基于语音频带与数据包长对齐的VoIP加密网络流量识别方法

赵俊舟 段 涛 李江龙 王平辉 陶 敬
（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应用在方便人们沟通联络的同时，也滋长了电信网络诈骗等网络违法犯罪活动，亟须研究VoIP加密网络流量识别技术，以检测和打击利用VoIP应用进行的网络黑灰产业。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加密网络流量，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用户语音通话安全，但是VoIP加密网络流量的传输模式仍可以被用于推断用户属性、用户身份，甚至通话内容等敏感信息，可以被用来识别VoIP加密网络流量并用于反诈治理。本文通过测量分析四种VoIP应用产生的加密网络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带与数据包长对齐的VoIP加密网络流量识别方法——VPrint。VPrint较已有的加密网络流量识别方法能更准确识别VoIP加密网络流量。以微信为例，VPrint在用户性别识别、用户身份识别、通话语种和短语识别任务上的$F1$值分别为0.77、0.99、0.88和0.92，较基线方法提升5%～76%。

关键词 互联网语音应用；加密网络流量识别；网络诈骗；反诈治理；数据安全
中图法分类号 TP309 **DOI号** 10.11897/SP.J.1016.2026.01227

Encrypted VoIP Network Traffic Recognition via Aligning Voice Spectra and Packet Length

ZHAO Jun-Zhou DUAN Tao LI Jiang-Long WANG Ping-Hui TAO Jing
（MOE KLINNS Lab, Xi'an Jiaotong University, Xi'an 710049）

Abstract With the rapid popularization of mobile devices and fast development of wireless networking technology, Voice over Internet Protocol (VoIP) applications represented by WeChat and Skype have become increasingly popular. VoIP applications are built on the Internet and transmit voice signals through IP packet-switched networks. During a voice call, the analog voice signal generated by the sender is compressed and encoded, and then packaged into data packets according to protocols such as TCP/IP, which are transmitted over the IP network to the destination. The receiver then reassembles and decodes the received data packets to restore the original voice signal, thus achieving voice communication over the Internet. While VoIP applications have greatly enhanced communication convenience in people's daily lives, they have also facilitated the proliferation of cybercrimes-particularly telecommunications fraud and online fraud-inflicting substantial harm on individuals. There is an urgent need to study VoIP encrypted network traffic identification technologies in order to efficiently detect and combat the cybercrimes

收稿日期：2025-06-26；在线发布日期：2026-03-06。本课题得到国家自然科学基金面上项目（No. 62272372）资助。**赵俊舟**，博士，副教授，主要研究领域为网络安全。E-mail：junzhou.zhao@xjtu.edu.cn。**段 涛**，博士研究生，主要研究领域为网络流量分析。**李江龙**，硕士研究生，主要研究领域为网络流量分析。**王平辉**，博士，教授，主要研究领域为网络安全。**陶 敬**（通信作者），硕士，研究员，主要研究领域为网络安全。E-mail：jtao@xjtu.edu.cn。

that exploit VoIP applications. To this end, this study collects and analyzes real-world VoIP network traffic generated by four popular VoIP applications in China, including WeChat, TIM, Tencent Meeting, and DingTalk. Although nowadays VoIP applications generally adopt private voice coding algorithms and encrypted communication to enhance their data security, the transmission patterns of VoIP encrypted network traffic traces may still leak side-channel information about user profiles, identities, and even private voice content, which can be possibly leveraged by networking administrators to recognize VoIP encrypted network traffic. Specifically, this work measures and analyzes the correlation between the encrypted network traffic transmission patterns of the four VoIP applications against user attributes (e. g., speaker's gender), speaking languages (e. g., Putonghua, English, etc.), user identities (e. g., a particular person of interest), and voice content (e. g., words relating to killing people, transfer money, etc.), and discovers a significant correlation between voice frequency and packet length. Specifically, we find that VoIP applications tend to transmit high-frequency voice signals using larger network packets, i. e., packets with larger bytes. Based on this observation, a VoIP encrypted network traffic identification method called VPrint is proposed for fingerprinting VoIP network traffic. VPrint mainly consists of two steps. Firstly, VPrint learns to align voice spectra with packet lengths in order to find the correlation between voice frequency and packet size. Secondly, VPrint extracts statistical packet features regarding to different groups of packets that are likely to correspond to different voice frequency bands such as low, medium, and high frequencies. This approach of feature extraction is proven to be more accurate and robust than existing encryption network traffic identification methods such as FS-Net, YaTC, and ET-BERT. 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 speaking language identification, and 0.92 for speaking phrase identification, which are improved by 5% to 76% compared with existing encrypted network traffic identification methods. We also show the robustness of VPrint against network jitter such as packets loss, disordering, and padding, showing its usefulness in detecting VoIP related cyber frauds.

Keywords voice over Internet protocol applications; encrypted network traffic identification; cyber fraud; anti-fraud governance; 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。截至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应用在方便人们联络沟通的同时，也带来严峻的网络犯罪问题。自2021年以来，国家反诈中心累计拦截诈骗电话69.9亿次，拦截涉案资金1.1万亿元[2]。VoIP应用在网络诈骗中扮演了越来越重要的角色，境内外违法犯罪分子利用VoIP应用进行电信网络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[3]，亟需针对VoIP黑灰产业进行监管与治理。因此有必要研究针对VoIP应用的电信网络诈骗检测技术，包括识别涉诈

VoIP加密网络流量、对涉诈人员进行身份画像（例如确定其性别、年龄等属性）等，为此，需要系统研究VoIP加密网络流量识别技术。

VoIP加密网络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[4-6]。而VoIP加密网络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[7-9]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），通过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，并不对外公开，因此语音编码算法对于流量分析者来说属于黑盒。此外，VoIP应用可能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，以抵抗第三方流量分析[10-12]。这些新挑战容易使White等人早期提出的VoIP加密流量识别方法变得不再可行，从而难以监管VoIP应用，因此有必要提出新的VoIP加密流量识别方法。

