中图法分类号:TP37　　**文献标识码**:A　　**文章编号**:1006-8961(2024)02-0369-13

论文引用格式:Zhang Y M, Chen K J, Ding J Y, Zhang W M and Yu N H. 2024. RoCC: robust covert communication based on cross-modal information retrieval. Journal of Image and Graphics, 29(02):0369-0381(张晏铭, 陈可江, 丁锦扬, 张卫明, 俞能海. 2024. 利用跨模态信息检索的鲁棒隐蔽通信. 中国图象图形学报, 29(02):0369-0381)[DOI:10.11834/jig.230504]

利用跨模态信息检索的鲁棒隐蔽通信

张晏铭,陈可江*,丁锦扬,张卫明,俞能海

中国科学技术大学网络空间安全学院,合肥　230000

摘　要:**目的**　隐蔽通信是信息安全领域的一个重要研究方向,现有基于多媒体数据流构建隐蔽信道的方法,未考虑网络传输时波动产生的数据包丢失问题。本文提出一种基于跨数据模态信息检索技术的对网络异常具有鲁棒性的隐蔽通信方法,同时可以满足高隐蔽性和高安全性的要求。**方法**　提出了一个名为RoCC(robust covert communication)的通用隐蔽通信框架,它基于跨模态信息检索和可证明安全的隐写技术。所提方法将直接通信和间接通信两种形式相结合。直接通信通过VoIP(voice over internet protocol)网络通话服务进行,传递实时生成的音频流数据,接收方可以通过语音识别将其还原为文本;而间接通信则借助公共网络数据库进行载密数据的传输,接收方通过文本语义相似度匹配的方式来还原完整语义的载密文本数据,这有助于解决网络数据包丢失和语音识别误差导致的文本语义丢失的问题。**结果**　经实验测试,本文方法在协议上具有更好的通用性,相对Saenger方法在丢包率抵抗能力方面提高了5%,所用隐写算法满足可证安全性。同时,RoCC的数据传输率有73～136 bps(bit per second),能够满足实时通信需要。**结论**　RoCC隐蔽通信框架综合可证明安全隐写、生成式机器学习方法和跨模态检索方法的优势,与现有的方法比较,具有更加隐蔽和安全的优势,并且是当前对数据传输丢包异常最鲁棒的模型。

关键词:信息隐藏;隐蔽通信;生成式模型;数据跨模态转换;可证明安全隐写;多媒体信息检索;相似度分析

RoCC: robust covert communication based on cross-modal information retrieval

Zhang Yanming, Chen Kejiang*, Ding Jinyang, Zhang Weiming, Yu Nenghai

School of Cyber Science and Technology, University of Science and Technology of China, Hefei　230000, China

Abstract: **Objective** Covert communication is a pivotal research area in the field of information security. A highly covert and secure covert channel for transmitting sensitive information must be developed to safeguard the privacy of communication users and prevent occurrences of eavesdropping on confidential data transmissions. Most methods build covert channels by tunneling multimedia streams. However, the problem of packet loss caused by fluctuations in network transmission is not considered. This study proposes a covert communication method that is robust to network anomalies and is based on cross-modal information retrieval and provably secure steganography. **Method** We propose a general covert communication framework named robust covert communication (RoCC), which is based on cross-modal information retrieval and provably secure steganography. Artificially generated information from artificial intelligence (AI) systems, including deep synthesis models, AI-driven artwork, intelligent voice assistants, and conversational chatbots, has emerged. These AI models can

收稿日期:2023-07-25;**修回日期**:2023-10-07;**预印本日期**:2023-10-14

*通信作者:陈可江　chenkj@ustc.edu.cn

基金项目:国家自然科学基金项目(62102386,U2336206,62002334,62072421)

Supported by:National Natural Science Foundation of China (62102386, U2336206, 62002334, 62072421)

synthesize multimodal data, such as videos, images, audio, and text. The practical application of provably secure steganography has become a reality as generative models make significant strides. Thus, we introduce generative models and provably secure steganography techniques into our framework, embedding secret messages within the cover text data. Furthermore, the domain of speech synthesis and recognition has witnessed the advent of numerous mature open-source models, facilitating seamless cross-modal conversion between speech and text. Our approach employs a combination of direct and indirect communication. In direct communication using voice over internet protocol (VoIP) network call service, real-time synthesized audio stream data are delivered, and the receiver can restore the text through voice recognition. Indirect communication uses a public network database for steganographic text data transmission. The receiver restores lost text semantics because of network packet loss and speech recognition errors via text semantic similarity matching. The entire communication process can be succinctly described as follows. Assuming that the sender of confidential data is Alice and the recipient is Bob, Alice and Bob share the same generative model and parameter settings for provably secure steganography. Alice embeds the confidential data into the generated text data using provably secure steganography techniques and publishes it on a publicly accessible and searchable network database. The only means of direct communication between the two parties is through VoIP network voice calls. Thus, the potential loss of network data packets is acknowledged. On the basis of the preserved semantic information, Bob performs cross-modal information retrieval from the public database and successfully locates the corresponding steganographic text data within the cover text. Subsequently, Bob recovers the confidential data from the steganographic texts by using the same generative model and parameter settings for steganography. **Result** The results of speech recognition experiments indicate that speech recognition often leads to semantic loss issues. The sentence error rate of the best model, standing at a mere 0.612 5, fails to meet the text recovery capability required for constructing covert channels through direct cross-modal transformations. Text similarity analysis experiments indicate that the best model can achieve a recall metric of 1.0, thereby theoretically enabling complete semantic information restoration. The experiment on combating network packet loss shows that RoCC achieves an impressive information recovery rate of 0.992 1 when the packet loss rate is 10% with a K value of 2. This finding demonstrates the exceptional resilience of RoCC to network anomalies and establishes it as the current state-of-the-art solution. In the experiment on real-time performance, we validate the high efficiency of the RoCC system in various components, such as speech synthesis and recognition, secure steganographic encoding and decoding, and text semantic similarity analysis. These results demonstrate the ability of RoCC to meet the real-time requirements of covert channel communication. In comparative experiments, RoCC is compared with eight representative methods. The results show that RoCC has outstanding advantages in terms of protocol versatility, robustness, and data steganography as provable security. Compared with the current robust model, RoCC shows increased resistance to packet loss rate by 5% in the antinetwork packet loss experiment. **Conclusion** The covert communication framework proposed in this study combines the advantages of provably secure steganography, generative machine learning methods, and cross-modal retrieval methods, making the covert communication process increasingly stealthy and secure. We also implement the first method of using semantic similarity to restore data communication lost due to an abnormal transmission process. After experimental verification, our framework meets the requirements of real-time communication in terms of performance, and the real-time transmission rate reaches 73～136 bps.

