一种基于码字分布特性的G.729A压缩语音流隐写分析方法

李松斌[1],孙东红[2],袁 键[1],黄永峰[1]
(1.清华大学电子工程系,北京 100084; 2.清华大学信息网络工程研究中心,北京 100084)

摘 要: 有学者提出了一种在压缩语音编码过程中进行QIM(Quantization Index Modulation)隐写的方法.该方法可用于在G.729A压缩语音流中高隐蔽性地嵌入秘密信息,研究其隐写分析方法很有必要.本文首先分析了QIM隐写对G.729A码流造成的显著性特征变化,发现该种隐写将使码流中LPC滤波器的量化索引(码字)发生转移,并导致码字分布的不均衡性及相关性特性发生改变.本文设计了统计模型,实现了对码字分布特性的量化特征抽取;结合支持向量机,本文构造了用于隐写检测的集成分类器系统.实验结果显示本文方法能够在低于30ms的时间内,获得超过98%的检测准确率,实现了对QIM隐写的快速有效检测.

关键词: 信息隐藏; 隐写分析; G.729A; 量化索引调制; 码字分布特性
中图分类号: TN918 **文献标识码:** A **文章编号:** 0372-2112 (2012) 04-0842-05
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.04.036

A Steganalysis Method for G.729A Compressed Speech Stream Based on Codeword Distribution Characteristics

LI Song-bin[1], SUN Dong-hong[2], YUAN Jian[1], HUANG Yong-feng[1]
(1. *Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*;
2. *Network Research Center, Tsinghua University, Beijing* 100084, *China*)

Abstract: An improved QIM (Quantization Index Modulation) steganography was proposed and it can be used for efficient information hiding in G.729A compressed speech stream. This paper wants to detect this type of steganography. This paper proves that such steganography will significantly change the imbalance and correlation distribution characteristics of codeword (quantization index) in the stream. And then it designs statistical models to extract the quantitative feature vectors of these characteristics. Combining the extracted vectors with the support vector machine, this paper constructs an ensemble classifier for detecting the QIM-based steganography in the G.729A speech stream. The experimental results show that the classifier can achieve up to 98% correct detection towards G.729A encoded speech stream with detection time less than 30 milliseconds.

Key words: information hiding; steganalysis; G.729A; quantization index modulation; codeword distribution characteristics

1 引言

G.729标准是ITU定义的VoIP(Voice Over IP)语音编码标准,其简化版本G.729A在VoIP得到广泛应用.这使G.729A压缩语音流成为一种潜在的极具威胁的信息隐藏载体,利用它进行隐蔽通信将对国家通信监管形成巨大威胁,研究其隐写分析方法很有必要.

当前在压缩编码语音中进行信息隐藏的方法可大致分为以下几类:其一是基于改写压缩语音流中的某些信息域的方法[1];其二是基于变换域的方法[2,3];其三是基于量化索引调制(Quantization Index Modulation,QIM)的方法[4].由于QIM隐写与语音编码过程中的矢量量化步骤紧密结合,几乎不增加任何编码延迟(前两种方法引入的延迟较大),因此非常适合用于在VoIP应用中建立隐蔽通信信道.据此,最近文献[5]针对低速率语音编码提出了一种改进的QIM信息隐藏方法,它首先基于互补邻居节点(Complementary Neighbor Vertex,CNV)算法对量化用的矢量码本进行优化划分,其后利用所得的分组码本进行QIM隐写.该方法对压缩编码语音所引入的附加失真非常小,导致对其进行隐写分析非常困难,本文试图攻克这一难题.

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[6~10].例如,文献[6]发现进行QIM信息隐藏会对

收稿日期:2011-05-12;修回日期:2011-11-14
基金项目:国家自然科学基金(No.60970148);国家863高技术研究发展计划(No.2011AA010704);清华大学自主科研计划(No.2011THZ0)

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结合预先确定的似然率参数实现对 QIM 隐写的检测.文献[7,8]的方法与此类似,其差异主要在于衡量局部不规则性的模型不同.文献[9]构造了图像块 DCT 系数直方图变化与机密信息长度之间的估计公式,实现了对 QIM 隐写嵌入率的估计;文献[10]则给出了一种根据矢量空间中码字的出现概率估计嵌入信息长度的方法.