为了提出可行的VoIP加密流量识别方法，本文在实验室环境中采集了微信①、TIM②、钉钉③、腾讯会议④四款流行VoIP应用产生的加密网络流量数据。测量发现，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带与数据包长的VoIP加密流量识别方法——VPrint。VPrint较已有的加密网络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。此外，VPrint只利用VoIP加密流量分组包长特征建模用户通话模式，克服了已有VoIP识别方法对语音编码算法参数的先验依赖。因此，VPrint同时适用于采用公开语音编码算法和采用私有语音编码算法的不同VoIP应用，使其在实际VoIP监管中具备更好的通用性。

本文主要贡献总结如下：

（1）本文系统测量分析了四款流行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。

（2）本文提出一种通过对齐语音频带与数据包长的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。

（3）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值较基线方法提升5%～76%。

本文的章节安排如下：第2节介绍VoIP加密流量识别技术的研究现状；第3节对实验室采集到的真实VoIP加密流量进行测量分析，发现VoIP加密流量的特有传输模式；第4节基于前一节的测量分析结果，设计一种新的VoIP加密流量识别方法——VPrint；第5节通过大量实验验证VPrint识别VoIP加密流量的有效性；第6节总结全文并给出未来研究方向。

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网络流量，然而随着加密技术的普及，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[6,14-17]。

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

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

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[9]、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[7]、敏感短语识别[8])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。

(1)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[12]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流量分类,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。

(2)基于语音建模的流量分析方法。Wright等人[7-9]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[7]、短语识别[8-9]。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加密流量识别技术在当前网络安全监管应用中主要面临以下挑战:(1)粗粒度统计流量特征无法建模通话语音内容与VoIP流量模式之间的关联;(2)细粒度音节流量特征依赖语言编码器参数,无法应对采用私有语音编码算法的当前VoIP应用。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。为了深入理解语音数据包与模拟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流行应用产生的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①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,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)。

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5.1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

① 语音内容为“你好”并重复30秒。

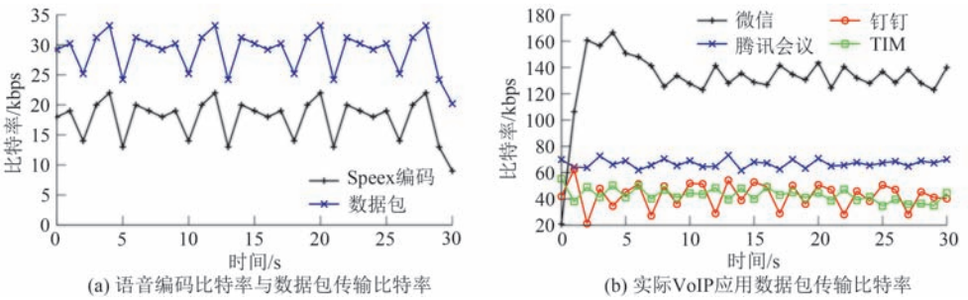

图1　语音编码及数据包传输比特率

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，编码后的语音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3.4 kHz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7 kHz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261.6 Hz、293.6 Hz、329.6 Hz、349.2 Hz、392 Hz、440 Hz、493.8 Hz。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。

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，在此略去。

根据本节的测量分析，可以得出以下结论：

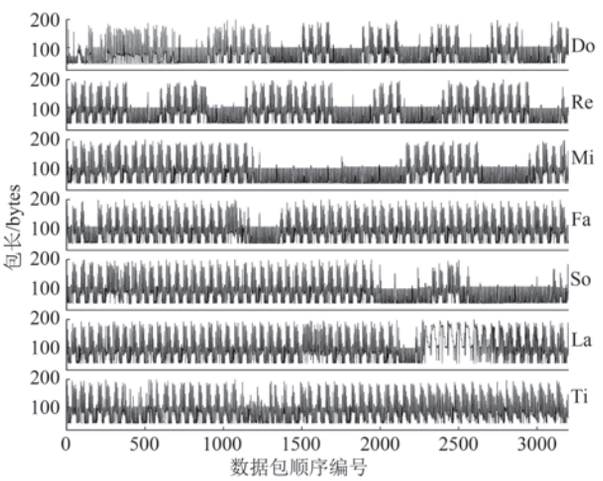

图2　七个音阶产生的VoIP加密流量包长分布

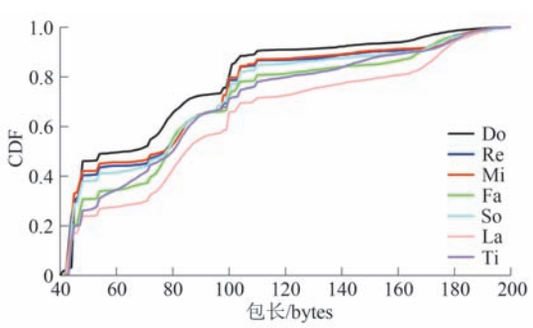

图3　七个音阶产生的VoIP加密流量包长分布

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加密流量传输模式？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。

