组. 假设原量化码书为 C，将其分为 C_1 和 C_2 两部分，满足 $C_1 \cap C_2 = \varnothing$ 且 $C_1 \cup C_2 = C$，C_1 和 C_2 分别代表比特“0”和“1”；当嵌入 0 时仅在分组码书 C_1 中选取最佳量化值，嵌入 1 时则仅在分组码书 C_2 中选取最佳量化值. 接收方根据所接收的量化结果中的索引值是属于 C_1 和 C_2 来恢复机密信息比特. 显然，这种方法实现简单，不增加计算量.

为了减少带宽消耗，VoIP 一般在发送端对语音进行低速率压缩编码然后传输. 因此，上述几类语音信息隐藏方法中第 3 种方法最适合用于在 VoIP 建立隐蔽信道，因为第 1 类方法嵌入后的秘密信息在进行压缩编码时将丢失，第 2 类方法的运算复杂度较高不适合在语音实时编码时使用，而第 4 类方法仅适用于 G.723.1. 文献[8]针对低速率语音编码提出了一种改进的基于 QIM 的信息隐藏方法，它的主要贡献在于可以保证原码书划分后每个码字和它最邻近码字属于不同的分组，从而使得嵌入机密消息后局部附加量化失真的极大值相对其它划分方式取得极小，减小了隐写带来的语音失真，提高了隐蔽性. 这使对其进行隐写分析非常困难，是当前在低速率压缩语音流中进行信息隐藏最先进的方法之一. 为此本文将以文献[8]提出的 QIM 信息隐藏的方法作为隐写检测目标.

当前 QIM 信息隐藏方法的隐写分析已有一些研究，但这些研究主要针对图像作为载体时的 QIM 隐写展开[9-12]. 文献[9]发现进行 QIM 信息隐藏会对载体图像的局部相关性引入相当强的扰动，通过引入 Gamma 分布对这种扰动进行建模并结合预先确定的似然率参数实现 QIM 嵌入的检测. 文献[10]观察到使用 QIM 嵌入机密信息会增加量化图像的不规则性(随机性)，通过引入“近似熵”对载体和载密图像的这种不规则性进行量化分析实现 QIM 嵌入的检测；文献[11]的方法与此类似，所不同的是该文使用基于核密度估计(Kernel Density Estimate，KDE)的方法对上述局部不规则性进行衡量. 文献[12]发现 QIM 嵌入扰乱了图像像素及 DCT 系数直方图，构造了直方图变化与机密消息长度之间的估计公式，实现了图像中 QIM 嵌入率的估计. 显然，这些方法都利用了 QIM 嵌入所引起的某一维度图像统计特征的显著变化进行隐写分析，因此对于语音流的 QIM 嵌入检测其难点也在于寻找并确定 QIM 嵌入后所引起的显著变化特征. 此外，一些盲检测方法也可用于对 QIM 隐写进行检测，例如文献[13]给出了一种基于 Mel 倒谱频率系数(MFCC)统计特征的音频信息隐藏盲检测方法. 该方法对于最低有效位隐写具有较好的检测效果，但对于 QIM 隐写其检测效果并不理想，其原因主要是压缩编码使语音产生很大的失真，直接从解码后语音采样值提取特征其实已经很难反映原始语音所蕴含的特征信息. 鉴于此，本文针对低速率语音编码中的 QIM 隐写给出了一种无需解码直接在压缩域提取特征的方法，在此基础上构建了基于机器学习理论的隐写检测器.

2 压缩域隐写检测特征提取

2.1 基本思想

VoIP 所使用的低速率语音编码标准主要是 G.729 和 G.723.1，这两种低速率语音编码器都使用了线性预测编码(LPC)方法，编码过程的核心步骤是对语音信号进行 LPC 分析以获得声道系统函数. 通常声道系统函数可由式(1)表示，

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{1}$$

其中 $a_1, a_2, \cdots, a_p$ 为语音信号的 p 阶 LPC 预测系数. 语音信号 $x(n)$ 可视为激励信号通过滤波器 $H(z)$ 获得，例如一般语音中的浊音可视为周期性脉冲激励得到(如图 1 的元音“o”)，而清音则由白噪声激励得到(如图 1 的清音“sh”). 不同音素发音时一

般具有不同的声道形态，据此可以推知，不同的音素发音时其声道系统函数也不同，所以在理想情况下应该对每个音素对应的语音片段分别进行 LPC 分析，每个音素的 LPC 预测系数刻画了该音素的量化发音特性. 获得 LPC 预测系数后的步骤是对其进行矢量量化，假设获得的量化矢量索引为 I，则对于音素 P 必有一个 I 与之对应，用符号 $P \mapsto I$ 表示这种关系并称 I 为 P 的量化特征索引.

