基于小数基音延迟相关性的自适应多速率语音流隐写分析

田晖[1),2),3)] 吴俊彦[1),2),3)] 严艳[1),2),3)] 王慧东[1),2),3)] 全韩彧[1),2),3)]

[1)](华侨大学计算机科学与技术学院，厦门 361021)

[2)](华侨大学厦门市数据安全与区块链技术重点实验室，厦门 361021)

[3)](华侨大学福建省大数据智能与安全重点实验室，厦门 361021)

摘要 网络语音流隐写分析是信息隐藏检测领域中的一个研究热点。针对自适应多速率语音流隐写检测问题，本文提出了一种基于小数基音延迟相关性的隐写分析方案。首先通过理论分析和实验对比验证了小数基音延迟相关性作为隐写特征的有效性；其次，摒弃了“手工”寻找特征的传统方式，通过采用深度神经网络获取编码参数的相关性，分别设计了基于局部相关性的检测模型、基于全局相关性的检测模型以及基于特征融合的检测模型；最后，以上述三种模型为基础，结合基于线性回归的多模型融合思想，给出了 7 种检测模式，即 3 种单一模型检测模式和 4 种多模型融合检测模式。通过大量的语音样本，对方案进行了性能评估，并与相关工作进行了实验对比分析。实验结果表明，方案中提出的各种检测模式均是可行和有效的，其中三模型融合检测模式整体性能最优。此外，本文工作填补了基于小数基音延迟隐写检测的空白，且较之已有方案对于各类基音延迟隐写方法在任意的嵌入率和样本长度下均具有更好的检测性能和更低的时间开销，从而实现了更为实时高效的检测。

关键词 隐写分析；深度学习；多元线性回归；网络语音流；自适应多速率语音编码；小数基音延迟

中图法分类号 TP309

Steganalysis of Adaptive Multi-Rate Speech Streams Based on the Correlation of Fractional Pitch Delay

TIAN Hui[1),2),3)] WU Jun-Yan[1),2),3)] YAN Yan[1),2),3)] WANG Hui-Dong[1),2),3)] QUAN Han-Yu[1),2),3)]

[1)](School of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

[2)] (Xiamen Key Laboratory of Data Security and Blockchain Technology, Huaqiao University, Xiamen 361021, China)

[3)](Fujian Key Laboratory of Big Data Intelligence and Security, Huaqiao University, Xiamen 361021, China)

Abstract Steganalysis of network speech streams is a research hotspot in the field of information hiding detection. Aiming at detecting steganography in adaptive multi-rate speech streams, this paper proposes a steganalysis scheme based on the correlation of fractional pitch delay. Firstly, through theoretical analysis and experimental comparison, the effectiveness of fractional pitch delay correlation as steganographic features is verified. Secondly, the traditional method of manual feature extraction is abandoned, and the correlation of coding elements is captured by using deep neural networks. Accordingly, a local correlation-based detection model, a global correlation-based detection model and a feature fusion-based detection model are respectively designed. Finally, based on the above three models, combined with the idea of multi-model fusion based on

本课题得到国家自然科学基金(No.61972168)、信息安全国家重点实验室开放课题(No.2019ZD09)资助.田晖(**通信作者**)，博士，教授，博士生导师，计算机学会高级会员(No. 19839S)，主要研究领域为网络与信息安全、数据安全、人工智能安全、信息隐藏及检测、数字取证等.E-mail:htian@hqu.edu.cn.吴俊彦，硕士，主要研究领域为信息隐藏及检测、深度学习.E-mail:wjy9754@stu.hqu.edu.cn.严艳，女，1997年生，硕士，主要研究领域为信息隐藏及检测、深度学习.E-mail:yyan@stu.hqu.edu.cn.王慧东，硕士，主要研究领域为信息隐藏及检测、联邦学习.E-mail:whdeast@qq.com.全韩彧，博士，讲师，主要研究领域为应用密码学、隐私保护.E-mail:quanhanyu@hqu.edu.cn.

linear regression, seven detection modes are given, i.e., three single-model detection modes and four multi-model fusion detection modes. Through a large number of speech samples, the performance of the proposed scheme is comprehensively evaluated, and compared with state-of-the-art works. The experimental results show that the presented various detection modes are feasible and effective, and the three-model fusion detection mode has the best overall performance. In addition, the work of this paper fills in the blank of the detection of steganography based on fractional pitch delay, and for various steganography methods based on pitch delay, it has better detection performance and lower time overhead than the existing steganalysis schemes at any embedding rate and sample length, thereby realizing more real-time and efficient detection.

Key words Steganalysis; Deep learning; Multiple-linear regression; Network speech stream; Adaptive multi-rate speech coding; Fractional pitch delay

1 引言

随着互联网的日益普及以及信息化程度的日趋提高，人们的信息安全意识也在不断提升。数字隐写（Steganography）技术是保护秘密信息的重要方法之一，它能够在不影响载体质量的情况下将秘密信息隐藏在常见媒体（如图像[1,2]，音频[3]，视频[4]以及网络协议[5,6]等）中并通过公共信道进行安全传播。尽管数字隐写技术能够作为保护信息安全的有效手段，但如果被恶意份子非法使用，会对网络空间安全造成严重威胁。为了应对这一挑战，以检测公开信道中可能潜藏的隐蔽信息为目标的隐写分析（Steganalysis）技术应运而生。

在移动即时通信高度发达的今天，IP 语音（Voice over Internet Protocol，VoIP）作为一种网络语音通信服务，在生产和生活的许多方面得到了广泛的应用。不仅如此，由于其具有实时性，动态性，灵活性和容量大等优点，VoIP 亦被视作信息隐藏的理想载体，受到了研究人员们的广泛关注。概括而言，基于 VoIP 的隐写技术大致可以分为两类，即基于协议的隐写[7-13]和基于语音载荷的隐写[14-15]。前者通过修改协议头部字段或调制数据报时序来加载信息，而后者则通过修改语音编码中的部分参数来隐藏信息。相较之下，后者因实时性高，隐蔽性强且嵌入容量大而备受关注。迄今为止，可用于 VoIP 的语音编码有 ITU G.711[16]， G729[17]，G723.1[18]，iLBC[19]和 SILK[20]，Speex[21]和 AMR（Adaptive Multi-Rate）[22]等。其中，AMR 编码是移动语音环境下使用范围最广的语音编码技术，不管是在 Android 和 iOS 移动操作系统还是在各种即时通信 APP（如微信和 iMessage 等）都有着广泛的应用。基于 AMR 语音的隐写及其分析很自然地成为 VoIP 信息隐藏的研究热点。与所有其他的基于代数码激励线性预测（Algebraic code-excited linear prediction，ACELP）语音编码类似，AMR 中有三类参数可用于隐藏信息，即线性预测参数（Liner Prediction Coefficient，LPC）[23]，自适应码本参数（Adaptive codebook，ACB）[24,25]和固定码本参数（Fixed codebook，FCB）[26,27]。其中，ACB 中的基音延迟参数由于难以准确预测，因而可通过调制基音延迟参数实现具有高不可感知性的信息嵌入。例如，文献[24]提出了一种修改闭环自适应码本的搜索规则来实现信息隐藏的方法；文献[25]提出了一种双层隐写方法，可通过修改偶数子帧的整数基音延迟搜索规则来嵌入秘密数据；文献[28]提出了一种基于自适应部分匹配的隐写方法，在保持整数参数不变的情况下，根据秘密信息和小数基音延迟的相似度自适应修改小数参数来实现隐写。

在 ACB 域隐写分析方面，研究者们主要采用了“特征提取+机器学习（支持向量机）”的检测模式，其重点是挖掘出恰当的特征向量以准确区分隐写前后自适应码本参数。代表性的工作如文献[29]采用校准的方式构建了整数基音延迟的二阶差分的马尔可夫转移概率矩阵（Calibration of the Markov transition probability matrix of the second-order difference of pitch delay, C-MSDPD）作为检测特征；文献[30]设计了整数基音延迟的奇偶贝叶斯概率特征（Parity Bayesian Probability, PBP）作为检测特征；文献[31]提出了基于整数基音延迟的奇偶分布特征（Probability Distribution of the Odevity for Pitch Delay, PDOPD），并融合降维后的 C-MSDPD 特征一并作为检测特征以提升检测性能等。尽管上述隐写分析方案取得了一定的检测效果，但是在实际的应用场景仍然面临着如下挑战。其一，随着隐写技术的深入发展，隐写者往往选择以低嵌入率的方式进行隐写以提升隐蔽性，因而隐写分析方案对于低嵌入率下的隐写依然需要较好

的适应性；其二，VoIP 是一种实时语音通信技术，要做到有效检测其中可能存在隐蔽通信，必须能够针对短时样本快速准确的给出检测结果。目前，现有方案对于 10 秒及更大长度的高嵌入率隐写样本检测精度已相当高，但是对于短时低嵌入率的样本的检测性能还不尽如人意。其中一个可能的原因是，上述方案均采用传统“手工”寻找特征的方式，而手工特征属于低层特征，对于目标的表征能力不足，从而最终导致了检测性能方面的缺陷。此外，现有方案在提取特征时均聚焦在整数基音延迟的相关性特性上，而这类特征并不完备，特别是对于某些基于小数基音延迟的隐写（如文献[28]），由于隐写后的整数基音延迟并无改变，从而难以有效检测。

有鉴于此，本文提出了基于小数基音延迟相关性的隐写分析方案。我们首先通过理论分析和实验对比验证了小数基音延迟的相关性及其相对于整数基音延迟具有更大的隐写敏感性，即较之整数基音延迟，小数基音延迟隐写后的变化更大；其次，我们摒弃了“手工”寻找特征的传统方式，而采用深度神经网络获取编码参数的相关性，并分别设计了基于局部相关性的检测模型、基于全局相关性的检测模型，以及综合利用上述两类相关性特征的融合检测模型；最后，以上述三种模型为基础，结合基于线性回归的多模型融合思想，共给出了 7 种可选的检测模式，即 3 种基于单一模型的检测模式和 4 种多模型融合检测模式。通过大量的语音样本，对本文方案进行了性能评估，并与相关工作进行了实验对比分析。实验结果表明，本文方案中各种检测模式均是可行和有效的，其中基于三类模型融合的检测模式整体性能最优。此外，本文工作首次成功实现了对小数基音延迟隐写的检测，且对于各类基音延迟隐写方法在任意的嵌入率和样本长度下较之已有方案均具有更好的检测性能和更低的时间开销，因而可以实现更为实时、高效的隐写检测。

本文其余部分组织结构如下：第 2 章介绍网络语音流隐写分析的背景及相关工作；第 3 章介绍本文提出的隐写分析方案，第 4 章对提出的方案进行实验和性能评估，第 5 章进行工作总结。

2 背景及相关工作

2.1 AMR编解码器

自适应多速率语音编码器 AMR（Adaptive Multi-rate）有着优异的压缩率，使其能够保持高通信效率和低频带资源占用，在移动设备端语音通话中有着广泛的应用。AMR 最常用的编码模式有 AMR-NB （ AMR-NarrowBand ） 和 AMR-WB （AMR-WideBand）两类。如图 1 所示，AMR 编码流程为[32]：首先通过高通滤波对输入的语音信号进行预处理，并对每帧进行一次线性预测分析得到线性预测滤波器系数；将上述系数转为线谱对，再通过预测式两级矢量量化器进行量化，即通过自适应码本搜索和代数码本搜索选择合适的码本来重构激励信号；随后，将激励信号作为线性预测合成滤波器的输入来合成语音。上述码本搜索的原理就是使得合成语音信号和原始语音信号之间感知加权误差最小。

