基于智能语音系统的隐蔽通信设计与实现

宋礼鹏[1)] 张瑞宸[1)] 刘文翔[1)] 王硕[1)] 张烨[1)] 任冠瑜[1)]

[1)](山东大学（威海） 机电与信息工程学院, 威海 264200)

摘 要 深度学习的发展促使对抗样本攻击的产生, 而传统的通讯加密技术和物理层安全技术都无法很好的解决隐私问题。当前诸多团队的成果，证实了选择性对抗样本具有信息隐藏的特点，可以很好的用于隐蔽通信的构建，且此种隐蔽通信模式区别与基于规则和密码学的传统隐蔽通信模式，是对隐蔽通信构建的一种新的启发。基于该特性，本团队将对抗样本攻击技术应用于隐蔽信道的构建。基于 DeepSpeech 深度学习开源语音转文本引擎，利用梯度下降产生语音选择性对抗样本。并在隐蔽模型和表层模型结合下, 实现隐蔽通信, 且效果良好。此种隐蔽通信模式区别与基于规则和密码学的传统隐蔽通信模式，是对隐蔽通信构建的一种新的启发，具有较高研究价值。

关键词 通讯加密技术 选择性对抗样本 信息隐藏 隐蔽通信

Design and Implementation of Covert Communication based on Intelligent Voice System

Lipeng Song[1)] Ruichen Zhang[1)] Wenxiang Liu[1)] Shuo Wang[1)] Ye Zhang[1)] Guanyu Ren[1)]

[1)](School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209)

Abstract The development of deep learning promotes the emergence of anti-sample attacks, while the traditional communication encryption technology and physical layer security technology can not solve the privacy problem. The results of many teams at present confirm that selective adversarial samples have the characteristics of hidden information, and can be used in the construction of covert communication. This covert communication mode is different from the traditional covert communication mode based on rules and cryptography, which is a new inspiration for the construction of covert communication. Based on this feature, the counter sample attack technique is applied to the construction of covert channel. An open source speech-to-text engine based on DeepSpeech deep learning uses gradient descent to generate speech selective adversarial samples. The covert communication is realized by the combination of covert model and surface model, and the effect is good. This kind of covert communication mode is different from the traditional covert communication mode based on rules and cryptography, which is a new inspiration for the construction of covert communication and has high research value.

Key words communication encryption technology; selective confrontation samples; information hiding; covert communication

---

宋礼鹏(通信作者)，男，1975年生，博士生导师，教授，主要研究领域为人工智能安全、软件漏洞分析.E-mail: slp880@sdu.edu.cn. 张瑞宸，男，2002年生，本科生，E-mail: 3033824368@qq.com. 刘文翔，男，2002年生，本科生，E-mail: 2425281352@qq.com. 王硕，男，2002年生，本科生，E-mail: 2404266246@qq.com.张烨，男，2002年生，本科生，E-mail: 1498792816@qq.com. 任冠瑜，男，2002年生，本科生，E-mail: guanyuren@mail.sdu.edu.cn.

1 引言

最近，深度神经网络（deep neural network, DNN）[1]在图像识别[2]、语音识别[3]、模式分析[4]和入侵检测[5]方面提供了更好的性能。尽管 DNN 效果较好，但很容易受到的对抗样本[6]的影响。在样本中添加一些不易于察觉的扰动信息，并指明扰动对于不同模型影响的结果，即选择性对抗样本，可以让特定参数条件下的神经网络模型做出错误的预测。

对抗样本在图像领域得到了广泛的研究。然而近年来，对于对抗样本的研究已经扩展到语音领域。音频领域的许多场景证实了对抗样本的潜在威胁。例如，由亚马逊开发的 Alexa [7]可以提供在用户经过认证后支持语音交互的服务，比如订购产品、获取信息和控制多个智能设备。然而，这样的语音交互可能存在一些问题。例如，它们可能会向其他用户泄露用户的个人隐私，或在听到电视或广播的声音后错误地订购一件物品。为了利用语音识别系统的这些弱点，许多研究提出了对抗样本，并通过向原始音频样本添加少量的噪声生成对抗样本。