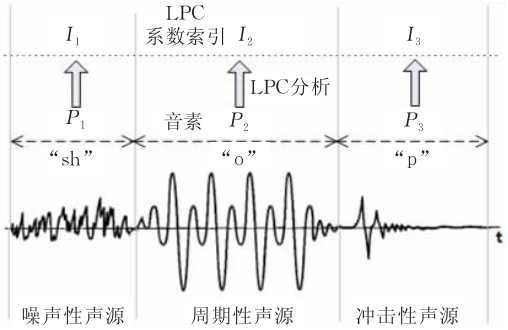


图 1 语音、音素序列及 LPC 系数索引序列关系示意图

音素在语言学中被称为音标，它是构成语言的基本单元，这些离散的基本单元根据一定的音素和文法规则或多或少地连缀成词语[14]，如图 1 中的单词"shop"的发音由 3 个音素构成；词语按照一定的句法形式构成完整的语言系统. 语言系统是存在某些统计规律的，例如，据统计英语中使用次数最多的字母是"e"，那么映射到语音上可以认为音素"e"的出现次数也最多；其次，英语中字母之间的组合排列方式是存在一定规律的，如"q"的后面大多数时候跟着"u"，那么映射到语音上可以认为音素之间的组合排列也存在一定的规律. 换句话说，一段语音中的各音素的出现是不均衡的，其次不同音素的出现存在相关性. 称上述特性为语音中的音素分布特性. 假设某段语音对应的音素序列为 $S=P_1P_2\cdots P_{n-1}P_n$，根据 $P \mapsto I$，它将有一个与之对应的量化特征索引序列：$S^*=I_1I_2\cdots I_{n-1}I_n$，如图 1 所示. 文献[8]给出的信息隐藏方法是在获得 LPC 系数的量化索引 I 时进行 QIM 隐写的. 显然，进行 QIM 隐写势必使序列 S^* 的某些量化索引值发生变化，例如对于音素 P_k，设其原量化索引为 I_m，进行 QIM 隐写后可能变为 I_{m+1}，S^* 中索引的改变将导致 S 中音素 P_k 发生相应的改变，如变为 P_{k+1}. 音素的改变将使 S 中的音素分布特性发生变化. 因此，如能够有效量化 S 中音素的分布特性，则通过比较 QIM 隐写前后该特征的变化即可实现隐写检测.

2.2 音素分布特性的量化统计模型

为便于设计量化统计模型，我们首先给出本文中音素这一概念的形式化描述. 本文将音素 P 用三元组 (p,s,t) 表示，其中 p 为音素的语言学符号即音标，s 为音标 p 的发音是具有一定时长的语音小片段，t 为 s 的时长. 根据语音学理论，音素 P 为语音的基本组成单位，且特定语言所包含的音素是有限的，如英语包含 40 个音素[14]，本文假设有一种虚拟语言 L，它包含有限个音素，这些因素构成集合：$B=\{P_1,P_2,\cdots,P_{n-1},P_n\}$. 基于上述假设，属于虚拟语言 L 的一段语音 S 可以根据 B 中的音素分解为多个小片段，即可将 S 切分为多个按时序排列的语音分片 $S=f_1f_2\cdots f_{m-1}f_m$，分片 f_k 实质上是音素 P_l 的发音，即存在 $f_k=s_l(k\in[1,m],l\in[1,n])$，据此可将语音片段 S 表示为音素序列：$S=P_kP_l\cdots P_xP_y(k,l,x,y\in[1,n])$. 显然，属于虚拟语言 L 的任意一段语音都可由上文方法获得其对应的音素序列.

如将音素 P 视为一个单词，那么相应的可将语音片段 S 视为一个文档. 据此，借鉴自然语言处理中的文档量化表示模型：文档向量空间模型，我们可用音素向量空间模型(Phoneme Vector Space Model，PVSM)作为音素序列的量化表示模型，如图 2 所示.

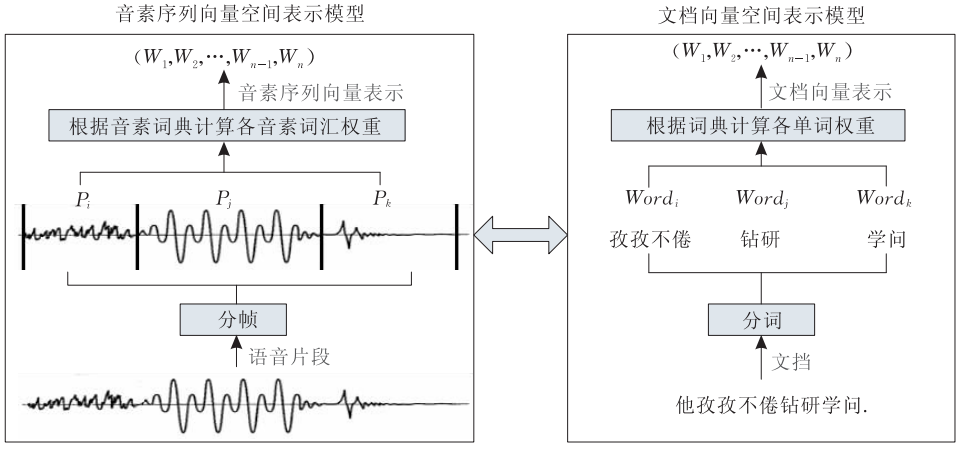


图 2 音素序列向量空间表示模型原理图

音素向量空间量化表示模型的正规定义如下.

定义 1.　虚拟语言 L 的音素集合 $B=\{P_1,P_2,\cdots,P_{n-1},P_n\}$，称 $P_i\in B$ 为音素词汇(Phoneme Word)，称 B 为语言 L 的音素词典，属于虚拟语言 L 的语音片段所包含的音素都在 B 中.

定义 2.　虚拟语言 L 的一段语音 S，通过查找音素词典，可切分为按时序排列的 N 个音素，称上述过程为基于音素的语音分帧.

定义 3.　设语音片段 S 分帧后所得的音素序列为 $S=P_kP_l\cdots P_xP_y$；根据音素词典 $B=\{P_1,P_2,\cdots,P_{n-1},P_n\}$ 可构造如下 n 维向量：$\boldsymbol{V}=\{W_1,W_2,\cdots,W_{n-1},W_n\}$ 对音素序列 S 进行量化表示，称 W_i 为音素词汇 P_i 的权重(它是与 P_i 在音素序列 S 中的分布相关的变量，其取值依据预先设定的计算规则求取)，称向量 $\boldsymbol{V}$ 对应的 n 维空间为音素向量空间，音素序列 S 可用该空间中的一个点表示；称上述定义构成的语音片段量化表示方法为音素向量空间量化表示模型，称 $\boldsymbol{V}$ 为 S 的音素向量.

