基于语音频谱与数据包长对齐的 VoIP 加密流量识别方法

赵俊舟[1)] 李江龙[1)] 段 涛[1)] 王平辉[1)] 陶 敬[1)]

[1)](西安交通大学智能网络与网络安全教育部重点实验室 西安 710049)

摘 要 随着智能手机等移动终端的迅速普及，以微信电话为代表的互联网语音（Voice over Internet Protocol, VoIP）应用日益流行。VoIP 应用在开放的 Internet 中传递涉及用户隐私的语音内容，保障用户个人数据安全至关重要。本文采集并分析了包括微信、TIM、腾讯会议、钉钉在内的四款流行 VoIP 应用在使用过程中产生的语音流量，发现尽管 VoIP 应用普遍采用私有语音编码算法、加密通信等手段保障安全，但是 VoIP 加密流量的传输模式仍有可能泄露用户属性、用户身份，甚至通话内容等敏感信息，存在隐私泄露风险。本文通过测量分析四种 VoIP 应用的加密流量传输模式与用户属性、通话内容等方面的关联关系，发现语音频率与数据包长存在明显的相关性，并基于该发现设计了一种语音频谱与数据包长对齐的 VoIP 加密流量识别方法——VPrint。VPrint 较已有的加密流量识别方法能更准确识别 VoIP 加密流量。以微信为例，VPrint 在用户性别识别、用户身份识别、通话语种识别和短语识别任务上的 F1 值分别为 0.77、0.99、0.88 和 0.92。本文研究结果表明微信等流行 VoIP 应用存在安全隐患，并建议相关厂商采取数据包填充等措施提升安全性，避免造成用户隐私泄露。

关键词 互联网语音应用；加密流量识别；隐私保护；数据安全

中图法分类号 *** DOI 号 ***

VoIP Encrypted Traffic Recognition via Aligning Voice Spectra and Packet Length

ZHAO Jun-Zhou[1)] LI Jiang-Long[1)] DUAN Tao[1)] WANG Ping-Hui[1)] TAO Jing[1)]

[1)](MOE KLINNS Lab, Xi'an Jiaotong University, Xi'an 710049)

Abstract With the rapid popularization of mobile devices, Voice over Internet Protocol (VoIP) applications represented by WeChat have become increasingly popular. VoIP applications transmit private user voice over the Internet, and their security is of vital importance. This work collects and analyzes the voice traffic generated by four popular VoIP applications, including WeChat, TIM, Tencent Meeting, and DingTalk. Although VoIP applications generally adopt private voice coding algorithms and encrypted communication to ensure security, the transmission patterns of VoIP encrypted traffic may still leak user identities and even private voice content, posing privacy leakage risks. Specifically, this work measures and analyzes the correlation between the encrypted traffic transmission patterns of the four VoIP applications and user attributes, voice content, etc., and discovers a significant correlation between voice frequency and packet length. Based on this finding, a VoIP encrypted traffic identification method called VPrint is designed, which aligns voice spectra with packet lengths. Compared with existing encrypted traffic identification methods, VPrint can identify VoIP encrypted traffic more accurately. Taking WeChat as an example, VPrint achieves an F1 score of 0.77 for user gender identification, 0.99 for user identification, 0.88 for call language iden-

收稿日期：***；修改日期：*** 本课题得到国家自然科学基金面上项目（62272372）资助.赵俊舟，副教授，主要研究领域为网络安全. E-mail: junzhou.zhao@xjtu.edu.cn. 李江龙，硕士研究生，主要研究领域为网络流量分析. E-mail: ljl522426@stu.xjtu.edu.cn. 段涛，博士研究生，主要研究领域为网络流量分析，网络隐私与安全. E-mail: duantao@stu.xjtu.edu.cn. 王平辉，教授，主要研究领域为网络流量分析与网络安全. E-mail: phwang@xjtu.edu.cn. 陶敬（**通信作者**），研究员，主要研究领域为网络安全. E-mail: jtao@xjtu.edu.cn.

tification, and 0.92 for phrase identification. This work shows that popular VoIP applications such as WeChat have security risks, and it is recommended that relevant companies take measures such as packet padding to enhance security and prevent possible user privacy leakage.

Key words Voice over Internet Protocol Applications; Encrypted Traffic Identification; Privacy Protection; Data Security

1 引言

随着移动网络覆盖率的不断提高和智能手机、平板电脑等移动终端的快速普及，以微信电话、QQ 语音、Skype 为代表的互联网语音（Voice over Internet Protocol，VoIP）应用日益流行。VoIP 应用构建在 Internet 基础之上，通过 IP 分组交换网络传输语音信号。在通话时，发送方产生的模拟语音信号经过压缩编码，按照 TCP/IP 等协议打包为数据包，经 IP 网络传输到目的地，然后接收方对收到的数据包进行重组、解码后恢复出原始语音信号，实现互联网上的语音通信。VoIP 应用因其方便快捷的使用方式以及低廉的使用成本，吸引了越来越多用户的使用，成为继移动电话之外人们使用的主要语音通讯工具。截止 2023 年底，我国 VoIP 用户数量已达 7.5 亿，占全国移动电话用户的 60% 以上；预计到 2025 年底，我国 VoIP 用户数量将达到 9 亿，占全国移动电话用户的 70% 左右[1]。

VoIP 应用的普遍使用同时带来严峻的网络安全隐患。一方面，用户语音通话内容属于高度隐私和机密，而 VoIP 应用在公开 Internet 上传递语音数据，容易受到侧信道攻击从而泄漏用户隐私。例如，White 等人[2]在 2007 年的研究中首次发现 VoIP 应用会泄漏用户语种信息（例如英语、汉语普通话等），在 2008 年的后续研究中进一步发现 VoIP 应用产生的数据包包长信息会泄漏语音通话内容[3]。经过近二十年的发展，SSL/TLS/SRTP/QUIC 等加密传输协议已广泛应用于 VoIP 协议设计[4]，数据包填充和数据包延时传输等主动防御技术[5-7]也已经提出并可以进一步加固 VoIP 应用，这些安全技术的应用使得当前的 VoIP 应用能更好地防范侧信道攻击。尽管如此，近几年来深度学习技术突飞猛进，使当前的机器学习模型具有更强的数据分析和理解能力。因此，有必要在当前新形势下仔细评估 VoIP 应用的安全性。另一方面，近几年来境内外违法犯罪分子利用 VoIP 应用进行电信诈骗等违法犯罪活动日益猖獗，给人民群众带来巨大的生命财产损失[8]，亟需对 VoIP 黑灰产业进行监管与阻断。鉴于此，当前迫切需要研究针对 VoIP 应用的电信诈骗检测技术，包括识别涉诈 VoIP 加密流量、对涉诈人员进行身份画像等。为解决以上问题，就需要系统研究 VoIP 加密流量识别技术。

VoIP 加密流量识别问题不同于传统加密网络流量识别问题。传统加密网络流量识别任务的目标是通过建模网络流量的传输和交互模式，识别网络服务类型、网络应用/协议种类、用户行为等粗粒度类别信息[9-11]。而 VoIP 加密流量识别任务旨在从加密语音通话流量中实现对通话语种、用户身份、甚至通话内容等细粒度信息的识别。VoIP 加密流量的模式不仅与通话内容相关，而且还与用户声纹特征、语音编码算法等相关，导致传统加密流量识别方法不再适用于 VoIP 加密流量识别任务。

White 等人[2-3,12]的早期研究依赖于 VoIP 应用使用的公开语音编码算法（即将模拟语音信号编码为数据包负载数据的过程），建立语音音节与数据包之间的关联关系，利用隐马尔可夫链等序列模型建模音节流量片段的模式，从而识别 VoIP 流量。然而，目前 VoIP 应用采用的语音编码算法往往由各厂商独立开发或高度定制，并不公开。同时，VoIP 应用通常会使用数据包填充和数据包延迟传输等主动防御手段来扰乱 VoIP 应用的流量模式[5-7]，以抵抗第三方流量分析。这些新挑战导致 White 等人早期提出的 VoIP 加密流量识别方法不再可行。

为全面评估当前 VoIP 应用的安全性，并提出可行的 VoIP 加密流量识别方法，本文在实验室环境采集了包括微信[1]、TIM[2]、钉钉[3]、腾讯会议[4]在内的四款流行 VoIP 应用产生的加密网络流量数据。研究发现，即使不同 VoIP 应用采用不同的私有语音编码算法以及加密通信技术，通过测量分析 VoIP 应用产生的加密流量传输模式与用户属性、通话内容等方面的关联关系，发现通话语音频率与数据包长存在明显的相关性：频率高的语音更容易产

[1]https://weixin.qq.com
[2]https://tim.qq.com
[3]https://www.dingtalk.com
[4]https://meeting.tencent.com

生大的数据包，而频率低的语音更容易产生小的数据包。基于该发现，本文设计了一种通过对齐语音频谱与数据包长的 VoIP 加密流量识别方法——VPrint。VPrint 较已有的加密流量识别方法能更准确识别 VoIP 加密流量，并且在用户属性识别、用户身份识别、通话语种识别、短语识别等任务上都优于基线方法。VPrint 只利用 VoIP 加密流量中包长分布特征，不依赖于具体的语音编码算法，因此具有一定的通用性。本文主要贡献总结如下：