当前诸多团队在智能语音攻击和防御领域都取得了一定成果，经过对相关领域论文研读讨论，并深入研究对抗样本特性后发现，现有的选择性对抗样本具有隐藏信息的特点，适用于涉及军事通信、自动电话窃听和隐蔽信道[8]的场景。例如，军事通信员需要发送特定的语音信息时，可以故意使敌人误解信息，同时使友军正确识别信息。

基于选择性对抗样本具有隐藏信息的特点，可以用于隐蔽通信，有安全性高的特点，团队提出将音频对抗样本攻击与隐蔽通信思想结合的方案。此种隐蔽通信模式区别与基于规则和密码学的传统隐蔽通信模式，是对隐蔽通信构建的一种新的启发，具有很高的研究价值。

团队通过音频对抗模型，将实际需要发送的内容，隐藏到音频文件。该方法具有如下特点：

1. 安全性高。有通讯双方同时掌握相同模型与模型参数时，才能正确的解析隐藏信息。
2. 隐蔽性强。即使中途遭到截获，也无法发现隐藏信息。

2 相关工作

隐蔽通信技术已经成为信息安全领域的一个研究热点，目前基于数字图像的隐蔽通信技术可以说已经比较完善和成熟，而语音隐蔽通信技术的相关研究却起步不久。近年来，语音模型的攻击技术被各团队不断完善改进，因此基于该对抗攻击技术的语音隐蔽通信技术开始逐步发展。

修改文本格式的文本信息隐藏方法，主要通过修改载体文本格式，实现信息隐藏。近些年，该类方法的改进大多是减小文字缩放比，并在载体文本中更加均匀的嵌入秘密信息，来增加含密文本的隐蔽性和鲁棒性，这类方法在视觉上具备较高的隐蔽性。但是，如果在文本传递过程中遇到修改文本格式或将文本重新录入的操作，就会丢失秘密信息。

Partala[9]首次尝试利用区块链作为媒介构建隐蔽通信信道，并提出区块链隐蔽信道（BLOCCE, blockchain covert channel）模型，将信息隐藏至交易地址的最后一位，并顺序使用对应循环生成的交易地址使其保证秘密信息的顺序性。此后，研究者针对该模型的不足进行改进提出改良的区块链隐蔽通信方法，并尝试降低其通信成本[10,11]。

Torki 提出了一种不需要人工改变源数据，只需重复执行隐藏算法的区块链隐蔽通信方案。Guo 等[12]与蓝怡琴等[13]通过结合多层可链接自发匿名群签名实现混合并引入新的椭圆曲线算法门罗币，在其区块链应用中构建隐蔽通信信道，利用了门罗币具备的高安全性以提高隐蔽信道的隐蔽性。

对于基于深度学习的隐蔽通信，Szegedy 等人[6]首先提出了一个对抗性样本，其中攻击者轻微地变换了图像。使用对抗性样本的主要目的是通过在原始图像中添加少量噪声来导致 DNN 出错；然而，人类无法区分原始图像和扭曲图像之间的差异。

基于深度学习的图像隐蔽通信算法，一般包括了基于卷积神经网络（Convolutional Neural Networks, CNN）和基于生成对抗网络（Generative Adversarial Networks, GAN)的两种信息隐藏算法，而基于 GAN 的算法通常采用了 CNN 的变种。很多学者基于信息隐藏和隐写分析之间的对抗特性，对深度学习的模型结构进行了有目标的优化工作，以实现对秘密数据的嵌入与提取。目前的研究现状显示，基于深度学习的图像信息隐藏方法相较于传统方法隐藏效果和性能指标更好。

在音频领域，“恶魔音乐”攻击，是指 2018 年

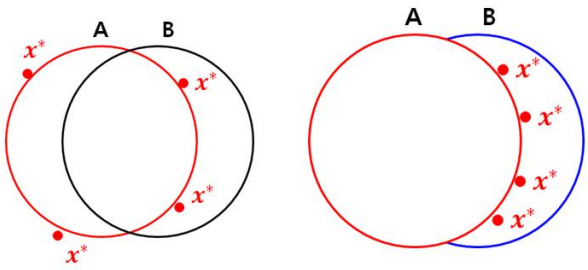
中国科学院信息工程研究所陈恺研究团队利用音乐为载体的音频对抗样本生成研究[15]，这种恶意音频样本将隐含的指令插入正常的音乐中，来达到攻击智能语音系统的目的。