本文音素 $P_i(1\le i\le n)$ 的权重 W_i 的计算规则如式(2)所示，

$$W_i=Count(P_i)/Sum(S) \tag{2}$$

其中 $Count(P_i)$ 表示音素词汇 P_i 在音素序列 $S=P_kP_l\cdots P_xP_y$ 的出现次数，$Sum(S)$ 表示 S 所包含的音素词汇总数. 据此，我们可计算出任一语音片段的音素向量 $\boldsymbol{V}$，它是一个 n 维向量.

如前文所述，音素在音素序列中的分布存在不均衡性和相关性，显然音素向量并没有体现音素分布的相关性特性. 为此，还必须设计相关性特性的量化统计模型. 根据语音产生模型，发音的基本单位为音素，发音过程实际上就是不断变换声道形态的过程，可将该过程视为离散时间随机过程$\{x(i),i>0\}$，$x(i)$ 表示音素发音时的声道形态，由于不同的声道形态对应不同的音素，因此可用音素来代表声道形态即取 $x(i)=P_k^i$，P_k^i 表示第 i 个时刻的声道正在发音素 P_k($P_k\in B$)的音. 据此，可将音素序列 $S=P_k^1P_l^2\cdots P_x^{N-1}P_y^N$ 视为声道状态转移序列. 根据语言学的统计规律，一般认为某个音素的出现仅与其前一个音素存在较大关联，鉴于此，本文假设下一个音素的出现仅与当前音素有关，即存在以下关系：

$$Pr(P^N/P^1P^2\cdots P^{N-1})=Pr(P^N/P^{N-1}) \tag{3}$$

据此可证，随机状态序列 $S=P_k^1P_l^2\cdots P_x^{N-1}P_y^N$ 为一阶马尔可夫链，即音素序列可视为声道(音素)状态转移一阶马尔可夫链. 显然，声道状态集合即音素集合 $B=\{P_1,P_2,\cdots,P_{n-1},P_n\}$. 根据上述性质，声道状态转移概率可用条件概率表示如下：

$$a_{i,j}=Pr(P_i/P_j),\ 1\le i,j\le n\ 且\sum_{j=1}^{M}a_{i,j}=1 \tag{4}$$

它表征了音素序列中各音素出现的相关性，可作为音素相关性的量化统计特征. 在实际计算时直接计算式(4)的条件概率较为困难，一般将其转化为联合概率进行计算，即根据条件概率公式将式(4)转化为式(5)：

$$a_{i,j}=Pr(P_i/P_j)=\frac{Pr(P_i,P_j)}{Pr(P_j)},\ P_i,P_j\in B \tag{5}$$

进行各音素间相关性的计算. 以 $a_{i,j}(1\le i,j\le n)$ 作为元素可获得一个 $n\times n$ 维的矩阵 $\boldsymbol{M}$，称该矩阵为音素状态转移矩阵. 显然，它量化不同音素出现的相关性.

综上，我们得到了音素分布不均衡性的量化表示(即音素向量 $\boldsymbol{V}$)以及音素分布相关性的量化表示(即音素状态转移矩阵 $\boldsymbol{M}$). 这两个不同角度量化特征必须进行融合，方能全面量化音素分布特性. 由于 $\boldsymbol{V}$ 和 $\boldsymbol{M}$ 的维度不同，我们首先对 $\boldsymbol{M}$ 进行降维操作，将其降为 n 维以便于和 $\boldsymbol{V}$ 进行融合. 对 $\boldsymbol{M}$ 降维后得到 n 维向量 $\boldsymbol{V}^*=\{R_1,R_2,\cdots,R_{n-1},R_n\}$，其中 $R_j(1\le j\le n)$ 的取值方法如下：

$$R_j=\max\{a_{1,j},a_{2,j},\cdots,a_{n-1,j},a_{n,j}\} \tag{6}$$

将 $\boldsymbol{V}^*$ 与 $\boldsymbol{V}$ 进行融合，获得融合向量 $\boldsymbol{H}=\{(W_1,R_1),(W_2,R_2),\cdots,(W_{n-1},R_{n-1}),(W_n,R_n)\}$ 作为音素分布特性的量化特征向量，下文称该向量为音素分布特征向量(Phoneme Distribution Feature Vector，PDFV).

2.3　分帧方法及音素集合的确定

上面，我们已经给出了音素分布特征的量化统计模型，但是要计算该量化特征，还必须针对不同的低速率编码标准确定音素集合以及分帧方法. G.729 和 G.723.1 是 ITU 为 VoIP 应用定义的低速率语音编码标准，因此，本文给出这两种编码器的音素集合和分帧方法，其它低速率编码器可类推.