- 本文系统测量分析了四款主流 VoIP 应用产生的加密流量的传输模式与用户属性、通话内容等方面的关联关系，发现通话语音频率与数据包长存在明显的相关性。
- 本文提出一种通过对齐语音频谱与数据包长的 VoIP 加密流量识别方法——VPrint。VPrint 较已有的加密流量识别方法能更准确识别 VoIP 加密流量。
- 实验表明 VPrint 在用户属性识别、用户身份识别、通话语种识别、短语识别等任务上都优于基线方法，F1 值较已有方法提升 5% 以上。
- 本文研究揭示了目前 VoIP 应用的安全隐患，建议相关厂商应采取数据包填充等措施加固应用的安全性，避免造成用户隐私泄露。

2　相关工作

网络流量识别是网络监管与安全分析的核心技术手段，被广泛应用在协议分析、服务识别、入侵检测、IoT 设备画像等场景。早期基于深度包检测（Deep Packet Inspection, DPI）[13] 的流量识别方法采用明文指纹匹配策略来识别流量内容。然而，随着内容加密技术的普及，基于 DPI 的方法逐渐失效。近年来，随着深度学习技术的发展，基于数据包长度、时序模式、交互行为等侧信道特征的加密流量识别技术取得显著进展[11,14-17]。

随着应用场景的扩展和网络管理需求的深化，加密流量识别的研究重点正逐步从基础服务类型识别向更精细化的分析维度演进，从而衍生出不同粒度的流量识别领域——粗粒度流量分类与细粒度流量信息解析[18]。粗粒度流量分类包括应用识别[19-20]、流量类型识别（如 VPN 流量与非 VPN 流量识别[21]、VoIP 流量识别[22] 等）、异常流量识别[23] 等。细粒度流量信息解析则关注加密流量中所包含的细粒度信息，例如识别用户与应用的交互行为（如点赞或发送消息等）[24-26]、用户浏览网站内容[14]、用户通话的内容[12]、通话用户的身份属性[27] 等。作为细粒度流量解析的一种，VoIP 加密流量识别主要针对 VoIP 应用传输语音信号所产生的 VoIP 流量，构建声学—流量特征关联，实现通话内容识别（如语种识别[2]、敏感短语识别[3]）或通话用户身份识别（如非法用户[27]）。现有 VoIP 加密流量识别按技术路线可分为基于特征工程的传统流量分析方法与基于语音建模的流量分析方法。

基于特征工程的传统流量分析方法。该类方法主要借鉴传统粗粒度流量识别框架，通过人工构建 VoIP 流量的多维统计特征集（如数据包长度分布、传输时间间隔、会话持续时间等）来实现 VoIP 流量指纹建模。例如，Wang 等人[7] 沿用 AppScanner 的特征工程范式，从不同语音内容的 VoIP 流量中提取时序特征构建指纹库，并引入随机森林与卷积神经网络实现内容分类，在加密环境下取得初步识别效果。为进一步简化特征工程复杂度，FS-Net[28]、DeepFinger[14] 等通用加密流量识别模型通过仅采用数据包长度、时序间隔等基础特征，构建 VoIP 流量内容编码器，实现了对 VoIP 流量内容模式的自动化表征。此类方法普遍存在特征表示维度受限的问题，人工特征工程不能构建语音与流量模式特征关联模型，难以捕捉 VoIP 流量模式中的声学特征，导致传统方法在 VoIP 加密流量识别任务中效果并不好。

基于语音建模的流量分析方法。Wright 等人[2-3,12] 开创性地提出基于语音建模的流量分析方法，根据语音编码算法参数设置将 VoIP 流量划分为多个片段并与语音音节相匹配，构建流量模式与语音音节的统计关联。随后，对不同音节流量模式提取统计特征构建音节关联模型，完成通话内容推断，实现语种识别[2]、短语识别[3,12]。为进一步揭露 VoIP 传输的信息泄露问题，Khan 等人[27] 基于相同方法，引入并构建声纹特征库，基于 SVM 方法实现 10 位通话用户 75% 的身份识别准确率。然而，当前流行 VoIP 应用普遍采用私有语音编码算法，导致同一语音内容在不同 VoIP 应用中传输的流量模式存在显著差异，阻碍了流量特征与语音音节间的关联解析。

综上所述，VoIP 加密流量识别技术主要面临以下挑战：1）基于特征工程的传统流量分析方法不能建模语音与流量之间的关联关系；2）私有语音编码算法导致已有的基于音节切割的方法失效。因

此，需要研究新的 VoIP 加密流量识别方法。

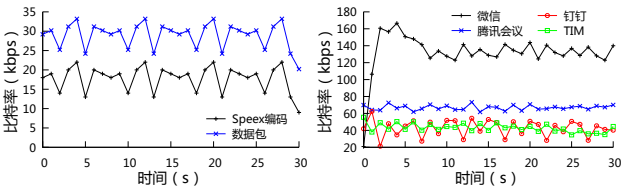
3 VoIP 加密流量测量分析

使用 VoIP 应用在 Internet 上进行语音通信时，发送方产生的模拟语音信号经过压缩编码，打包为数据包并经网络传输到接收方，接收方对收到的数据包进行重组与解码，恢复出原始语音信号。为了理解语音数据包与语音信号之间的关系，以及语音数据包在网络中的传输模式，本节对微信、TIM、钉钉和腾讯会议四款应用产生的 VoIP 加密流量进行测量分析。

3.1 语音编码与数据包传输模式

VoIP 应用使用语音编码算法将模拟语音信号压缩编码为数据包负载，语音编码算法直接影响 VoIP 应用的性能，包括音质、延迟和网络带宽适应性等。为了能够在有限的网络带宽中传输高质量的语音信号，常用的语音编码算法（例如 AMR 编码、LPC 编码、ISAC 编码、SILK 编码、Speex 编码等）都属于可变比特率（Variable Bit Rate, VBR）编码[29]。VBR 编码会根据模拟语音信号（例如低音、高音、清音、浊音、背景音等）及网络环境采用不同的编码比特率（也称码率）进行编码，以平衡通话音质和网络延迟，使用户具有良好的体验。

本文使用 Speex 编码对一段 30 秒语音[5]进行编码，得到编码数据比特率以及数据包传输比特率的关系，结果如图 1(a) 所示。可以看到，即使网络环境稳定，由于语音信号变化，语音编码结果的比特率也会变化。当语音编码数据作为数据包负载传输时，数据包传输比特率也呈现相同的变化规律：编码比特率大时，数据包传输比特率大（即数据包大），反之数据包传输比特率小（即数据包小）。



(a) 语音编码比特率与数据包传输比特率　(b) 实际 VoIP 应用数据包传输比特率

图 1　语音编码及数据包传输比特率

本文在实验室环境中采集了微信、TIM、钉钉和腾讯会议四种应用产生的 VoIP 加密流量（详见第 5.1 小节），发送方的通话语音为与之前实验相同的 30 秒语音，得到数据包传输比特率变化情况如图 1(b) 所示。可以看到所有应用产生的数据包比特率都随时间波动，因此可以推测出这四种应用使用的语音编码都属于 VBR 编码。尽管四种应用采用的具体语音编码算法未公开，但是由数据包比特率的显著差异可以推测它们使用了不同的语音编码算法，其中微信的数据包传输比特率最大，其次为腾讯会议，TIM 和钉钉则最小。

根据本节的测量分析，可以得出以下结论：四种应用采用的语音编码算法都属于 VBR 编码，编码后的 VoIP 数据包大小会随语音信号变化。

3.2 声音频率与数据包传输模式

语音编码算法往往会对不同的声音频率采用不同的编码策略（即分频带编码策略），以尽可能保留声音中的细节，保障通话音质。例如，AMR-NB 窄带编码算法会舍弃频率高于 3.4kHz 的声音，实现传统电话的语音通话质量，而 AMR-WB 宽带编码算法会对最高 7kHz 的声音进行编码，实现高清语音通话[30]。

为了研究不同频率声音的 VoIP 加密流量传输模式的差异，本文使用微信分别采集了由音乐中七个音阶构成的语音产生的流量数据。以国际标准音 A440 为基准，七个音阶 Do、Re、Mi、Fa、So、La、Ti 对应的声音频率分别为 261.6Hz、293.6Hz、329.6Hz、349.2Hz、392Hz、440Hz、493.8Hz。七个音阶的语音（语音时长均为 61 秒）产生的数据包序列模式如图 2 所示。