语音-文本对抗样本攻击，是指谷歌大脑人工智能科学家 Carlini，及其团队所研究的语音对抗性样本生成方式。该攻击方式是在智能语音识别系统识别音频文件过程中，引导语音识别模型将语音样本错误的识别为任意指定文本[16]。

经过对上述文章及相关工作的重点分析，团队决定将音频对抗样本攻击与隐蔽通信思想结合，实现通过针对音频转文本引擎设计的有目标攻击，将用户发送的内容，隐藏入一个音频文件之中。只有通讯双方同时掌握相同模型与模型参数时，才能正确的解析隐藏信息，即使中途遭到窃听者有目的的截获，也无法发现并破译隐藏信息。

3 背景

对抗样本是通过故意对数据集中输入样例添加难以察觉的摄动，使模型以高置信度给出一个错误的输出。即对抗音频样本只需要在一段音频上做微小的扰动，分类器就会以很高的置信度将音频错误转写，甚至被转写成一段指定的文本（不是音频正确转写的文本）。出现这种情况的原因是，神经网络很容易受到“欺骗”。在图 1(a) 中，模型 A 是具有神经网络的目标模型。对应的线是目标模型 A 的决策边界。如果样本在目标模型 A 的边界内，则该样本会被目标模型 A 正确分类。图 1(b) 是被目标模型 B 正确分类但被目标模型 A 错误分类的选择性对抗样本的示例。在图 1(b)中，选择性对抗样本 x* 在目标模型 B 的决策边界内，但偏离了目标模型 A 的决策边界。



(a) 对手的例子。(b) 朋友安全对抗的例子。

图 1：可转移性示例：单个敌方目标模型 A 和友方目标模型 B。

对抗样本具有可迁移性是 Szegedy 等人[6]在文献中首次提出的,对抗样本的可迁移性是指对抗样本被 A 模型错误分类, 也同样可以被 B 模型错误分类，对抗样本的迁移属性意味着攻击者可以不用直接接触基础模型，而选择攻击一个机器学习模型使样本被错误分类。Szegedy 等人[6]在相同数据集上研究了不同模型的迁移性，此外，还在数据不相交子集上训练相同或不同模型并研究其间的迁移性问题，但不足的是，其实验成果都是在 MNIST 数据集上实现的。Goodfellow 等人[17]提出对抗样本在不同模型间的泛化能力是对抗干扰与模型的向量高度一致所致，所以当训练相同的任务时，对手可以在不同的模型上学习相似的函数。这种泛化特性意味着，若敌人要对模型进行攻击，无需访问目标模型，仅将自身模型训练产生的对抗样本送至目标模型中即可实现。

团队通过进一步研究和探讨，发现对抗样本的可迁移性可以很好的用于隐蔽通信的构建。因此，需要一个新的体系架构，它存在一个无需访问的目标模型和一个用于攻击的自身模型，并使用自身模型来生成对抗样本。

4 模型

在本节中，将详细介绍选择性音频对抗样本和信息隐藏模型架构。

4.1选择性音频对抗样本

在图 2(a)中，模型 A 是基于神经网络的隐蔽模型。图中曲线是该隐蔽模型的决策边界。若样本在隐蔽模型决策边界内，则表示该样本被隐蔽模型正确分类。反之，分类错误。在图 2(b)中，模型 B 是基于神经网络的保护模型，与模型 A 的分类决策相同。其中样本 x*在模型 A 和模型 B 决策边界之间的情况，表示样本 X*被模型 A 分类错误而被模型 B 分类正确。这就引出了选择性对抗样本的概念。

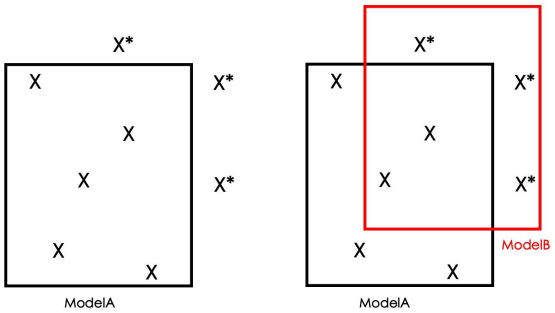


图 2 单个选择性对抗样本 x，隐蔽模型 A 和保护模型 B

图 2(b)显示的选择性对抗样本 x*在模型 A 的决策边界之外，同时，x*在模型 B 的决策边界之内。但值得注意的是 x*应尽可能地靠近模型 A 的决策边界，这意味着原始样本 x 与选择性对抗样本 x*的距离度量较近。此时，进一步在决策边界上收敛到最小距离，则得到的选择性对抗样本噪声最小，且具有选择对抗性。