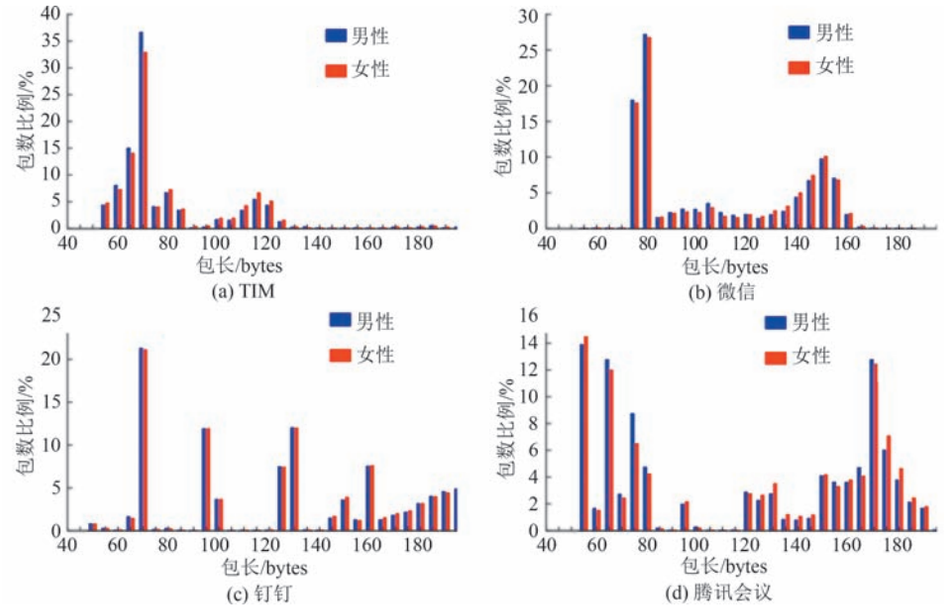

图4 男女VoIP加密流量数据包包长分布

从图4首先可以观察到，来自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，这与中的测量结果一致。其次，对比同一种应用中不同流量样本的整体包长分布，发现男女流量包长分布是高度相似的。这表明，直接在原始包长空间提取流量统计特征作来差异化VoIP用户属性是不可靠的。与之相反，对比同一种应用中男女流量样本在不同包长区间的分布情况，发现存在一种比较稳定的现象：在包长较小的区间（例如包长小于100字节），男性的包数比例往往略高于女性；而在包长较大的区间（例如包长大于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。

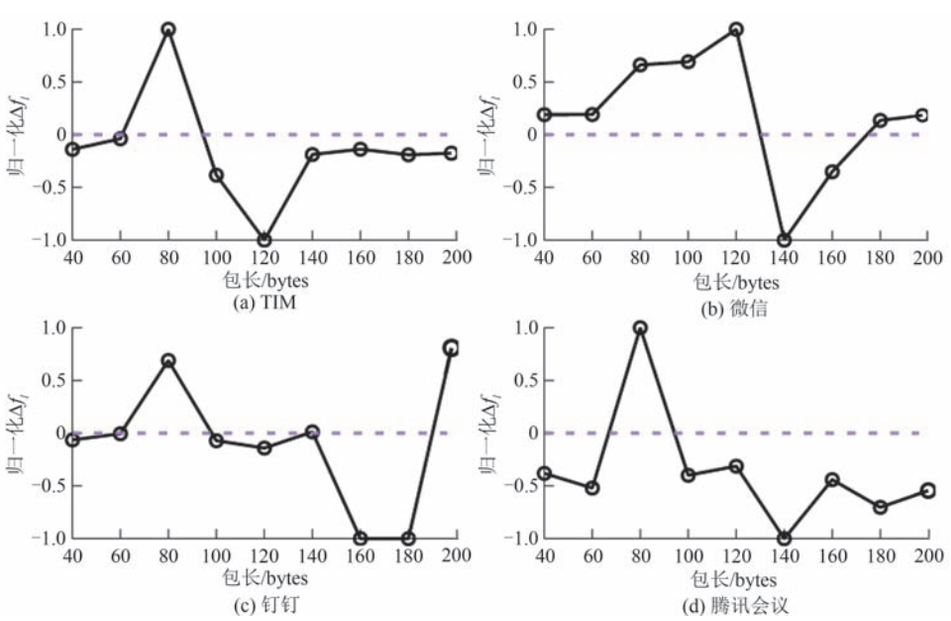

图5 归一化男女数据包比例差异

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3.2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、VoIP加密流量包长分布以及包长分布随时间的变化[①]。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。

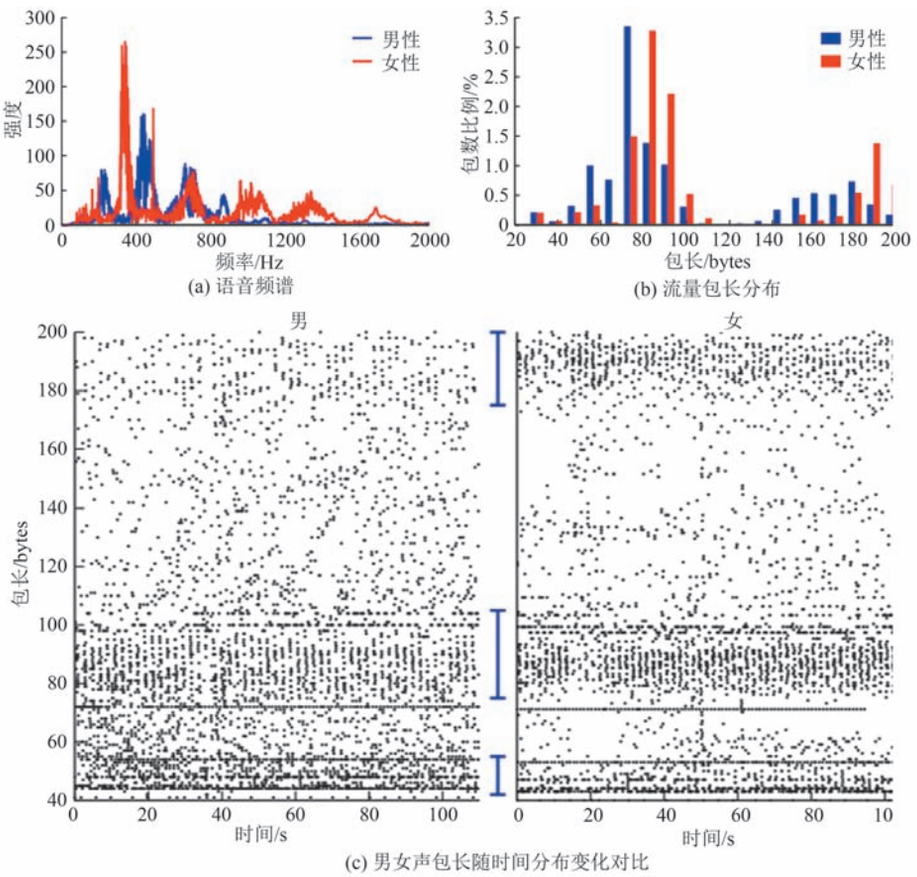

图6　男女语音频谱及VoIP加密流量包长分布

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380 Hz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。对比图6(c)也可以清楚地观察到：(1)数据包集中在几个主要的包长区间中，并且包长区间的边界与原始音频的频谱分布有关。(2)在不同包长区间内，男女性数据包的差异显著，小包长区间中男性数据包分布更密集，而大包长区间中女性数据包分布更密集。这些发现表明，利用音频频谱与VoIP流量包长对应关系，分离并提取不同包长区间数据包能有效差异化来自不同性别属性的流量。