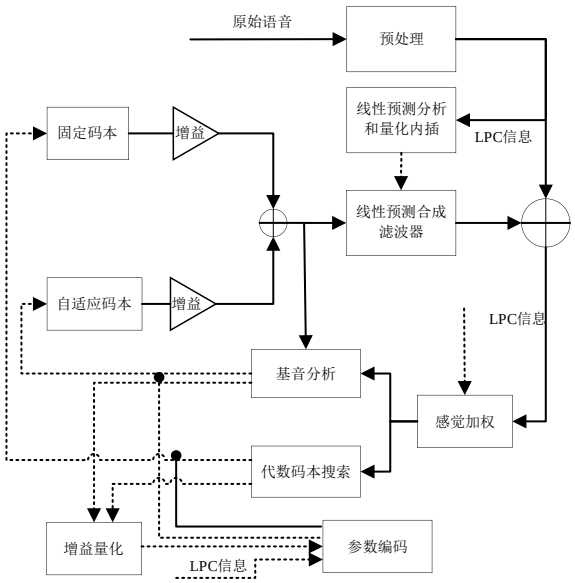


图 1 ACELP 编码框架[32]

自适应码本参数是编码后的主要参数之一，包括基音延迟和增益。其中，基音延迟有着整数和小数部分。编码过程通过自适应码本搜索算法来得到最佳的自适应码本索引。自适应码本搜索是在每个子帧上进行的，它包括闭环基音搜索和自适应码本矢量计算。后者通过在小数基音延迟内插过去的激励来得到。此外，闭环基音搜索通过最小化加权均方误差来决定每一子帧的最佳基音延迟，即使得 $R(k)$最大，其计算公式为：

$$R(k)=\frac{\sum_{n=0}^{39}x(n)y_k(n)}{\sqrt{\sum_{n=0}^{39}y_k(n)y_k(n)}}, \tag{1}$$

其中，x 是目标函数，$y_k(n)$是滤波器在基音延迟为 k 时的输入。一旦确定了最佳整数基音延迟，就需要

测试并确定最佳整数基音延迟周围的小数基音延迟，并选择最大化内插归一化相关的小数基音延迟。以 AMR-NB 为例，基音延迟间的关系如式（2）所示：

$$T_1,T_3 \in \begin{cases} [18,24], & T_{OP} \le 21, \\ [T_{OP}-3,T_{OP}+3], & 21 \le T_{OP} \le 140, \\ [137,143], & T_{OP} > 140, \end{cases}$$

$$T_2,T_4 \in \begin{cases} [18,27], & T_{OP} \le 23, \\ [T_{OP}-5,T_{OP}+4], & 21 \le T_{OP} \le 140, \\ [134,143], & T_{OP} > 140, \end{cases} \quad (2)$$

$$T_F \in \begin{cases} 0, & T_1,T_3 > 94, \\ [-3/6,-2/6,...,3/6], & \text{Otherwise}, \end{cases}$$

其中，T_{OP} 为开环基音延迟，T_i 为第 i 子帧的闭环整数基音延迟，T_F 为小数基音延迟。

2.2 相关工作

基音周期具有很强的不稳定性，在不同个体谈话之间有着较大的差异，因而，在自适应码本搜索时很难做到准确预测。正是基于这一特性，修改基音延迟参数对合成语音的质量影响较小，因而基音延迟可作为理想的信息隐藏载体。迄今为止，已有很多学者对基音延迟隐写进行了深入的研究。例如，Huang 等人[24]提出了一种基于 G.723.1 编码器下修改基音延迟的隐写方法，该隐写方法根据嵌入秘密信息的奇偶性，在闭环基音分析时修改基音延迟的搜索范围来嵌入秘密信息。据实验统计，该隐写方法每帧可嵌入 4bits 的秘密信息。严书凡等人[25]认为在基于基音延迟的隐写过程中如果只修改第二四子帧会减少语音失真，因而提出了一种基于二层嵌入的隐写方法，其原理是：在第一层嵌入时，根据秘密信息的奇偶性，在第一三子帧的闭环搜索范围内选择第二四子帧的基音延迟 T_2 和 T_4，因此在第一层每帧可以嵌入 2bits 的秘密信息；为了提升嵌入容量，在第二层嵌入中，通过调整第一层 T_2 和 T_4 的取值直至它们的异或值等于待嵌入的第 3bit 的秘密信息。该隐写方法每帧可以嵌入 3bits 的秘密信息，较之文献[24]的方法具有更好的感知透明性。此外，Liu 等人[28]提出了一种基于小数基音延迟的自适应匹配的隐写方法，该方法通过 m 序列加密秘密信息，每子帧计算一次加密后的秘密信息和小数基音延迟的相似度，选择合适的阈值自适应调整小数基音延迟参数，最后使用缓存覆盖策略，在保持整数参数不变的情况下提升抗检测能力。该隐写方法可达到每帧最多 8bits 的嵌入容量，并具有较高的隐写透明性和抗检测能力。

为检测基于基音延迟的隐写，研究者们提出了一些基于机器学习的隐写分析方案，其基本思路是通过手工提取整数基音延迟相关性特征对机器学习分类器（SVM）进行训练，因而其核心是如何获得准确的检测特征。比如，Ren 等人[29]研究发现隐写操作会破坏整数基音延迟二阶差分的相关性，于是构建了基于相邻帧二阶差分值的马尔科夫转移矩阵（the Markov transition probability matrix of the second-order difference of pitch delay， MSDPD）作为检测特征，并引入校准方法得到了性能更好的 C-MSDPD 特征。Liu 等人[30]考虑到隐写操作会破坏闭环基音延迟的奇偶关联性，利用当前子帧整数基音延迟的奇偶性和下一子帧的奇偶性的条件概率来描述相关性变化，并提出了基音延迟奇偶性贝叶斯概率 PBP 作为检测特征。文献[31]研究发现由于秘密信息的随机分布特性隐写后的基音延迟奇偶性会趋于一致，当嵌入率越高奇偶分布的一致性越高，因而提出一种基音延迟奇偶分布特征 PDOPD，同时还发现 C-MSDPD 特征将阈值设定在(-1,1)时候的检测性能最好，于是构建了 PDOPD 和降维 C-MSDPD 的混合特征用于检测，取得了较好的检测效果。

尽管上述“手工特征+机器学习”的隐写分析方案取得了一定的检测效果，但研究表明它们仍然存在着一些不足，特别是手工特征属于低层特征，对于目标的表征能力不足，从而最终导致了检测性能方面的缺陷，比如对于短时低嵌入率样本的检测性能相对较低；另外，由于现有方案采用的都是基于整数基音延迟的相关性特征，因而对于仅改变小数基音延迟的隐写方法显得无能为力。鉴于此，本文对基音延迟特性进行了更深入的研究，发现不仅是整数基音延迟，事实上小数基音延迟在隐写后也发生了不同程度的改变，且变化较之前者更为敏感；另外，对于语音帧而言，编码参数的局部相关性和全局相关性也是实现高效隐写分析的重要突破口。基于这些考虑，本文提出了一种新的基音延迟隐写分析方案，该方案聚焦小数基音延迟相关性，引入深度学习网络用于自动学习和抽取以获得高表征性的检测特征，并结合特征融合和基于线性回归的多模型融合策略，给出了多种可行的深度神经网络

检测模式。

3 本文方案

3.1 小数基音延迟参数相关性分析

目前基于基音延迟的隐写方法大致可分为两类：其一是通过修改闭环基音延迟参数的搜索范围来实现隐写，例如 Huang 等人的隐写方法[24]和严书凡等人的双层隐写方法[25]，这类方法的隐写操作不仅直接改变了整数基音延迟参数值，而且之后搜索最佳小数部分时也间接性改变了小数基音延迟值；其二是通过改变对语音质量影响较小的小数基音延迟实现隐写嵌入，例如 Liu 等人提出的自适应隐写方法[28]，这类方法改变了小数基音延迟参数值，却不会改变整数基音延迟。因而，现有的基于整数基音延迟统计特性的隐写分析方案[29,30,31]无法有效检测上述第二类隐写方法。

算法 1. 小数基音延迟提取算法.

输入：第 i 帧自适应码本索引值向量 $I_i = [I_{i,1}, I_{i,2}, I_{i,3}, I_{i,4}]$

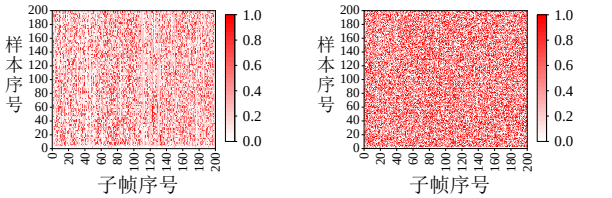
输出：第 i 帧小数基音延迟参数向量 $F_i = [f_{i,1}, f_{i,2}, f_{i,3}, f_{i,4}]$

For j = 1 to 4 **do**
  If j % 2 == 0 **Then**
    If I_j < 463 **Then**
      $l_{i,j} = (I_j + 5)/6 + 17$.
      $f_{i,j} = I_j - l_{i,j} * 6 + 105$.
    Else
      $f_{i,j} = 0$.
    End if
  Else
    $temp = (I_j + 5)/6 - 1$.
    $f_{i,j} = I_j - 3 - 6 * temp$.
  End If
End For
Return $F_i = [f_{i,1}, f_{i,2}, f_{i,3}, f_{i,4}]$.

以 AMR-NB 编码为例，对于长度为 t ms 的语音样本，总帧数为 $T = t/20$，其中每帧 20ms，包含 4 个子帧，即相应包含 4 个基音延迟参数。小数基音延迟的提取过程如算法 1 所示，对每一帧计算其小数基音延迟参数向量 $F_i = [f_{i,1}, f_{i,2}, f_{i,3}, f_{i,4}]$ (i = 1, 2, …, T)，其中，$f_{i,j}$ 代表第 i 帧中的第 j 个子帧的小数基音延迟值。进而，对于整个语音样本，可得到其数基音延迟参数矩阵 $\mathcal{L}$ 如式（3）所示：

$$\mathrm{L} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_T \end{bmatrix} = \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,4} \\ f_{2,1} & \ddots & & f_{2,4} \\ \vdots & & \ddots & \vdots \\ f_{T,1} & f_{T,2} & \cdots & f_{T,4} \end{bmatrix}. \tag{3}$$

为了分析整数和小数基音延迟隐写前后的变化情况，以 1 秒长度的中英文 AMR-NB 编码语音样本各 100 个和使用 Huang 等人的隐写方法[24]在 100%嵌入率下对应得到隐写样本为实验对象，绘制了基音延迟参数变化热图，如图 2 所示。



(a) 整数基音延迟的变化热图　(b) 小数基音延迟的变化热图

图 2　隐写前后基音延迟参数的变化热图。其中，隐写前后基音延迟的变化通过归一化后的差值来衡量，每个色块代表一个差值，其颜色越深，表示变化越大。

从中不难看出，整数和小数基音延迟参数在隐写后均发生不同程度的改变，而小数基音延迟参数的变化更加明显，其原因是在编码过程中小数基音延迟参数会随着整数基音延迟参数的变化而变化，且相较于后者，前者取值范围更小，从而对于隐写操作带来的变化也更为敏感。有鉴于此，与现有工作普遍采用整数基音延迟的特征不同，本文选取隐写后变化更大的小数基音延迟参数作为挖掘对象。