语音中每个音素的持续时间是不等长的，例如浊音“o”可能持续 50 ms 以上，浊爆破音“b”则可能仅持续 10 ms，而且随着发音人及语速的不同其持续时长更是千变万化. 因此，音素的持续时长是很难事先确定的，这导致将一段语音进行基于音素的分帧甚为困难. 但是，本文利用低速率语音编码器都是对语音进行分帧处理这一事实解决这一问题. 例如，G.729 以 10 ms 为单位对语音进行分帧并对每帧计算一次 LPC 预测系数(即估计一次声道发音参数)，这意味着 G.729 认为在 10 ms 的短时内声道的形态是稳定的；假设不同的声道形态对应不同音素发音，

那么可以认为 G.729 中每帧对应一个音素或者是一个音素的一部分.根据对实际语音的统计,英语中音素的持续时长均值远大于10ms,这印证了上述结论的正确性.为此以 10ms 为界限,本文将时长不超过 10ms 的音素称为 α 类,反之为 β 类.作为一种近似,对于 α 类音素其时长设为 G.729 的帧长 l,对于 β 类音素设其时长为 $n\times l(n>1)$即 β 类音素包含多个 G.729 帧.我们发现 β 类音素发音时的信号波形一般具有周期性特征,例如图 1 中的音素"o"包含了 4 个明显的周期,此时一个周期的信号已可反映声道特征,因此对于 β 类音素在 G.729 中可视为对其声道参数进行了多次重复估计.鉴于此,本文认为对于 β 类音素,可分成 n 个帧分别进行 LPC 分析.综合上述分析,本文认为每个 G.729 帧可近似地跟一个音素对应(对于 β 类音素,可能连续几个帧都对应相同的音素),据此,对 G.729 压缩语音流直接以其原有的帧结构进行分帧即可.由于 G.729 对每个帧的 LPC 预测系数采用分级矢量量化,每个音素 P 的量化特征索引 $I=(i^1,i^2,i^3)$,其中 i^1 有 128 种取值,i^2 和 i^3 都有 32 种取值,因此,索引 I 共有 $128\times32\times32=131072$ 种取值,这意味着音素集合包含了 131072 个音素.音素集合太大,在音素序列的长度较小时不易凸显其统计特性(例如,设音素序列的长度为 100,此时音素集合中 99%以上的音素都将不在音素序列中出现,这将导致量化统计特征中很多元素的值为 0),因此必须对量化特征索引 I 进行降维.由于,一级矢量 i^1 与所有的 LPC 系数的量化有关其重要性超过了 i^2 和 i^3,而且 QIM 隐写是在 3 个分裂矢量量化时分别进行的,因此本文近似地取 $I=i^1$,即取 i^1 作为 G.729 的音素集合 B 中元素 P 的量化特征索引,据此可得 $B=\{i_1^1,i_2^1,\cdots,i_{127}^1,i_{128}^1\}$.所以,对于 G.729 其音素向量 $\boldsymbol{V}$ 与音素状态转移向量 $\boldsymbol{V}^*$ 都是 128 维向量,而融合向量 $\boldsymbol{H}$ 为 256 维向量.对于 G.723.1,基于类似的分析,仍可基于其压缩语音流的原始帧结构进行分帧并近似地取其第 1 个分裂矢量作为音素的量化特征索引,此时其音素集合 $B=\{i_1^1,i_2^1,\cdots,i_{255}^1,i_{256}^1\}$,对应的音素向量 $\boldsymbol{V}$ 与音素状态转移向量 $\boldsymbol{V}^*$ 都是 256 维向量,而融合向量 $\boldsymbol{H}$ 为 512 维向量.

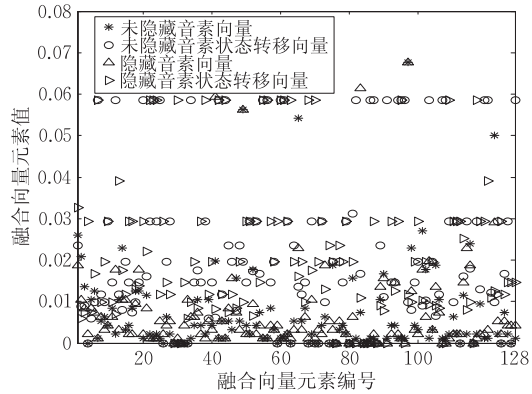
确定音素集合及分帧方法后,对于给定的压缩语音片段可方便地计算其融合特征向量 $\boldsymbol{H}$.图 3 给出了 QIM 隐写对融合特征向量扰动情况的分析结果.其中,图 3(a)是一段长度为 10s 的 G.729 压缩语音流片段及其使用文献[8]的方法进行 QIM 隐写后的融合特征向量 $\boldsymbol{H}$ 对比图,从该图可以看出隐写前后融合特征向量重合的点极少,这说明隐写前后融合特征向量的变化幅度较大.为了量化分析隐写对融合特征向量的扰动程度,本文引入向量变化率(Vector Variation Rate,VVR)对向量的改变进行衡量.设对某个压缩语音流片段,其在隐写前后计算所得的融合特征向量为 $\boldsymbol{H}_1$ 和 $\boldsymbol{H}_2$,VVR 定义为 $\boldsymbol{H}_1$ 中取值发生变化的子向量的比例,定义如下:

$$VVR=\sum_{i=1}^{N}\tau_i\Big/\sum_{i=1}^{N}\mu_i \tag{7}$$

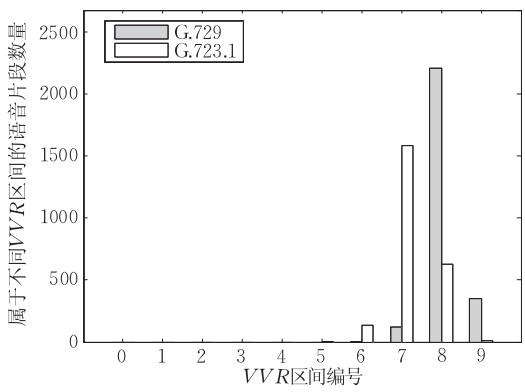
其中 N 为融合特征向量维数,μ_i 和 τ_i 定义如下:

$$\mu_i=\begin{cases}1, & a_i\neq0\\ 0, & \text{否则}\end{cases},\quad \tau_i=\begin{cases}1, & a_i\neq0 \text{ 且 } a_i\neq b_i\\ 0, & \text{否则}\end{cases} \tag{8}$$