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区间流量模式，男性数据包在小包长区间更密集，而女性数据包在大包长区间更密集。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加密网络流量识别算法(VoIP Encrypted Network 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。

① 语音内容同前，VoIP加密流量采集自微信。

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等重要的VoIP安全监管任务。VPrint算法整体框架如图7所示。

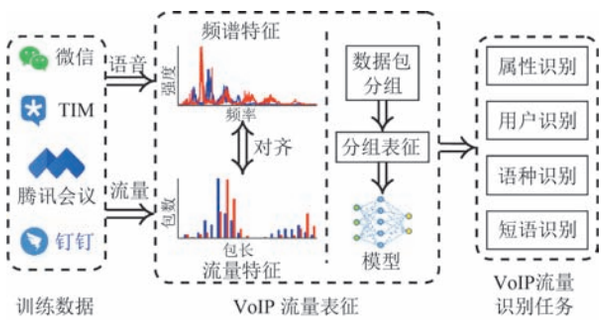

图7 VPrint整体框架图

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。

需要注意的是，VPrint的设计考虑了对VoIP网络监管方（例如，公安、司法等机关）的实际可用性，并通过以下约束避免了恶意攻击者对该技术的滥用。一方面，VPrint以VoIP流量为识别目标，需要部署方预先从开放世界准确提取VoIP流量；另一方面，VPrint的训练同时依赖VoIP流量和与其对应的原始音频，需要部署方预先获得目标用户的语音数据。实际应用中，网络监管方可以联合VPrint服务提供商和执法部门准确获得上述信息，从而实现对VPrint的安全部署。

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严格的一对一关系，这给数据包分组带来困难。为解决该问题，本节提出一种语音频带和数据包长对齐算法，实现按包长进行数据包分组，其伪代码如算法1所示。

算法1. 语音频带与数据包长对齐算法

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$G_1, \cdots, G_K$;
2. REPEAT
3. $U \leftarrow \Phi$;
4. FOR $i \leftarrow 1$ to $|D|$ DO
5. $(l_i^{(1)}, \cdots, l_i^{(K)}) \leftarrow \text{PeakFinder}(T_i, \{G_k\})$;
6. $(f_i^{(1)}, \cdots, f_i^{(K)}) \leftarrow \text{TopFFTFreqs}(V_i)$;
7. 构建频率和包长关联关系$\left\{(f_i^{(k)}, V_i^{(k)})\right\}_{k=1}^{K}$;
8. FOREACH $pkt \in T_i$ DO
9. $k^* \leftarrow \text{argmin}_k \left| pkt.len - l_i^{(k)} \right|$;

10. $U \leftarrow U \cup \left\{ \left(f_i^{(k^*)}, pkt.len \right) \right\}$;
11. END FOREACH
12. END FOR
13. $C \leftarrow \text{KMeansClustering}(U, K)$;
14. $(G_1, \cdots, G_K) \leftarrow \text{Update}(C)$;
15. UNTIL 数据包分组$\{G_k\}$稳定
16. 返回数据包分组$\{G_k\}_{k=1}^{K}$。

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$G_k=[l_{min}^{(k)}, l_{max}^{(k)}], k=1, \cdots, K$，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2-15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13行对这些点进行K-means聚类，得到数据包的K个类簇。第14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。

需要注意的是，算法1仅用于模型训练阶段，这一过程对齐VoIP流量样本和与之对应的语音样本，学习语音频带和数据包长的近似对应关系。实际部署应用中，网络监管者可以预先提取嫌疑人语音数据，并利用后文5.1节中所搭建的流量采集系统获取嫌疑人VoIP流量样本（因为很多诈骗分子往往是惯犯，在公安等部门留有犯罪记录，例如审讯录音等），从而学习可靠的音频-包长对应关系。在推理阶段，模型使用上述音频-包长对应关系提取数据包分组统计特征，并识别VoIP内容，而不再需要任何语音样本。这在提高系统推理效率的同时保障了其在实际安全监管部署中的可用性。下一节给出每个数据包分组统计特征的详细计算方法。

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$\{l_1, \cdots, l_9\}$；

（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9个区间的比例值作为特征（第10个区间可通过前9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。

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

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	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

个包含三层卷积神经网络的机器学习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。

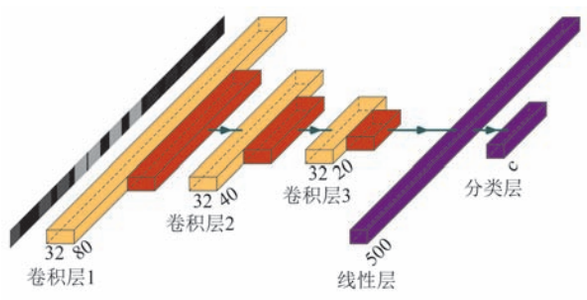

图8 模型架构及参数（$K=4$）

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

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

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

本节在实验室环境中采集真实的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。本节的实验包含两部分：一是数据采集实验（第5.1～5.3节），二是VoIP加密流量识别实验（第5.4～5.6节）。注意每部分实验分别在不同的实验环境中进行，以下分别对这两部分实验进行详细说明。

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[①]，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。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）。

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）是一个公开语音数据集，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，本文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9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。

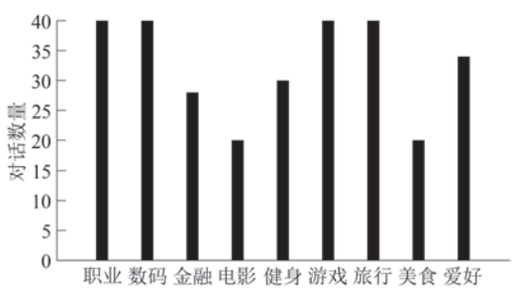

图9 Scenarios多场景语音数据集

① https://vb-audio.com/Cable

（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 Putonghua 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秒～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秒～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：

（1）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。

（2）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。

（3）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。

（4）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。

（5）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。

（6）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 32 GB GPU，操作系统为Ubuntu 22.04 LTS。深度学习模型使用PyTorch 2.4构建。