Key words: information hiding; covert communication; generative model; data cross-modal conversion; provable security steganography; multimedia information retrieval; similarity analysis

0 引　言

隐蔽通信(covert communication)是信息安全领域一个重要的研究方向,为了保护通信用户的隐私数据和防止窃听机密数据传输事件的发生,需要构建一个隐蔽性强、安全性高(李凤华 等,2022)的网络隐蔽信道(covert channel, CC)完成敏感数据的传输。

Tian等人(2020)系统总结了现有的隐蔽通信方法和技术,将构建网络隐蔽信道的关键技术分为通信内容层(基于信息隐写)和传输网络层(基于代理和匿名通信技术)两个方面。虽然传输网络层的隐蔽信道技术较为成熟,但是其部署难度较高,因其依赖网络协议的特点,也更容易受到窃听者或审查者

（下文统称攻击者）针对协议的攻击。这类通信技术的安全性依赖于协议本身加密的安全性，虽然通过加密保护了明文信息的机密性，但它们传输的数据很容易被识别为加密流量，从而引起攻击者的注意，因此损失隐蔽性。而对于通信内容层的隐蔽信道技术，一般在数据层面应用隐写技术嵌入需要传输的机密数据，具有较高的隐蔽性，但是传统的隐写算法受限于经验安全，易受到隐写分析的攻击（郎荣玲等，2004），因此安全性不足。

通信内容层的隐蔽通信方法，相比于传输网络层通常具有更高的隐蔽性，其中部分方法的数据隐写算法不依赖于特定的通信协议，又具有相对更大的可扩展性。此类方法目前国内研究较少，国外的工作主要选择常见的流数据应用作为部署的对象，如 VoIP 网络语音通话服务，YouTube 流视频媒体平台，WebRTC（web real-time communications）网络音视频通话软件等。例如 FreeWave（Houmansadr 等，2013）将客户的互联网流量调制成声音信号，通过 VoIP 连接传输。SkypeLine（Kohls 等，2016）利用直接序列扩频（direct-sequence spread spectrum，DSSS）为基础的隐写技术在 VoIP 语音流中隐藏信息。这两种方法未曾考虑网络波动异常状况对数据传输过程的影响，当网络延迟较大和出现较多网络数据包丢失等情况时，隐蔽信道则会失效。Saenger 等人（2020）利用许多客户端接口中可用的语音活动检测功能来产生假沉默数据包（silence packets），以该数据包作为隐藏数据的载体完成信息隐蔽传输。此方法考虑了网络波动问题对数据传输可能造成的错误，但当丢包率在 5% 时隐蔽信道依然会失效。Peng 等人（2021）针对 PCM（pulse code modulation）编解码器设计了一种主动语音周期检测算法，用于检测 VoIP 数据包中是否携带活跃或不活跃的语音数据，并根据逻辑混沌映射生成的随机序列随机选择 VoIP 流中的数据嵌入位置，在 VoIP 数据流中隐写机密数据。该方案对网络波动异常不够鲁棒。

另外有使用流媒体平台进行隐蔽传输的方法，如 CovertCast（McPherson 等，2016）将机密数据编码为图像序列并通过 YouTube 等直播流媒体服务传输给用户。此方案安全性依靠于流媒体平台的 HTTPS（hypertext transfer protocol secure）协议的安全性，但是缺乏隐蔽性，对消息接收用户不做限制，而且长时间的无意义视频是一种容易识别的异常行为。此外，还有基于网络音视频通话服务的方法，如 Protozoa（Barradas 等，2020）通过钩取 WebRTC 堆栈，用 IP（internet protocol）数据包的有效载荷替换编码的视频帧数据来实现隐蔽信道。此方法与 CovertCast 一样，安全性依赖于 WebRTC 网络传输协议对数据的加密，但是当攻击者设法获取数据流加密的密钥，可以对加密的数据流进行解密时，则可以完整获得传输的机密数据。而对于通信端，被机密数据替换的视频帧本身是无意义的图像，存在易被识别的异常行为，因此该方法欠缺隐蔽性。Stegozoa（Figueira 等，2022）在 Protozoa 的基础上应用视频隐写技术将机密数据嵌入到 WebRTC 视频信号中，防止通过直接视频内容检查检测到隐蔽的有效载荷。这个方法改进了 Protozoa 的隐蔽性，引入视频隐写技术在有意义的视频中嵌入人类难以识别的机密数据而不改变视频质量，符合行为安全要求，使得攻击者更难检测隐蔽通信的过程，但是所使用的隐写方法仍然存在被隐写分析攻击的可能性。

Balboa（Rosen 等，2021）在 TLS（transport layer security）层拦截传出的网络流量并将其重写为嵌入数据。为了避免引入与预期应用程序行为的任何可区分的差异，Balboa 只重写与通信各方之间预先共享的外部指定流量模型匹配的流量。流量模型捕获网络流量的某些子集（例如，音频流服务器流的音乐的某些子集），发送方使用此模型将传出的数据替换为指向模型中相关位置的指针，并在释放的空间中嵌入数据。然后，接收器提取数据，在将数据传递给应用程序之前，用模型中的原始数据替换指针。Balboa 提供了一种可以部署在任意受 TLS 保护的协议或者应用程序上构建隐蔽信道的框架，具有很大的可扩展性，在音频流和网络流量中构建隐蔽信道时，应用层面上维持程序行为不变，确保了行为安全性。但是其安全性依赖于 TLS 层的加密算法，对于有权限获得密钥的攻击者，这种通信将丧失安全性。

同时，在生成模型逐渐普及的时代，人工智能生成的信息变得更加常见，比如深度合成模型、AI（artificial intelligence）绘画、智能语音助手、智能聊天机器人等，多种模态数据包括视频、图像、音频、文本都可以由 AI 模型合成，并应用隐写术嵌入秘密消息（张卫明 等，2022）。AI 生成的数据量不断增长，并与正常人类生成的数据相互交错混合，想要区分数据的来源，技术要求高而且代价高昂。同时，随着

生成模型的进步，可证明安全隐写（Hopper 等，2002）已经投入实际应用。Chen 等人（2022）基于音频生成模型 Waveglow 设计了基于可逆采样的信息嵌入方法，实现可证明安全的音频隐写方法。Ding 等人（2023）则提出基于分布式副本的可证明安全的隐写方法，可以应用到图像、音频、文本生成模型，分别实现可证明安全的图像、音频、文本隐写。

接着，根据人类语言语义的特点，在语音通信过程中丢失部分数据信息时，语音信号仍然可以保留大部分语义信息。因此，考虑借助成熟的语音合成和识别技术，利用这个规律构造基于文本到语音的模态转换的鲁棒隐蔽信道。但是实验表明现有语音识别方法无法完全准确识别语音中的语义信息，因此该方案鲁棒性不足。此外，如果只采用文本单模态数据作为载体进行隐蔽通信，会出现不能满足隐蔽信道实时性要求的问题，另外大量文本的端到端传输容易被攻击者识别，会损害通信行为的安全性。其次，如果采用音频模态数据作为载体构建隐蔽信道，实验表明其不具备抵抗网络传输过程信息丢失的鲁棒性。

因此，现有隐蔽通信方法无法同时满足强隐蔽性、高安全性和强鲁棒性的要求。针对上述方案存在的缺陷和利用现有的可证安全隐写及跨模态技术成果，本文提出一种利用跨模态信息检索技术的强鲁棒性隐蔽通信方法。假设机密数据发送方为 Alice，接收方为 Bob，Alice 与 Bob 共享相同的生成模型及生成数据的设置参数，Alice 会将需要传输的机密数据通过可证安全隐写技术嵌入到生成的文本数据，并发布到公开可检索的网络数据库中。双方只能通过存在网络数据包丢失的 VoIP 网络语音通话方式进行直接通信。Bob 根据保留下来的语义信息，到公共数据库中进行跨模态信息检索，可以找到对应的载密数据，接着用相同的生成模型恢复出隐写文本中的机密数据。实验结果表明，所提方案满足强隐蔽性、高安全性以及在网络异常状态下进行消息传递的强鲁棒性，且实时传输率为 73～136 bps，缓冲扩展的传输率可以到达 300 kbps 以上。