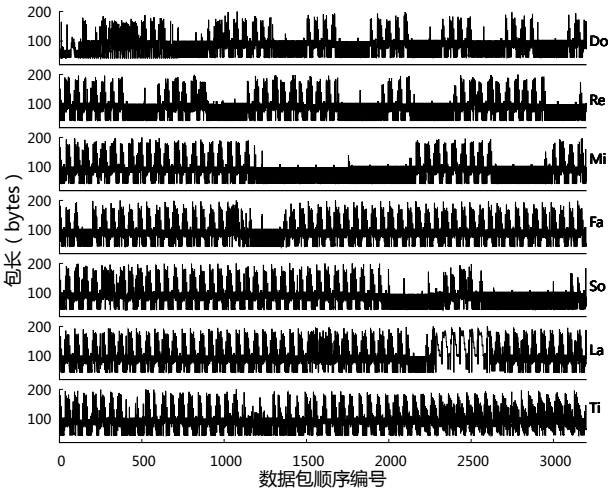


图 2　七个音阶产生的 VoIP 加密流量包长时序分布

从图 2 可以看到不同音阶的语音呈现出显著不同的流量模式。音阶 Do 的频率最低，对应在流

[5]语音内容为“你好”并重复 30 秒。

量中有大量包长小于 100 字节的数据包；音阶 Re 的频率次最低，对应在流量中同样有大量包长小于 100 字节的数据包。而对于高频率音阶 La 和 Ti，从流量中可以观察到存在大量包长大于 100 字节的数据包。这个规律在图 3 中给出的数据包长累积概率分布（CDF）统计结果中同样十分明显。对于其他应用进行相同的测量分析，可以观察到类似的流量模式，由于篇幅限制，不再赘述。

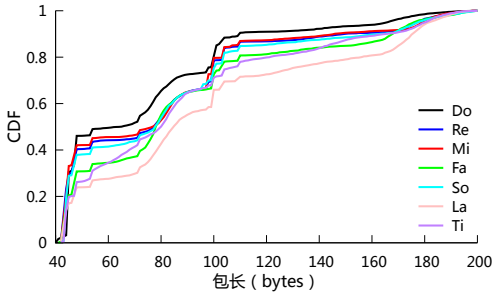


图 3　七个音阶产生的 VoIP 加密流量包长分布

根据本节的测量分析，可以得出以下结论：VoIP 应用对于不同频率的声音会采用不同的编码策略，总的来说，低频声音倾向于采用包长小的数据包传输，高频声音倾向于采用包长大的数据包传输。

3.3　用户属性与数据包传输模式

每个人都有独特的语音特征，即“声纹”，那么每个人是否也具有独特的 VoIP 加密流量模式？为简化问题，本节分析用户属性（例如性别、年龄等）异同造成的流量模式差异，并重点关注通话用户的性别属性。为此，本文采集了 1000 个 VoIP 加密流量样本，其中男女比例为 1 ： 1，并且保持语音内容相同。分别统计四种应用男女流量样本的数据包包长分布，得到结果如图 4 所示。

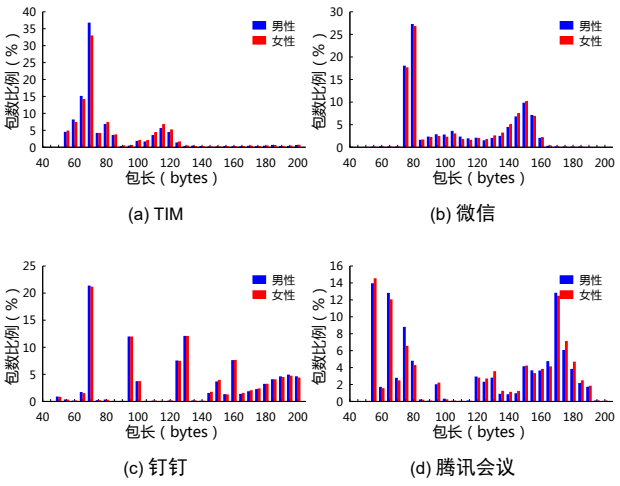


图 4　男女 VoIP 加密流量数据包包长分布

从图 4 可以看到，对于同一种应用，男女流量样本的包长分布相似，而不同应用的数据包包长分布存在显著差异，原因是不同应用的语音编码算法不同，这与图 1(b) 中的测量结果一致。观察不同应用产生的男女流量样本的包长分布情况，发现存在一种比较稳定的现象：分布中对应包长较小的区间（包长小于 100 字节），男性的包数比例往往略高于女性；而对应包长较大的区间（包长大于 120 字节），女性的包数比例往往略高于男性。为了更清楚地描述这一细微差异，用 $f_{男,l}$ 和 $f_{女,l}$ 分别表示男女流量样本中包长为 l 的数据包所占比例，用

$$\Delta f_l \triangleq f_{男,l} - f_{女,l}$$

表示包长为 l 的男女数据包比例差异。将 Δf_l 进一步归一化到区间 $[-1, 1]$，得到结果如图 5 所示。

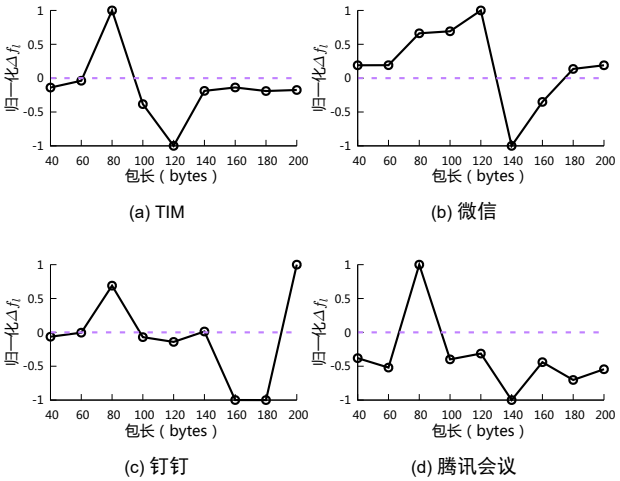


图 5　归一化男女数据包比例差异

从图 5 中可以清楚看到存在如下规律：在四种应用中，随着包长 l 的增大，Δf_l 都出现了从正到负的变化，说明随着包长的增大，在一定包长范围内，男女数据包比例发生消长变化，从男性数据包占主导变化为女性数据包占主导。产生这个现象的可能原因是女性声音往往比男性声音存在较多的高频分量，而根据第 3.2 小节中的发现，VoIP 应用倾向于对高频声音采用包长大的数据包传输，因此 VoIP 应用对女性声音倾向于产生包长大的数据包。

为了验证该论断，图 6 给出了包含相同语音内容的男女声音频谱，以及 VoIP 加密流量包长分布[6]。从图 6(a) 中的频谱分布可以看到，男声往往比女声有更强的低频分量，而女声比男声有更强的高频分量；从图 6(b) 中的包长分布可以看到，在一定包长

[6]语音内容同前，VoIP 加密流量采集自微信。

范围内，男声往往比女声有更多的小数据包，而女声往往比男声有更多的大数据包。

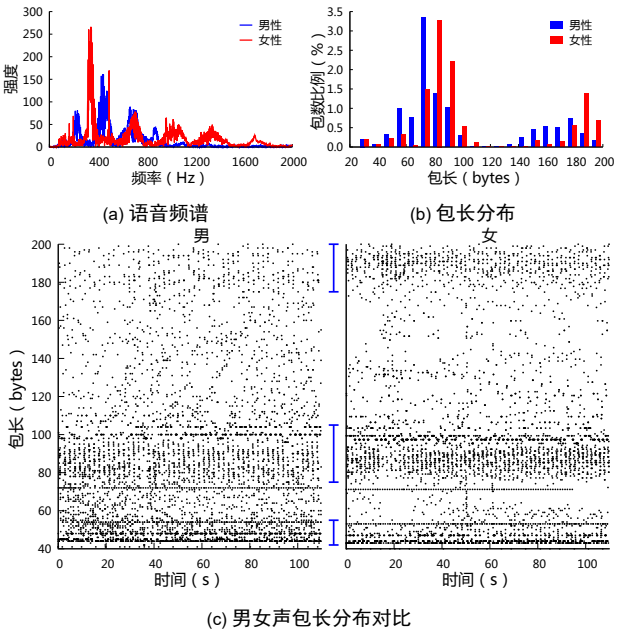


图 6 男女声语音频谱及 VoIP 加密流量包长分布

此外，图 6(b) 中数据包主要集中的包长范围和图 6(a) 中声音的主要频率分量存在一定的对应关系。例如，女性 380Hz 附近的频率分量对应包长为 80 字节左右的数据包，而女性高频分量对应包长为 190 字节左右的数据包。这一现象在图 6(c) 也可以清楚的观察到：数据包集中在几个主要的包长区间中，这类似于声音集中在几个主要的频带范围内。以上这些发现可以被用来由 VoIP 加密流量的传输模式识别说话人的性别属性。

根据本节的测量分析，可以得出以下结论：VoIP 应用对于不同性别的声音会呈现不同的流量模式，总的来说，对于较小包长的数据包，男性的包数比例往往略高于女性，而对于较大包长的数据包，女性的包数比例往往略高于男性；此外，音频频谱与数据包长分布间存在一定的频带对应关系。