在语音处理中，通常将语音信号划分为4个频带，分别代表低频、次低频、次高频和高频，以优化编码和降噪[35]。本文测试了数据包分组算法（算法1）中的分组数量K的不同取值，发现当$K=4$时VoIP加密流量识别性能最优，并且由算法1得出在不同任务及流量数据集上的数据包分组结果，如表3所示。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。数据包分组的有效性将在5.6.5节的消融实验和5.7.2节的超参数敏感分析实验中进行详细验证。

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，包括基于统计特征的流量表征方法：

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、VCF；基于深度学习的流量表征方法：DeepFinger、FS-Net；基于负载表征的方法：ET-BERT、YaTC。

（1）HMM[8]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。

（2）VCF[12]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。

（3）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。

（4）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。

（5）ET-BERT[36]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。

（6）YaTC[37]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[38]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：

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

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

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

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

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

本节从实际VoIP安全监管场景出发，在用户性别属性识别、用户身份识别、通话语言识别和通话短语识别四个典型的VoIP监管识别任务中评估VPrint与已有方法的性能，并通过消融分析来验证数据包分组特征提取的有效性。

5.6.1 性别识别

性别识别任务旨在通过加密VoIP流量识别通话用户的性别属性，该任务可应用于针对VoIP诈骗犯罪嫌疑人的性别识别。本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。

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	0.662	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**

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%，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表现稍差于识别TIM用户，这可能是由于微信采用了较TIM更强的流量防御措施。与之相反，VPrint在微信和TIM上保持最稳定的性能，表明了VPrint在实际安全监管中具备对流量防御措施一定程度的适用能力。

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，这是因为VPrint不依赖于VoIP应用的语音编码算法，而是通过音频-包长对齐来提取数据包分组特征，使其具备对不同语音编码算法的通用性。

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

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

本节实验结果表明VPrint在一定程度上可应用于通话用户的性别属性识别，并表现出优于已有方法的识别精度和对不同语音编码算法的通用性。

5.6.2　用户识别

给定来自不同通话人的加密VoIP流量样本，用户识别任务旨在通过VoIP流量识别通话人的身份，该任务可用于在网络环境中追踪目标嫌疑人。本实验基于{WeChat，TIM}-speakers流量数据集，每个数据集分别包含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

本节实验证明，VPrint能有效应用于通话用户的身份识别，实现超0.99的$F1$值。

5.6.3　语种识别

语种识别任务旨在通过加密VoIP流量识别通话用户所使用的语言语种，该任务可以应用于协助分析网络罪犯的国籍、所处地域等信息。本实验基于{WeChat，TIM}-Languages流量数据集，共包括12种语种。以下主要分析WeChat-Languages流量数据集上的结果（如表9和表10所示），在TIM-Languages流量数据集上的结果与之类似，故在此略去。

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

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

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。

本节实验结果表明，VPrint可以有效应用于通话用户的语种类型识别，在不同语种中实现0.81～0.95的$F1$值。

5.6.4 短语识别

短语识别任务旨在通过加密VoIP流量识别通话中的关键短语，该任务对检测VoIP电信诈骗和非法VoIP交易犯罪具有重要应用价值。本实验基于{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

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

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。

本节实验结果表明，VPrint可以有效应用于VoIP内容短语识别，并在一些与违法犯罪相关的关键短语上实现0.76～0.97的$F1$值。

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

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

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。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。

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**

5.7　敏感性分析与鲁棒性评估

基于上述实验设置，本文进一步分析模型结构和超参数设定对VPrint性能的影响，并评估VPrint在网络抖动和主动防御场景下的鲁棒性。

5.7.1　模型结构敏感性分析

本节实验对比不同模型结构对VPrint性能的影响，基于WeChat-speakers和WeChat-Languages数据集。VPrint保持原始VoIP加密网络流量特征提取模块，分别采用卷积神经网络(CNN)、多层感知机(MLP)、支持向量机(SVM)、随机森林(RF)、朴素贝叶斯网络(NB)和XGBoost作为所提取流量特征的识别模型，实验结果如图10所示。

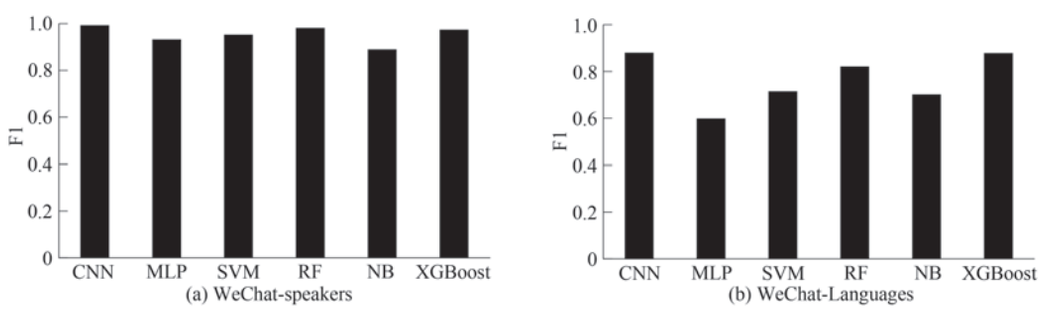

图10　VPrint采用不同流量识别模型的性能对比

对比图10(a)和图10(b)中不同模型的$F1$值，在不同的数据集上采用卷积神经网络的VPrint始终表现出最高性能，其次是基于树模型的随机森林和XGBoost，并与最高性能接近。这表明，VPrint的性能优势主要得益于所提出的VoIP加密网络流量特征提取方法，能有效地建模流量中的声学差异，而不依赖复杂模型的表征学习能力。

5.7.2　超参数敏感性分析

第5.6.5节的消融实验表明，VPrint的性能优势来源于本文所设计的VoIP数据包分组特征提取方法。本节实验对数据包分组特征提取模块中的超参数K(即数据包分组的组数)进行敏感性分析，对比不同K对VPrint性能的影响。实验基于TIM-Hello和WeChat-Languages数据集，K设置在[1,8]区间，结果如图11所示。