本文的主要创新点包括：1）提出利用跨模态数据构建隐蔽信道的思想。具体而言，借助模态转化中语义基本不变的特点，将跨模态语义信息作为载体，进行隐蔽信息传递。基于上述思想，本文提出了一个名为 RoCC 的全新隐蔽通信框架，通过模态转化来克服信道对数据的有损传输。2）以文本语音转化为例，将可证安全隐写方法引入到跨模态语义信息中，实现安全的信息传输。同时，还提出了利用语义相似度分析方法来解决音频语义提取精度不够的问题。本文所提隐蔽通信方法克服现有技术方案改变协议行为模式和依靠流量加密的缺陷，具有高隐蔽性、高安全性和语音协议间的通用性。3）最后，在实验上对所提方案进行跨模态信息恢复准确率、抗信息丢失鲁棒性、性能效率分析和横向对比实验，证明了本方案与现有方案对比中具有高隐蔽性、高安全性的优势，且性能效率上符合行为安全性，与正常用户行为难以区分，抗信息丢失的信息恢复准确率在 1 k 大小的候选池中 Recall@1 指标可达 0.992 以上，Recall@2 指标可达到 1.0。实时传输率约有 100 bps，在实际应用中可以实现小数据量的传输，而在缓冲扩展模式下可以达到 300 kbps 以上，可以短时间内传输文件数据。

1 相关知识

1.1 隐写术

1.1.1 信息隐写

随着安全通信需求的不断增长，传统的加密通信方式已经愈发难以满足要求。而隐写术则是一种以数字图像、音频、视频或文本等载体为基础，将秘密信息巧妙地嵌入其中以实现秘密通信的技术。与传统的密码学相比，隐写术不仅能够保护秘密数据的安全，更能够隐藏秘密通信的存在，使其具备更高的通信隐蔽性。形式化表达为，一个隐写系统（stegosystem）Σ_D 具有通道分布 D（偏差为 P_e），这个系统可以用一个概率算法三元组表示 $\Sigma_D=\left(KeyGen_D, Encode_D, Decode_D\right)$。

1）$KeyGen_D(1^\lambda)$ 接受长度为 λ 的输入，生成一个共享密钥 $K\in\{0,1\}^k$，长度为 k，用于编码与解码过程。

2）$Encode_D(K,m,H)$ 接受密钥 K，秘密消息 $m\in\{0,1\}^*$，以及一个通道历史 H 作为输入，返回载密对象 $s=s_1\|s_2\|\cdots\|s_l$，长度为 l。

3）$Decode_D(K,s,H)$ 接受密钥 K，载密对象 $s\in\{0,1\}^*$，以及一个通道历史 H 作为输入，返回从 s 中提取的秘密消息 m。

与密码学一样，隐写术需要满足 Kerckhoffs 原理（Kerckhoffs，1883），即攻击者知道除密钥以外的任何信息。

1.1.2　可证安全隐写

语言隐写术（linguistic steganography，LS）是最常用的隐写术之一。受益于生成模型，Kaptchuk 等人（2021）研究了使用生成模型作为隐写采样器，因为它们代表了最著名的近似人类交流的技术，以生成模型作为隐名采样器生成的文本，可证明与诚实模型生成的正常本文不可区分。Hopper 等人（2002）提出了隐写安全的复杂性理论定义，该定义通过区分 oracle（O_D）和 $Encode_D$ 输出的概率博弈来建立。当其对所有概率多项式时间（probabilistic polynomial time，PPT）的敌手 A_D 都满足以下条件，称隐写系统是对选择载密对象攻击（chosen hiddentext attacks，CHA）安全的

$$\left|Pr\left[A_D{}^{Encode_D(K,\cdot,\cdot)}=1\right]-Pr\left[A_D{}^{O(\cdot,\cdot)}=1\right]\right|<negl(\lambda)\tag{1}$$

式中，O_D 是一个从数据分布 D 随机采样的 oracle。

目前，先进的可证安全隐写技术是 Ding 等人（2023）提出的基于分布式副本的 Discop。以文本生成任务为例，文本生成利用计算语言学和人工智能知识自动生成近似人类编写的文本。自回归语言模型，如 GPT 系列模型（Brown 等，2020）是这个领域最有代表性的生成模型，可以给定前一个上下文（context）$x_{<t}$ 预测下一个 token 的概率分布 $P^{(t)}=Pr\left[x_t\middle|x_{<t}\right]$。

对于预测分布，采用随机采样策略。随机抽样生成文本的整个过程为：首先，使用伪随机数生成器（pseudo-random number generator，PRNG）生成一系列伪随机数 $\boldsymbol{r}=\left\{r^{(0)},r^{(1)},\cdots\right\}$，满足在区间 $[0,1)$ 上的均匀分布，即 $r^{(t)}\sim U[0,1)$。对于每个时间步 t，生成模型 M 预测 $P^{(t)}$，然后每个在词汇表 V 中的 token 根据 $P^{(t)}$ 分配一个在 $[0,1)$ 中的左闭右开区间。然后，使用一个伪随机数 $r^{(t)}$，并选择对应于 $r^{(t)}$ 所处区间的 token 作为下一个标记 x_t，并将其附加到 context 中。重复这个过程，直到达到终止条件（例如，生成的令牌序列的长度达到预设的最大长度）。

Discop 的原理是，在生成过程中，可以构建生成模型预测的概率分布的多个副本（即区间分配方案），称为“分布副本”（distribution copies），并使用“分布副本”的索引来表达信息。例如，假设只有两个标记，“a”和“b”，概率分别为 0.4 和 0.6。将 [0，0.4）和 [0.4，1.0）赋值给“a”和“b”，或者将 [0.6，1）和 [0，0.6）赋值给“a”和“b”。因为每个令牌在几个“分布副本”中的概率是相同的，这意味着这些“分布副本”的分布是相同的。通过这种方式，发送方可以创建多个“分布副本”，并根据消息决定从哪个副本中采样令牌。只要发送方和接收方处于相同的设置下，包括 PRNG、种子、生成模型和上下文，就可以同步它们的所有状态，其中种子是用于初始化 PRNG 的数字（或向量），可以视为密钥的一部分。相应地，接收方可以通过确定从哪个“分布副本”中采样令牌来提取消息。因此，Discop 从原理上满足可证明安全的理论条件，详细证明过程可见 Ding 等人（2023）的论文。

1.2　语音合成与识别

文本转语音（text-to-speech，TTS）的目的是合成给定文本的自然的、可理解的语音。基于神经网络的 TTS 采用（深度）神经网络作为语音合成的模型主干。一些端到端模型，如 Tacotron（Wang 等，2017）和 FastSpeech（Ren 等，2019）简化了文本分析模块，直接将字符/音素序列作为输入，并使用梅尔频谱（Mel spectrogram，MS）简化了声学特征，从而加快了语音合成的过程。在语音识别方面，一些语音识别应用已经非常成熟，如微软亚洲研究院发布的语言处理模型 SpeechT5（Ao 等，2022）集成了语音合成与识别等功能。成熟的语音合成和识别技术为跨模态转换带来了很多便利。

1.3　文本语义相似度

文本语义相似度是自然语言处理（natural language process，NLP）领域的重要研究方向，并且拥有许多成熟的研究成果。基本的衡量文本差异的方法之一，即编辑距离（Levenshtein distance，LD）（Li 和 Liu，2007），是文本 a 转换为文本 b 所需要的最少单字符编辑操作次数，形式化定义为

$$lev_{a,b}(i,j)=\begin{cases}\max(i,j) & \min(i,j)=0\\ \min\begin{cases}lev_{a,b}(i-1,j)+1\\ lev_{a,b}(i,j-1)+1\\ lev_{a,b}(i-1,j-1)+1_{(a_i\neq b_j)}\end{cases} & \text{其他}\end{cases}\tag{2}$$

式中，a,b 为两个字符串，a_i 为 a 的前 i 个字符，b_j 为 b 的前 j 个字符，$lev_{a,b}(i,j)$ 是 a_i 与 b_j 之间的编辑距离。

对于传统方法，如搜索引擎的基本技术之一，以词频—逆文本频率（term frequency-inverse document frequency，TF-IDF）（Salton 和 Buckley，1988）作为词向量进行文本信息检索。两段文本之间的相似度，可以用词向量之间的距离衡量。