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,对于语音编码中 QIM 隐写的检测其难点也在于寻找并确定 QIM 隐写后所引起的显著性特征变化.

2　QIM 隐写引起的显著性特征变化分析

G.729A 采用基于合成-分析法的线性预测编码(AbS-LPC)方法.LPC 分析所获得的 LPC 合成滤波器如式(1)所示:

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

其中 a_i 为滤波器系数.由于语音信号仅具有短时平稳特性,因此编码时要将语音分割为较小时长的帧(G.729A 设定帧长为 10ms),对每一帧分别进行 LPC 分析.

G.729A 编码时首先对每帧求解最优的 LPC 预测系数,并将其转换为 LSF(线谱频率)系数,然后使用 3 个分裂矢量(f_1,f_2,f_3)对 LSF 系数进行矢量量化.每个 f_i $(i=1,2,3)$有一个对应的码书 $L_i(i=1,2,3)$其码字空间为$\{v_i^1,v_i^2,\cdots,v_i^{|L_i|}\}$,其中 v_i^k 表示码书 L_i 的第 k 个码字,$|L_i|$为 L_i 中码字数量.量化时为每个分裂矢量 f_i 从码书 L_i 中选择一个最优码字,以使量化后的 LSF 系数失真最小.码书 L_1、L_2、L_3 分别包含 128、32、32 个码字.

QIM 隐写在 f_i 挑选最优码字时进行.以码书 L_1 为例,文献[5]进行 QIM 隐写时首先使用 CNV 算法将码书 L_1 均分为 L_{11}和 L_{12}两部分,满足:

$$L_{11}\cap L_{12}=\varnothing \text{ 且 } L_{11}\cup L_{12}=L_1 \tag{2}$$

L_{11}和 L_{12}分别代表比特"0"和"1";当嵌入 0 时仅在分组码书 L_{11}中选取最佳量化值,嵌入 1 时则仅在分组码书 L_{12}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 L_{11}还是 L_{12}来恢复机密信息比特.

显然,QIM 隐写将使 f_i 的量化索引(码字)发生变化:正常量化时为码书中的第 i 个码字进行隐写后可能转移为第 j 个码字.这种变化将可能使码流中的 3 个量化索引序列(Quantization Index Sequence,QIS)中码字的分布特性发生改变.显然,如能有效量化该种改变,则可据此进行隐写分析.

3　码字分布特性的量化统计模型

QIM 隐写将可能导致码字的出现频率及码字出现的相关性发生改变.本文称上述特性为码字分布的不均衡性及相关性特性.下面,给出上述两类特性的量化特征提取方法,并通过对大量样本的统计证明 QIM 隐写将导致这两类特性发生显著性改变.

将码流中某个分裂矢量对应的量化索引序列简记为 $S=c_1\cdots c_j\cdots c_N$,其中 $c_j(j\in[1,N])$表示按时序排在第 j 个位置的码字,L 为码书包含有限个码字,c_j 所能取的值设为 $v_i(v_i\in L)$.S 中码字分布的不均衡性特性用式(3)所示的码字分布不均衡性特征向量进行量化表示:

$$\boldsymbol{H}=(h_1,h_2,\cdots,h_n) \tag{3}$$

其中 n=|L|为码书 L 所包含的码字数量,$h_i(i\in[1,\mathrm{n}])$表示 S 中码字取值为 v_i 的概率,定义为:

$$h_i=\sum_{j=1}^{N}Pr_{i/j}\cdot Pr_j=\frac{1}{N}\sum_{j=1}^{N}Pr_{i/j} \tag{4}$$