可以观察到，VPrint的性能随着K取值的逐步增大呈现先增长后下降的趋势。当$K=4$时，VPrint表现出最佳性能。如3.3节所讨论的，这是

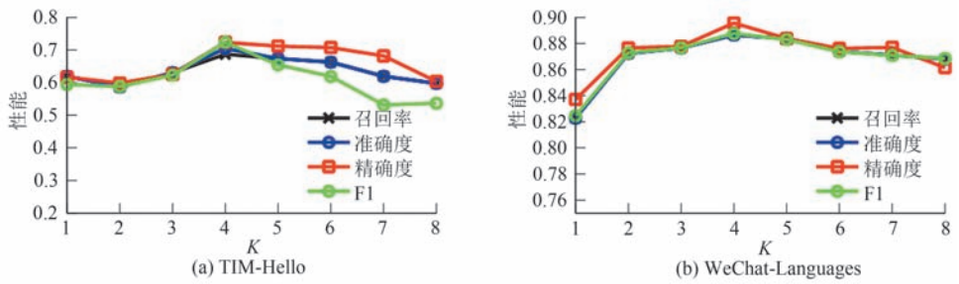

图11 不同超参数K对VPrint性能的影响分析

由于人类语音信号中的信息主要集在低频、次低频、次高频和高频四个频带[35]，而VoIP流量数据包的包长与所编码语音的频率相关。因此设置$K=4$有助于分离VoIP流量中来源于不同音频频段的语义信息，从而提取到分辨能力最佳的VoIP流量分组特征。与之相反，过小的K会导致不同频段语音对应的VoIP流量混杂在一起，生成语义混乱的流量特征；而过大的K则会导致过度的频段分割，破坏了所提取流量特征的语义完整性。

5.7.3 网络抖动环境中的性能评估

实际VoIP通话中，网络会话会因网络故障、网络拥塞等原因出现抖动。考虑到VoIP基于UDP协议，本节实验评估不同算法在数据包乱序传输故障（乱序）和数据包丢失故障（丢包）场景中的性能。实验基于WeChat-speakers和WeChat-Languages数据集，模拟不同强度网络异常，对比VPrint与性能最佳的基线方法（DeepFinger和VCF）的$F1$值，实验结果如图12和图13所示。

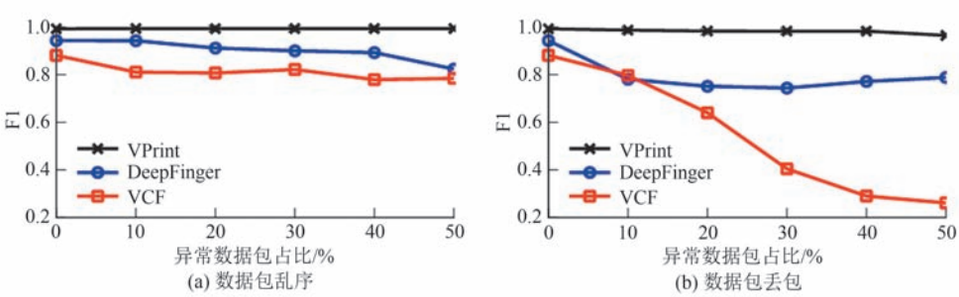

图12 不同算法在WeChat-speakers数据集网络异常场景性能分析

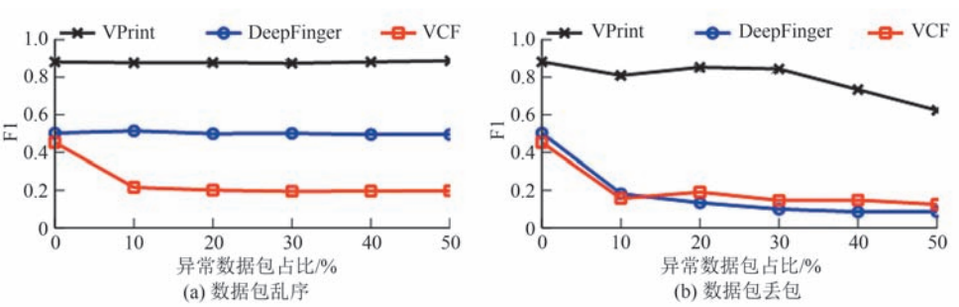

图13 不同算法在WeChat-Languages数据集网络异常场景性能分析

对比图12和图13中不同网络环境下各算法的$F1$值，可以观察到，VPrint始终保持最高性能。在数据包乱序场景中，VPrint的$F1$值不受故障强度影响。这是由于本文所提出的VoIP流量特征仅依赖VoIP数据包的分组统计信息，与数据包序列的顺序无关。相反，DeepFinger和VCF在两个数据集上分别表现出不同程度的性能下降。在数据包丢包场景中，由于数据包丢失必然导致流量特征变稀疏，可以观察到，随着丢包强度的增大，所有算法的$F1$值均逐渐下降。尽管如此，VPrint仍然表现出最高的$F1$值和最低的$F1$值下降量，表明在网络抖动场景中VPrint具备最佳稳定性。

5.7.4 主动防御场景中的性能评估

数据包填充[39]、数据包注入[40]和数据包延迟发

送[10]等流量主动防御技术可能被网络服务提供商使用以对抗流量分析技术。本节实验评估不同算法在流量主动防御场景中的性能表现，探究VPrint对不同防御措施的敏感性。

由于所采集真实VoIP流量中主动防御措施无法直接测量，实验基于WeChat-speakers和WeChat-Languages数据集，模拟不同强度的防御策略，同时测量不同强度防御所引入的网络带宽消耗，量化评估安全防御能力和带宽开销间的平衡。考虑到VPrint及基线方法中的流量特征不涉及数据包时间属性，实验略去对数据包延迟发送的分析，只评估数据包填充和数据包注入策略，实验结果如图14和图15所示。