BERT 系列模型（Devlin 等，2019）实现了当前最好的文本相似度任务的效果，BERT 模型将文本嵌入到高维向量空间中，通过衡量文本的嵌入向量之间的距离，来判断文本语义之间的相似度。

2 威胁模型

多媒体隐蔽流（multimedia covert streaming，MCS）工具的一般系统模型如图 1 所示。它代表两个用户，一个作为客户端（Alice），另一个作为代理（Bob）。客户端位于互联网开放区域，代理位于受国家级别对手控制的限制区域。攻击者能够观察、存储、干扰和分析其管辖范围内的所有网络数据，并阻止区域外用户对区域内网络服务的普遍访问。审查策略可以基于目的 IP 地址或目的域名、通信中使用的协议（例如，BitTorrent 或 Tor），或列入黑名单的内容（例如，通过关键字和图像过滤）。

MCS 工具旨在使客户端能够通过利用以下 3 个条件，与代理端构建实时性的隐蔽信道：1）由受信任的限制区域内部用户（Bob）操作的代理服务器的合作；2）由加密流服务（例如 Skype）组成的运营商应用程序，其流量由攻击方授权越过审查区域边界；3）攻击方允许的限制区域内用户可以访问的特定网站或平台，这类网站或平台不提供用户间隐私交流功能，所发布的信息为透明公开。

客户端和代理用户需要在各自的本地计算机上运行 MCS 软件，通过运营商应用程序管理的媒体流来创建隐蔽信道。这个信道将允许客户端在受限网络区域中访问机密数据。在此，对威胁模型做出以下假设：

1）通信双方仅能通过语音通话方式进行直接沟通，例如常用的 VoIP 服务。这是网络受审查地区内常见的被允许的通信方式之一。同时也可以通过文本形式进行通信，例如邮件和聊天室，但通信频率会受到限制（行为合理的要求），并且审查者有权随意审查所有通信内容。

2）通信双方可以访问多个（用户匿名的）公开网络数据库，例如公开的文字论坛、社交媒体平台等。这些数据库会受到审查者的监控，但由于经济因素和服务质量的考虑，审查者不会禁止正常数据的发布，即只禁止敏感或有害的数据发布。通信双方可以使用相似的文本进行搜索，以查找对方发布的文本数据。然而，在这些网络数据库下，无法安全地进行点对点的信息传输，因为所有的通信行为在平台上都是透明可见的，直接的通信将会失去隐蔽性。

3）通信双方事先共享相同的生成模型和隐写参数设置，但彼此不知道对方在隐写时所使用的 context。审查者没有相同的模型，因此该模型和参数设置为隐蔽通信的密钥。通信双方需要在不暴露该模型和参数的情况下完成隐蔽通信过程。

4）语音通信过程中，允许网络存在波动等常见的异常状态，即存在数据包丢失，语音信息会因丢包率的不同存在相应程度的丢失。这是现实条件下经常会发生的事情，因为网络语音服务通常是基于 UDP（user datagram protocol）协议进行数据传输，一方面是为了降低通话延迟，另一方面是为

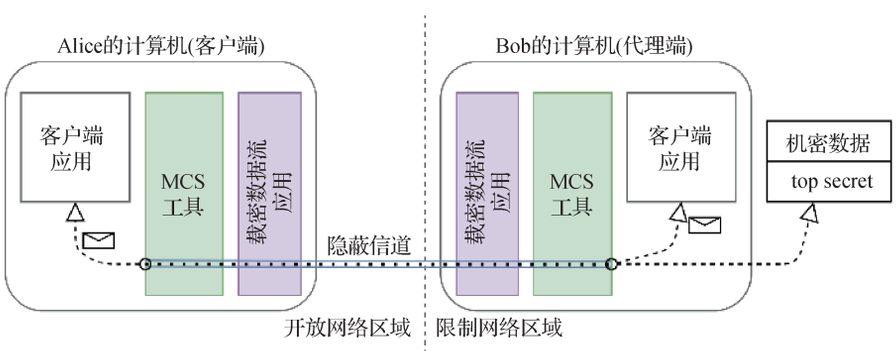

图 1 抗审查隐蔽通信的多媒体隐蔽流工具的系统模型

Fig. 1 System model of a multimedia covert streaming tool for censorship-resistant covert communication

了减轻 SIP(session initiation protocol)服务器的通信负担,因此语音通话过程中很可能存在信息丢失的问题。

3 系统框架

本节详细描述系统设计的框架结构、工作流程,以及每个模块的具体作用与要求。

3.1 总体框图

本文的系统框架如图 2 所示,整体设计由两个通信用户 Alice 和 Bob、一条直接通信的 VoIP 数据流通道以及一条借由公开网络数据库(open network database,OND)连接的间接信道组成。

每个通信方计算机段会安装 RoCC 应用程序,包括 4 个组件:RoCC 客户端(client)/RoCC 代理端口(socket),负责与用户数据交互;可证明安全隐写组件(provably secure steganography,PSS),这里采用 Discop 技术实现,包含编码器(encoder)和解码器(decoder),负责生成载密文本与提取机密数据;音频处理组件(audio processor,AP),其中包含语音合成模型和语音识别模型(speech recognition,SR),分别负责实时合成音频流数据并通过 VoIP 通道发送,以及将接收到的音频流数据转换为文本 m';搜索引擎组件(search engine,SE),包含相似度分析模块(similarity analyzer,SA)用以分析识别的文本 m' 与搜索引擎获取的文本池(text pool,*TP*)两两之间的语义相似度,找出原文本从而还原文本语义信息,以及内容发布模块(publisher,Pu)负责将载密文本通过正常用户行为发布到 OND 中。

通信双方由 VoIP 通话方式取得直接联系,该通道在部分时间存在数据包丢失,并受攻击者的审查,攻击者可以解密 VoIP 数据流,对其内容进行窃听,且可以对数据流进行隐写分析,试图找出其中隐藏的机密数据。此外,Alice 和 Bob 还可以借由受审查的公开网络数据库进行数据传输,但由于平台的局限性,无法隐蔽地完成点对点数据传输,因此该通信过程只能采取间接通信形式。即 Alice(Bob)在 OND 上以匿名形式发布公开的数据内容,让网络内容用户可见,接着,Bob(Alice)设法找出对应的数据并从网络上拉取至本地。整个间接通信过程保持双方匿名,因此具有很高的隐蔽性。