4 VPrint：VoIP 加密流量识别算法

基于第 3 节的测量分析结果，本节提出一种 VoIP 加密流量识别算法（**VoIP Encrypted Traffic Fingerprinting**，简称 VPrint），解决 VoIP 应用语音编码算法未知情况下的 VoIP 加密流量识别问题。

4.1 方法概述

VPrint 利用上一节发现的 VoIP 加密流量数据包传输模式与语音编码算法、语音频率及说话人属性之间的关系，提取加密流量中的关键特征，并构建神经网络分类模型，实现说话人属性识别、身份识别、语种识别和短语识别。VPrint 算法整体框架如图 7 所示。

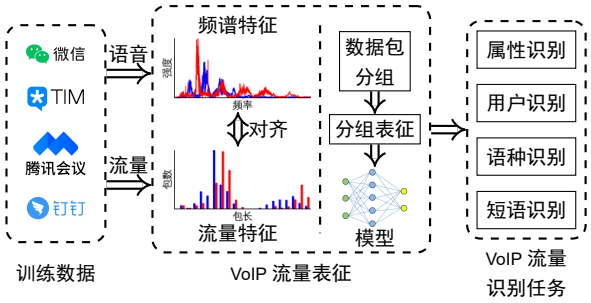


图 7 VPrint 整体框架图

VPrint 同时使用 VoIP 应用产生的流量数据和输入语音数据作为训练数据，学习 VoIP 加密流量的表征。在进行流量表征时，通过对齐语音频谱特征和流量包长特征，VPrint 能获得较好的 VoIP 流量表征向量。具体而言，VPrint 首先对数据包进行分组，分组时会融合语音频谱特征和 VoIP 流量特征，并且数据包分组是一个无监督学习过程；然后，VPrint 获取每个数据包分组的统计特征，将不同分组的特征拼接，形成整个流量的特征向量；最后，VPrint 使用一个神经网络模型，以流量特征向量作为输入，在不同下游任务上训练该模型。在测试阶段，VPrint 不再使用语音数据，利用学到的与任务相关的数据包分组策略和神经网络模型对 VoIP 流量进行识别，解决不同的 VoIP 流量识别任务。

4.2 VoIP 加密流量特征提取

鉴于数据包长与语音频率之间的相关性，一种简单直接的 VoIP 加密流量特征提取方法是将每个流量样本表示为包长序列，然后训练包长序列到样本标签之间的映射关系。然而，这种流量表征方法容易受数据包序列长度的影响，为了使不同长度流量样本的表征能够对齐，需要进行数据包序列截断或填充；另外，即使进行流量样本长度对齐，也不能保证语音频带在数据包长表征空间中实现对齐，从而可能导致流量表征失效。

为了解决以上问题，使流量样本的数据包长在语音频带上保持对齐，本节提出一种通过对数据包长进行自动分组，使不同流量样本的每个包长分组近似与一个语音频带对应（例如高频、中高频、中低频、低频等），从而实现流量表征的语义对齐，进而实现较好的 VoIP 加密流量表征，提升在多种下游流量识别任务中的性能。以下分别详细介绍如何

进行数据包分组以及对每个数据包分组进行表征。

4.2.1　数据包分组

对数据包进行分组源于以下基本想法：类似于每个人的语音都具有独特的频谱特征，并且语音的低频、中频、高频等频带都是刻画个人语音频谱特征的重要成分，因此，如果可以对 VoIP 加密流量数据包进行相应的分组，使每个分组对应一个语音频带，那么基于这些数据包分组来刻画 VoIP 加密流量就可以得到能反映关键语音特征的流量表征，将有助于进行 VoIP 加密流量识别。

尽管第 3 节的测量分析已经发现数据包长与语音频率存在关联性，但数据包长并不是和语音频率存在一对一的关系，这给数据包分组带来困难。为解决该问题，本节提出一种融合样本语音频谱特征的 VoIP 加密流量数据包分组算法，其伪代码如算法 1 所示。

算法 1: 融合语音频谱特征的数据包分组算法

输入: 语音及流量样本集合 $D=\{(V_i,T_i)\}$，其中 V_i 和 T_i 分别表示第 i 条语音样本和对应的流量样本；

输出: 数据包分组 $\{G_k\}_{k=1}^{K}$。

1 初始化 $G_1,\dots,G_K$；　// 将整个包长范围均匀划分为 K 个区间
2 **repeat**
3　$U\leftarrow\varnothing$;
4　**for** $i\leftarrow 1$ **to** $|D|$ **do**
　　// 提取每个分组出现数量最多的包长
5　　$(l_i^{(1)},\dots,l_i^{(K)})\leftarrow$ `PeakFinder`$(T_i,\{G_k\})$;
　　// 提取语音样本主要频率分量
6　　$(f_i^{(1)},\dots,f_i^{(K)})\leftarrow$ `TopFFTFreqs`(V_i);
7　　构建频率和包长近似关联关系 $\{(f_i^{(k)},l_i^{(k)})\}_{k=1}^{K}$;
　　// 为每个数据包关联一个语音频率
8　　**foreach** $pkt\in T_i$ **do**
9　　$k^*\leftarrow\arg\min_k|pkt.len-l_i^{(k)}|$;
10　　$U\leftarrow U\cup\{(f_i^{(k^*)},pkt.len)\}$;
　　// 对点集 U 聚类并更新数据包分组
11　$C\leftarrow$ `KMeansClustering`(U,K);
12　$(G_1,\dots,G_K)\leftarrow$ `Update`(C);
13 **until** 数据包分组 $\{G_k\}$ 稳定;
14 **返回** $\{G_k\}_{k=1}^{K}$;

数据包分组将数据包按包长划分为 K 个互不重叠的包长区间 $G_k=[l_{min}^{(k)},l_{max}^{(k)}]$，$k=1,\dots,K$，其中 $l_{min}^{(k)}$ 和 $l_{max}^{(k)}$ 分别是数据包分组 G_k 的包长下限和上限，并且希望属于分组 G_k 的数据包与某个语音频带相关联。算法 1 联合语音样本的频谱分布实现对数据包的近似分组。首先，将数据包分组的初始状态设定为整个包长范围的均匀划分（第 1 行），然后通过对数据包分组不断迭代更新（第 2–12 行），得到稳定的数据包分组（即数据包分组的变化量小于一个设定的阈值）。其中，第 5 行获取每条流量样本在 K 个数据包分组中的峰值包长（即出现数量最多的数据包长），第 6 行获取对应的语音样本的前 K 个主要频率分量，并构建包长与频率的关联关系（第 7 行）。第 8–10 行通过为流量样本中的每个数据包关联一个语音频率，进而可以将每个数据包表示为二维平面中的点，每个点用频率和包长表示。第 11 行对这些点进行 K-means 聚类，得到数据包的 K 个类簇。第 12 行计算每个类簇的包长上下限，进而更新对应的数据包分组。重复以上步骤，直至数据包分组稳定，于是便得到数据包分组与语音频带的近似对应关系。

4.2.2　数据包分组表征

得到数据包分组后，接下来可以对每个分组进行表征，用来代表每个语音频带的特征。然后，将所有分组的表征拼接起来，得到整个 VoIP 加密流量的表征。每个数据包分组的特征包括基础统计量、方向差异、分布密度和流量强度四类统计特征，构成一个 20 维向量。以下详细给出这四类统计特征的计算方法。

基础统计量：计算数据包分组包长序列的方差（维度 1）、标准差（维度 2）、均值（维度 3）和中位数（维度 4），这些统计量可有效表征数据包长的分布特性。

方向差异特征：采用分位点分割法计算前后方向方差（维度 5–8），该特征主要体现流量序列的时间稳定性与前后关联性。具体计算步骤为：

(1) 确定数据包长序列的第一三分位数 Q_1 和第二三分位数 Q_2 作为分割点；
(2) 在每个分位点处，将序列划分为前向子序列（包含该点之前的数据）和后向子序列（包含该点之后的数据），共四个子序列；
(3) 分别计算四个子序列的方差作为特征值。

分布密度特征：通过十等分数据包分组并统计每个等分的包长分布（维度 9–18），该特征用于表征更细粒度的语音频带。令 l_{max},l_{min} 分别表示数据包分组包长序列的最大和最小值。

(1) 将 $[l_{min},l_{max}]$ 区间十等分，产生九个内部分界点 $\{l_1,l_2,\dots,l_9\}$；
(2) 计算每个子区间 $[l_k,l_{k+1}]$ 内数据包数量占总数据包数量的比例；

(3) 取前九个区间的比例值作为特征（第十个区间可通过前九个推算）。

流量强度特征：记录总数据包数量（维度 19）和单位时间数据包数量（维度 20），分别表征通信量和传输速率。

每个数据包分组提取的四类统计特征如表 1 所示，最后将所有分组的表征拼接起来得到整个 VoIP 加密流量的表征。

表 1 每个数据包分组包含的统计特征

维度	特征	说明
1 ~ 4	基础统计量	方差、标准差、均值、中位数
5 ~ 8	方向差异	基于三分位点的前/后向子序列方差
9 ~ 18	分布密度	十等分区间包长分布比例
19 ~ 20	流量强度	总包数、单位时间包数