其中,Pr_j 表示从 S 中选择第 j 个位置码字的概率,$Pr_{i/j}$代表 c_j 取值为 v_i 的条件概率,其定义为:

$$Pr_{i/j}=\begin{cases}1, & \text{if } \ s_j=v_i\\ 0, & \text{else}\end{cases} \tag{5}$$

为了量化分析码字分布之间的相关性,本文将 $S=c_1\cdots c_j\cdots c_N$ 视为一个随机序列,其中 $c_j(j\in[1,N])$表示码字随机变量在第 j 个时刻的状态.为便于计算,假设每个码字的出现仅与其前一个码字有关,据此可用一阶马尔科夫链对 S 进行建模,以此可用状态转移概率对码字出现的相关性进行量化表示.各状态(码字)间的转移概率可用式(6)计算:

$$Pr_{\alpha/\beta}=Pr(c_{j+1}=\alpha/c_j=\beta)\ ,\ \alpha,\beta\in L \tag{6}$$

直接计算式(6)的条件概率较为困难,故将其转化为联合概率进行计算,如式(7)所示:

$$Pr_{\alpha/\beta}=\frac{Pr(c_{j+1}=\alpha,c_j=\beta)}{Pr(c_j=\beta)},\ \alpha,\beta\in L \tag{7}$$

使用式(7),对于码字序列 S,可获得一个$|L|^2$ 维的状态转移矩阵 A_{ij},其中元素 e_{ij}表示L 中第i 个码字转移到第 j 个码字的概率.显然,该特征量化了相邻码字出现的相关性,但其维度太高难以实用.为此,采用选维技术对 A_{ij}进行降维处理,对每个当前状态仅统计它具有最大转移概率的下一状态对应的转移概率,降维后将获得用于衡量码字相关性的码字分布相关性特征向量,定义如下:

$$\boldsymbol{T}=(p_1,p_2,\cdots,p_n)\ ,\ n=|L| \tag{8}$$

其中,p_i 表示 L 中第 i 个码字与其最可能伴随状态之间转移概率,定义如下:

$$p_i = \max\{e_{i1}, e_{i2}, \cdots, e_{i|L|},\} \tag{9}$$

为了证明 QIM 隐写将导致码字分布特性发生显著性改变.本文定义了向量变化率(Vector Variation Rate, VVR):设对向量 $\boldsymbol{V}$ 进行某种操作将使 $\boldsymbol{V}$ 中某些维的取值发生改变,并得到新向量 $\boldsymbol{V}^*$,VVR 定义为 $\boldsymbol{V}$ 中取值发生变化的子向量的比例,表示如下:

$$\mathrm{VVR} = \sum_{i=1}^{N} \tau_i \Big/ \sum_{i=1}^{N} \mu_i \tag{10}$$

其中 N 为向量 $\boldsymbol{V}$ 的维数,μ_i 和 τ_i 定义如下:

$$\mu_i = \begin{cases} 1 & \text{如 } a_i \neq 0 \\ 0 & \text{其他} \end{cases}, \tau_i = \begin{cases} 1 & \text{如 } a_i \neq 0 \text{ 且 } a_i \neq b_i \\ 0 & \text{其他} \end{cases} \tag{11}$$

其中 a_i 和 b_i 分别为 $\boldsymbol{V}$ 和 $\boldsymbol{V}^*$ 中第 i 维子向量的取值.设未隐写时抽取的码字分布不均衡性特征向量为 $\boldsymbol{H}$,隐写后为 $\boldsymbol{H}^*$,则据式(10)可计算 QIM 隐写对 $\boldsymbol{H}$ 的扰动情况.用同样的方法可计算 QIM 隐写对码字分布相关性特性向量 $\boldsymbol{T}$ 的扰动情况.为便于统计 $\boldsymbol{H}$ 和 $\boldsymbol{T}$ 的扰动幅度,将 VVR 的值域分为 10 个区间:$d_i = [i \times 0.1, (i+1) \times 0.1]$,其中 i 取值为 0 至 9.本文对 2674 个不同发音人语音片段计算了隐写前后向量 $\boldsymbol{H}$ 和 $\boldsymbol{T}$ 向量的 VVR 值,并统计了所得 VVR 值属于区间 d_i 的语音文件数量占所有文件的比例,结果如图 1 所示.