其中 a_i 和 b_i 分别为 $\boldsymbol{H}_1$ 和 $\boldsymbol{H}_2$ 中第 i 维子向量的取值.显然,VVR 的值越大,隐写对融合特征向量的扰动幅度越大.将 VVR 的值域分为 10 个区间:$d_i=[i\times0.1,(i+1)\times0.1)$,其中 i 取值为 0～9;本文对实验部分所涉及的 2674 个不同语音片段使用 G.729 及 G.723.1 分别计算了其 VVR 值,图 3(b)统计了计算所得 VVR 值属于区间 d_i 的语音文件数量.从图 3(b)可以看出对于 G.729 和 G.723.1,文件对应的向量变化率值都超过 0.5,这意味着至少有一半以上的



(a) 隐写前后融合特征向量变化示例



(b) 隐写前后融合特征向量VVR值分布

图 3 进行 QIM 隐写对融合特征向量造成的扰动

融合特征向量中的子向量在隐写前后的取值发生了改变；对于 G.729，所有文件的 *VVR* 均值为 0.86，对于 G.723.1 该值为 0.68. 因此可以认为本文所提取的特征对隐写是非常敏感的——隐写将导致该特征发生显著性变化. 这对隐写检测非常有利.

3　基于机器学习的隐写检测

假设有一个未知是否存在 QIM 隐写的压缩语音片段 S，隐写检测的目标即判定 S 是否存在 QIM 隐写. 假设通过对 S 进行处理所抽取的可用于隐写检测的特征向量为 $\boldsymbol{t}$，则隐写检测过程可用式(9)表示：

$$y=f(\boldsymbol{t}),\ y\in\{+1,-1\} \tag{9}$$

其中函数 f 为隐写检测器其输出结果即为检测结果，若 $y=+1$，表示 S 不存在隐写，否则存在隐写. 显然函数 f 是一个二值分类器，隐写检测过程实质上是分类过程：假设 $y=+1$ 时 S 属于未隐写类(称为 cover 类)，$y=-1$ 时 S 属于隐写类(称为 stego 类)，则隐写检测就是将未知类别的样本 S 分为 cover 类或 stego 类. 对于分类问题，基于机器学习的分类方法是当前主流，本文也采用这种方法. 对于未知类别的压缩语音片段，本文基于机器学习的隐写检测过程如图 4 所示. 显然，隐写检测的关键是确定特征向量 $\boldsymbol{t}$ 和分类器 f.

图 4　压缩语音片段隐写检测过程

在文献[13]中，特征向量的提取必须首先对压缩语音片段进行解码，其后基于解码获得的语音数据计算基于 MFCC 的统计特征向量，这种特征提取方法需要进行解码操作，甚为耗时. 上一节中我们介绍了本文的特征提取方法，该方法不需要对压缩语音进行解码，直接在压缩域抽取特征向量，计算速度较快. 为此，本文将上节给出的音素分布特性量化向量 $\boldsymbol{H}$ 作为特征向量 $\boldsymbol{t}$.

关于分类器的设计，现有研究中，不同的对象分类识别系统有不同的训练方法，这些方法大致可分为两大类：判别法(discriminative approach)和生成法(generative approach). 判别法可以灵活地选择用来识别的特征，检测速度也较快，为此本文采用基于判别法的分类器. 在判别型分类器中，由于支持向量机(Support Vector Machine，SVM)较适合小样本训练的情况，本文考虑到训练时间和训练样本量，使用支持向量机作为分类器. SVM 分类器是一种监督学习分类器，它是通过使用某些已标注类别的样本进行训练获得的. 对于特征向量 $\boldsymbol{t}$，分类器 f 的训练和预测步骤如下：

(1) 获取尽可能多的 cover 类别低速率压缩编码语音片段，并使用 QIM 嵌入方法(分组码本使用文献[8]算法进行优化划分)进行隐写以获得 cover 类别中每个样本对应的 stego 样本，并做好标注；

(2) 抽取上一步骤所获得的两类样本的特征向量 $\boldsymbol{t}$，标记每个向量的类别；

(3) 训练分类器：使用上一步骤获得已标记类别的特征向量集合训练分类器，获得分类器 f；

(4) 使用分类器 f 对未知类别样本进行隐写检测：对于未知类别样本首先抽取特征向量 $\boldsymbol{t}$，将 $\boldsymbol{t}$ 作为分类器 f 的输入，分类器输出即为隐写检测结果. LIBSVM 是一个优秀的 SVM 工具，本文基于 LIBSVM 进行分类器的训练和预测.

4　实验及讨论

本文选择 G.729 和 G.723.1 作为实验测试所用的低速率语音编码器，并采用文献[8]给出的方法作为隐写算法. 本文针对两种编码器分别进行了本文隐写检测方法的性能测试，并与文献[13]给出的隐写检测方法进行了比较.

为了阐明算法具有较好的普适性，本文选择不同发音人的多个语音片段组成语音样本库. 所用语音片段样本包含 4 个种类，分别是中文男声(Chinese Man，CM)，包含 500 个语音片段；中文女声(Chinese Woman，CW)，包含 532 个语音片段；英文男声(English Man，EM)，包含 818 个语音片段；英文女声(English Woman，EW)，包含 824 个语音片段. 语音片段总计 2674 个. 每个语音片段的时长为 10 s，采样率为 8000 Hz，对每个采样点用 16 bit 进行量化，用 PCM 格式存储.