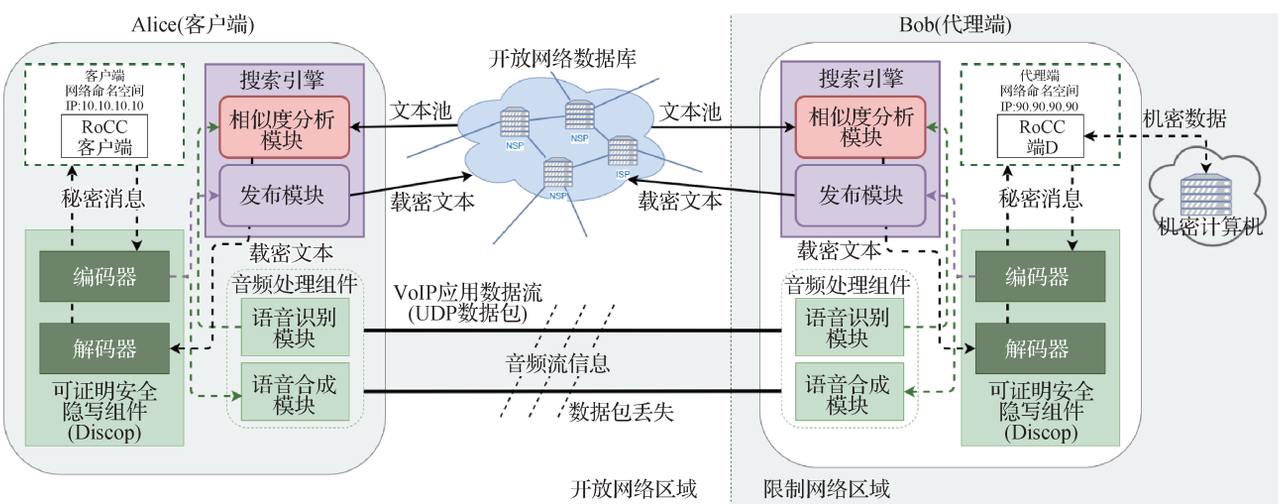

图 2 RoCC 的系统框图

Fig. 2 Framework of RoCC system

3.2 工作流程

3.2.1 总体流程

系统的整体流程分为两条通信方式进行描述,称语音传输的 VoIP 数据流通道为信道 A,称借由 OND 进行间接数据传输的通道为信道 B。整个 RoCC 的隐蔽信道则是由 A 结合 B 构成,两条信道同时传输消息,延迟在可控范围内。下面以 Alice 向 Bob 发送秘密消息为例,Bob 向 Alice 发送机密数据的流程则与之类似。

信道 A:Alice 经由 RoCC client 输入需要传输的秘密消息(secret message)记为 m,经由 PSS 的编码器生成的载密文本(stegotext)记为 c。接着,c 输入到

AP，生成音频流（audio stream）记为st。st经由VoIP数据流通道传输至Bob的计算机，由其内的AP识别转换为文本c'，由于语音识别的误差和网络传输过程可能丢失信息，c和c'的语义存在差异。最后c'输入到SE的SA，还原出文本c。载密文本c经过PSS的解码器解码提取m。此时，Bob通过client接收到秘密消息m，或者经由代理端口RoCC Socket，自动化访问机密计算机（secret computer，SC）获取目标机密数据。

信道B：Alice输入的秘密消息m，经过PSS的编码器生成载密文本c后，转入SE的Pu，自动向OND发布完整的载密文本c，并记发布完成时间为t_b，即从Pu传输数据开始至OND发回确认反馈的时间。此时Alice端在信道B的操作结束。Bob端，当从信道A接收到c'之后，设音频流信号有效时间为t_a，则Bob端的SE会获取$t_w = \max(t_a, t_b)$时间窗口内新发布的数据内容到本地，然后通过SA分析还原出真正的c，最后输入PSS的解码器提取秘密消息m，将其送入RoCC Socket完成全部通信流程。

时间t_a通常会比t_b更长，所以Bob端只需要取$t_w = t_a$，不需要得知t_b的具体数值，如果出现OND网络异常，那么Alice端尝试重传即可，这里假设通信所需要的网络服务是可用状态。引入参数t_w可以对系统性能提供有效的优化，可以解决实际ODB文件池过大，增加文本语义分析时间，导致传输效率不足的问题。

对于信道A和信道B的通信过程，可以形式化表述为

信道A:

1）$c \leftarrow Encoder_{\mathrm{dis}}(m)$;

2）$st \leftarrow AP_{\mathrm{TTS}}(c)$;

4）$c' \leftarrow AP_{\mathrm{ASR}}(as)$;

7）$m \leftarrow Decoder_{\mathrm{dis}}(c)$。

信道B:

3）$ODB_{\mathrm{TP}} \leftarrow SE_{\mathrm{Pu}}(c)$;

5）$TP \leftarrow SE(ODB, t_w)$;

6）$c \leftarrow SE_{\mathrm{SA}}(c, TP)$。

其中，$Encoder_{\mathrm{dis}}$和$Decoder_{\mathrm{dis}}$分别为可证明安全隐写Discop模块的编码及解码函数；AP_{TTS}和AP_{ASR}分别为音频处理模块的语音合成及识别函数；SE_{Pu}是搜索引擎的发布模块，负责将生成的载密文本c发布到ODB，$SE(ODB, t_w)$将ODB上t_w时间窗口内的文本集TP下载到本地，SE_{SA}通过相似度分析从TP中还原有损载密文本c'的原文本c。

另外，为了加速大文件数据传输的过程，在原通信方式上进行扩展，称为缓冲扩展模式。此模式下，在通信双方构建直接通信之前，可以将需要传输的大文件先用隐写组件进行编码，获得文本集合$\boldsymbol{MS}$，在构建直接通信之后可以直接以音频流传输$\boldsymbol{MS}$的前缀语义信息，即不需要完整传输所有文本，只需要传输其中前$q(<5)$个句子的文本。同时将$\boldsymbol{MS}$上传至OND。如此可以节省隐写组件对大量数据进行编码的时间，通信的时间消耗主要为语音通话，传输率即可得到很大的提高。

3.2.2　可证安全隐写

可证安全隐写组件由编码器和解码器组成，这里采用Discop作为实现的技术，当然可以选择其他隐写技术，但若不是可证明安全隐写，则安全性会有损失，易遭受隐写分析攻击。更详细地，Discop的安全性依赖于通信双方共享的生成模型M，随机数生成器P_{RNG}，以及初始化设置s_{eed}的保密性，整个隐写的密钥可以记为一个三元组$(M, P_{\mathrm{RNG}}, s_{\mathrm{eed}})$，当密钥不被攻击者获得，则整个通信过程理论上是安全的。值得指出的是，本系统的安全性与Protozoa等方法有所不同，本方案的通信不依赖流量加密所带来的安全性，因为国家级攻击者有权限解密任何网络供应商提供的数据流服务的加密流。

3.2.3　语音通话

Alice和Bob之间使用VoIP网络通话直接通信，比如Skype服务。音频处理模块包含TTS和SR两个模块，从隐写模块输出载密文本c到TTS模型，实时合成为音频流st，并通过VoIP数据流通道发送，接着被对方的语音识别模块接收，因网络数据包丢失和语音识别模型导致语义有损，识别得到的载密文本信息为c'。

因为隐蔽信道既要求安全性，又要求隐蔽性，所以在通信过程中的行为必须符合正常行为模式，又称为满足行为安全性。需要达到的要求包括如下：1）隐蔽信道的通话内容需要接近正常人类交流的内容，且不包含敏感内容和可疑信息；2）隐蔽信道的通话行为反应时间要与正常用户的通话行为反应时间

不可区分，即能够抵抗时间延迟模式的分析攻击；3）VoIP 数据流能够抵抗攻击者的隐写分析攻击。

为了满足要求 1），采用能够生成接近人类正常文本的 GPT 系列模型作为可证安全隐写的生成模型 M，通过精心构造 context，可以使得文本内容符合正常对话逻辑。为了满足要求 2），实际部署系统时，选择采用高性能的 SpeechT5 模型作为 TTS 模块和 SR 模块的实现，经过实验检验，该模型可以实现平均 0.336 s 生成 1 个 token，且每个 token 的语音识别延迟在 0.166 s 以内，足够接近人类正常通话行为模式，实验结果详见 4.4.1 节。对于要求 3），Alice 与 Bob 之间传递的语音内容 st 是从隐写模块生成的载密文本 c 合成语音得到的，因此 VoIP 数据流内本身并没有嵌入秘密消息，故满足该要求。