4.3 VoIP 加密流量识别模型

得到 VoIP 加密流量的特征向量后，本文使用一个包含三层卷积神经网络的模型聚合各数据包分组的特征，并用于下游流量识别任务。图 8 给出了本文所使用的神经网络模型的基本架构和参数。模型共包含三层卷积和池化操作以及两层线性层。卷积层的激活函数使用 ReLU 函数，卷积核数量为 32（即通道数），每次激活后均使用池化步长 2 进行最大池化处理，最后通过两个线性层作为分类器。

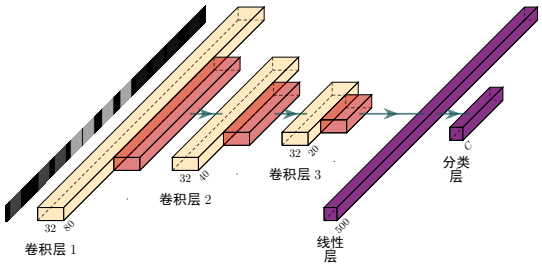


图 8 模型架构及参数（$K=4$）

VoIP 加密流量识别任务可以描述为一个多分类任务（见表 2 ），采用 Softmax 函数将网络输出映射为多类概率分布，对应的交叉熵损失函数可表示为：

$$L=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}y_i^{(c)}\log p_i^{(c)}$$

其中 N 表示样本总数，C 为类别总数，$y_i^{(c)}\in\{0,1\}$ 表示样本 i 在类别 c 上的真实标签，$p_i^{(c)}$ 表示模型预测样本 i 属于类别 c 的概率。

表 2 VoIP 加密流量识别任务

任务	类别集合	C	说明
性别识别	{男, 女}	2	识别说话人的性别
用户识别	{用户$_1$, …, 用户$_C$}	7	识别说话人
语种识别	{语言$_1$, …, 语言$_C$}	12	识别通话语言类型
短语识别	{短语$_1$, …, 短语$_C$}	6	识别短语

5 数据采集与实验结果分析

本节采集真实的 VoIP 加密流量数据，并进行流量识别实验，通过与基线方法对比，评估 VPrint 的性能表现。

5.1 VoIP 加密流量数据采集

由于缺少公开的 VoIP 加密流量数据集，本文在实验室环境中搭建了一个多场景 VoIP 加密流量采集和标注系统。系统主要包括两台运行 Windows 操作系统的个人电脑（PC），分别模拟通话的双方。每台 PC 均安装有待测试的 VoIP 应用、Wireshark 抓包工具以及虚拟声缆驱动 VB-CABLE[7]，可以将系统播放声直接作为麦克风输入，避免环境音干扰。两台 PC 都与互联网连接，并分别登录待测 VoIP 应用的两个不同账号。

为方便生成流量数据，系统使用预先录制的音频文件产生语音信号。采集流量时，其中一台 PC 发出语音通话请求，另一台 PC 接听。然后，发送方在本地播放语音音频文件，并通过 VB-CABLE 将音频信号输入发送方麦克风，开始语音通话。同时，接收方执行抓包程序，完成 VoIP 加密流量捕获。此外，通话双方可以同时播放语音文件，模拟语音交互，并且在双方本地同时抓包。以上采集过程可以通过脚本实现自动化控制，以便于进行大规模 VoIP 加密流量数据采集。

5.2 语音数据集

本文使用了以下三个公开及三个私有语音数据集作为 VoIP 应用的语音输入。

(1) **多场景语音数据集（Scenarios）**是一个公开数据集，包含多个主题的语音通话[31-32]。该数据集覆盖了多种真实的语音通话场景，使用语音剪切工具将用户对话分为两段独立语音，以模拟通话双方各自的语音输入。共包括 8 个场景（如图 9 所示），产生自 60 位男性和 50 位女性用户，共有 292 个语音样本，每个样本通话时长为 2 分钟。

[7]https://vb-audio.com/Cable

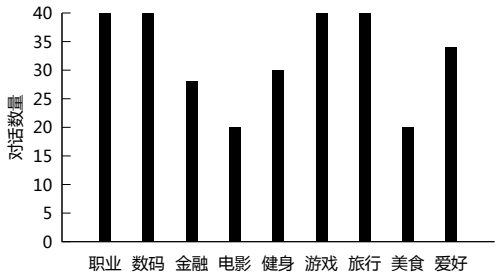


图 9　Scenarios 多场景语音数据集话题场景对话数量

(2) **问候语音数据集（Hello）**是一个公开数据集，由原始语音数据集 MAGICDATA Mandarin Chinese Read Speech Corpus[33]使用语音剪切工具剪切出多个简单短语“你好”的语音构成。数据集共包含 100 位男性用户产生的 200 个语音样本以及 100 位女性用户产生的 200 个语音样本。对每个语音样本重复 60 次“你好”，形成时长为 30 ~ 50 秒的语音信号。

(3) **唤醒词语音数据集（Hi-Mia）**是一个公开数据集，来源于开源数据集 AISHELL-WakeUp-1[34]。该数据集和 Hello 数据集类似，语音内容均为“你好，米亚”。该数据集共包含 131 位男性和 123 位女性通话用户，本文选取了其中 250 个男性语音样本和 250 个女性语音样本。每个语音样本重复 30 次“你好，米亚”，形成时长为 30 ~ 60 秒的语音信号。

(4) **多用户数据集（Speakers）**是一个私有数据集，由实验室真人志愿者采集和 AI 自动生成。该数据集包含 7 种通话用户身份，其中有 5 名真人志愿者（3 男 2 女）和 2 个 AI 智能语音（1 男 1 女）。每名用户采集 2 个语音样本（共 14 个语音样本），内容包括“你好”及“再见”，语音时长均为 1 秒。

(5) **多语种数据集（Languages）**是一个私有数据集，由 AI 生成的多语种语音数据。该数据集包含 12 种通话语言，每种语言生成 2 个语音样本，语音内容为简单短语“你好”和“再见”，语音时长均为 1 秒，共计 24 个语音样本。语种包括法语、德语、俄语、拉丁语、西班牙语、葡萄牙语、印地语、汉语普通话、孟加拉语、日语、英语以及阿拉伯语。

(6) **多短语数据集（Phrases）**是一个私有数据集，由 AI 生成的敏感短语语音样本。语音内容为 6 个敏感词汇，包括“转账”、“破坏”、“杀了”、“交易”、“洗钱”和“病毒”，每个词汇对应一个语音样本，共包含 6 个语音样本，语音时长均为 1 秒。

5.3　VoIP 加密流量数据集

将以上语音数据集作为流量采集系统的语音输入，共得到 24 个 VoIP 加密流量数据集。各流量数据集详细采集过程如下：

- 使用 Scenarios 语音数据集在不同应用中重复采集 3 次，得到流量数据集 {WeChat, TIM, WeMeet, DingDing}-Scenarios。
- 使用 Hello 语音数据集在不同应用中重复采集 3 次，得到流量数据集 {WeChat, TIM, WeMeet, DingDing}-Hello。
- 使用 Hi-Mia 语音数据集在不同应用中重复采集 4 次，得到数据集 {WeChat, TIM, WeMeet, DingDing}-Hi-Mia。
- 使用 Speakers 语音数据集，每个用户的语音重复播放 30 次，然后在不同应用中重复采集 170 次（每个用户有 170 个流量样本），得到流量数据集 {WeChat, TIM, WeMeet, DingDing}-Speakers。
- 使用 Phrases 语音数据集，每个语音重复播放 30 次，然后在不同应用中重复采集 120 次（每个短语有 120 个流量样本），得到流量数据集 {WeChat, TIM, WeMeet, DingDing}-Phrases。
- 使用 Languages 语音数据集，每个语音重复 30 次，然后在不同应用中重复采集 270 次（每种语言有 270 个流量样本），得到流量数据集 {WeChat, TIM, WeMeet, DingDing}-Languages。

5.4　实验设置

本文实验所用服务器包含 2 块 Intel(R) Xeon(R) Silver 4316 CPU 以及 1 块 NVIDIA Tesla V100 32GB GPU，操作系统为 Ubuntu 22.04 LTS。深度学习模型使用 PyTorch 2.4 构建。实验中使用的四种 VoIP 应用及版本号分别为：TIM 3.5.0、钉钉 7.6.15、微信 3.9.12 和腾讯会议 3.28.0（均为 Windows 版）。

表 3 给出了数据包分组算法（算法 1 ）在不同识别任务及流量数据集上的数据包分组结果，该分组的有效性将在第 5.6.5 小节的消融实验中进行验证。后续实验中在对应区间划分上提取各区间内流量序列特征，构建流量样本的特征向量。