为了进一步验证小数基音延迟参数作为分析对象的有效性，同时考虑到语音信号具有短时不变性和长时相关性，我们对语音编码参数的局部相关性和全局相关性进行了更深入的分析和研究。其中，局部相关性包含帧内相关性和连续帧相关性，全局相关性包含交叉帧相关性和交叉词相关性。以 AMR-NB 编码语音为例，小数基音延迟的上述相关性具体表现如下：

（1）连续帧中小数基音延迟参数间相关性：编码中，相邻帧只间隔 20ms，相当于一个单词中的一个音素的长度。考虑到一个单词有多个音素，而相邻音素是相关的。因此，相邻单词之间具有音素相关性，引申为编码流中连续帧的参数之间具有相关性。我们将连续帧参数之间的相关性称为连续帧相关性。

（2）相同帧内小数基音延迟参数间相关性：

AMR 语音每帧有四个子帧，每子帧搜索一次基音延迟参数，即每帧有四个参数。基音延迟的二四子帧是在一三子帧的搜索范围内找到的，因此相同帧内的参数之间具有相关性。我们将每帧中四个参数之间的相关性称为帧内相关性。

（3）交叉帧内小数基音延迟参数间相关性：由于一个单词有多个音素，当前音素不仅由前一个音素决定，还受到之前出现过的音素的影响，因此一个单词之间不相邻音素之间也具有相关性，引申为编码流中的交叉帧具有相关性，我们称之为交叉帧相关性。

（4）交叉词间小数基音延迟参数间相关性：语音编码流本质上是由句子生成的，考虑到上下文相关性，因此在语音编码流中，一个词的基音延迟参数不仅由同一词的其他参数决定，也受整个语境中其他词的参数所决定。我们将一个句子中不相邻词之间的相关性称为交叉词相关性。

为了评估隐写前后上述相关性变化的程度，我们引入马尔可夫转移矩阵（如式 4 所示）对小数基音延迟参数进行相关性分析。

$$P(f_{i,k}=\alpha, f_{j,l}=\beta)=\frac{P(f_{i,k}=\alpha)\cdot P(f_{j,l}=\beta)}{P(f_{j,l}=\beta)}, \quad (4)$$

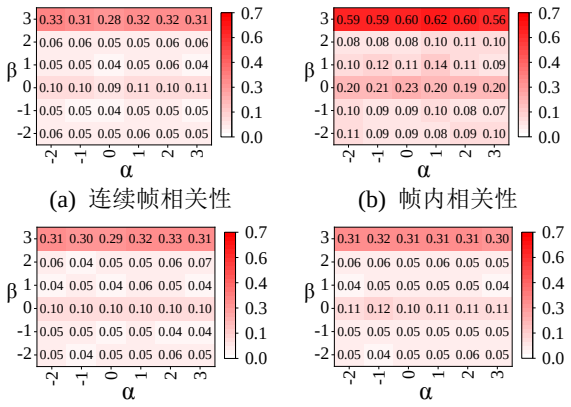


图 3 隐写前后样本转移概率差值矩阵的相关性热图，即隐写样本和对应正常样本的转移概率矩阵之差。

其中，$f_{i,k}(f_{j,l})$代表第 $i(j)$帧中的第 $k(l)$子帧的小数基音延迟参数值；$\alpha,\beta\in$ {-2, -1, 0, 1, 2, 3}为小数基音延迟值。我们采用共计 2000 秒长度的中英文 AMR-NB 12.2kb/s 模式编码语音样本(中英文各 1000 秒，共计 400000 个子帧)和使用 Huang 等人的隐写方法[24]在 100%嵌入率下对应得到隐写样本为实验对象，对小数基音延迟隐写前后的相关性差异进行分析。对于连续帧相关性，设置 $j-i=1, k=l=1$；对于帧内相关性，设置 $i=j, k=1, l=2$；对于交叉帧相关性，设置 $j-i=2, k=l=1$；对于交叉词相关性，设置 $j-i=50, k=l=1$。我们使用了相关性热图来对统计结果进行可视化展示，如图 3 所示。从中可以看出，参数之间的相关性在秘密信息嵌入前后发生了明显的变化，其中，帧内相关性的变化最为明显。由此我们得到启发，通过对小数基音延迟相关性的挖掘可为隐写分析提供有力依据。

3.2 基于深度神经网络的隐写分析模型

为了有效捕获语音流中编码参数的相关性特征以进行准确的隐写分析，本文首先设计了三种基于深度神经网络的隐写分析模型，即局部相关性检测模型(Local Correlation-Based Detection Model, LCDM)，全局相关性检测模型(Global Correlation-Based Detection Model，GCDM)和特征融合检测模型(Feature Fusion-Based Detection Model，FFDM)。

3.2.1 局部相关性检测模型

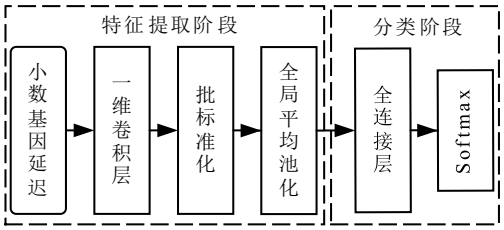


图 4 基于局部相关性的检测模型

AMR 编码器在进行基音延迟参数搜索时，每个语音帧的第二四子帧的基音延迟参数搜索范围由第一三子帧的基音延迟参数决定，这表明基音延迟有着很强的局部相关性。鉴于卷积神经网络的卷积层具有很强的局部感知性，我们首先设计了一个基于卷积神经网络的局部相关性检测模型(LCDM)，其网络结构如图 4 所示，包括一维卷积层，批标准化层(Batch Normalization, BN)，全局平均池化(Global Average Pooling, GAP)以及两个用于分类的全连接层。

假设卷积层的输入为 X，输出为 Y，一维卷积过程可形式化描述为：

$$Y=\mathrm{H}(W\circ X+b), \quad (5)$$

其中，$\mathcal{H}$ 是非线性双曲正切函数，$\circ$是对应向量元素相乘运算，W 是滤波器，b 是偏置。我们以小数基音延迟参数矩阵 $\mathcal{L}$ 作为模型的输入，经过一维卷积后提取得到参数的局部特征 $\mathcal{F}_C$，即：

$$\mathrm{F}_C=f_{Conv}(\mathrm{L},S_C), \quad (6)$$

其中，$f_{Conv}(\cdot)$表示一维卷积层所使用的函数，S_C 是

该层的网络参数集合。为了调整数据分布，提升模型收敛速度，我们对卷积后的输出进行批标准化操作，得到标准化特征 $\mathcal{F}_{BN}$，即：

$$\mathrm{F}_{BN} = f_{BN}\left(\mathrm{F}_{C}, S_{BN}\right), \tag{7}$$

其中，$f_{BN}(\cdot)$表示批标准化层所使用的函数，S_{BN} 是该层的网络参数集合。为了减少后续网络层次所需的参数量，我们使用全局平均池化层进行特征降维，通计算每个特征空间的平均值作为输出。假设有 N_{FS} 个大小为 $n\times m$ 的特征空间，其中第 i 个（i = 1, 2, …, N_{FS}）特征空间 V_i 表示为：

$$V_i = \begin{bmatrix} \lambda_{1,1} & \lambda_{1,2} & \cdots & \lambda_{1,m} \\ \lambda_{2,1} & \ddots & & \lambda_{2,m} \\ \vdots & & \ddots & \vdots \\ \lambda_{n,1} & \lambda_{n,2} & \cdots & \lambda_{n,m} \end{bmatrix}, \tag{8}$$

其全局平均特征 G_i 可表述为：

$$G_i = \frac{1}{n\cdot m}\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_{i,j}. \tag{9}$$

进一步将归一化后的特征 $\mathcal{F}_{BN}$ 送入全局平均池化层可得到输出 $\mathcal{F}_{Pool}$ 为：

$$\mathrm{F}_{Pool} = f_{Pool}\left(\mathrm{F}_{BN}, S_{Pool}\right), \tag{10}$$

其中，$f_{Pool}(\cdot)$表示全局平均池化层所使用的函数，S_{Pool} 表示该层的网络参数集合。模型的分类部分由全连接层和 softmax 组成。其中，全连接层的作用为将上述步骤得到的“分布式特征表示”映射到样本标记空间，即：

$$\mathrm{O}_{FC} = f_{FC}\left(\mathrm{F}_{Pool}, S_{FC}\right), \tag{11}$$

其中，$\mathcal{O}_{FC}$ = [$\mathcal{O}_1$, $\mathcal{O}_2$]是全连接层的输出。$f_{FC}(\cdot)$表示全连接层所使用的函数，S_{FC} 代表该层的网络参数集合。最后，我们使用 softmax 函数给出隐写分析的预测结果 $\mathcal{R}$，即：

$$\mathrm{R} = f_{S}\left(\mathrm{O}_{FC}, S_{Out}\right), \tag{12}$$

其中，$f_S(\cdot)$表示 softmax 函数，S_{Out} 代表该层的网络参数集合。通过上式可得到输出值 $\mathcal{R} = [r_0, r_1]$，其中 r_0 为正常样本的概率值，r_1 为隐写样本的概率值，且 r_0+ r_1 = 1。进一步，可求得给定样本 s 的预测标签 C_s 如下：

$$C_s = \begin{cases} 1, & \text{If } r_1 \geq r_0, \\ 0, & \text{Otherwise.} \end{cases} \tag{13}$$

其中，C_s = 0 时为正常样本，C_s = 1 时为隐写样本。此外，我们使用交叉熵损失函数(Cross-Entropy Loss)用作判断实际标签与模型输出的接近程度，模型通过最小化预测值和真实标签值间的损失值来反向调整各层网络参数以进行训练。其中，交叉熵损失函数定义为：

$$loss = \frac{\sum_{i=1}^{n} -(y_i * \log(r_{i,1}) + (1-y_i) * \log(r_{i,0}))}{n}, \tag{14}$$

其中，n 为样本总数，$r_{i,1}$ 是第 i 个样本被预测为隐写（正类）样本的概率，$r_{i,0}$ 则是该样本被预测为正常（负类）样本的概率，$y_i \in$ {0,1}为第 i 个样本的真实标签值，即 y_i = 1(0)为正（负）类样本。