3.2.4　相似度分析

由语音识别模块输出的语义有损文本 c'，输入到相似度分析模块 SA，并与从网络公开数据库 OND 中搜索获取的文本池 TP 进行相似度比对。这里的实现采用结合式相似性度量，计算式为

$$f_{\text{Similarity}}(t_1,t_2) = \alpha \times Cos\_Sim(\boldsymbol{M}_{\text{emb}}(t_1),\boldsymbol{M}_{\text{emb}}(t_2)) + \frac{(1-\alpha)\times(1-lev(t_1,t_2))}{\max(len(t_1),len(t_2))} \tag{3}$$

式中，α 为动态参数，用以设置两种度量公式的权重比例，$len(t_1)$ 和 $len(t_2)$ 分别表示文本 t_1 和 t_2 的长度，$lev_{t_1,t_2}(len(t_1),len(t_2))$ 为编辑距离，参考式(2)，$\boldsymbol{M}_{\text{emb}}(t_1)$ 和 $\boldsymbol{M}_{\text{emb}}(t_2)$ 分别表示文本 t_1 和 t_2 经过 BERT 系列模型计算的嵌入向量。采用的度量公式，即向量之间的余弦相似度，计算式为

$$Cos\_Sim(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}\times\boldsymbol{y}}{|\boldsymbol{x}|\times|\boldsymbol{y}|} \tag{4}$$

式中，$\boldsymbol{x},\boldsymbol{y}$ 是等维向量，$|\boldsymbol{x}|$，$|\boldsymbol{y}|$ 为向量的模。

3.2.5　信息检索

信息检索过程由搜索引擎 SE 模块完成，文本池 TP 由 SE 从 OND 获取，设大小为 T_{size}，经过文本语义相似度分析，排序分析结果，以 K 个相似度分数最高的匹配结果作为返回。算法 1 文本搜索具体计算过程如下：

输入：文本池 TP，识别文本 m'，参数 K；

输出：隐写本文候选表 $\boldsymbol{CL}$(size = K)。

1）优先队列 q = {}；

2）$\boldsymbol{emb}_{m'} = Embedding_{\text{Model}}(m')$；

3）$\boldsymbol{emb}_{TP} = Embedding_{\text{Model}}(TP)$；

4）for $\boldsymbol{emb}$, mc in $\boldsymbol{emb}_{TP}$:

5）　$\boldsymbol{q}$. put($f_{\text{Similarity}}(\boldsymbol{emb}_{m'}, emb, m', mc), mc$)；

6）　if $\boldsymbol{q}$. size > K:

7）　　$\boldsymbol{q}$. get()；

8）$\boldsymbol{CL}$ = {}；

9）for $\boldsymbol{emb}$, mc in $\boldsymbol{q}$:

10）　$\boldsymbol{CL}$. put(mc)；

11）return $\boldsymbol{CL}$。

算法 1 输入候选文本池 TP，识别文本 m'，参数 K，输出 K 个候选文本列表。算法过程，维护优先队列 q，维持大小在 K 以内，对文本 m' 和每个候选池的文本 mc 计算其嵌入向量 $\boldsymbol{emb}$，遍历所有候选文本通过 Similarity 函数计算相似度得分，进入以相似度得分排序的优先队列 $\boldsymbol{q}$，超过 K 大小时，取出得分低的元素，最后将相似度得分最高的 K 个文本并入 $\boldsymbol{CL}$，作为结果返回。

4　实验结果

4.1　跨模态检索

本节是跨模态方法的准确率实验结果，所用指标为 SER（sentence error rate）和 CER（characters error rate），SER 即识别出错的句子数除以句子总数，CER 的值为 $(S+D+I)/N$，其中，S 是字符替换的个数，D 是字符删除的个数，I 是字符插入的个数，C 是正确的字符数，N 是引用中的字符数（$N = S + D + C$）。文本相似度检测的准确率所用指标为 $Recall@K = TP@k/(TP@k + FN@k)$，其中，$TP@k$ 是前 k 个返回结果中正确分类的正样本数，$FN@k$ 是前 k 个返回结果中错误分类的正样本数。

4.1.1　语音识别准确率

总共测试了 14 种开源模型的语音识别准确率，取 5 种模型作为代表性结果。测试所用中文数据集为 THCHS-30（T-30），英文数据集为 Librispeech Asr Datasets（LA）和 Discop 生成的载密文本合成语音 Stegotext Dataset（SD）。

结果如表 1 所示。其中，Uniasr 具体指模型 uniasr_8k_common_vocab8358，8 k 为参数，词汇表大小为 8 358，是中英语言通用模型。Para_large 即 Paraformer_large 模型，相比于 Paraformer 模型采用了更

表1　语音识别准确率

Table 1　Recognition accuracy of the speech recognition models

模型	SER	CER	数据集
Paraformer(Gao等,2023)	1.0	0.093 0	LA
Uniasr(Gao等,2020)	1.0	0.210 9	LA
Paraformer(Gao等,2023)	0.981 2	0.213 4	T-30
Uniasr(Gao等,2020)	0.914 5	0.126 0	T-30
Para_large(Gao等,2023)	**0.612 5**	**0.040 7**	T-30
Transformer(Gao等,2023)	0.981 2	0.215 0	T-30
SpeechT5(Ao等,2022)	1.0	0.146 7	SD

注:加粗字体表示各列最优结果。

深更大的模型结构,在多个中文公开数据集上取得SOTA(state-of-the-art)效果。Transformer模型,即Transformer-LM,是基于Transformer(Vaswani等,2017)的decoder架构构建。具有代表性的预训练模型,依然存在很高的词错率,而句错率最低的只有0.612 5,因此如果采用直接跨模态方法构建隐蔽信道,必然存在文本语义丢失的问题。因语义丢失而导致秘密消息提取失败。

4.1.2　文本相似度分析

相似度任务的实验设置,文本池的大小 T_{size} = 1 000,测试了3种开源文本嵌入模型的Recall指标,这里设置一个阈值参数threshold,表示排除文本tokens数目差距超过该阈值的候选对象,是一种启发式的剪枝策略,实验设置threshold = 5。

实验结果如表2所示,其中bert-mt表示bert-base-nli-mean-tokens模型,bert-cl表示bert-base-nli-cls-token模型,mulLM表示paraphrase-multilingual-MiniLM-L12-v2模型,3个模型均由Reimers和Gurevych(2019)提出,α参数是调节相似度函数的权重,详见算法1。

可以看出,mulLM模型可以在*Recall*@1上达到0.992 7,*Recall*@2上达到1.0,即用语音缺失文本c'在SA的前2个返回结果中可以找到原文本c,因此文本语义的恢复率可以达到100%。

表2　文本相似度分析准确率

Table 2　Text similarity analysis accuracy

模型	α = 0		α = 0.5		α = 1	
	Recall@1	*Recall*@2	*Recall*@1	*Recall*@2	*Recall*@1	*Recall*@2
bert-mt	0.970 3	0.977 4	0.969 1	0.976 2	0.935 8	0.954 8
bert-cl	0.970 3	0.977 4	0.969 1	0.976 2	0.939 4	0.958 4
mulLM	**0.992 7**	**1**	**0.992 7**	**1**	**0.982 9**	**0.993 9**

注:加粗字体表示各列最优结果。

4.2　鲁棒性实验

本节中,对比本文方案与StegoTTS方法(Chen等,2022),Saenger等人(2020)方法,以及朴素的基于直接跨模态数据转换构建隐蔽信道(direct cross-model,DCM)的方法,对网络异常状态消息传输的鲁棒性。其中,DCM是基于文本语音跨模态的隐蔽通信方案,其思想是发送方在文本中嵌入秘密消息,通过语音合成获得对应的音频数据,在语音通话信道上传输,接收方应用语音识别将音频数据还原回文本数据,再对文本数据提取秘密消息。如表3所示,DCM因为语音识别效果的不完美,导致语义的丢失成为无法避免的问题,因此缺失鲁棒性。对于StegoTTS方案,因为音频隐写基于自回归模型,当消息序列出现丢失,则提取秘密消息就会失败,对于Saenger的方案,因为在通话静默期进行隐蔽传输,所以丢包率较低时,并不会破坏隐蔽信道功能。可以看到,$RoCC_{k=2}$方案对抗丢包现象,具有很强的鲁棒性,在k = 2时,比Saenger的方案高出至少5%的