我们称没有进行信息隐藏的压缩语音片段为未隐写类(C 类)，否则称其为隐写类(S 类). 不同类别发音人的语音片段编码所得的 C 类及其对应的 S 类压缩语音流片段构成进行分类器进行训练和预测时的数据集. 由于本文已将隐写检测问题转化为分类问题，因此本文采用式(10)定义的分类准确率 *Precision* 对检测算法的性能的进行评估：

$$Precision=\frac{\hat{\lambda}+\hat{\theta}}{\lambda+\theta} \tag{10}$$

其中 λ 和 θ 是数据集中的 C 类和 S 类样本的个数，$\hat{\lambda}$ 和 $\hat{\theta}$ 则是被分类器准确判定类别的 C 类和 S 类样本

的个数.

4.1 低速率语音编码器 QIM 隐写检测结果

对语音样本库中 CM 中的每个 PCM 格式存储的语音片段使用 G.729 编码器进行压缩编码,获得没有进行信息隐藏的 500 个 G.729 压缩语音流片段,由于 G.729 的帧长为 10 ms,因此每个片段包含 1000 个 G.729 帧,这些压缩语音片段组成未隐写类别(C 类)样本.使用文献[8]介绍的 CNV 算法方法对 G.729 进行矢量量化时的 3 个分裂矢量码本进行优化划分,获得进行 QIM 嵌入的分组码本.再次对每个 PCM 格式的语音样本进行基于 G.729 标准的编码压缩,并且,在对每个帧的 LPC 系数进行矢量量化时使用 QIM 机制嵌入机密信息,获得包含隐藏信息的 500 个 G.729 压缩语音流片段,这些压缩语音片段组成隐写类别(S 类)样本.C 类及其对应的 S 类压缩语音流片段构成进行分类器训练和预测时的 CM 数据集.同理可得 CW、EM 和 EW 数据集.这 4 个数据集的所有样本构成混合(Hybrid)数据集.因此,本文在 5 个不同的数据集上评估了算法性能.

用类似的方法获得使用 G.723.1 作为低速率语音编码器时,进行检测算法性能评估的数据集.由于每个语音片段的长度为 10 s, G.723.1 的帧长为 30 ms,因此每个 G.723.1 压缩语音片段包含 333 个帧.

对上述的每个数据集,选择 75%的 C 类样本及其对应的 S 类样本,组成该种类分类器的训练样本库,剩余的 25%样本组成测试样本库用于评估训练所得分类器的分类准确性.表 1 给出了测试结果,表 1 中列 PDFV 是使用本文方法获得的隐写检测结果,列 MFCC 是使用文献[13]的方法获得的隐写检测结果.从测试结果看本文方法在 5 个测试数据集上均优于文献[13]的方法,在语音片段时长为 10 s 时,对于两种低速率语音编码标准,本文方法检测准确率均超过 98%,而文献[13]的方法对于 G.723.1 基本上无法有效检测:对 5 个数据集检测准确率均低于 60%.

表 1 语音片段时长为 10 s 时的测试结果

数据集名	使用 G.729 的结果/%		使用 G.723.1 的结果/%	
	PDFV	MFCC	PDFV	MFCC
CM	100.00	94.00	98.40	49.60
CW	100.00	88.72	96.80	52.26
EM	100.00	80.00	98.22	54.63
EW	100.00	77.43	97.87	56.55
Hybrid	99.98	86.70	98.62	52.76

上面获得的测试结果所用的语音片段的时长为 10 s.本文面向的是 VoIP 中低速率编码的压缩语音流的隐写检测;VoIP 中的语音流是实时流,进行隐写检测前必须进行流的存储.为了达到较快检测以及减少存储的数据量,显然达到可以接受的隐写检测准确率时,我们希望所需要存储的语音流时长越短越好.为此,我们在下文对语音片段时长与隐写检测的性能进行了评估.

4.2 压缩语音流时长对隐写检测结果的影响

为了评估语音片段时长对隐写检测结果的影响,首先根据不同的低速率编码器的帧长,对数据集中的 10 s 长度的语音片段进行截短处理.对于 G.729,由于其帧长为 10 ms,10 s 长度的语音片段总共包含了 1000 帧,截取前 $N(0<N\leq1000)$ 个帧编码所需的采样值,构成时长为 $0.01\times N$ s 的新的 CM、CW、EM、EW 和 Hybrid 数据集.对这些新的数据集进行分类器的训练并测试分类准确性.表 2 给出了不同语音片段时长时(N 取不同值)的检测结果.

表 2 压缩语音流时长变化时的 G.729 QIM 隐写检测结果

时长/s	CM 的检测结果/%		CW 的检测结果/%		EM 的检测结果/%		EW 的检测结果/%		Hybrid 的检测结果/%	
	PDFV	MFCC	PDFV	MFCC	PDFV	MFCC	PDFV	MFCC	PDFV	MFCC
0.10	69.16	53.60	65.22	52.26	67.84	57.07	66.20	52.91	74.33	58.37
0.15	78.14	59.60	80.35	57.89	78.42	63.41	74.75	56.80	83.38	61.21
0.20	85.42	58.80	87.59	57.52	85.57	63.66	81.67	58.50	89.60	61.73
0.40	94.61	66.40	94.73	60.53	94.25	67.07	93.56	60.44	95.92	66.29
0.80	99.40	67.60	98.21	65.41	99.02	77.07	98.66	62.86	99.14	69.51
1.60	99.90	77.60	100.00	67.67	99.87	75.12	99.75	67.23	99.85	75.11
3.20	100.00	87.20	100.00	73.68	100.00	77.32	99.93	74.76	100.00	78.92
4.80	100.00	89.60	100.00	81.95	100.00	75.59	100.00	75.24	100.00	81.54
6.40	100.00	89.60	100.00	86.84	100.00	76.83	100.00	80.34	100.00	84.45
8.00	100.00	94.00	100.00	88.35	100.00	77.07	100.00	81.30	100.00	86.92