3.2.2 全局相关性检测模型

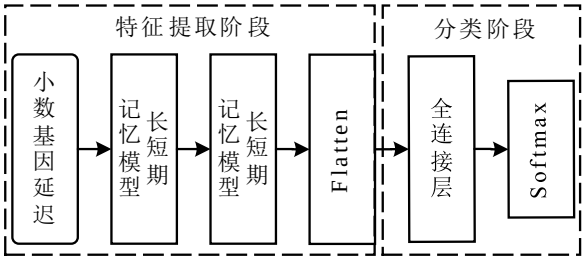


图 5 基于全局相关性的检测模型

语音信号不仅具有短时不变性还具有长时相关性，因此本文还引入了基于循环神经网络的全局相关性检测模型(GCDM)。同时，为了避免传统循环神经网络中可能会出现的梯度消失问题，我们在该模型中引入长短期记忆模型单元层(Long Short-Term Memory, LSTM)来处理时间序列。GCDM 的网络结构，如图 5 所示，利用 LSTM 根据时间序列捕获全局参数之间的联系，即全局相关性，包含了两个 LSTM 层，一个 FLATTEN 层以及两个用于分类的全连接层。LSTM 单元的结构包括输入门、输出门、遗忘门和记忆单元。记忆单元常用于保存和控制长期信息，输入门将输入信息选择性记录到记忆单元中，遗忘门将记忆单元中的信息选择性的遗忘，输出门决定了记忆单元的输出部分。上述过程可形式化描述为：

$$\begin{aligned} I_t &= \sigma\left(w_i \cdot \left[H_{t-1}, X_t\right] + b_i\right), \\ F_t &= \sigma\left(w_f \cdot \left[H_{t-1}, X_t\right] + b_f\right), \\ O_t &= \sigma\left(w_o \cdot \left[H_{t-1}, X_t\right] + b_o\right), \\ M_t &= F_t \cdot M_{t-1} + I_t \cdot \tanh\left(w_c \cdot \left[H_{t-1}, X_t\right] + b_c\right), \end{aligned} \tag{15}$$

其中，σ 是 sigmoid 函数，X_t 为当前第 t 时刻的输入，X_t 经过输入门（遗忘门，输出门）得到的输出 I_t (F_t,O_t)，H_{t-1} 为第 t-1 时刻的隐藏状态向量序列，w_i(w_f, w_o)分别为输入门（遗忘门，输出门）的权重，b_i (b_f, b_o)分别为输出门（遗忘门，输出门）的偏置。M_t 为当前时刻的记忆单元状态，M_{t-1} 代表第 t-1 时刻的

记忆单元状态，tanh 为激活函数。

我们使用小数基音延迟参数矩阵 $\mathcal{L}$ 作为模型的输入，为了提升特征表征参数相关性的能力，我们利用两个 LSTM 层来提取参数间的全局特征，首先经过第一层后得到输出 $\mathcal{F}_{RNN\text{-}1}$：

$$\mathrm{F}_{RNN-1} = f_{LSTM}\left(\mathrm{L}, S_{RNN-1}\right), \tag{16}$$

其中，$f_{LSTM}(\cdot)$表示 LSTM 层使用的函数，$S_{RNN\text{-}1}$ 表示第一层的网络参数集合，经过第二个 LSTM 层提取特征的过程可描述为：

$$\mathrm{F}_{RNN-2} = f_{LSTM}\left(\mathrm{L}, S_{RNN-2}\right), \tag{17}$$

其中，$\mathcal{F}_{RNN\text{-}2}$ 为经过第二层 LSTM 的输出，$S_{RNN\text{-}2}$ 表示第二层的网络参数集合。接着，经过 Flatten 层将输出特征展平成一维空间，其过程可形式化描述为：

$$\mathrm{F}_{Fla} = f_{Fla}\left(\mathrm{F}_{RNN-2}\right), \tag{18}$$

其中，$f_{Fla}(\cdot)$表示特征展平操作。经过展平处理后的特征 $\mathcal{F}_{Fla}$ 将送入全连接层和 softmax 函数进行分类，其过程与 LCDM 模型类似，在此不再赘述。

3.2.3 特征融合检测模型

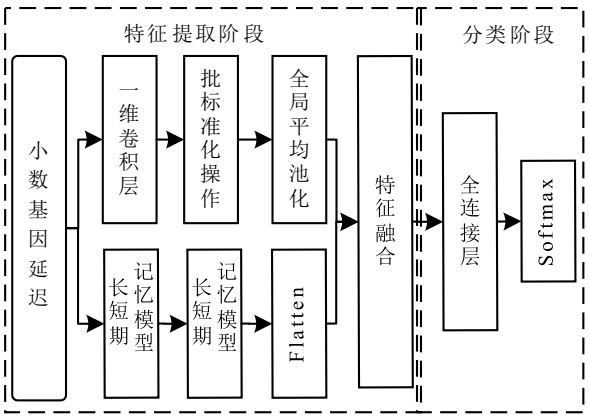


图 6 基于特征融合的检测模型

考虑到上述两个检测模型各有侧重，为取长补短，实现优势互补，我们设计了第三个检测模型，即特征融合检测模型(FFDM)，其网络结构如图 6 所示。在该模型中，我们将小数基音延迟参数经过 LCDM 和 GCDM 提取的特征在中间层进行特征融合。为了保持特征的完整性，我们选择了特征拼接的方式进行特征融合，其过程可表述如下：

$$\mathrm{F}_{M} = f_{concat}\left(\mathrm{F}_{Pool}, \mathrm{F}_{Fla}\right), \tag{19}$$

其中，$f_{concat}(\cdot)$表示特征拼接的融合操作，$\mathcal{F}_{Pool}$ 和 $\mathcal{F}_{Fla}$ 分别为 LCDM 和 GCDM 提取的特征，$\mathcal{F}_{M}$ 为融合后的特征。最后，融合特征 $\mathcal{F}_{M}$ 经过全连接层和 softmax 函数处理后输出预测结果。

3.3基于线性回归的多模型融合

根据机器学习的常规经验，通过融合多个不同的模型或可进一步提高分类性能。鉴于此，以上述三种基本模型为基础，根据选取参与融合模型的不同，共有七种可选的检测模式，如表 1 所示，即 3 种基于单一模型的检测模式和 4 种多模型融合检测模式。

表 1 可选的检测模式

方案	模型选择		
	LCDM	GCDM	FFDM
1	√	×	×
2	×	√	×
3	×	×	√
4	√	√	×
5	√	×	√
6	×	√	√
7	√	√	√

对于模型融合而言，有许多可选的方式，如加权投票法和加权平均法等。它们通过给各分类器赋予不同的权重，并将加权结果作为预测结果，以提高分类性能。然而，在本文工作中，不同的检测条件（如嵌入率、样本长度、隐写方法和样本语种）下，各模型的最优权重组合也各不相同。如采用传统的人工调参方式不仅过程繁琐，且难以找到不同条件下的最佳模型权重组合。鉴于此，本文采用在数据分析领域得到广泛应用的多元线性回归方式进行模型融合。多元线性回归模型是一种形式简单，易于建模，且可解释性很强的数理统计模型，它通过多个自变量的最优线性组合来进行预测。假设对 k 个模型进行融合，对于任意的第 i 个语音样本，基于多元线性回归的预测概率值为：

$$f(x_i) = x_i w + b, \tag{20}$$

其中，$x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,k}]$为不同模型判断第 i 个样本为正类的概率值向量，$x_{i,j}$ 为第 j 个模型判断该样本为正类的概率值，$j = 1, 2, \ldots, k$；$w^{\mathrm{T}} = [w_1, w_2, \ldots, w_k]$为待求的各模型权重向量，$b$ 为随机误差项。多元线性回归常采用均方误差作为损失函数，即：

$$e = \sum_{i=1}^{n}(y_i - f(x_i))^2 = \sum_{i=1}^{n}(y_i - x_i w - b)^2, \tag{21}$$

其中，n 为样本数量，$y_i \in \{0,1\}$为第 i 个样本的真实标签。将向量 w^{T} 和 b 合并记为新的权重向量 $W^{\mathrm{T}} = [w_1, w_2, \ldots, w_k, b]$，相应地概率值向量将增加一维常量，即 $X_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,k}, 1]$，则损失函数可写为：

$$e=(Y-\mathbf{X}W)^{\mathrm{T}}(Y-\mathbf{X}W), \tag{22}$$

其中 $Y^{\mathrm{T}}=[y_1, y_2, \ldots, y_n]$是样本集的标签向量，$\mathbf{X}$ 为多模型输出的概率值矩阵，即：

$$\mathbf{X}=\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}=\begin{bmatrix} x_{1,1}, & x_{1,1}, & \cdots & x_{1,k},1 \\ x_{2,1}, & x_{2,2}, & \cdots & x_{2,k},1 \\ \vdots & \vdots & \vdots & \vdots\ \vdots \\ x_{n,1}, & x_{n,2}, & \cdots & x_{n,k},1 \end{bmatrix}. \tag{23}$$

进一步，可通过最小二乘法可求得权重向量的最优解 W^*为：

$$W^*=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}Y. \tag{24}$$

最终在检测阶段，对于某个待测的语音样本 s，其概率值向量 $Xs=[x_{s,1}, x_{s,2}, \ldots, x_{s,k}, 1]$，则基于多元线性回归的预测概率值为：

$$f(X_s)=X_sW^*. \tag{25}$$

从而，多个神经网络融合后给出的分类结果 C_s 为：

$$C_s=\begin{cases}1, & \text{If}\left(f(X_s)\ \geq 0.5\right),\\ 0, & \text{Otherwise.}\end{cases} \tag{26}$$

其中，当 $C_s=1$ 时判断为隐写样本，当 $C_s=0$ 时判断为正常样本。

4 实验分析与性能评估

4.1 实验设置和性能衡量指标

实验采用清华大学研究团队提供的公开语音数据集[33,34]。该数据集中的语音样本均来自互联网，共包含 41 小时的中文语音和 72 小时的英文语音。为便于进行实验验证，我们对数据集进行了如下预处理：

（1）构建载体样本库（cover sample set）：对公开数据集中的原始音频进行 PCM 单声道编码，采样频率为 8000Hz，量化位数为 16bit。我们将编码后的语音以 0.1s 为步长，将原始语音分割成 0.1s，0.2s 到 1s 不等的长度，每种时长的语音样本集均包含 12500 个中文语音样本和 12500 个英文语音样本；对各语音样本进一步采用 AMR-NB 12.2 kbps 速率模式进行编码，得到载体语音样本。

（2）构建载密样本库（stego sample set）：对于每个 AMR 语音样本，分别采用 Huang 等人提出的隐写方法[24]（记为 S_1），严书凡等人提出的隐写方法[25]（记为 S_2）和 Liu 等人提出的隐写方法[28]（记为 S_3）生成对应的载密样本。其中，对于 1s 秒时长的语音样本，以 10%为步长，生成从 10%到 100%共十种不同的嵌入率下精确控制嵌入秘密信息的载密样本；对于 0.1s，0.2s 到 0.9s 时长的语音样本，生成 100%嵌入率的载密样本。表 2 和表 3 分别给出了 1s 时长不同嵌入率条件下及 0.1s，0.2s 到 1s 时长且 100%嵌入率条件下，各方法的隐写容量及带宽。此外，考虑到实际应用中隐写方法普遍采用先加密再隐藏的操作过程，上述秘密信息均通过均匀分布随机生成以确保与密文形式同分布。

（3）构建实验样本集（experimental sample set）：在以下所有隐写分析实验中，对于既定条件（即给定的语种、样本时长、隐写方法和嵌入率），组合 12500 个正常载体样本和对应的隐写样本构成实验样本集，并按 4:1 的比率将样本集分割成训练集和测试集。

本节将进行的实验包括两个部分，即本文提出方案的性能分析实验以及与以下现有研究工作的性能对比实验[1]，包括 Ren 等人的方案（记为 C-MSDPD）[29]， Liu 等人的方案（记为 PBP）[30]和 Tian 等人的方案（记为 HYBIRD）[31]等。其中，C-MSDPD，PBP 和 HYBIRD 三种对比隐写分析方案中使用的 SVM 均基于 python 语言的机器学习工具包 Scikit-learn(sklearn)[2]实现，它包含了数据预处理、训练模型、评价指标计算等各模块，本次使用的版本号 0.24.2。在 SVM 的超参数设置中，我们将“kernel”设置为“rbf”，并将“gamma”设置为“scale”，其余参数均使用默认设置。本文方案均基于 python 语言的神经网络框架 keras[3]实现，它是一个模型级的程序库，为开发深度学习模型提供了高层次的构建模块，本次使用的版本号为 2.6.0。经过调参后，我们使用的训练批尺寸设置为 256，训练回合为 20，采用交叉熵为损失函数，各模型的超参数初始化如下：LCDM 中，设置卷积核数量为 64，大小为 3，步长为 1，填充大小为“same”；在 GCDM 中，使用双层 LSTM 结构，维度分别为 64 和 32，返回隐藏状态；在 FFDM 中，使用 concat 特征拼接方式，分类前经过带有 relu 的全连接层，输出维度为 32。模型的分类部分均选用带有 softmax 的全连接分类层，输出的特征维度为 2。此外，多模型融合检测

1 相关源码和模型参见 GitHub 地址：https://github.com/junono97/SM-MDNNF.