表3　抗网络数据包丢失鲁棒性

Table 3　Robust against network packet loss

方法	秘密消息恢复率				
	0%	1%	2%	5%	10%
DCM	0.229 7	–	–	–	–
StegoTTS	1.0	–	–	–	–
Saenger等人(2020)	1.0	1.0	1.0	–	–
$RoCC_{k=2}$	**1.0**	**1.0**	**1.0**	**1.0**	**0.992 1**

注:加粗字体表示各列最优结果,"–"表示提取消息失败,下标k表示召回率指标*Recall*@k的前k个相关结果参数。

丢包率抵抗能力。而且当丢包率超过10%时，依然可以通过取更大的k值，增加消息恢复率。

4.3 实时性实验

本节主要验证整个系统的性能效率能否满足隐蔽信道对实时性的要求。从3个组件的效率进行实验，包括语音合成与识别、可证安全隐写和文本相似度分析。

4.3.1 语音合成与识别效率

对于语音合成，测试1 000个句子的合成时间，总tokens数目为17 144，平均单位时间合成tokens数为2.97。对于语音识别，测试1 000个句子的识别时间，总tokens数为16 799，平均单位时间识别tokens数为6.06。所以音频处理组件的性能满足语音流实时通信的要求。

4.3.2 可证安全隐写效率

如表4所示，Tokens指生成文本的tokens个数。Bits指嵌入文本的秘密消息的比特数。可证安全隐写的编解码效率随着生成文本的tokens长度不同而有所差别，平均单位时间生成tokens数为47。

4.3.3 文本相似度分析效率

在改变文本池大小$T_{\text{size}} \in \{100, 200, 400, 800\}$，在$K$值取1和5时，进行相似度分析效率测量，如表5所示，检测单文本对象的相似度平均时间在0.02 s以内，表明信息检索时间非常短，对通信整体过程的性能的影响可以忽略不计。

4.4 对比讨论

本节将RoCC系统与现有的具有代表性的多媒体数据流构建隐蔽信道的方法进行多个角度的对比分析，如表6所示。其中，“标准协议”是指那些不改变协议行为的方法。“未修改二进制”表示那些在底层运行时，未修改目标协议实现的二进制程序。“不需要加密”表示那些方法不依赖于加密。“应用隐写”是指利用隐写方法的方法。“通用性”是指那些为支持各种应用程序/协议提供框架的方法，可以扩展其

表4 可证安全隐写效率

Table 4 Provably secure steganography efficiency

Tokens	编码器时间/s	解码器时间/s	Bits
100	4.0	4.0	293
200	**4.0**	**5.0**	**541**
400	8.0	8.0	956
600	13.0	14.0	1 506
800	17.0	17.0	2 019
1 000	21.0	22.0	2 515

注：加粗字体表示隐写嵌入的最高效率。

表5 文本相似度分析效率

Table 5 Text similarity analysis efficiency

T_{size}	K=1		K=5	
	总时间/s	平均时间/s	总时间/s	平均时间/s
100	0.164 4	0.001 6	0.165 7	0.001 7
200	0.584 6	0.002 9	0.588 2	0.002 9
400	2.421 0	0.006 1	2.447 2	0.006 1
800	9.831 4	0.012 3	9.881 9	0.012 3

表6 多媒体数据流隐蔽信道方案对比

Table 6 Comparison of covert channel schemes for multimedia data streams

方法	标准协议	未修改二进制	不依靠加密	应用隐写	通用性	可证明安全	传输率/bps
Freewave(Houmansadr等,2013)	×	×	×	×	×	×	19 k
SkypeLine(Kohls等,2016)	√	×	√	√	×	×	64
Saenger's(Saenger等,2020)	√	√	×	√	×	×	0.9～2.5 k
Peng's(Peng等,2021)	√	√	√	√	×	×	0.1～20 k
CovertCast(McPherson等,2016)	×	×	×	×	×	×	–
Protozoa(Barradas等,2020)	×	×	×	×	×	×	160～1 400 k
Stegozoa(Figueira等,2022)	×	×	×	√	×	×	2.6～11.4 k
Balboa(Rosen等,2021)	√	√	×	×	√	×	**0.14～8 M**
RoCC(本文)	√	√	√	√	√	√	73～136/310 k

注：加粗字体表示传输率最优结果，CovertCast方法的传输率与网络实时速率相关，“–”表示省略具体数值，“√”和“×”表示是否满足该条件。

他协议。“可证明安全”是指那些达到最优计算安全性的方法。“传输率”表示方法的传输速率。需要特别说明的是，CovertCast的方案通过网站视频流传输数据，主要与网络速度相关，所以表6内不标明传输率。可以看到，在维持协议行为不变，不修改底层二进制程序这两个条件同时满足时，只有Saenger的方法、Peng的方法、Balboa和RoCC。这两个条件决定了通信的隐蔽性，其他方案或者改变协议行为，或者修改底层二进制代码，使得隐蔽通信过程出现异常行为模式，易被攻击者识别。在不依靠加密这个条件上，只有SkypeLine、Peng的方法和RoCC满足，这意味着当攻击者可以解密数据流审查其内容时，依靠加密流的隐蔽信道则会暴露，失去安全性。SkypeLine、Saenger的方法、Peng的方法、Stegozoa和RoCC都应用了隐写术，可以在加密无效时提供更进一步的数据保护。在通用性上，只有Balboa和RoCC达到这个条件，即这两个方案的设计均非针对特定协议，Balboa可以适用于所有受TLS保护的通信协议，RoCC可以适用于所有语音通信协议。在可证明安全条件上，只有RoCC满足，因为RoCC是第1个将可证安全隐写技术应用到构建隐蔽信道上，同时确保了隐蔽性和安全性。在传输率上，RoCC的实时传输率为70～140 bps，满足少量秘密消息的传输需求，而在缓冲扩展模式下传输率达到300 kbps以上，可以用来传输少量文件数据等。

5 结 论

本文研究了现有多媒体数据流隐蔽通信方法的不足，并提出了一种名为RoCC的隐蔽通信方法，具备高隐蔽性、高安全性和强鲁棒性。RoCC是第1个以跨模态方法构建隐蔽信道的工作，直接跨模态方案的缺点在于语音识别无法完全还原文本语义。为解决通信过程和跨模态模型的缺陷导致的文本语义丢失问题，通过文本语义的相似度分析，实现了在常见网络异常状态下的鲁棒通信。相较于现有的具有鲁棒性的Saenger方法，RoCC的丢包率抵抗能力提高了5%以上。此外，本文是第1个将可证安全隐写技术引入构建隐蔽信道的工作。面对国家级攻击者时，RoCC能够在无需流量加密的情况下保持高隐蔽性和高安全性。在长时间实时通信中，RoCC的传输率可达到约100 bps，适合用于少量秘密消息的传输。通过缓冲扩展模式，传输率可达到300 kbps以上，适用于少量文件数据的传输。未来的研究方向可以将跨模态技术扩展至图像、视频等数据模态。例如，引入文本图像检索技术，实现以载密图像为数据传输媒介、以文本作为数据索引的隐蔽信道。

参考文献（References）

Ao J Y, Wang R, Zhou L, Wang C Y, Ren S, Wu Y, Liu S J, Ko T, Li Q, Zhang Y, Wei Z H, Qian Y, Li J Y and Wei F R. 2022. Speecht5: unified-modal encoder-decoder pre-training for spoken language processing[EB/OL].[2023-07-25]. https://arxiv.org/pdf/2110.07205.pdf