为了更直观地比较两种方法的性能,图 5 给出了 5 个数据集的平均检测准确率与语音片段时长的关系图.从该图可以看出,随着语音片段时长的增加,隐写检测准确率也随之提升;本文方法在任一时长下其检测准确率均优于文献[13]的方法;在语音片段时长为 0.40s 时本文方法已能够达到有效检测(检测准确率已经超过 90%),而此时文献[13]的方法仍不超过 70%.因此,对于 G.729,在语音片段时

长较小时本文方法性能远优于文献[13]；在语音片段时长较大时(超过 4.8 s)，本文达到 100%的隐写检测准确率，这一点是文献[13]无法达到的.

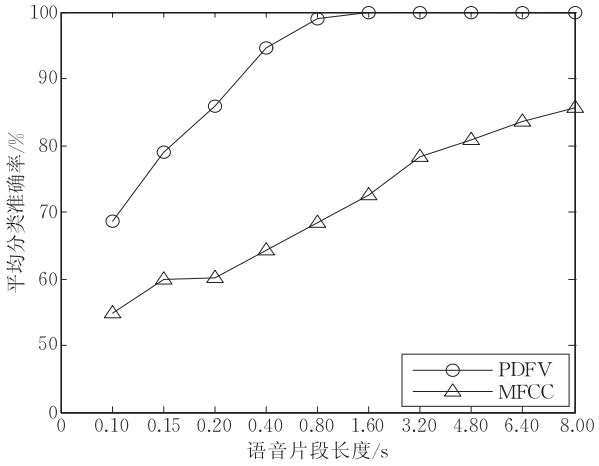


图 5　5 个数据集的 G.729 QIM 隐写检测平均准确率

由于 G.723.1 的帧长为 30 ms，10 s 长度的语音片段总共包含了 333 帧，仍截取前 $N(0<N\leq 333)$ 个帧编码所需的采样值，构成时长为 $0.03\times N$ s 的新的 CM、CW、EM、EW 和 Hybrid 数据集. 对这些新的数据集进行分类器的训练并测试分类准确性. 表 3 给出了不同语音片段时长时(N 取不同值)的检测结果. 为了更好地比较两种方法的性能，图 6 给出了 5 个数据集的平均检测准确率与语音片段时长的关系图. 从该图可以看出，随着语音片段时长的增加，本文方法的隐写检测准确率也随之提升，但是文献[13]的方法其检测准确率一直低于 60%(可以认为无法对隐写作出检测). 其原因可能是 G.723.1 每 30 ms 的采样值采用文献[8]的 QIM 隐写方法仅嵌入 3 bit 秘密信息，嵌入率太低导致解码后的语音采样值序列并不因隐写而产生较大的改变，这使基于采样值序列统计的特征对隐写不够敏感，从而导致检测率低. 但是本文方法是压缩域方法，不考察解码后的语音数据，因此仍能获得较好的隐写检测准确率：在语音片段时长较大超过 6 s 时，本文方法检测准确率超过 90%.

表 3　压缩语音流时长变化时的 G.723.1 QIM 隐写检测结果

时长/s	CM 的检测结果/%		CW 的检测结果/%		EM 的检测结果/%		EW 的检测结果/%		Hybrid 的检测结果/%	
	PDFV	MFCC	PDFV	MFCC	PDFV	MFCC	PDFV	MFCC	PDFV	MFCC
0.30	52.79	56.80	45.95	54.14	44.74	50.24	45.69	47.82	49.71	48.80
0.60	55.48	52.40	46.52	51.88	49.75	50.98	50.60	50.73	53.96	50.45
1.20	57.48	56.80	51.12	47.74	56.47	52.44	60.74	55.10	68.37	48.73
2.40	71.85	54.40	64.56	51.50	80.01	52.44	74.27	52.18	80.18	52.47
3.60	87.62	51.20	80.45	51.13	85.81	53.91	87.10	53.64	88.78	52.02
4.50	89.42	50.00	85.90	56.39	90.95	53.17	89.44	53.88	91.70	51.79
6.00	92.61	57.20	90.97	53.38	93.82	51.95	94.23	53.40	95.12	53.44
7.50	96.10	50.00	94.73	57.89	95.90	52.68	97.02	52.43	97.23	52.32
9.00	97.90	51.20	95.86	53.38	97.31	49.02	97.63	55.10	98.09	53.14

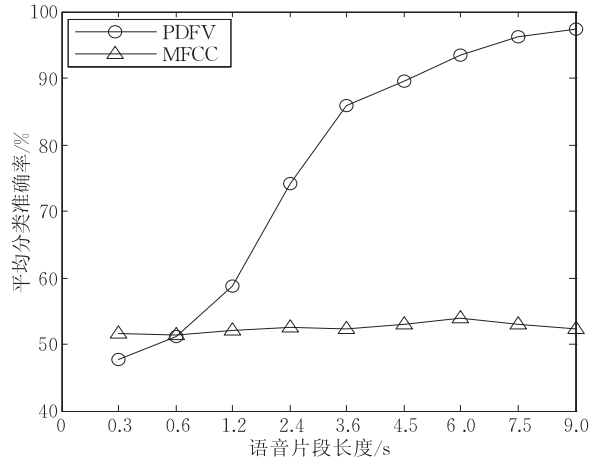


图 6　5 个数据集的 G.723.1 QIM 隐写检测平均准确率

根据上述实验，本文方法对于两种典型的低速率语音编码器中的 QIM 隐写均能有效检测，检测性能远优于时域特征抽取方法.