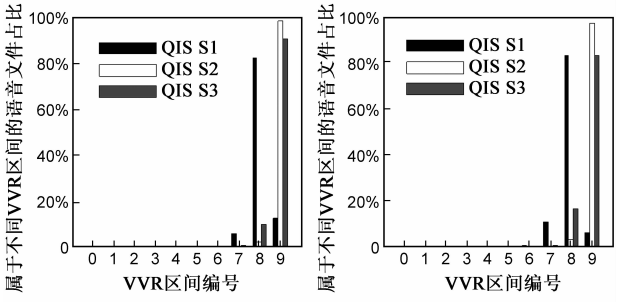

图1 QIM隐写对码字分布特性的改变幅度统计

从图 1 可以看出 3 个码字序列的 $\boldsymbol{H}$ 向量变化率值都超过 0.7,这意味着 $\boldsymbol{H}$ 中超过 70% 的子向量在隐写前后的取值发生了改变.对于 $\boldsymbol{T}$ 向量该值为 60%.显然,隐写确实导致了码字分布特性发生显著性变化.这对隐写检测非常有利.

4 基于分类器集成的隐写检测过程

本文隐写分析的目标是判定码字序列 S 是否存在 QIM 隐写,其判别结果只有"是"(本文称为 stego 类别)和"否"(本文称为 cover 类别)两类.因此,隐写检测过程实质上是分类过程.本文对于未知类别样本的类别判定过程如图 2 所示.

包含两个主要步骤:其一是提取码字序列 S 的特征向量,其二是利用所获得特征向量基于分类器进行类别检测.分类器一般采用有监督学习的方法获得即通过使用某些已标注类别的样本进行训练获得分类器,本文采用支持向量机(Support Vector Machine, SVM)作为分类器.

上一节中我们介绍了提取码字分布的不均衡性及相关性特征向量 $\boldsymbol{H}$ 和 $\boldsymbol{T}$ 的方法,$\boldsymbol{H}$ 和 $\boldsymbol{T}$ 必须进行融合才能得到完整的表征码字分布特性的特征向量.但是,由于 $\boldsymbol{H}$ 和 $\boldsymbol{T}$ 分属不同维度,对其进行融合很容易造成相互干扰.为此,本文对码字序列提取了 3 类特征向量,分别是单一维度特征 $\boldsymbol{H} = (h_1, h_2, \cdots, h_n)$ 和 $\boldsymbol{T} = (p_1, p_2, \cdots, p_n)$ 以及融合特征 $\boldsymbol{M} = (h_1, p_1, h_2, p_2, \cdots, h_n, p_n)$,首先使用这三类特征分别进行类别检测,然后再对检测结果进行综合.如图 2 所示.最终的检测结果为:

$$Y = \mathrm{sgn}(y_1 + y_2 + y_3) \tag{12}$$

其中 Y 及 $y_i (i = 1,2,3)$ 的取值均为正 1 或负 1,为正 1 时表示样本存在 QIM 隐写,否则为负 1.G.729A 压缩码流中包含 3 个码字子序列,只要任意一个子序列的分类结果 $Y = 1$ 则可判定码流中存在 QIM 隐写.