2 Scikit-learn: the machine learning in Python, available at: https://scikit-learn.org/stable.

3 Keras: the Python deeping learning API, available at: https://keras.io/.

表 2 1 秒不同嵌入率下各隐写方法的隐写容量及带宽

隐写方法	性能指标	嵌入率（%）									
		10	20	30	40	50	60	70	80	90	100
S_1	隐写容量(bits)	20	40	60	80	100	120	140	160	180	200
	隐写带宽(bits/帧)	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0
S_2	隐写容量(bits)	15	30	45	60	75	90	105	120	135	150
	隐写带宽(bits/帧)	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3.0
S_3	隐写容量(bits)	40	80	120	160	200	240	280	320	360	400
	隐写带宽(bits/帧)	0.8	1.6	2.4	3.2	4.0	4.8	5.6	6.4	7.2	8.0

表 3 不同样本长度下 100%嵌入率时各隐写方法的隐写容量及带宽

隐写方法	性能指标	样本长度（s）									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
S_1	隐写容量(bits)	20	40	60	80	10	120	140	160	180	200
	隐写带宽(bits/帧)	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
S_2	隐写容量(bits)	15	30	45	60	75	90	105	120	135	150
	隐写带宽(bits/帧)	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
S_3	隐写容量(bits)	40	80	120	160	200	240	280	320	360	400
	隐写带宽(bits/帧)	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0

模式中的多元线性回归算法均选用 sklearn 工具包中的 LinearRegression 算法库进行实现。

为了衡量隐写分析方案的检测性能，我们使用检测准确率(Accuracy，ACC)和单个语音样本平均检测时间(T_{AVG})作为衡量指标。ACC 指的是正确分类的样本占所有样本的比率，若 ACC 越高，检测性能就越好，其定义如下：

$$ACC = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}, \tag{27}$$

其中，N_{TP} 为正确分类为正类（隐写）样本的数量；N_{TN} 为正确分类为负类（未隐写）样本的数量，N_{FP} 为将负类样本错误分类为正类样本的数量；N_{FN} 为将正类样本错误分类为负类样本的数量。

T_{AVG} 为单个语音样本平均检测时间，反映了隐写分析算法的检测效率，其计算公式如下：

$$T_{AVG} = \frac{\sum_{i=1}^{K}\left(\sum_{j=1}^{N}\sum_{l=1}^{S} t_{l,j}\right)}{K \cdot N}, \tag{28}$$

其中，$t_{i,j}$ 为第 j 个样本在 l 个隐写方法下的检测时间，S 为检测方法数目，N 为测试样本总数目，K 为测试总次数。为减少测试误差，本文采用 20 次测试取平均值的方法给出结果，即 K=20。本文隐写检测实验在一台 Dell 工作站上进行，其主要配置参数如下：Intel Xeon E3-1225 v5 CPU 3.30GHz, 2×8GB DDR4-2133MHz 内存， 500GB 7200RPM SATA 硬盘，以及 Linux 系统（Ubuntu 20.04.3 LTS 64 位, kernel version 5.11.0-27-generic）。

4.2 多种检测模式的性能对比和分析

如前所述，根据选取参与融合模型的不同，共有七种可选的检测模式，如表 1 所示，前 3 种为基于单一模型的检测模式，第 4 到 6 种为双模型融合检测模式，第 7 种为三模型融合检测模式。我们在 1 秒长度的中英文样本集下分别测试了各模式对于 S_1，S_2 和 S_3 三种方法的检测准确率，其实验结果如表 4 和表 5 所示。从中可以看出：

（1）从单模型模式的检测结果来看，各模型在单独检测时亦能取得较好的检测性能；模式 1(LCDM)的检测性能整体上优于模式 2(GCDM)，如在中文样本集下检测 10%至 60%嵌入率的 S_1 方法时，前者的检测准确率比后者高 0.18%至 4.36%，说明隐写操作对于参数的局部相关性的影响大于全局相关性；模式 3(FFDM)的检测性能整体上优于 LCDM，如在中文样本集下检测 10%至 100%嵌入率的 S_3 方法时，前者的检测准确率比后者高 0.40%至 3.06%，说明特征融合方法对两类特征“取长补短”后有了更好的检测效果。

（2）双模型融合模式（模式 4 至模式 6）的检测性能整体上优于基于单模型的检测模式。例如在中文样本集下检测嵌入率为 10%至 60%的 S_1 方法（嵌入率为 10%至 90%的 S_3 方法）时，双模型融合模式的最佳检测准确率相比 FFDM 提升了 0.02%至

表 4 1 秒中文样本集下各检测方案对三种隐写方法的准确率对比（%）

隐写方法	检测方案	嵌入率（%）									
		10	20	30	40	50	60	70	80	90	100
	1	70.88	88.86	95.74	99.04	99.10	99.84	99.98	**100.00**	**100.00**	**100.00**
	2	69.50	84.50	94.62	98.24	98.88	99.66	99.96	**100.00**	**100.00**	**100.00**
	3	73.12	89.14	96.52	99.08	99.54	99.88	**100.00**	**100.00**	**100.00**	**100.00**
S_1	4	71.14	88.98	95.82	99.14	99.08	99.84	**100.00**	**100.00**	**100.00**	**100.00**
	5	**75.00**	90.68	96.84	99.16	99.54	99.88	99.98	**100.00**	**100.00**	**100.00**
	6	73.54	89.68	96.50	99.10	99.56	**99.92**	**100.00**	**100.00**	**100.00**	**100.00**
	7	74.92	**90.90**	**97.26**	**99.26**	**99.58**	99.88	**100.00**	**100.00**	**100.00**	**100.00**
	1	83.28	95.46	99.14	99.68	99.92	**99.96**	**99.98**	**100.00**	**100.00**	**100.00**
	2	82.32	94.50	97.60	99.56	99.88	99.90	**99.98**	99.98	**100.00**	**100.00**
	3	83.70	95.92	99.16	99.72	99.92	**99.96**	**99.98**	**100.00**	**100.00**	**100.00**
S_2	4	85.36	95.46	99.14	99.72	**99.96**	99.92	**99.98**	**100.00**	**100.00**	**100.00**
	5	83.62	96.64	**99.42**	99.70	99.92	**99.96**	99.96	**100.00**	**100.00**	**100.00**
	6	85.34	96.22	99.22	99.72	99.90	**99.96**	99.96	**100.00**	**100.00**	**100.00**
	7	**85.38**	**96.70**	**99.42**	**99.76**	99.94	**99.96**	99.96	**100.00**	**100.00**	**100.00**
	1	59.28	68.30	74.04	81.66	87.24	90.82	92.70	96.02	97.70	98.34
	2	58.78	68.02	77.06	83.64	89.76	93.20	94.56	95.10	98.62	98.26
	3	60.00	68.70	77.10	83.38	89.60	93.24	94.46	96.82	98.64	**99.28**
S_3	4	59.74	69.82	77.76	83.82	89.88	93.42	94.54	96.40	98.60	98.56
	5	**60.86**	69.72	78.20	84.12	89.96	93.56	95.52	97.00	98.72	**99.28**
	6	59.56	69.20	78.12	84.26	90.04	93.78	95.26	96.84	98.74	**99.28**
	7	60.58	**70.08**	**78.52**	**84.34**	**90.08**	**93.84**	**95.54**	**97.04**	**98.76**	**99.28**

表 5 1 秒英文样本集下各检测方案对三种隐写方法的准确率对比（%）

隐写方法	检测方案	嵌入率（%）									
		10	20	30	40	50	60	70	80	90	100
	1	75.66	89.56	96.30	99.42	99.44	99.82	99.92	**100.00**	**100.00**	**100.00**
	2	70.82	87.74	95.06	98.56	99.20	99.80	99.86	**100.00**	**100.00**	**100.00**
	3	75.72	90.34	96.84	99.58	99.76	99.86	**99.98**	**100.00**	**100.00**	**100.00**
S_1	4	76.06	89.82	96.26	99.46	99.46	99.80	99.92	**100.00**	**100.00**	**100.00**
	5	77.14	**91.90**	97.88	**99.62**	99.76	**99.88**	**99.98**	**100.00**	**100.00**	**100.00**
	6	76.18	90.68	97.24	99.58	**99.80**	**99.88**	**99.98**	**100.00**	**100.00**	**100.00**
	7	**77.34**	91.82	**97.96**	**99.62**	**99.80**	**99.88**	**99.98**	**100.00**	**100.00**	**100.00**
	1	85.22	95.90	99.02	99.82	**99.92**	99.94	**100.00**	**100.00**	**100.00**	**100.00**
	2	83.56	94.50	98.28	99.38	99.84	99.86	99.94	99.98	**100.00**	**100.00**
	3	85.32	96.08	99.00	99.82	99.90	99.94	**100.00**	99.98	**100.00**	**100.00**
S_2	4	85.40	95.74	99.02	**99.86**	**99.92**	99.92	**100.00**	**100.00**	**100.00**	**100.00**
	5	86.56	**96.72**	**99.12**	**99.86**	99.90	**99.96**	**100.00**	**100.00**	**100.00**	**100.00**
	6	85.76	96.30	99.08	99.82	99.90	**99.96**	**100.00**	**100.00**	**100.00**	**100.00**
	7	**86.58**	**96.72**	99.10	**99.86**	99.90	**99.96**	**100.00**	**100.00**	**100.00**	**100.00**
	1	59.90	69.02	73.92	81.30	86.44	89.82	93.84	96.12	98.00	98.26
	2	59.78	68.44	78.88	84.40	89.90	93.46	93.84	95.88	97.56	98.18
	3	60.64	70.30	78.86	84.96	90.22	93.56	**95.28**	**97.74**	**99.22**	**99.02**
S_3	4	60.04	70.30	78.94	84.76	89.84	93.56	93.94	96.16	98.06	98.48
	5	60.58	**70.82**	79.32	85.16	90.46	93.78	95.26	97.74	99.20	**99.02**
	6	**60.78**	70.42	79.38	85.40	90.40	93.90	**95.28**	**97.74**	**99.22**	99.00
	7	60.74	70.76	**79.44**	**85.42**	**90.54**	**94.10**	95.26	97.70	99.20	**99.02**

1.88% (0.10%至 1.12%)。这说明模型融合在一定程度上能够综合各单模型之所长，特别是在低嵌入率下（60%及以下）能够达到了比单一模型更好的检测效果。