4.2 模型架构

该模型架构的前提条件对隐蔽模型和保护模型的所有参数已知。信息隐藏模型架构如下图所示:

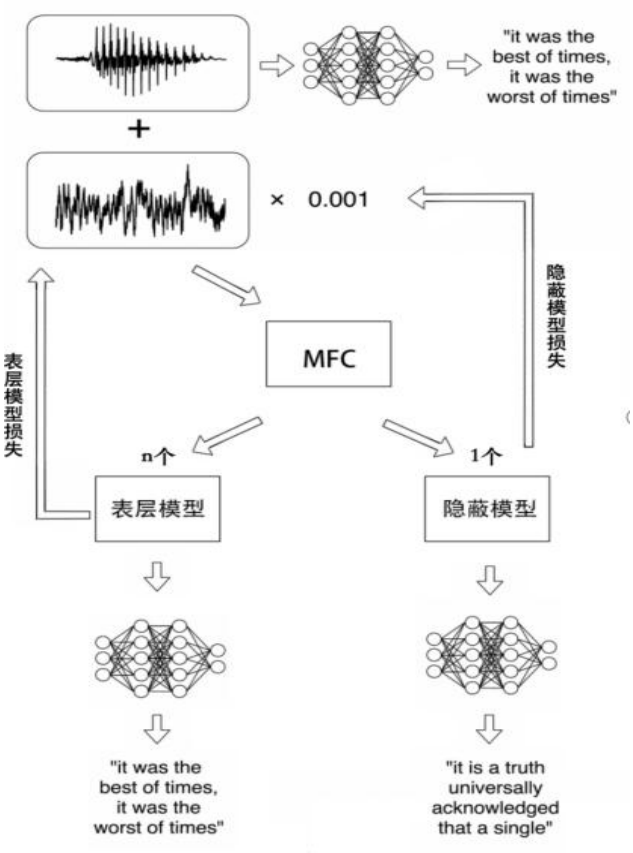


图 3　生成选择性对抗样本的模型架构图

从图（3）中可以看到生成选择性对抗样本的架构由 n 个保护模型（表层模型）Mp_i(1<i<=n)和一个隐蔽模型 Mh 构成，隐蔽模型同样可以扩展为 n 个，但考虑到产品实用性，一个专有隐蔽模型即可。

整体模型架构的输入为原始样本 x，初始噪声 delta 以及正确分类短语标签 org_phrase 和隐蔽分类短语标签 target_phrase。输入后，架构随机生成初始噪声，加入原始样本，生成初始对抗样本，而后进行梅尔倒谱，在保护模型和隐蔽模型中进行前向传播。梅尔倒谱（MFC）用作预处理步骤以降低输入 x 的维数。MFC 以每秒 50 帧的速度分割波形，并将每帧映射到每个频域。

Mp_i 和 Mh 在生成选择性对抗样本的过程中保持模型不变（生成过程中只需要模型预训练参数和模型结构）。Mp_i 和 Mh 得到音频数组输入后，Mp_i 输出所有保护者模型的分类概率，Mh 输出隐蔽模型的分类概率。其次使用 CTC 损失函数将分类概率矩阵和标签对齐，得到每个 Mpi 模型损失值和 Mh 的损失值，最后将 Mp_i 加权损失值和 Mh 的损失值进行加权后，反馈给加入的噪声形成闭环，进一步调整噪声直到满足预定的选择对抗程度。

如 4.1 节选择性对抗样本所示，该模型架构最终目的是在保证各个保护模型 Mp_i 分类正确且隐蔽模型 Mp 分类错误的前提下，不断缩减原始样本 x 和选择对抗样本 x^*之间的距离。从而可以设置目标函数，设 $f^p(x^*)$ 为保护模型的识别行为，$f^h(x^*)$为隐蔽模型的识别行为,org_phrase 是原始短语，target_phrase 是目标隐藏短语。目标函数如下：

$$x^*: \ \text{argmin}(x,x^*) \text{ s.t. } f^p_i(x^*) = \text{org\_phrase and } f^h(x^*) = \text{target\_phrase}$$

$$x^* = x + \text{delta}$$

针对目标函数中使保护模型和隐蔽模型分别正确分类给定的目标短语，并需要噪声 delta 尽可能小，加入的噪声调整需要损失函数的反馈，损失函数总反馈函数如下：

$$\text{Loss}_T = C_{dis}\,\text{Loss}_{distortion} + C_{Mp} \sum_{i=0}^{m} w_i \ \text{Loss}(Mp_i) + C_{Mh}\,\text{Loss}(Mh)$$