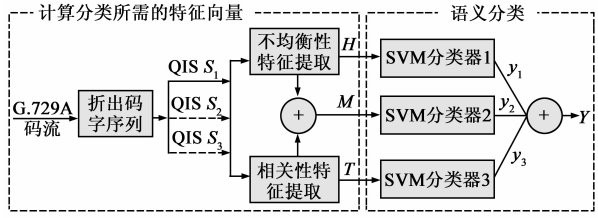

图2 G.729A压缩码流隐写检测过程

5 实验及讨论

为了证明本文方法具有较好的普适性,在多个语音样本数据集上进行了实验.数据集首先根据发音人类别的不同,分为以下四类,分别是中文男声(简记为 CM)、中文女声(CW)、英文男声(EM)以及英文女声(EW),这四类数据集均包含多个发音人对应的语音片段,数量分别为 500、532、818 和 824.将这四类数据集进行混合将得到第五个数据集(Hybird).数据集中语音样本时长均为 10s,以 PCM 格式存储.对每个数据集中的语音片段进行压缩编码获得 cover 类别和 stego 类别的压缩语音流样本.在每个数据集中,取 75% 的 cover 类别压缩码流及其对应的 stego 类别压缩码流组成训练集,剩余的 25% 组成测试集.

本文使用检测准确率 C 对隐写检测的有效性进行衡量,定义如下:

$$C = (E^* + Q^*)/(E + Q) \tag{13}$$

其中 E 和 Q 是预测集中的 cover 类别和 stego 类别的压缩码流样本个数,E^* 和 Q^* 则是被集成分类器准确判定类别的 cover 类别和 stego 类别的个数.本文将分析不同码流时长时的检测算法性能,由于样本数据集中每

个语音片段的时长为 10s,编码后将包含 1000 个 G.729A 帧,因此当对时长为 $t(0<t<10)$s 的压缩码流进行分析时,截取前 $t/0.01$ 个帧组成的码流进行隐写分析.

对 3 个码字序列 S_1、S_2 和 S_3 的实验结果如表 1 所示.由表 1 可知在码流长度大于 3.2s 时,对于五个数据集 3 个码字序列的检测准确率均超过 98%.从实验还可以看出本文方法在 G.729A 码流长度较小时对码字序列 S_1 的检测性能略低于码字序列 S_2 和 S_3.其原因主要是码字序列 S_1 的码书空间较大为 128 而 S_2 和 S_3 仅为 32,这导致码流时长较小时无法充分凸显码字分布的统计特性;在码流长度逐步增大后这种性能差异逐步减少.

为了分析本文检测算法的效率,本文统计了时长为 3.2s 时五个数据集的测试集进行隐写检测的平均耗时情况,如图 3 所示.从图中可以看出对单一样本的检测耗时低于 30ms(测试用机器处理器主频为 2.27GHz).因此本文约可在 30ms 时间内判定码流中是存在 QIM 隐写,基本实现了接近实时的快速隐写检测,这对 VoIP 实时流的 QIM 隐写检测是非常有利的.

表 1　隐写检测结果

时长 (s)	CM 数据集			CW 数据集			EM 数据集			EW 数据集			Hybrid 数据集		
	QIS S1	QIS S2	QIS S3	QIS S1	QIS S2	QIS S3	QIS S1	QIS S2	QIS S3	QIS S1	QIS S2	QIS S3	QIS S1	QIS S2	QIS S3
0.10	69.85%	90.62%	79.94%	76.13%	92.20%	81.49%	75.52%	89.73%	86.43%	77.42%	92.91%	85.88%	71.25%	93.46%	86.69%
0.20	77.76%	96.31%	91.62%	81.92%	96.33%	90.13%	77.86%	96.39%	91.75%	82.26%	97.37%	91.63%	78.66%	97.66%	92.88%
0.40	78.76%	98.30%	97.51%	87.82%	98.87%	97.37%	79.50%	98.53%	97.62%	86.88%	98.90%	97.86%	85.78%	98.60%	97.66%
0.80	91.68%	99.60%	99.80%	95.13%	99.53%	99.81%	96.42%	99.69%	99.22%	90.52%	99.63%	99.82%	93.23%	99.78%	99.89%
1.60	99.60%	100%	100%	98.68%	100%	99.91%	98.43%	100%	100%	96.78%	100%	100%	98.43%	99.98%	100%
3.20	100%	100%	100%	100%	100%	100%	98.71%	100%	100%	99.92%	100%	100%	99.76%	100%	100%
4.80	100%	100%	100%	100%	100%	100%	99.96%	100%	100%	100%	100%	100%	99.85%	100%	100%
6.40	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99.94%	100%	100%
8.00	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