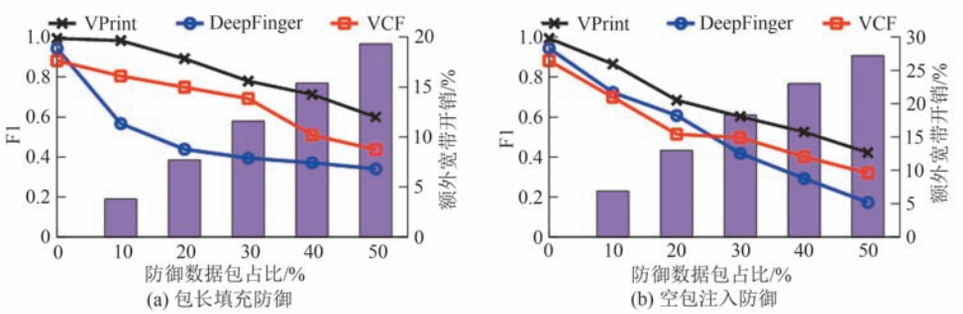

图14　不同算法在WeChat-speakers数据集主动防御场景性能分析

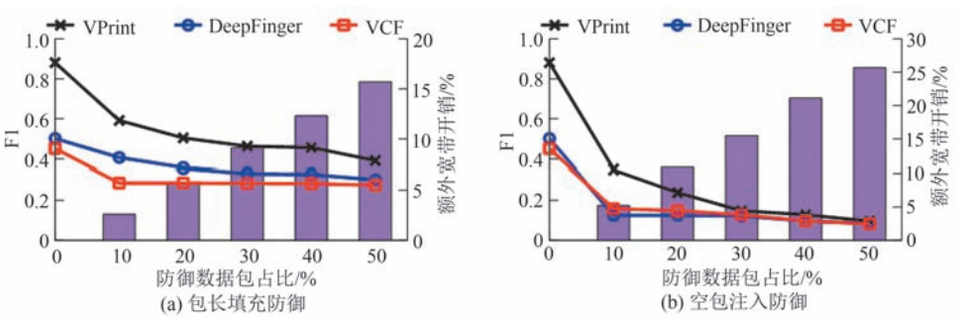

图15　不同算法在WeChat-Languages数据集主动防御场景性能分析

可以观察到，在不同防御措施中，随着防御强度（即防御数据包占比）的增大，所有方法的$F1$值均逐渐下降。相较于基线方法，VPrint的$F1$值下降量最小，表明其受防御措施影响最小。这是由于VPrint提取数据包分组统计特征建模VoIP流量，能有效弱化防御数据包对流量整体模式的扰动。

对比不同的防御措施，可以发现空包注入比包长填充具有更好的防御能力。这是因为包长填充只能改变VoIP流量的包长统计特征，而空包注入能同时改变VoIP流量的交互模式和包长统计特征。此外，测量不同防御强度带来的带宽开销可以发现，所有防御措施均会严重影响网络服务质量。例如，50%的包长填充会增加15.7%～19.3%的带宽开销，而50%的空包注入会增加25.7%～27.2%的带宽开销。

在实际部署中，网络安全监管方可以与VPrint服务商协调调整流量主动防御措施，减小主动防御措施对VoIP安全监管的影响，同时避免潜在攻击者滥用VPrint算法来非法侵犯用户隐私。

6　总　结

VoIP应用使用日益广泛，利用VoIP应用进行网络犯罪也日益猖獗，如何识别VoIP加密流量是一个亟待研究的问题。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带与数据包长对齐的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。

本文立足于VoIP安全监管和反诈治理，一方面，本文的研究发现有助于理解加密VoIP流量模式，将其应用于对电信诈骗、非法交易等网络犯罪的

取证和制止。另一方面，本文的研究发现也揭露了现有VoIP应用的安全隐患，恶意分析者可能利用VoIP加密流量窥探用户隐私信息。启发VPrint服务商进一步设计隐私安全并且部署高效的流量主动防御措施，保障用户通话安全。

未来工作可以围绕VoIP流量高效建模、表征模型优化和实际部署增强三个方面展开探索。这包括：(1)探索不同音频频段与VoIP流量分组间的细粒度映射关系，提高流量分组特征的建模能力；(2)通过采用Transformer等更先进的模型架构表征VoIP流量特征，提升VPrint识别准确率；(3)探索在高噪声环境、多人重叠语音、长时非重复对话、动态带宽调整等更复杂、开放世界中VPrint的有效性和增强方法，扩展VPrint可部署性。

参 考 文 献

[1] 2025 China VoIP network telephone industry panoramic research and future trend analysis report. China Market Research Website, 2025 (in Chinese)
(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. 中国市场调研网, 2025)

[2] The Ministry of Public Security: Over the past five years, a total of 1.945 million cases of telecommunications and online fraud crimes have been cracked. People.cn, 2024 (in Chinese)
((公安部:五年来共破获电信网络诈骗犯罪案件194.5万起. 人民网, 2024.) http://society.people.com.cn/n1/2024/0527/c1008-40244357.html)

[3] New VoIP device fraud: A single data cable turned overseas calls into local ones, and someone was defrauded of over 240, 000 yuan. Chengdu Economic Daily, 2025 (in Chinese)
((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. 成都商报, 2025.) https://static.cdsb.com/micropub/Articles/202501/cacadac5faca3163a2d2b3c71fa9852f.html)

[4] Shapira T, Shavitt Y. FlowPic: A generic representation for encrypted traffic classification and applications identification. IEEE Transactions on Network and Service Management, 2021, 18(2): 1218-1232

[5] Wright C V, Coull S E, Monrose F. Traffic morphing: an effincient defense against statistical traffic analysis//Proceedings of the 14th Annual Network and Distributed Systems Symposium. San Diego, USA, 2009, 1-14

[6] Ding L, Yuefei Z, Bin L, et al. Survey of side channel attack on encrypted network traffic. Chinese Journal of Network & Information Security, 2021, 7(4): 114-130

[7] Wright C V, Ballard L, Monrose F, et al. Language identification of encrypted VoIP traffic: Alejandra y roberto or Alice and Bob?//Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium. Boston, USA, 2007, 1-12

[8] Wright C V, Ballard L, Coull S E, et al. Spot me if you can: Uncovering spoken phrases in encrypted VoIP conversations//Proceedings of the IEEE Symposium on Security and Privacy. Oakland, USA, 2008, 35-49

[9] White A M, Matthews A R, Snow K Z, et al. Phonotactic reconstruction of encrypted VoIP conversations: Hookt on fon-iks//Proceedings of the IEEE Symposium on Security and Privacy, Oakland, USA, 2011, 3-18

[10] Nasr M, Bahramali A, Houmansadr A. Defeating DNN-Based traffic analysis systems in Real-Time with blind adversarial perturbations//Proceedings of the 30th USENIX Security Symposium. Online, 2021, 2705-2722