5　总　结

本文对在低速率语音编码过程中的 QIM 隐写给出了高效的检测方法. 本文发现一段语音中的音素其分布存在不均衡性和相关性，据此本文提出了一种基于压缩域的隐写检测特征抽取方法，并结合支持向量机构建了隐写检测分类器. 与基于时域的特征抽取方法相比，本文方法不仅具有较高的检测准确率，而且节省了压缩语音的解码时间，实现了对压缩语音流的快速隐写检测. 本文方法借鉴了文档的向量空间表示方法及其分类模型，正是利用这些方法所蕴含的深刻思想建立了本文的隐写检测算法. 本文方法为隐写检测提供了一种新的思路.

参　考　文　献

[1]　Mazurczyk W, Lubacz J. Steganography of VoIP streams//

Proceedings of the 3rd International Symposium on Information Security. Monterrey, Mexico, 2008: 1001-1018

[2] Huang Y, Xiao B, Xiao H. Implementation of covert communication based on steganography//Proceedings of the 4th International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Harbin, China, 2008: 1512-1515

[3] Li X, Yu H H. Transparent and robust audio data hiding in cepstrum domain//Proceedings of the IEEE International Conference on Multimedia and Expo. New York, USA, 2000: 397-400

[4] Wang C T, Chen T S, Chao W H. A new audio watermarking based on modified discrete cosine transform of MPEG/Audio Layer III//Proceedings of the IEEE International Conference on Networking, Sensing and Control. Taipei, China, 2004: 265-277

[5] Wu S, Huang J, Huang D, et al. Effciently self-synchronized audio watermarking for assured audio data transmission. IEEE Transactions on Broadcasting, 2005, 51(1): 69-76

[6] Chen B, Wornell G W. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. IEEE Transactions on Information Theory, 2001, 47(4): 1423-1443

[7] Huang Y, Tang S, Yuan J. Steganography in inactive frames of VoIP streams encoded by source codec. IEEE Transactions on Information Forensics and Security, 2011, 6(2): 296-306

[8] Xiao Bo, Huang Yongfeng, Tang Shanyu. An approach to information hiding in low bit-rate speech stream//Proceedings of the IEEE Global Communications Conference. New Orleans, USA, 2008: 1940-1944

[9] Malik H. Statistical modeling of footprints of QIM steganography//Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME 2010). Singapore, 2010: 1487-1492

[10] Malik H, Subbalakshmi K P, Chandramouli R. Nonparametric steganalysis of QIM data hiding using approximate entropy. IEEE Transactions on Information Forensics and Security, 2012, 7(2): 418-431

[11] Malik H. Steganalysis of QIM steganography using irregularity measure//Proceedings of the 10th ACM Workshop on Multimedia and Security. Oxford, UK, 2008: 149-158

[12] Wu Qinxia, Li Weiping, Yu Xiao Yi. Revisit steganalysis on QIM-based data hiding//Proceedings of the 5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Kyoto, Japan, 2009: 929-932

[13] Liu Qingzhong, Sung Andrew H, Qiao Mengyu. Temporal derivative-based spectrum and mel-cepstrum audio steganalysis. IEEE Transactions on Information Forensics and Security, 2009, 4(3): 359-368

[14] Quatieri F Thomas. Discrete-Time Speech Signal Processing: Principles and Practice. NJ, USA: Prentice Hall PTR, 2002

LI Song-Bin, born in 1981, Ph. D., associate researcher. His research interests include multimedia information processing and information hiding.

HUANG Yong-Feng, born in 1967, Ph. D., professor. His research interests include multimedia network security and next generation Internet.

LU Ji-Cang, born in 1985, Ph. D. candidate. His research interests include steganography and steganalysis.

Background

Quantization Index Modulation (QIM) steganography was proposed by Chen B et al. from Massachusetts Institute of Technology in 2001. The QIM steganography hides the secret information during the scalar or vector quantization process which is the necessary step in most digital audio and video (image) compression standards. By connecting with the coding process closely, the additional distortion caused by the QIM steganography is so little that perception of the steganography is very hard. Therefore, detection of the QIM steganography is a very challenging work.

Undoubtedly, attack of the QIM steganography has attracted many researchers' interest. However, most prior works mainly focus on detection of the QIM steganography during the image compression coding, and the research achievements can not be directly applied to audio which is significantly different from image. Therefore, this paper studies the detection method of the QIM steganography in low bit-rate audio compression coding.

This paper proposes a novel method for QIM steganography detection. Based on the speech generation and compress coding theory, this paper firstly gives deep analysis of the possible significant feature degradation caused by the QIM steganography in audio compressed stream. And then, it presents the statistical models to extract the significantly changed features in the compressed domain. By combining the extracted features with supervised learning classifier, this paper builds a high performance detector towards the QIM steganography in low bit-rate compressed audio stream.

The research work in this article has been partially supported by the National Natural Science Foundation of China under Grant Nos. 60970148, 61271392. Under these supports, we aims to build the whole theory system of streaming-media-based information hiding and we have published dozens of papers at many high impact journals in this area.