其中 C_{dis}、C_{Mp}、C_{Mh} 分别是在总损失中噪声，保护模型和隐蔽模型损失的权重分布，可以使生成的选择性对抗样本有更注重某一方面的特性。W_i 是不同保护模型的损失权重，可以根据不同保护模型攻击的难易程度来调整权重的分布。

三种损失函数的选择：对于噪声的损失偏差 $\text{Loss}_{distortion}$，团队采用的是 L2 范数对音频数组进行距离之间的度量。对于 $\text{Loss}(Mp_i)$和 $\text{Loss}(Mh)$，团队采用的是具有自动对齐特点的 CTC 损失函数。

5 实验

在 DEEPSPEECH 模型上对一段英文音频进行隐藏信息“this is a test”的测试，并多次重复以观

察隐藏效果和耗时情况。示范音频使用 github 上 N.Carlini 的 carlini/audio_adversarial_examples 中的音频样本，该样本是内容为英文语句“without the dataset the article is useless” 102.2kb 大小的 wav 文件，实验在安装有 Python 3.6，tensorflow-gpu==1.15.4 的 UBUNTU 20.04 LTS 下进行，耗时情况测试中使用 CUDA11.4 调用显示适配器英伟达 RTX2070 SUPER 进项硬件加速，并统计每一次测试时成功识别（模型第一次正确识别出隐藏信息）耗时、完全收敛（CTCloss < 0.005 时认为完全收敛）耗时、60000 轮迭代耗时这三组数据。

表 1　不同要求下识别耗时 单位 秒/S

序号	成功识别耗时	完全收敛耗时	60000th 迭代耗时
1	38.33	357.39	423.62
2	37.12	345.46	368.00
3	36.51	344.41	375.94
4	32.38	352.22	369.32
5	35.25	370.01	398.83

由上表可知，模型成功识别耗时较短以上五组平均成功识别耗时仅为 35.92s，而平均的完全收敛耗时则达到 353.90s，两者差异较大主要由最小化噪声这一过程所需时间长导致，由此可知在不考虑噪声情况下可以快速制作隐藏信息的音频，但此时的音频带有人耳可听见的类似电流的“刺刺”声，若要除去此类噪声需要进行大量迭代所需时间过长，目前在快速减小噪声方面仍有很大改进空间。

5.1　添加噪声对原始音频影响

接下来对于完全收敛音频中的噪声做进一步的分析评估，隐藏信息前后音频样本的波形图如下：

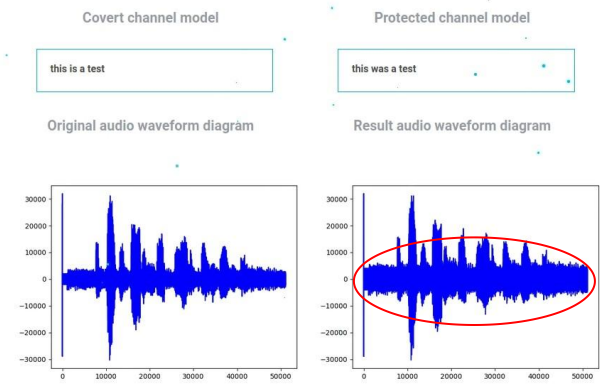


图 4　干扰前后音频波形变化

如图 4 所示，隐藏前后样本波形图基本一致，尤其高振幅部分肉眼难见显著差异，但低振幅部分出现略微增多和模糊的现象，该现象对于人耳听觉影响为音频听感发糊，能略微听到电流声般的噪音，为进一步探究对于人耳听觉影响，选取 10 位志愿者进行双盲实验评估，分别选取短英文音频、长英文音频、白噪声各十组在其中隐藏短语“this is a test”，每组均运行 350 秒左右后完全收敛。