（3）三模型融合模式（模式 7）的检测性能整体上略优于双模型融合模式。不过，同时也可以看到，随着嵌入率的增加，单一模型的检测准确率也会随之提高；对于嵌入率不小于 50%的 S_1 和嵌入率不小于 40%的 S_2 方法，FFDM 的检测准确率已达到 99.8%以上，此时多元回归融合模型检测性能提

升空间已经非常小。因而，对于高嵌入率的 S_1 和 S_2 方法的检测，可直接选取 FFDM 作为检测模型。然而，在检测 S_3 方法时候，三模型融合模式在各种嵌入率下整体上都略优于其他两类检测模式。

（4）我们以 100%嵌入率的中英文样本集为实验对象，进一步测试了不同检测模式下的单个样本平均检测时间，其统计结果如表 6 所示。整体而言，随着采用模型数量的增多，其对应模式的单个样本平均检测时间会随之增加。然而，即使是用时最多的三模型融合模式，其单个样本平均检测时间也不超过 0.75ms，能够有效实现实时检测。

综上所述，本文提出的 7 种模式不管从检测准确率还是检测效率方面均能达到较好的检测性能，但综合考虑对于 S_3 方法及低嵌入率下的 S_1 和 S_2 方法的检测性能，建议选取三模型融合模式作为最佳检测方案。

4.3 与相关工作的性能对比和分析

为全面分析和评估不同隐写分析方案的检测性能，我们分别使用 1 秒长度下 10%至 100%不同嵌入率的实验样本集和 100%嵌入率下 0.1 秒至 1 秒不同长度的实验样本集进行了对比实验。本节参与对比实验的隐写分析方案包括：①本文中提出的三模型融合检测模式（即模式 7），也称为以小数基音延迟参数(Fractional Pitch Delay Parameters)作为输入的多深度神经网络融合隐写分析模型(Steganalysis Model Based on Multi-Depth Neural Network Fusion, SM-MDNNF)，记为 SM-MDNNF+FPDP；②C-MSDPD[29]；③PBP[30]；④HYBIRD[31]；⑤以整数基音延迟参数(Integer pitch delay parameters)作为输入的 SM-MDNNF，记为 SM-MDNNF+IPDP。图 7 和图 8 给出了各种隐写分析方案对于不同嵌入率下的 1 秒长度语音样本集的实验结果；图 9 和图 10 给出了各种隐写分析方案对于 100%嵌入率下不同长度语音样本集的实验结果。从中可以得出以下结论：

（1）在样本长度相同的情况下，各类隐写分析方案的检测准确率均随着嵌入率的增加而提高，其原因是随着秘密信息的嵌入量越大，隐写前后的相关性特征差异越大，从而使得不管是 SVM 还是神经网络都能更好地分辨它们的不同。并且，较之传统方案，本文提出的隐写分析方案在检测已有隐写方法时的检测性能均显著提升。以中文样本集上的实验结果（如表 7 所示）为例，在检测隐写方法 S_1 时，SM-MDNNF+FPDP 在 10%嵌入率（隐写带宽为 0.4bits/帧）时的检测准确率可达 74.92%，当嵌入率超过 20%（隐写带宽为 0.8bits/帧）时，其检测准确率达到 90.90%以上，当嵌入率超过 70%（隐写带宽超过 2.8bits/帧）时，其检测准确率达到 100%，然而传统隐写分析方案在 10%嵌入率下的最佳检测准确率不足 56%，在 100%嵌入率（隐写带宽为 4.0bits/s 帧）下也仅能达到 77.48%，性能明显低于 SM-MDNNF+FPDP；对于隐写方法 S_2，SM-MDNNF+FPDP 在 10%嵌入率（隐写带宽为 0.3bits/帧时）下能够达到 85.38%的检测准确率，当嵌入率超过 20%（隐写带宽超过 0.6bits/帧）时，其检测准确率可达到了 96.70%以上，当嵌入率超过 80%（隐写带宽超过 2.4bits/帧）时，其检测准确率达到 100%，然而传统隐写分析方案在 10%嵌入率下的最佳检测准确率不足 54%，在 100%嵌入率（隐写带宽为 3.0bits/帧）时的检测准确率也只能达到 76.74%，性能亦明显低于 SM-MDNNF+FPDP；对于隐写方法 S_3 而言，由于该方法只改变了小数基音延迟而保持整数参数不变以提升抗检测性能，因而仅使用整数基音延迟参数特征的三种传统隐写分析方案的检测准确率在 55%±(5%)的区间内浮动，即无法有效检测，然而 SM-MDNNF+FPDP 在 20%嵌入率（隐写带宽为 1.6bits/帧）时的检测准确率即可达到 70.08%，当嵌入率为 50%（隐写带宽为 4.0bits/帧）时达 90.08%，在 100%嵌入率（隐写带宽为 8.0bits/帧）时可达 99.28%，能够很好的检测该类隐写方法。

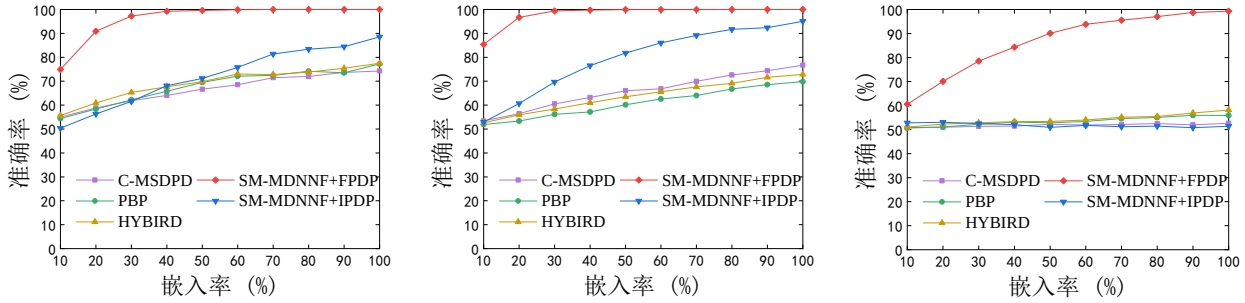
（2）对于给定的隐写方法，各类隐写分析方案的检测性能随着样本长度的增加而提高，其原因

表 6 不同检测模式的单个样本平均检测时间(ms)

检测模式	0.1s	0.2s	0.3s	0.4s	0.5s	0.6s	0.7s	0.8s	0.9s	1.0s
1	0.023	0.024	0.025	0.026	0.027	0.028	0.029	0.030	0.031	0.031
2	0.199	0.216	0.223	0.241	0.271	0.289	0.309	0.320	0.330	0.351
3	0.199	0.223	0.242	0.261	0.267	0.286	0.319	0.336	0.351	0.361
4	0.222	0.240	0.248	0.273	0.298	0.317	0.337	0.344	0.361	0.388
5	0.228	0.252	0.267	0.280	0.293	0.312	0.348	0.365	0.382	0.385
6	0.404	0.434	0.465	0.501	0.537	0.580	0.628	0.650	0.681	0.712
7	0.427	0.457	0.489	0.527	0.564	0.608	0.657	0.680	0.712	0.744

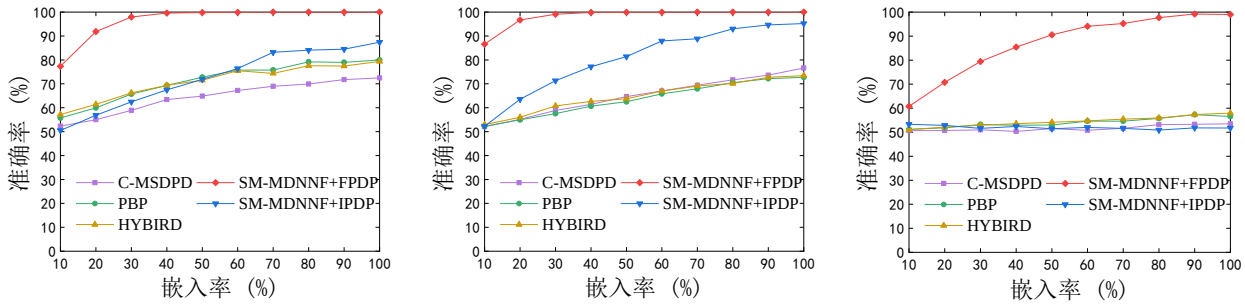
是样本越长，用于隐写分析的数据就越多，从而使

得各类隐写分析方案的检测准确率都会随之增高；



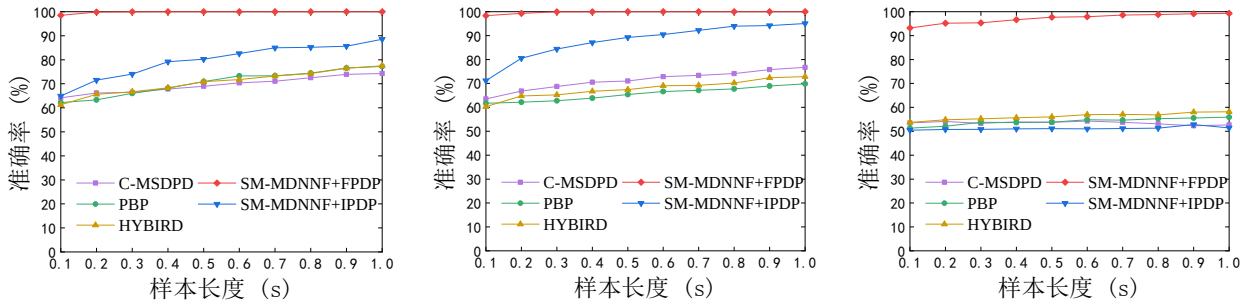
(a) 对于方法 S_1 的检测准确率　(b) 对于方法 S_2 的检测准确率　(c) 对于方法 S_3 的检测准确率

图 7 中文样本集上不同嵌入率下各隐写分析方案对三种隐写方法的检测准确率



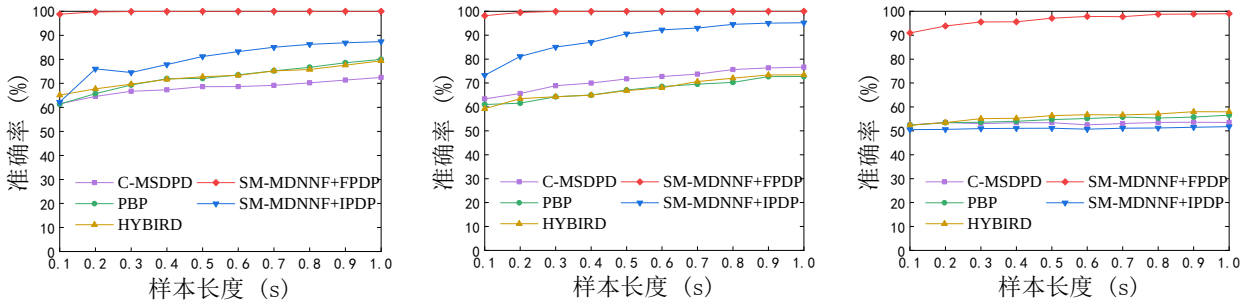
(a) 对于方法 S_1 的检测准确率　(b) 对于方法 S_2 的检测准确率　(c) 对于方法 S_3 的检测准确率