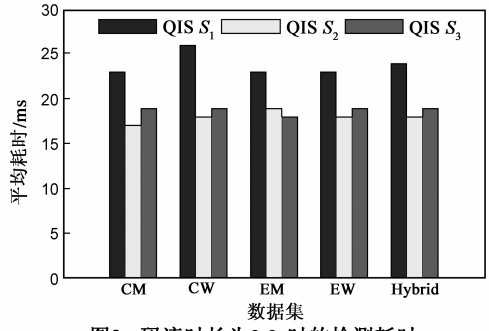

图3　码流时长为3.2s时的检测耗时

6　总结

本文对 G.729A 编码过程中的 QIM 隐写的检测算法进行了研究.本文发现 QIM 隐写将导致码字分布的不均衡性及相关性统计特性发生改变,并据此构造了隐写检测算法.最后,通过对大量语音样本的实验,证实了本文方法的有效性.

参考文献

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

[2] 谭良,吴波,刘震,周明天.一种基于混沌和小波变换的大容量音频信息隐藏算法[J].电子学报,2010,38(8):1812 – 1818.
TAN Liang, WU Bo, LIU Zhen, ZHOU Ming-tian. An audio information hiding algorithm with high-capacity which based on chaotic and wavelet transform [J]. Acta Electronica Sinica, 2010, 38(8): 1812 – 1818. (in Chinese)

[3] 白剑,景晓军,杨榆,徐迎晖,钮心忻,杨义先.语音信息隐藏中的 AERA 算法[J].电子学报,2005,33(9):1541 – 1544.
BAI Jian, JING Xiao-jun, YANG Yu, XU Ying-hui, NIU Xin-xin, YANG Yi-xian. AERA algorithm on speech hiding system [J]. Acta Electronica Sinica, 2005, 33(9): 1541 – 1544. (in Chinese)

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

[5] Xiao Bo, Huang Yongfeng, Shanyu Tang. An approach to information hiding in low bit-rate speech stream[A]. IEEE GlobeCom2008 [C]. New Orleans, USA, 2008. 1940 – 1944.

[6] Malik Hafiz. Statistical modeling of footprints of QIM

steganography [A]. 2010 IEEE International Conference on Multimedia and Expo[C]. IEEE, 2010. 1487 – 1492.

[7] Malik Hafiz, K P Subbalakshmi, R Chandramouli. Nonparametric steganalysis of QIM data hiding using approximate entropy [A]. Proc SPIE, Vol 6819 [C]. CA, USA, 2008. 681914 – 681921.

[8] Malik Hafiz. Steganalysis of QIM steganography using irregularity measure [A]. Proceedings of the 10th ACM Workshop on Multimedia and Security[C]. ACM, 2008. 149 – 158.

[9] Qinxia Wu, Weiping Li, Xiao Yi Yu. Revisit steganalysis on QIM-based data hiding [A]. The 5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing[C]. IEEE, 2009. 929 – 932.

[10] 吴▮,张勇,李岳楠,牛夏牧. VQ 域信息隐藏检测算法[J]. 电子学报, 2005, 33(B12): 2549 – 2551.
WU Di, ZHANG Yong, LI Yue-nan, NIU Xia-mu. Detection of VQ Based Information Hiding [J]. Acta Electronica Sinica, 2005, 33(B12): 2549 – 2551. (in Chinese)

作者简介

李松斌 男, 1981 年生于福建漳州. 2010 年毕业于中国科学院声学研究所获工学博士学位. 现为清华大学电子工程系博士后. 主要研究方向为多媒体信息处理、多媒体内容安全、隐蔽通信等.

E-mail: lisb@mail.tsinghua.edu.cn

孙东红 女, 1974 年生于黑龙江哈尔滨, 博士. 现为清华大学网络中心助理研究员. 主要研究方向为网络与信息安全, 从事该领域的体系结构、技术标准、应用技术、应急响应、系统研发、平台建设等方面开展研究和开发工作.