5.2　对于信息隐藏效果的评估

短音频组	长音频组	白噪声
1.this is a test	1.this is a test	1.this is a test
2.this is a test	2.this is a test	2.this is a test
3.this is a test	3.this is a test	3.this is a test
4.this was sist	4.this is a test	4.this is a test
5.this is a test	5.this is a test	5.this is a test

图 5　音频识别结果

如图 5 所示，除短音频组有一个样本识别出错外其余所有样本均被模型成功识别出隐藏信息“this is a test”，信息隐藏效果良好。

问卷中包含干扰前音频 20 组，要求志愿者判断音频是否经过干扰，问卷提供干扰，未干扰，无差异三个选项，

结果显示：

9 名志愿者不能够准确分辨出干扰前后音频差异（正确选出 20%以上的干扰前后音频）。

8 名志愿者不能够大致分辨出干扰前后音频差异（正确选出 10%以上的干扰前后音频）。

同时有 9 人表示音频前后差异极小，难以分辨。

综上所述，目前模型已能够做到制造人耳难以分辨是否干扰的音频样本，但此类样本制造时间过长，在减少噪音消减时间和进一步减小低频噪音方面仍有很大改进空间。

6 总结

选择性对抗样本具有信息隐藏的特点，基于这一特点，团队将对抗样本攻击技术应用于隐蔽信道的构建。该方法基于 DeepSpeech 语音转文本引擎，

使用梯度下降生成语音选择性对抗样本，将实际需要发送的内容，隐藏到音频文件中。采用隐蔽模型与表面模型相结合的方法实现了信息隐藏，取得了较好的效果。这种信息隐藏模式不同于传统的基于规则和密码学的信息隐藏方式，为信息隐藏的构建提供了新的启示，具有很高的研究价值。

目前本项目研究存在产生对抗样本时间过长、噪音较大的问题。未来将着力于提高对抗样本产生速度、缩短噪音消减时间和进一步减弱低频噪音。

参 考 文 献

[1] J. Schmidhuber, “Deep learning in neural networks: An overview,” Neural networks, vol. 61, pp. 85–117, 2015.

[2] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in International Conference on Learning Representations, 2015.

[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.

[4] R. Collobert and J. Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in Proceedings of the 25th international conference on Machine learning, pp. 160– 167, ACM, 2008.

[5] S. Potluri and C. Diedrich, “Accelerated deep neural networks for enhanced intrusion detection system,” in Emerging Technologies and Factory Automation (ETFA), 2016 IEEE 21st International Conference on, pp. 1–8, IEEE, 2016.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in International Conference on Learning Representations, 2014.

[7] “https://developer.amazon.com/alexa”

[8] M. Smeets and M. Koot, “Covert channels,” research report for RPI university of Amsterdam MSc in System and Network Engineering, 2006

[9] PARTALA J. Provably secure covert communication on blockchain[J]. Cryptography, 2018,2(3)：18.

[10] 宋上，彭伟．BLOCCE+：一种改进的基于区块链的隐蔽通信方法[J]，重庆理工大学学报(自然科学)，2020,34(9):238-244.

[11] ZHANG L J, ZHANG Z J, WANG W Z, et al. A covert communication method using special bitcoin addresses genera ted by vanitygen[J]. Computers, Materials & Continua, 2020, 65(1): 597-616.

[12] GUO ZZ, SHI L C, XU M Z, et al. MRCC: a practical covert channel over monero with provable security[J]. IEEE Access, 2021, 9:31816-31825.

[13] 蓝怡琴，张方国，田海博．利用门罗币实现隐蔽通信[J]. 西安电子科技大学学报，2020,47(5)：19-27.

[15] XuejingYuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, ShengzhiZhang, Heqing Huang, Xiaofeng Wang, and Carl A. Gunter, "CommanderSong: ASystematic Approach for Practical Adversarial Voice Recognition," InUSENIX Security Symposium, pages 49–64, 2018.

[16] N. Carlini and D. Wagner, "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text," 2018 IEEE Security and Privacy Workshops (SPW), 2018, pp. 1-7, doi: 10.1109/SPW.2018.00009.

[17] Tramèr F，Papernot N，Goodfellow I，et al. The space of transferable adversarial examples. arXiv preprint arXiv: 1704.03453，2017