图 8 英文样本集上不同嵌入率下各隐写分析方案对三种隐写方法的检测准确率



(a) 对于方法 S_1 的检测准确率　(b) 对于方法 S_2 的检测准确率　(c) 对于方法 S_3 的检测准确率

图 9 中文样本集上不同样本长度下各隐写分析方案对三种隐写方法的检测准确率



然而，较之传统方案，本文提出的隐写方案在短样本长度下较之已有方案的检测性能均显著提升。以英文样本集上的实验结果（如表 8 所示）为例，当检测隐写方法 S_1 时，SM-MDNNF+FPDP 在 0.1 秒语音样本长度下的检测准确率可达 98.80%，当语音样本长度超过 0.6 秒时，SM-MDNNF+FPDP 的检测

准确率就可达到 100%，然而传统隐写分析方案在 0.1 秒语音样本长度下的最佳检测准确率为 65.12%，在 1 秒语音样本长度下也仅达到 79.96%，性能明显低于 SM-MDNNF+FPDP；对于隐写方法 S_2，SM-MDNNF+FPDP 在 0.1 秒样本长度下能够达到 98.16%的检测准确率，当语音样本长度超过 0.5 秒时，SM-MDNNF+FPDP 的检测准确率就可达到 100%，然而传统隐写分析方案在 0.1 秒语音样本长度下的最佳检测准确率仅为 63.38%，在 100%嵌入率下的检测准确率也只能达到 76.64%，性能亦明显低于 SM-MDNNF+FPDP；对于隐写方法 S_3 而言，如前所述，使用整数基音延迟参数特征的三种传统隐写分析方案无法有效检测该类隐写方法，然而 SM-MDNNF+FPDP 在 0.1 秒语音样本长度下的检测准确率可达 90.94%，当语音样本长度达到 0.3 秒及以上时，检测准确率可达到 95.56%以上，能够有效检测该类隐写方法。

（3）即使同样使用整数基音延迟参数相关特征，在绝大部分情况下（当 1s 语音长度且嵌入率不低于 50%，或长度为 0.2 到 1s 且嵌入率为 100%时），本文提出的基于多深度神经网络融合方案能取得较之传统手工提取特征的隐写分析方案对于基于整数基音延迟的隐写方法（S_1 和 S_2）更好的检测性能。例如，在 1 秒中文语音样本长度和 100%嵌入率的条件下，对于隐写方法 S_1，SM-MDNNF+ IPDP 的检测准确率为 88.52%，而传统隐写分析方案的最佳检测准确率为 77.48%，相较之下前者高出 11.04 个百分点；对于隐写方法 S_2，SM-MDNNF+IPDP 的检测准确率为 95.06%，而传统隐写分析方案的最佳检测准确率为 76.74%，相比之下高出了 18.32 个百分点。这再次表明了利用深度学习模型所捕获参数间的相关性能够更为全面地描述语音隐写特性。

（4）SM-MDNNF+FPDP 的检测性能明显优于 SM-MDNNF+IPDP，不仅表现在 SM-MDNNF+FPDP 可有效检测 SM-MDNNF+IPDP 无能为力的 S_3 方法，而且在检测 S_1 和 S_2 方法时其检测性能也显著优于 SM-MDNNF+IPDP。具体来说，以中文样本集上的实验结果（如表 7 所示）为例，检测 $S_1(S_2)$ 方法时，在 1 秒语音样本长度下，根据嵌入率的不同（10%到 100%）检测准确率提升 11.48%到 35.68%(4.94%到 36.02%); 以英文样本集上的实验结果（如表 8 所示）为例，检测 $S_1(S_2)$方法时，在 100%嵌入率下，根据语音样本长度的不同（0.1 到 1 秒）检测准确率提升 12.60%到 36.58%(4.82%到 24.92%)。上述实验结果再次表明，小数基音延迟参数能够更好地表征隐写前后自适应码本域相关性的变化。

表 8 英文样本集上不同样本长度下各隐写分析方案对三种隐写方法的检测准确率对比（%）

隐写方法	隐写分析方案	样本长度（s）									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
S_1	(a) C-MSDPD	61.50	64.54	66.70	67.32	68.62	68.66	69.18	70.20	71.36	72.48
	(b) PBP	61.36	65.74	69.34	71.90	72.08	73.54	75.22	76.68	78.56	79.96
	(c) HYBIRD	65.12	67.74	69.68	71.66	72.78	73.34	75.16	75.80	77.58	79.36
	(d) SM-MDNNF+IPDP	62.22	76.04	74.52	77.84	81.16	83.24	85.04	86.24	86.90	87.40
	(e) SM-MDNNF+FPDP	98.80	99.74	99.90	99.94	99.98	100.00	100.00	100.00	100.00	100.00
	(e)−MAX(a,b,c)	33.68	32.00	30.22	28.04	27.20	26.46	24.78	23.32	21.44	20.04
	(e)−(d)	36.58	23.70	25.38	22.10	18.82	16.76	14.96	13.76	13.10	12.60
S_2	(a) C-MSDPD	63.38	65.64	68.88	69.96	71.72	72.72	73.70	75.64	76.34	76.64
	(b) PBP	61.00	61.54	64.24	64.94	67.08	68.50	69.54	70.28	72.68	72.72
	(c) HYBIRD	59.24	63.48	64.24	64.90	66.84	68.00	70.62	72.02	73.38	73.42
	(d) SM-MDNNF+IPDP	73.24	81.14	85.06	87.04	90.58	92.22	93.00	94.54	95.06	95.18
	(e) SM-MDNNF+FPDP	98.16	99.52	99.96	99.96	100.00	100.00	100.00	100.00	100.00	100.00
	(e) − MAX(a,b,c)	34.78	33.88	31.08	30.00	28.28	27.28	26.30	24.36	23.66	23.36
	(e) − (d)	24.92	18.38	14.90	12.92	9.42	7.78	7.00	5.46	4.94	4.82
S_3	(a) C-MSDPD	52.46	53.48	53.02	53.38	53.40	52.52	53.06	53.52	53.58	53.52
	(b) PBP	52.40	53.40	53.60	54.00	54.70	55.14	55.74	55.38	55.74	56.58
	(c) HYBIRD	52.34	53.46	55.12	55.28	56.38	56.76	56.66	57.08	58.00	57.98
	(d) SM-MDNNF+IPDP	50.48	50.58	50.94	51.06	51.10	50.68	51.10	51.20	51.48	51.76
	(e) SM-MDNNF+FPDP	90.94	93.84	95.56	95.62	97.12	97.84	97.76	98.76	98.84	99.02
	(e)−MAX(a,b,c)	38.48	40.36	40.44	40.34	40.74	41.08	41.10	41.68	40.84	41.04
	(e)−(d)	40.46	43.26	44.62	44.56	46.02	47.16	46.66	47.56	47.36	47.26

表 9 1 秒不同语种样本下 SM-MDNNF+FPDP 对于三种隐写方法的检测准确率对比（%）

隐写方法	语种	嵌入率（%）									
		10	20	30	40	50	60	70	80	90	100
S_1	中文	74.92	90.90	97.26	99.26	99.58	99.88	100.00	100.00	100.00	100.00
	英文	77.34	91.82	97.96	99.62	99.80	99.88	99.98	100.00	100.00	100.00
S_2	中文	85.38	96.70	99.42	99.76	99.94	99.96	99.96	100.00	100.00	100.00
	英文	86.58	96.72	99.10	99.86	99.90	99.96	100.00	100.00	100.00	100.00
S_3	中文	60.58	70.08	78.52	84.34	90.08	93.84	95.54	97.04	98.76	99.28
	英文	60.74	70.76	79.44	85.42	90.54	94.10	95.26	97.70	99.20	99.02

（5）在低嵌入率下，SM-MDNNF+FPDP 在英文样本集上的检测准确率往往会高于在中文样本集上的检测准确率。如表 9 所示，对于 1 秒语音长度和 10%嵌入率条件下，对中文样本集下 S_1 (S_2)方法的检测准确率为 74.92%(85.38%)，而在英文样本集下的检测准确率为 77.34%(86.58%)，高出了 2.42(1.2) 个百分比。之所以出现这种现象，可能与两种语言的字母、语法和音系等不同有关[33, 34]。特别是音系可能影响最大，因为汉语有 412 个音节，而英语则由 20 个元音和 28 个辅音组成。换言之，较之英语，汉语音系更为复杂[33, 34]。然而，随着嵌入率的增加，隐写前后的检测特征差异愈来愈大，而语种的影响也将随之降低。因此，在较高嵌入率下，SM-MDNNF+FPDP 在中、英文样本集上的检测准确率基本持平，并且受初始参数的随机性引起误差的影响，还可能偶尔出现在英文样本集上的准确率略低于中文样本集上准确率的情况，但其差异一般较小（不超过 0.32%）。

（6）我们以 100%嵌入率的中英文样本集为实验对象，进一步测试了不同检测方案的单个样本平均检测时间，其统计结果如表 10 所示。整体而言，随着样本长度的增加，不同检测方案检测单个样本的时间开销都会随之增加。此外，不难看出，SM-MDNNF+FPDP 的检测时间开销（0.427ms 至 0.744ms）要远小于现有的三种方案，分别仅占 CMSDPD 对应样本长度下检测时间开销的 2.60%至 6.91%，PBP 时间开销的 14.03%至 35.83%，Hybrid 时间开销的 15.08%至 33.85%。这充分说明本文方案较之传统隐写方案在检测实时性方面具有明显的优势。

综上所述，本文提出的各种检测模式均是可行和有效的，其中基于三类模型融合的检测模式整体性能最优。特别是本文研究工作首次实现了对小数基音延迟隐写方法的成功检测，且对于各类基音延

表 10 不同隐写检测方案的单个样本平均检测时间(ms)

检测方案	0.1s	0.2s	0.3s	0.4s	0.5s	0.6s	0.7s	0.8s	0.9s	1.0s
CMSDPD	6.185	8.545	10.933	13.468	15.938	18.418	20.984	23.388	25.931	28.596
PBP	1.192	1.637	2.096	2.551	3.002	3.478	3.924	4.397	4.852	5.300
HYBIRD	1.262	1.657	2.072	2.459	2.859	3.272	3.653	4.062	4.481	4.931
SM-MDNNF+FPDP	**0.427**	**0.457**	**0.489**	**0.527**	**0.564**	**0.608**	**0.657**	**0.680**	**0.712**	**0.744**

迟隐写方法在任意的嵌入率和样本长度下较之已有方案均具有更好的检测性能和更少的时间开销。

5 结论

基于 AMR 的语音服务不管是在 Android 和 iOS 移动操作系统还是在各种即时通信 APP（如微信和 iMessage 等）中都有着广泛的应用。基于 AMR 的隐写及其检测是当前信息隐藏检测领域中的一个研究热点。针对基于基音延迟隐写的高效检测问题，提出了一种新的基音延迟隐写分析方案。该方案聚焦小数基音延迟相关性，并引入深度学习网络对特征进行自动学习和抽取以获得高表征性的检测特征及模型，并结合特征融合和基于线性回归的多模型融合策略，给出了多种可行的深度神经网络检测模式。通过大量的语音样本，对提出的方案进行了性能评估，并与相关工作进行了比较分析。实验结果表明本文提出隐写方案是可行和有效的，本文提出的各种检测模式均是可行和有效的，其中基于三类模型融合的检测模式整体性能最优。特别是本文研究工作首次实现了对小数基音延迟隐写方法的成功检测，且对于各类基音延迟隐写方法在任意的嵌入率和样本长度下较之已有方案均具有更好的检测性能和更少的时间开销。