[11] Vaidya T, Walsh T, Sherr M. Whisper: A unilateral defense against VoIP traffic re-identification attacks//Proceedings of the 35th Annual Computer Security Applications Conference. San Juan. Puerto Rico, USA, 2019, 286-296

[12] Wang C. Kennedy S, Li H, et al. Fingerprinting encrypted voice traffic on smart speakers with deep learning//Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks, Linz, Austria, 2020, 254-265

[13] Sikos L F. Packet analysis for network forensics: A comprehensive survey. 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//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Toronto, Canada, 2018, 1928-1943

[15] Wang S, Yu B, Wu M. MVCM car-following model for connected vehicles and simulation-based traffic analysis in mixed traffic flow. 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. Soft Computing, 2020, 24(3): 1999-2012

[17] Wu D, Wang X, Qiao Y, et al. NetLLM: Adapting Large Language Models for Networking//ACM SIGCOMM. Sydney, Australia, 2024, 661-678

[18] Wu C, Ni S, et al. An Encrypted Video Recognition Method Based on the Transmission Characteristics of HTTP/3. Chinese Journals of Computers, 2024, 47(7): 1640-1664 (in Chinese)
(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. 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//Proceedings of the 37th Annual Computer Security Applications Conference. Virtual, USA, 2021, 1025-1038

[20] JiangI M, Li Z, Fu P, et al. Accurate mobile-app fingerprinting using flow-level relationship with graph neural networks. 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//Proceedings of the International Conference on Collaborative Computing: Networking, Applications and Worksharing. Hangzhou, China, 2022, 437-456

[22] Choudhury P, Kumar K P, Nandi S, et al. An empirical approach towards characterization of encrypted and unencrypted VoIP traffic. 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. 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// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017, 335-344

[25] Fu Y, Xiong H, Lu X, et al. Service usage classification with encrypted internet traffic in mobile messaging apps. 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// Proceedings of the 16th IEEE International Conference on Data Mining. Barcelona, Spain, 2016, 877-882

[27] Khan L A, Baig M , Youssef A M. Speaker recognition from encrypted VoIP communications. Digital Investigation, 2010, 7(1): 65-73

[28] Lu C, He L, Xiong G, et al. FS-Net: A flow sequence network for encrypted traffic classification//Proceedings of the IEEE Conference on Computer Communications. Paris, France, 2019, 1171-1179

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

[30] Adaptive Multi-Rate audio codec. 2025. https://en.wikipedia.org/wiki/Adaptive_Multi-Rate_audio_codec

[31] DatasetASR-CTeleCSC. 2025. https://magichub.com/datasets/mandarin-chinese-conversational-speech-corpus-telephony/

[32] DatasetASR-MultiDeviCCSC. 2025. https://magichub.com/cn/datasets/mandarin-chinese-conversational-speech-corpus-mutiple-devices/

[33] Magicdata Putonghua Chinese Read Speech Corpus. 2025. https://www.openslr.org/68

[34] Aishell-WakeUp-1 Chinese and English Wake-up Words Speech Database. 2025 (in Chinese)
(AISHELL-WakeUp-1 中英文唤醒词语音数据库. 2025. https://www.aishelltech.com/wakeup_data)

[35] Ueda K, Nakajima Y. An acoustic key to eight languages/dialects: Factor analyses of critical-band-filtered speech. Scientific Reports, 2017, 7: 1-4

[36] Lin X, Xiong G, Gou G, et al. ET-BERT: A contextualized datagram representation with pre-training transformers for encrypted traffic classification//Proceedings of the ACM Web Conference. Lyon, France, 2022, 633-642

[37] Zhao R, Zhan M, Deng X, et al. Yet another traffic classifier: A masked autoencoder based traffic transformer with multi-level flow representation//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 5420-5427

[38] He K, Chen X, Xie S, et al. Masked Autoencoders Are Scalable Vision Learners//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2022, 15979-15988

[39] Luo X, Zhou P, Chan E W, et al. HTTPOS: Sealing information leaks with browser-side obfuscation of encrypted flows//Proceedings of the the Network and Distributed System Security Symposium. San Diego, USA, 2011, 1-20

[40] Gong J, Wang T, Zero-delay lightweight defenses against website fingerprinting//Proceedings of the 29th USENIX Conference on Security Symposium. Berkeley, USA, 2020, 717-734

ZHAO Jun-Zhou, Ph. D., associate professor. His research interests mainly focus on network security.

DUAN Tao, Ph. D. candidate. His research interest is network traffic analysis.

LI Jiang-Long, M. S. candidate. His research interest is network traffic analysis.

WANG Ping-Hui, Ph. D., professor. His research interest is network security.

TAO Jing, M. S., researcher. His research interest is network security.

Background

With the rapid popularization of mobile devices and fast development of wireless networking technology, Voice over Internet Protocol (VoIP) applications have become increasingly popular. VoIP applications are built on the Internet and transmit voice signals through IP networks. During a voice call, the analog voice signal generated by the sender is compressed and encoded, and then packaged into data packets according to protocols such as TCP/IP, which are transmitted over the IP

network to the destination. The receiver then reassembles and decodes the received data packets to restore the original voice signal, thus achieving voice communication over the Internet. While VoIP applications have greatly enhanced communication convenience in people's daily lives, they have also facilitated the proliferation of cybercrimes, inflicting substantial harm on individuals. This work studies the VoIP encrypted network traffic classification problem which is related to the extensively studied encrypted traffic classification problem. However, VoIP encrypted traffic classification is considered to be more challenging than encrypted traffic classification as it aims to achieve fine-grained classification. In the literature, VoIP encrypted traffic classification requires the knowledge of voice encoding algorithms to model the relationship between voice units and traffic segments. However, nowadays, the voice encoding algorithms adopted by VoIP application providers are usually not open to public, which makes existing methods no longer applicable. This work investigates four popular VoIP applications in China, i. e., WeChat, TIM, DingTalk, and Tencent Meeting. By collecting a large amount of real traffic data in the lab, this work measures and analyzes the correlation between the encrypted network traffic transmission patterns of the four VoIP applications and user attributes, voice content, etc., and discovers a significant correlation between voice spectra 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 network 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 identification, 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.

This work was supported in part by the National Natural Science Foundation of China (62272372).