5.5　基线方法与评估指标

本文共选取六个加密流量识别方法作为基线进行对比实验。

表 3 数据包分组结果（$K=4$）

任务	应用	数据包分组			
		G_1	G_2	G_3	G_4
性别识别	TIM	[40, 70]	[70, 100]	[100, 150]	[150, 200]
	微信	[54, 78]	[78, 112]	[112, 123]	[123, 200]
	钉钉	[48, 87]	[87, 117]	[117, 172]	[172, 200]
	腾讯会议	[28, 88]	[88, 114]	[114, 138]	[138, 200]
用户识别	TIM	[40, 70]	[70, 100]	[100, 150]	[150, 200]
	微信	[54, 78]	[78, 112]	[112, 123]	[123, 200]
语种识别	TIM	[52, 67]	[67, 102]	[102, 141]	[141, 200]
	微信	[55, 74]	[74, 108]	[108, 128]	[128, 200]
短语识别	TIM	[54, 78]	[78, 105]	[105, 145]	[145, 200]
	微信	[55, 78]	[78, 111]	[111, 148]	[148, 200]

- **HMM**[3]是早期基于语音建模的 VoIP 加密流量识别方法，直接使用语音音节（基本发音单元）的包长序列作为特征输入，构建 HMM 关联语音音节与包长度序列信息，从而实现 VoIP 加密流量的通话用户与通话内容识别。
- **VCF**[7] 是 Wang 等人于 2020 年针对智能音箱产生的 VoIP 加密流量设计的指纹提取方法，VCF 使用了 11 层卷积神经网络进行训练。
- **DeepFinger**[14]基于卷积神经网络构建特征提取器，利用数据包负载字节信息，基于堆叠自动编码器和一维卷积神经网络构建流量分类模型。
- **FS-Net**[28]是一种基于时间序列建模的通用加密流量识别框架，使用数据包长序列作为输入并基于 Bi-GRU 循环神经网络构建流量特征提取器，构建了网络流序列的时序表征。
- **ET-BERT**[35]基于 BERT 的通用网络流表征学习模型，利用数据包负载字节信息，适用于加密流量识别多领域下游任务。
- **YaTC**[36]是一种加密流量通用识别模型，将网络流转换为灰度图像，利用 Masked AutoEncoder (MAE) 框架[37]学习流量表征。

本文使用准确率、召回率、精确率以及 F1 值作为算法性能评估指标，各指标计算公式如下：

$$\text{准确率} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

$$\text{召回率} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{精确率} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{F1} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

其中 TP (True Positive) 和 TN (True Negative) 分别表示正确分类的正样本和负样本数量，FP (False Positive) 和 FN (False Negative) 分别表示错误分类的正样本和负样本数量。

5.6 实验结果与分析

5.6.1 性别识别

本节实验的目的是评估不同加密流量识别方法在区分通话用户性别属性任务上的性能表现。实验使用了三个语音数据集在 TIM 和微信应用中产生的 VoIP 加密流数据集：{TIM, Wechat}-Hello、{TIM, Wechat}-Hi-Mia 和 {TIM, Wechat}-Scenarios，实验结果分别如表 4 和表 5 所示。

可以看到，基于 VoIP 加密流量识别通话用户性别属性是比较难的任务，所有方法的识别性能指标均低于 0.8。相较于基线方法，本文所提的 VPrint 方法在所有数据集上的性能表现均呈现出显著的优势。VPrint 在性别识别任务上的 F1 值在 0.68 ~ 0.77 之间，相比于最佳基线方法提升了 5% ~ 10%，已经显著优于随机猜测，而其他基线方法只是略微优于随机猜测。此外，对比分析 TIM 和微信两种应用产生的 VoIP 加密流量，总的来说，可以看到基线方法在识别微信用户的性别属性性能表现稍差于识别 TIM 用户的性别属性，说明微信在防御已有方法的性别识别能力上稍强于 TIM。然而，本文方法 VPrint 在微信用户性别识别任务上的表现与已有基线方法不同，在三个语音数据集上的 F1 值表现反而优于 TIM，表明微信在防御本文方法进行性别识别的能力上并未增强。

为进一步验证 VPrint 的通话用户性别识别能力，表 6 给出了 VPrint 在 VoIP 加密流量数据集 {TIM, Wechat, Dingding, Wemeet}-Hi-Mia 上的性能表现。尽管四种应用采用了不同的语音编码算法，VPrint 性别识别的 F1 值均高于 0.68，表明 VPrint 具有较好的鲁棒性，不依赖于 VoIP 应用的语音编码算法。

本节实验表明，目前的 VoIP 应用可能泄露用户的性别属性，存在用户隐私泄露的风险。

5.6.2 用户识别

本节实验评估不同加密流量识别方法在区分通话用户身份任务上的性能表现，基于 {Wechat, TIM}-speakers 流量数据集。注意 Speakers 语音数据集共包含 7 名用户，其中 2 名为虚拟 AI 用户。以下主要分析 Wechat-speakers 流量数据集上的实验结果（如表 7 和表 8 所示），在 TIM-speakers 流量数据集上的实验结果类似，在此略去。

表 4　TIM 性别识别实验结果

方法	TIM-Hello				TIM-Hi-Mia				TIM-Scenarios			
	召回率	精确率	准确率	F1 值	召回率	精确率	准确率	F1 值	召回率	精确率	准确率	F1 值
HMM	0.5641	0.4783	0.5192	0.5177	0.5392	0.5225	0.5392	0.4235	0.5147	0.4368	0.4607	0.4667
VCF	0.5266	0.6817	0.6110	0.5980	0.6695	0.6861	0.6695	0.6418	0.5392	0.5640	0.5392	0.5513
DeepFinger	0.6739	0.5744	0.6739	0.5693	0.6840	0.7525	0.7119	0.7166	0.5723	0.5835	0.5850	0.5778
FS-Net	**0.6920**	0.5845	0.5985	0.6337	0.6342	0.6745	0.6589	0.6537	0.6022	0.6135	0.6080	0.6078
ET-BERT	0.5025	0.4936	0.5025	0.4980	0.5930	0.5924	0.5930	0.5927	0.6645	0.6559	0.6645	0.6602
YaTC	0.5625	0.5518	0.5625	0.5540	0.6449	0.6363	0.6449	0.6352	0.6675	0.6605	0.6675	0.6577
VPrint	0.6869	**0.6862**	**0.6869**	**0.6831**	**0.7542**	**0.7726**	**0.7610**	**0.7633**	**0.7436**	**0.7407**	**0.7436**	**0.7401**

表 5　微信性别识别实验结果

方法	Wechat-Hello				Wechat-Hi-Mia				Wechat-Scenarios			
	召回率	精确率	准确率	F1 值	召回率	精确率	准确率	F1 值	召回率	精确率	准确率	F1 值
HMM	0.4227	0.5543	0.5112	0.4796	0.4607	0.4678	0.4607	0.4601	0.3475	0.5557	0.3475	0.4276
VCF	0.6157	0.6239	0.6157	0.5287	0.6524	0.6620	0.6524	0.5927	0.6695	0.6298	0.6695	0.5900
DeepFinger	0.6423	0.6380	0.6423	0.6396	0.6931	0.6494	0.6494	0.6705	0.6737	0.6399	0.6737	0.6023
FS-Net	0.5014	0.5151	0.5014	0.5072	0.5872	0.5717	0.5872	0.5564	0.6228	0.5407	0.6228	0.5704
ET-BERT	0.5224	0.5253	0.5224	0.5238	0.5618	0.5633	0.5619	0.5625	0.5932	0.5846	0.5932	0.5887
YaTC	0.5398	0.5306	0.5398	0.5327	0.5625	0.5518	0.5625	0.5540	0.6314	0.5959	0.6314	0.6066
VPrint	**0.7628**	**0.7046**	**0.7235**	**0.7325**	**0.7787**	**0.6909**	**0.7320**	**0.7322**	**0.7754**	**0.7696**	**0.7754**	**0.7707**

表 6　VPrint 在 {TIM, Wechat, Dingding, Wemeet}-Hi-Mia 流量数据集进行用户性别属性识别的实验结果

VoIP 应用	召回率	精确率	准确率	F1 值
TIM	0.7542	0.7726	0.7610	0.7633
微信	0.7787	0.6909	0.7320	0.7322
钉钉	0.6942	0.6935	0.6940	0.6938
腾讯会议	0.6820	0.6834	0.6820	0.6827

表 7　不同算法在 Wechat-Speakers 流量数据集上的用户识别实验结果

方法	召回率	精确率	准确率	F1 值
HMM	0.5423	0.5344	0.5423	0.5383
VCF	0.8819	0.8890	0.8819	0.8816
DeepFinger	0.9444	0.9505	0.9444	0.9436
FS-Net	0.8024	0.8200	0.8025	0.8028
ET-BERT	0.2872	0.2474	0.2872	0.2320
YaTC	0.4979	0.5097	0.5021	0.4976
VPrint	**0.9931**	**0.9934**	**0.9931**	**0.9931**

表 7 给出了不同方法的通话用户身份识别结果，可以看到 VPrint 在所有指标上均优于已有方法，用户识别 F1 值相较于最优基线方法提升 5%，较早期的 VoIP 加密流量识别方法 HMM 提升 84%。这种性能提升源于 VPrint 融合利用了语音频谱与数据包长信息以及对数据包分组，优于传统方法仅单一利用包长或负载字节信息。表 8 进一步分析了使用 VPrint 方法进行通话人识别的结果，可以看到 VPrint 能准确识别通话人。