值得指出的是，尽管提出的方案以 AMR 语音流作为应用背景进行描述，由于基于 ACELP 的语音编码过程基本类似，本方案还可扩展应用于其他各类 ACELP 编码（如 G729， G723.1 和 iLBC 等）语音流。

参考文献

[1] Cheddad A, Condell J, Curran K, Mc K. Digital image steganography: survey and analysis of current methods. Signal Processing, 2010, 90(3): 727–752

[2] Wu You-Qing, Guo Yu-Tang, Tang Jin, Luo Bin, Yin Zhao-Xia. Reversible Data Hiding in Encrypted image using apaptive Huffman encoding strategy. Chinese Journal of Computers, 2021, 44(4): 846-858 (in Chinese)
(吴友情, 郭玉堂, 汤进, 罗斌, 殷赵霞. 基于自适应哈夫曼编码的密文可逆信息隐藏算法. 计算机学报,2021, 44(4): 846-858)

[3] Yan D Q, Wang R D, Yu X M, Zhu J. Steganography for mp3 audio by exploiting the rule of window switching. Computers & Security, 2012, 31(5): 704–716

[4] Li Song-Bin, Wang Lin-Rui, Liu Peng, Huang Yong-Feng. A hevc information hiding approach based on motion vection space encoding. Chinese Journal of Computers, 2016, 39(7): 1450-1463(in Chinese)
(李松斌, 王凌睿, 刘鹏, 黄永峰. 一种基于运动矢量空间编码的HEVC信息隐藏方法. 计算机学报, 2016, 39(7): 1450-1463)

[5] Shah M K, Patel S B. Network based packet watermarking using tcp/ip protocol suite//Proceedings of the 2011 Nirma University International Conference on Engineering. Ahmedabad, India, 2011: 1–5

[6] Mazurczyk W, Szaga P, Szczypiorski K. Using transcoding for hidden communication in ip telephony. Multimedia Tools and Applications, 2014, 70(3): 2139–2165

[7] Murdoch S J, Lewis S. Embedding Covert Channelsinto TCP/IP. //Proceedings of the 7th International Workshop on Information Hiding. Berlin,Germany, 2005: 247-261

[8] Mazurczyk W, Szczypiorski K. Covert Channels in SIP for VoIP Signalling//Proceedings of International Conference on Global e-Security. Berlin, Germany,2008: 65-72

[9] Huang Y F, Tang S Y. Covert voice over internet protocol communications based on spatial model. Science China Technological Sciences, 2016, 59(1): 117-127

[10] Liang C, Wang X, Zhang X. A payload-dependent packet rearranging covert channel for mobile VoIP traffic. Information Sciences, 2018 (465): 162-173

[11]Mazurczyk W. Lost audio packets steganography: the first practical evaluation. Security and Communication Networks, 2012, 5(12): 1394-1403

[12] Mazurczyk W, Szaga P, Szczypiorski K. Using transcoding for hidden communication in IP telephony. Multimedia Tools and Applications, 2014, 70(3): 2139-2165

[13] Tian H, Sun J, Chang C C, et al. Hiding Information into Voice-over-IP Streams Using Adaptive Bitrate Modulation. IEEE Communications Letters, 2017, 21(4): 749-725

[14] Tian H, Qin J, Guo S, et al. Improved adaptive partial matching steganography for Voice over IP. Computer Communications, 2015, 70(C): 95- 108

[15] Peng X, Huang Y, Li F. A steganography scheme in a low-bit rate speech codec based on 3D-sudoku matrix//Proceedings of IEEE International Conference on Communication Software and Networks. New York, USA,2016: 13-18

[16] Harada N, Kamamoto Y, Moriya T, Hiwasaki Y, Ramalho M, Netsch L, Stachurski J, Miao L, Taddei H, Qi F. Emerging itu-t standard

g.711.0—lossless compression of g.711 pulse code modulation// Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. Dallas, USA, 2010: 4658–4661

[17]Benyassine A, Shlomot E, Su H Y, Massaloux D, Lamblin C, Petit JP. ITU-t recommendation g.729 annex b: a silence compression scheme for use with g.729 optimized for v.70 digital simultaneous voice and data applications. IEEE Communications Magazine, 1997, 35(9): 64–73

[18] Xu T T, Yang Z. Simple and effective speech steganography in g.723.1 low-rate codes//Proceedings of the 2009 International Conference on Wireless Communications Signal Processing. Nanjing, China, 2009: 1–4

[19] Andersen S V, Kleijn W B, Hagen R. iLBC-a linear predictive coder with robustness to packet losses//Proceedings of the Speech Coding, 2002, IEEE Workshop Proceedings. Ibaraki, Japan, 2002: 23-25.

[20]Goudarzi M, Sun L F, Ifeachor E. Modelling speech quality for nb and wb silk codec for voip applications//Proceedings of the Fifth International Conference on Next Generation Mobile Applications, Services and Technologies. Cardiff, UK, 2011: 42–47

[21]Xie Xiao-Gang, Cai Jun, Chen Qi-Chuan, Ou Jian-Lin. Design and implementation of voip system based on speex codec. Application Research of Computers, 2007, 24(12): 320-323(in Chinese)
(谢晓钢,蔡骏,陈奇川,欧建林.基于Speex语音引擎的VoIP系统设计与实现.计算机应用研究, 2007, 24(12): 320-323)

[22] Barany P, Bharatia J, Bontu C. Communications using adaptive multi-rate codecs. US7072336B2. Washington, USA 2006-07-04

[23] Xiao B, Huang Y F, Tang S Y. An approach to information hiding in low bit-rate speech stream//Proceedings of the IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference. New Orleans, USA, 2008: 1–5

[24] Huang Y F, Liu C, Tang S Y, Bai S. Steganography integration into a low-bit rate speech codec. IEEE Transactions on Information Forensics and Security, 2012, 7(6): 1865–1875

[25] Yan S F, Tang G M, Sun Y F. Steganography for low bit-rate speech based on pitch period prediction. Application Research of Computers, 2015, 32(6): 1774-1777 (in Chinese)
(严书凡,汤光明,孙怡峰.基于基音周期预测的低速率语音隐写.计算机应用研究, 2015, 32(6): 1774-1777)

[26] Miao H B, Huang L S, Chen Z L, Yang W. A new scheme for covert communication via 3g encoded speech. Computers & Electrical Engineering, 2012, 38(6): 1490–1501

[27] Geiser B, Vary P. High rate data hiding in acelp speech codecs// Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, USA, 2008: 4005–4008

[28] Liu X K, Tian H, Huang Y F, Lu J. A novel steganographic method for algebraic-code-excited-linear-prediction speech streams based on fractional pitch delay search. Multimedia Tools and Applications, 2019, 78(7): 8447–8461

[29] Ren Y Z, Yang J, Wang J, Wang L. AMR steganalysis based on second-order difference of pitch delay. IEEE Transactions on Information Forensics and Security, 2017, 12(6): 1345–1357

[30] Liu X K, Tian H, Liu J, Li X, Lu J. Steganalysis of adaptive multiple-rate speech using parity of pitch-delay value//Proceedings of the Security and Privacy in New Computing Environments. Tianjin, China, 2019, 282–297

[31] Tian H, Huang M, Chang C C, Huang Y F, Lu J, Du Y Q. Steganalysis of adaptive multi-rate speech using statistical characteristics of pitch delay. Journal of Universal Computer Science,2019, 25(9): 1131–1150

[32] Salami R, Laflamme C, Adoul J P. 8 kbit/s ACELP coding of speech with 10 ms speech-frame: A candidate for CCITT standardization//Proceedings of ICASSP'94 IEEE International Conference on Acoustics, Speech and Signal Processing. Adelaide, Australia, 1994:II/97-II100

[33] Lin Z N, Huang Y F, Wang J L. RNN-SM: fast steganalysis of voip streams using recurrent neural network. IEEE Transactions on Information Forensics and Security, 2018, 13(7): 1854–1868

[34] Yang H, Yang Z, Huang Y. Steganalysis of voip streams with cnn-lstm network//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. New York, United States, 2019: 204-209

TianHui,Ph.D., professor, doctoral supervisor. His research interests include network and information security, big data security, steganography & steganalysis, artificial intelligencesecurity, cloud security and digital forensics.

Wu Jun-Yan,M.S. candidate. Her research interests include information hiding and detection and deep learning.

Yan Yan, M.S. candidate. Her research interests include information hiding and detection and deep learning.

Wang Hui-Dong, M.S. candidate. His research interests include information hiding and detection and federated learning.

Quan Han-Yu, Ph.D., lecturer. His research interests include applied cryptography and privacy protection.

Background

Steganography and steganalysis of network speech streams are research hotspots in the field of information hiding. Adaptive Multi-Rate (AMR) coding is the most widely used speech coding technology in the mobile speech scenarios. Moreover, it has a wide range of applications in not only mobile operating systems like Android and iOS but also various instant messaging apps (such as WeChat and iMessage). This paper mainly focuses on the detection of

steganography based on pitch delay in AMR speech streams. In the AMR coding, the pitch period has strong instability, and there are large differences between the voices of different individuals. Therefore, it is difficult to achieve accurate prediction during adaptive codebook search. It is precisely based on this feature that modifying the pitch delay parameters has less impact on the quality of synthesized speech, so the pitch delay parameters can be used as ideal information hiding carriers. At present, the detection methods for pitch delay-based steganography usually adopt the method of constructing manual features combined with machine learning. Although they can achieve a certain detection performance, research shows that they still have some shortcomings. The detection accuracy of machine learning-based detection methods depends on the design of manual features, and the characterization ability of the target is usually insufficient, which ultimately leads to defects in detection performance, such as relatively low detection performance for short-term or low-embedding speech samples. In addition, since the existing schemes all employ the statistical characteristics of integer pitch delays, they are powerless for the steganography methods that only change the decimal pitch delay.

Therefore, this paper proposes a steganalysis scheme based on the correlation of fractional pitch delay. Firstly, through theoretical analysis and experimental comparison, the effectiveness of fractional pitch delay correlation as steganographic features is verified. Secondly, the traditional method of manual feature extraction is abandoned, and the correlation of coding elements is captured by using deep neural networks. Accordingly, a local correlation-based detection model, a global correlation-based detection model and a feature fusion-based detection model are respectively designed. Finally, based on the above three models, combined with the idea of multi-model fusion based on linear regression, seven detection modes are given, i.e., three single-model detection modes and four multi-model fusion detection modes. Through a large number of speech samples, the performance of the proposed scheme is comprehensively evaluated, and compared with state-of-the-art works. The experimental results show that the presented various detection modes are feasible and effective, and the three-model fusion detection mode has the best overall performance. In addition, the work of this paper fills in the blank of the detection of steganography based on fractional pitch delay, and for various steganography methods based on pitch delay, it has better detection performance and lower time overhead than the existing steganalysis schemes at any embedding rate and sample length, thereby realizing more real-time and efficient detection.

The research work in this paper was partially funded by the National Natural Science Foundation of China (No.619 72168) and the Open Project of the State Key Laboratory of Information Security (No.2019ZD09).

The work in this paper aims to achieve steganalysis based on pitch delay more effectively, even for speech samples with low embedding rates or/and short duration. Our research group has long focused on research regarding steganography and steganalysis in network speech streams, and has published a series of papers in high-level journals and academic conferences at home and abroad.