Barradas D, Santos N, Rodrigues L and Nunes V. 2020. Poking a hole in the wall: efficient censorship-resistant internet communications by parasitizing on WebRTC//Proceedings of 2020 ACM SIGSAC Conference on Computer and Communications Security. Virtual Event, USA: ACM: 35-48 [DOI: 10.1145/3372297.3417874]

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D. 2020. Language models are few-shot learners//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 1877-1901

Chen K J, Zhou H, Zhao H Q, Chen D D, Zhang W M and Yu N H. 2022. Distribution-preserving steganography based on text-to-speech generative models. IEEE Transactions on Dependable and Secure Computing, 19(5): 3343-3356 [DOI: 10.1109/TDSC.2021.3095072]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. Bert: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2023-07-25]. https://arxiv.org/pdf1810.04805.pdf

Ding J Y, Chen K J, Wang Y F, Zhao N, Zhang W M and Yu N H. 2023. Discop: provably secure steganography in practice based on “Distribution Copies”//2023 IEEE Symposium on Security and Privacy (SP). San Francisco, USA: IEEE: 2238-2255 [DOI: 10.1109/SP46215.2023.10179287]

Figueira G, Barradas D and Santos N. 2022. Stegozoa: enhancing WebRTC covert channels with video steganography for internet censorship circumvention//Proceedings of 2022 ACM on Asia Conference on Computer and Communications Security. Nagasaki, Japan: ACM: 1154-1167 [DOI: 10.1145/3488932.3517419]

Gao Z F, Zhang S L, Lei M and McLoughlin I. 2020. Universal ASR: unifying streaming and non-streaming ASR using a single encoder-

decoder model [EB/OL]. [2023-07-25]. https://arxiv.org/pdf/2010.14099.pdf

Gao Z F, Zhang S L, McLoughlin I and Yan Z J. 2023. Paraformer: fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition [EB/OL]. [2023-07-25]. https://arxiv.org/pdf/2206.08317.pdf

Hopper N J, Langford J and Von Ahn L. 2002. Provably secure steganography//Proceedings of the 22nd Annual International Cryptology Conference Santa Barbara. California, USA: Springer: 77-92 [DOI: 10.1007/3-540-45708-9_6]

Houmansadr A, Riedl T J, Borisov N and Singer A C. 2013. I want my voice to be heard: IP over Voice-over-IP for unobservable censorship circumvention//20th Annual Network and Distributed System Security Symposium. San Diego, USA: The Internet Society: 861-878

Kaptchuk G, Jois T M, Green M and Rubin A D. 2021. Meteor: cryptographically secure steganography for realistic distributions//Proceedings of 2021 ACM SIGSAC Conference on Computer and Communications Security. Virtual Event, Republic of Korea: ACM: 1529-1548 [DOI: 10.1145/3460120.3484550]

Kerckhoffs A. 1883. La cryptographie militaire. Journal des Sciences Militaires, IX: 5-38

Kohls K, Holz T, Kolossa D and Pöpper C. 2016. SkypeLine: robust hidden data transmission for VoIP//Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security. Xi'an, China: ACM: 877-888 [DOI: 10.1145/2897845.2897913]

Lang R L, Xia Y, Zhi Y and Dai G Z. 2004. Analysis and evaluation of several typical steganalysis algorithms. Journal of Image and Graphics, 9(2): 249-256 (郎荣玲, 夏煜, 郅艳, 戴冠中. 2004. 几类典型隐写术分析算法的分析与评价. 中国图象图形学报, 9(2): 249-256) [DOI: 10.3969/j.issn.1006-8961.2004.02.023]

Li F H, Li C Y, Guo C, Li Z F, Fang L and Guo Y C. 2022. Survey on key technologies of covert channel in ubiquitous network environment. Journal on Communications, 43(4): 186-201 (李凤华, 李超洋, 郭超, 李子孚, 房梁, 郭云川. 2022. 泛在网络环境下隐蔽通道关键技术研究综述. 通信学报, 43(4): 186-201) [DOI: 10.11959/j.issn.1000-436x.2022072]

Li Y J and Liu B. 2007. A normalized Levenshtein distance metric. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6): 1091-1095 [DOI: 10.1109/TPAMI.2007.1078]

McPherson R, Houmansadr A and Shmatikov V. 2016. CovertCast: using live streaming to evade internet censorship. Proceedings on Privacy Enhancing Technologies, 2016(3): 212-225 [DOI: 10.1515/popets-2016-0024]

Peng J H, Jiang Y J, Tang S Y and Meziane F. 2021. Security of streaming media communications with logistic map and self-adaptive detection-based steganography. IEEE Transactions on Dependable and Secure Computing, 18(4): 1962-1973 [DOI: 10.1109/TDSC.2019.2946138]

Reimers N and Gurevych I. 2019. Sentence-BERT: sentence embeddings using siamese BERT-networks [EB/OL]. [2023-07-25]. https://arxiv.org/pdf/1908.10084.pdf

Ren Y, Ruan Y J, Tan X, Qin T, Zhao S, Zhao Z and Liu T Y. 2019. FastSpeech: fast, robust and controllable text to speech//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 3171-3180

Rosen M B, Parker J and Malozemoff A J. 2021. Balboa: bobbing and weaving around network censorship//The 30th USENIX Security Symposium. Virtual Event: USENIX Association: 3399-3413

Saenger J, Mazurczyk W, Keller J and Caviglione L. 2020. VoIP network covert channels to enhance privacy and information sharing. Future Generation Computer Systems, 111: 96-106 [DOI: 10.1016/j.future.2020.04.032]

Salton G and Buckley C. 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5): 513-523 [DOI: 10.1016/0306-4573(88)90021-0]

Tian J, Xiong G, Li Z and Gou G P. 2020. A survey of key technologies for constructing network covert channel. Security and Communication Networks, 2020: #8892896 [DOI: 10.1155/2020/8892896]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang Y X, Skerry-Ryan R J, Stanton D, Wu Y H, Weiss R J, Jaitly N, Yang Z H, Xiao Y, Chen Z F, Bengio S, Le Q, Agiomyrgiannakis Y, Clark R and Saurous R A. 2017. Tacotron: towards end-to-end speech synthesis [EB/OL]. [2023-07-25]. https://arxiv.org/pdf/1703.10135.pdf

Zhang W M, Wang H X, Li B, Ren Y Z, Yang Z L, Chen K J, Li W X, Zhang X P and Yu N H. 2022. Overview of steganography on multimedia. Journal of Image and Graphics, 27(6): 1918-1943 (张卫明, 王宏霞, 李斌, 任延珍, 杨忠良, 陈可江, 李伟祥, 张新鹏, 俞能海. 2022. 多媒体隐写研究进展. 中国图象图形学报, 27(6): 1918-1943) [DOI: 10.11834/jig.211272]

作者简介

张晏铭，男，硕士研究生，主要研究方向为隐蔽通信。E-mail：azesinter@mail.ustc.edu.cn

陈可江，通信作者，男，副研究员，主要研究方向为信息隐藏与人工智能安全。E-mail：chenkj@ustc.edu.cn

丁锦扬，男，硕士研究生，主要研究方向为信息隐藏、人工智能安全和隐私保护。E-mail：source@mail.ustc.edu.cn

张卫明，男，教授，主要研究方向为信息隐藏、数字水印、对抗样本、深度伪造与检测。E-mail：zhangwm@ustc.edu.cn

俞能海，男，教授，主要研究方向为图像处理、信息隐藏和数据安全。E-mail：ynh@ustc.edu.cn