表 8　VPrint 在 Wechat-Speakers 流量数据集上的用户识别实验结果

通话用户	召回率	精确率	准确率	F1 值
Male_u1	0.9990	0.9990	0.9990	0.9990
Male_u2	0.9990	0.9990	0.9990	0.9990
Male_u3	0.9990	0.9990	0.9990	0.9990
Male_AI	0.9990	0.9565	0.9880	0.9778
Female_u4	0.9667	1.0000	0.9800	0.9831
Female_u5	0.9990	1.0000	0.9990	0.9995
Female_AI	0.9840	1.0000	0.9990	0.9919

本节实验表明，目前的 VoIP 应用可能泄露用户的身份信息，而且只要能获得足够多目标用户的样本，那么现有基于深度学习的方法（例如 VCF、DeepFinger、FS-Net、VPrint 等）就可以准确识别目标用户，存在用户隐私泄露的风险。

5.6.3　语种识别

本节实验评估不同加密流量识别方法在区分通话语种类型任务上的性能表现，实验基于 {Wechat, TIM}-Languages 流量数据集，共包括 12 种语种。以下主要分析 Wechat-Languages 流量数据

集上的实验结果（如表 9 和表 10 所示），在 TIM-Languages 流量数据集上的实验结果与之类似，在此略去。

表 9 不同算法在 Wechat-Languages 流量数据集上的通话语种识别结果

方法	召回率	精确率	准确率	F1 值
HMM	0.2112	0.3052	0.2110	0.2496
VCF	0.4668	0.4659	0.4668	0.4522
DeepFinger	0.5150	0.5260	0.5150	0.5021
FS-Net	0.1839	0.1551	0.1839	0.1568
ET-BERT	0.4337	0.4750	0.4337	0.4297
YaTC	0.3390	0.3564	0.3390	0.3418
VPrint	**0.8802**	**0.8871**	**0.8802**	**0.8811**

表 10 VPrint 在 Wechat-Languages 流量数据集上的通话语种识别结果

通话语种	召回率	精确率	准确率	F1 值
法语	0.8333	0.8333	0.8333	0.8333
德语	0.9167	0.9167	0.9167	0.9167
俄语	0.9090	0.9090	0.9090	0.9090
拉丁语	0.9474	0.9990	0.9474	0.9730
西班牙语	0.8333	0.7692	0.8333	0.8000
葡萄牙语	0.8947	0.8500	0.8930	0.8718
印地语	0.9980	0.7500	0.8500	0.8571
汉语普通话	0.8462	0.9990	0.8490	0.9167
孟加拉语	0.8462	0.7857	0.8592	0.8148
日语	0.9167	0.9990	0.9225	0.9565
英语	0.7619	0.8889	0.7620	0.8205
阿拉伯语	0.9286	0.8667	0.9240	0.8966

表 9 的结果表明 VPrint 方法在语种识别任务中比基线方法展现出显著优势，F1 值达到 0.88，较次优方法 DeepFinger 提升 76%，较传统 HMM 方法提升了 2.5 倍，表明 VPrint 融合语音频率特征和数据包长特征的有效性。表 10 进一步给出 VPrint 在 12 种语言上识别结果性能指标。VPrint 在德语、俄语、拉丁语、日语、汉语普通话等语种的识别 F1 值均高于 0.9，而在某些语种（例如西班牙语）的识别 F1 值为 0.8。这种对不同语种的识别差异可能源于不同语言的频谱特性差异，如何针对特定语种设计更好的识别方法可以作为未来研究工作。

本节实验表明，目前的 VoIP 应用可能泄露通话语种信息，存在用户隐私泄露的风险。

5.6.4 短语识别

本节实验评估不同加密流量识别方法在识别通话短语任务上的性能表现，实验基于 {Wechat, TIM}-Phrases 流量数据集，共包括 6 类短语。以下主要分析 Wechat-Phrases 流量数据集上的实验结果，如表 11 和表 12 所示。在 TIM-Phrases 流量数据集上的实验结果与之类似，在此略去。

表 11 不同算法在 Wechat-Phrases 流量数据集上的通话短语识别结果

方法	召回率	精确度	准确率	F1 值
HMM	0.2510	0.4523	0.2510	0.3228
VCF	0.5610	0.5667	0.5610	0.5388
DeepFinger	0.5366	0.5608	0.5366	0.5437
FS-Net	0.2414	0.2375	0.2414	0.2344
ET-BERT	0.6232	0.6861	0.6232	0.6303
YaTC	0.4301	0.4402	0.4301	0.4228
VPrint	**0.9180**	**0.9294**	**0.9180**	**0.9159**

表 12 VPrint 在 Wechat-Phrases 流量数据集上的通话短语识别结果

通话短语	召回率	精确度	准确率	F1 值
转账	0.6250	1.0000	0.7640	0.7692
破坏	0.9286	0.8125	0.9230	0.8667
杀了	0.9615	0.8065	0.9540	0.8772
交易	0.9524	1.0000	0.9540	0.9756
洗钱	0.9583	1.0000	0.9620	0.9787
病毒	1.0000	0.9546	0.9840	0.9767

表 11 表明 VPrint 在短语识别任务中的性能表现远优于已有基线方法，识别 F1 值达 0.92，较次优基线方法提升 46%，较早期的 VoIP 加密流量识别方法提升 1.9 倍，表明 VPrint 融合语音频率特征和数据包长特征的有效性。表 12 进一步给出了 VPrint 在 6 种短语识别任务上的性能表现。VPrint 在某些短语识别任务（例如“洗钱”等）上的 F1 值高于 0.97，而在某些短语识别任务（例如“转账”）上的 F1 值为 0.77 左右，这种对不同短语的识别差异可能源于不同短语发音的频谱特性差异（例如发音以爆破音为主与摩擦音为主的短语在频谱特性上差异较大），如何针对特定短语设计更好的识别方法可以作为未来研究工作。

本节实验表明，目前的 VoIP 应用可能泄露敏感的通话内容信息，存在用户隐私泄露的风险。

5.6.5 消融实验

本节实验目标是通过消融实验验证数据包分组方法（算法 1 ）的有效性。为此，消融实验考虑了几种不同设置：不采用数据包分组、随机数据包分组，以及使用不同数量的数据包分组等。在选择

表 13　TIM-Hi-Mia 数据集性别识别消融实验

数据包分组	召回率	精确率	准确率	F1 值
不分组	0.5980	0.3577	0.5980	0.4476
随机分组	0.6373	0.6277	0.6373	0.6015
$G_4\_G_3$	0.6823	0.6742	0.6823	0.6782
$G_4\_G_2$	0.7231	0.7142	0.7230	0.7186
$G_4\_G_1$	0.7126	0.7535	0.7336	0.7325
$G_4\_G_3\_G_2$	0.7485	0.7522	0.7490	0.7503
$G_4\_G_3\_G_1$	0.7120	0.7120	0.7120	0.7120
$G_4\_G_2\_G_1$	0.7638	0.7442	0.7640	0.7539
$G_4\_G_3\_G_2\_G_1$	0.7542	0.7726	0.7610	0.7633

表 14　TIM-Hello 数据集性别识别消融实验

数据包分组	召回率	精确率	准确率	F1 值
不分组	0.5714	0.3000	0.5714	0.3934
随机分组	0.6087	0.6020	0.6087	0.5904
$G_4\_G_3$	0.6081	0.7541	0.6081	0.6732
$G_4\_G_2$	0.6260	0.6650	0.6260	0.6449
$G_4\_G_1$	0.6082	0.6384	0.6082	0.6229
$G_4\_G_3\_G_2$	0.8603	0.5912	0.6147	0.7008
$G_4\_G_3\_G_1$	0.7081	0.6305	0.6299	0.6671
$G_4\_G_2\_G_1$	0.6610	0.6692	0.6514	0.6651
$G_4\_G_3\_G_2\_G_1$	0.6869	0.6862	0.6869	0.6831

表 15　TIM-Scenarios 数据集性别识别消融实验

数据包分组	召回率	精确率	准确率	F1 值
不分组	0.5957	0.7368	0.5957	0.6588
随机分组	0.6596	0.6596	0.6596	0.6596
$G_4\_G_3$	0.6451	0.6854	0.6450	0.6646
$G_4\_G_2$	0.6842	0.6930	0.6842	0.6886
$G_4\_G_1$	0.7010	0.6995	0.6956	0.7002
$G_4\_G_3\_G_2$	0.7023	0.7054	0.7024	0.7038
$G_4\_G_3\_G_1$	0.7052	0.7173	0.7057	0.7112
$G_4\_G_2\_G_1$	0.7234	0.7303	0.7235	0.7268
$G_4\_G_3\_G_2\_G_1$	0.7436	0.7407	0.7436	0.7401

不同的数据包分组时，由于 G_4 分组包含了大数据包的流量数据，与语音中的高频频段相关，分组组合均保留 G_4 分组进行，依次对比不同分组与 G_4 分组的组合特征效果。表 13 、表 14 、表 15 分别给出了不分组、随机分组以及分组 G_1，G_2，G_3 分别与 G_4 组合后，在 TIM-{Hi-Mia、Hello、Scenarios} 流量数据集上的性别分类结果。

实验结果表明，随着更多分组的使用，分类评估指标整体呈上升趋势，其中使用所有分组的评估指标均高于其他分组组合，说明完整的分组有助于分类器识别 VoIP 流量中的语音信息。算法 1 得到的分组实验结果优于随机分组和不分组的实验结果。此外，随机分组的实验结果相比于不分组的实验结果也存在明显优势。以上消融实验结果证明了本文数据包分组的合理性和有效性。

6　总结

VoIP 应用使用日益广泛，其安全性受到人们越来越多的重视。本文通过在实验室环境中采集四种 VoIP 应用产生的语音流量数据，系统测量分析了 VoIP 加密流量的传输模式与用户属性、语音内容等方面之间的关联关系，并提出了一种语音频谱与数据包长对齐的 VoIP 加密流量识别方法——VPrint。本文在真实数据上系统评估了 VPrint 在四种 VoIP 加密流量识别任务上的表现，包括性别识别、用户识别、语种识别和短语识别。实验结果表明，VPrint 较已有的加密流量识别方法能更准确地识别 VoIP 加密流量。

本文的研究同时揭示了目前流行 VoIP 应用存在的安全隐患，有可能泄露用户属性、用户身份以及通话内容等用户隐私信息。尽管目前学术界已经提出了多种增强网络通信安全性的技术，例如数据包填充、数据包延时发送等，但是本文测试的四种 VoIP 应用中并没有发现使用这些防御技术，因此暴露出安全隐患。产生该结果的一种可能原因是这些 VoIP 应用提供厂商往往更关心用户的使用体验（例如通话质量、延时等），而忽视了对数据安全的要求。

本文的研究结果有助于理解并改进 VoIP 应用的安全性以及用于打击利用 VoIP 应用进行电信诈骗及相关黑灰产业。未来工作可以进一步改进 VPrint 识别准确率，例如采用 Transformer 等模型架构，以及在开放世界环境中测试评估 VoIP 应用的安全性。

参考文献

[1] 2025 年中国 VoIP 网络电话机行业全景调研及未来趋势分析报告[R]. 中国市场调研网, 2025.

[2] WRIGHT C V, BALLARD L, MONROSE F, et al. Language identification of encrypted VoIP traffic: Alejandra y Roberto or Alice and Bob? [C]//USENIX Security. 2007.

[3] WRIGHT C V, BALLARD L, COULL S E, et al. Spot me if you can: Uncovering spoken phrases in encrypted VoIP conversations[C]//IEEE SP. 2008.

[4] The role of VoIP encryption in securing communications[EB/OL]. Ce-

bod Telecom, 2025. https://www.cebodtelecom.com/voip-encryption-secure-communications.

[5] NASR M, BAHRAMALI A, HOUMANSADR A. Defeating DNN-Based traffic analysis systems in Real-Time with blind adversarial perturbations[C]//USENIX Security. 2021.

[6] VAIDYA T, WALSH T, SHERR M. Whisper: A unilateral defense against VoIP traffic re-identification attacks[C]//Proceedings of the 35th Annual Computer Security Applications Conference. 2019.

[7] WANG C, KENNEDY S, LI H, et al. Fingerprinting encrypted voice traffic on smart speakers with deep learning[C]//Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks. 2020.

[8] 新型 VoIP 设备电诈：一根数据线让境外电话变本地，有人被骗 24 万余元[EB/OL]. 红星新闻, 2025. https://news.ifeng.com/c/8g1wLLbqdv0.

[9] SHAPIRA T, SHAVITT Y. Flowpic: A generic representation for encrypted traffic classification and applications identification[J]. IEEE Transactions on Network and Service Management, 2021, 18(2): 1218-1232.

[10] WRIGHT C V, COULL S E, MONROSE F. Traffic Morphing: An effincient defense against statistical traffic analysis[C]//NDSS. 2009.

[11] DING L, YUEFEI Z, BIN L, et al. Survey of side channel attack on encrypted network traffic[J]. Chinese Journal of Network & Information Security, 2021, 7(4): 114-130.

[12] WHITE A M, MATTHEWS A R, SNOW K Z, et al. Phonotactic reconstruction of encrypted VoIP conversations: Hookt on fon-iks[C]//IEEE SP. 2011.

[13] SIKOS L F. Packet analysis for network forensics: A comprehensive survey[J]. Forensic Science International: Digital Investigation, 2020, 32(1): 1-12.

[14] SIRINAM P, IMANI M, JUAREZ M, et al. Deep fingerprinting: Undermining website fingerprinting defenses with deep learning[C]//ACM CCS. 2018.

[15] WANG S, YU B, WU M. MVCM car-following model for connected vehicles and simulation-based traffic analysis in mixed traffic flow[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(6): 5267-5274.

[16] LOTFOLLAHI M, JAFARI SIAVOSHANI M, SHIRALI HOSSEIN ZADE R, et al. Deep packet: A novel approach for encrypted traffic classification using deep learning[J]. Soft Computing, 2020, 24 (3): 1999-2012.

[17] WU D, WANG X, QIAO Y, et al. NetLLM: Adapting large language models for networking[C]//ACM SIGCOMM. 2024.

[18] 吴桦, 倪珊珊, 罗浩, 等. 一种基于 HTTP/3 传输特性的加密视频识别方法[J]. 计算机学报, 2024, 47(7): 1640-1664.

[19] PHAM T D, HO T L, TRUONG-HUU T, et al. Mappgraph: Mobile-app classification on encrypted network traffic using deep graph convolution neural networks[C]//Proceedings of the 37th Annual Computer Security Applications Conference. 2021.

[20] JIANG M, LI Z, FU P, et al. Accurate mobile-app fingerprinting using flow-level relationship with graph neural networks[J]. Computer Networks, 2022, 217: 1-12.

[21] XU H, LI S, CHENG Z, et al. VT-GAT: A novel VPN encrypted traffic classification model based on graph attention neural network[C]//International Conference on Collaborative Computing: Networking, Applications and Worksharing. 2022.

[22] CHOUDHURY P, KUMAR K P, NANDI S, et al. An empirical approach towards characterization of encrypted and unencrypted VoIP traffic[J]. Multimedia Tools and Applications, 2020, 79(1): 603-631.

[23] DONG S, XIA Y, PENG T. Network abnormal traffic detection model based on semi-supervised deep reinforcement learning[J]. IEEE Transactions on Network and Service Management, 2021, 18(4): 4197-4212.

[24] LIU J, FU Y, MING J, et al. Effective and real-time in-app activity analysis in encrypted internet traffic streams[C]//ACM SIGKDD. 2017.

[25] FU Y, XIONG H, LU X, et al. Service usage classification with encrypted internet traffic in mobile messaging apps[J]. IEEE Transactions on Mobile Computing, 2016, 15(11): 2851-2864.

[26] FU Y, LIU J, LI X, et al. Service usage analysis in mobile messaging apps: A multi-label multi-view perspective[C]//IEEE ICDM. 2016.

[27] KHAN L A, BAIG M S, YOUSSEF A M. Speaker recognition from encrypted VoIP communications[J]. Digital Investigation, 2010, 7(1): 65-73.

[28] LIU C, HE L, XIONG G, et al. FS-Net: A flow sequence network for encrypted traffic classification[C]//Proceedings of the IEEE Conference On Computer Communications. 2019.

[29] G N B, ANEES M, G T Y. Speech coding techniques and challenges: a comprehensive literature survey[J]. Multimedia Tools and Applications, 2024, 83: 29859-29879.

[30] Adaptive multi-rate audio codec[EB/OL]. 2025. https://en.wikipedia.org/wiki/Adaptive_Multi-Rate_audio_codec.

[31] Asr-ctelecsc[EB/OL]. 2025. https://magichub.com/datasets/mandarin-chinese-conversational-speech-corpus-telephony/.

[32] Asr-multideviccsc[EB/OL]. 2025. https://magichub.com/cn/datasets/mandarin-chinese-conversational-speech-corpus-mutiple-devices/.

[33] MAGICDATA mandarin chinese read speech corpus[EB/OL]. 2025. https://www.openslr.org/68.

[34] AISHELL-WakeUp-1 中英文唤醒词语音数据库[EB/OL]. 2025. https://www.aishelltech.com/wakeup_data.

[35] LIN X, XIONG G, GOU G, et al. ET-BERT: A contextualized datagram representation with pre-training transformers for encrypted traffic classification[C]//Proceedings of the ACM Web Conference. 2022.

[36] ZHAO R, ZHAN M, DENG X, et al. Yet another traffic classifier: A masked autoencoder based traffic transformer with multi-level flow representation[C]//AAAI. 2023.

[37] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]//CVPR. 2022.