面向微控制单元的高效语音隐私保护编码器

蔡栋琪[1,2],王尚广[1,2*],张泽凌[1,2],马 骁[1,2],徐梦炜[1,2]
(1. 北京邮电大学计算机学院,北京 100876;2. 网络与交换技术全国重点实验室,北京 100876)

摘 要: 语音是现有嵌入式移动设备广泛使用的一种输入接口. 尽管现有的云端服务提供商提供了强大的语音语言理解(Spoken Language Understanding,SLU)服务,但也对用户隐私造成了极大的威胁. 为此,基于信息解耦的隐私保护编码器被提出,以在不影响SLU功能的前提下,从语音信号中移除敏感信息. 然而,这些编码器往往需要较高的内存和复杂的计算,因而在资源受限的小型设备上难以实际应用. 本文基于大量实验观察到了一个关键现象,即SLU依赖于整个语句的全局信息,而隐私敏感词的识别则多为局部信息依赖. 利用这一观察,我们提出了一个面向语音意图理解的高效编码器(SImpLe ENCodEr designed for efficient privacy-preserving SLU offloading,SILENCE)系统. 我们在STM32H7微控制单元上实现了该系统,并在不同的攻击场景下评估了其效果. 实验结果表明:SILENCE在语音意图提取任务的性能和隐私保护能力上可与传统隐私保护编码器媲美,同时实现了高达53.3倍的速度提升和134.1倍的内存占用减少,首次在内存仅有1 MB的微控制单元上实现了隐私保护的SLU服务.

关键词: 语音语言理解(SLU);资源受限终端;隐私保护;微控制单元;语音意图提取;内存优化
基金项目: 国家自然科学基金(No.62032003, No.U21B2016, No.62425203);中国科协青年人才托举工程(No.2023QNRC001)
中图分类号: TP31;TP36 **文献标识码:** A **文章编号:** 0372-2112(2025)08-2601-13
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20241154

Efficient and Privacy-Preserving Spoken Language Understanding for Resource-Constrained Microcontroller Unit

CAI Dong-qi[1,2], WANG Shang-guang[1,2*], ZHANG Ze-ling[1,2], Ma Xiao[1,2], XU Meng-wei[1,2]
(1. *Department of Computer Science*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*;
2. *State Key Laboratory of Networking and Switching Technology*, *Beijing* 100876, *China*)

Abstract: Speech input is increasingly adopted as an intuitive interface for various embedded mobile devices. Cloud-based solutions provide powerful speech language understanding (SLU) capabilities but introduce privacy risks, as sensitive information may be processed remotely. To address these concerns, disentanglement-based encoders have been developed to strip sensitive data from audio signals, allowing SLU without compromising privacy. However, such encoders are often memory-intensive and computationally demanding, limiting their practicality on resource-constrained devices. Based on extensive experiments, this paper observes a key phenomenon: SLU relies on global information from the entire sentence, whereas the recognition of privacy-sensitive words predominantly depends on local information. We implemented simple encoder designed for efficient privacy-preserving SLU offloading (SILENCE) on an STM32H7 microcontroller and evaluated its performance under various privacy threat scenarios. Results demonstrate that SILENCE provides competitive speech intent classification accuracy and privacy protection compared to more complex encoders. Simultaneously, it achieves a speedup of up to 53.3 times and a reduction in memory footprint by 134.1 times, marking the first time that privacy-preserving SLU services have been realized on a microcontroller with only 1 MB of memory.

Key words: spoken language understanding (SLU); resource-constrained devices; privacy-preserving; microcontroller unit; speech intent classification; memory efficient
Foundation Item(s): National Natural Science Foundation of China (No. 62032003, No. U21B2016, No. 62425203); Young Elite Scientists Sponsorship Program by CAST under Grant (No.2023QNRC001)

收稿日期:2024-12-23;录用日期:2025-04-15;责任编辑:覃怀银
*通讯作者:王尚广

1 引言

随着移动端设备的智能化演进，越来越多的语音数据正在被上传到云端，以运行在线的语音语言理解（Spoken Language Understanding，SLU）任务处理[1]，其中比较典型的应用场景包括在文字输入不便的轻量级控制单元上配备语音助手[2,3]，例如家庭音响[4]、智能手表[5]、远程医疗传感器[6]以及智能工厂传感器[7]等等．然而，如图1（a）所示，将原始语音信号暴露于云端引发了一系列隐私问题[8]．研究显示，为了提升Siri意图识别的准确性，Apple下游承包商会定期监听用户的语音录音（https://www.theguardian.com/technology/2019/jul/26/apple-contractors-regularly-hear-confidential-details-on-siri-recordings），其中包括私人谈话、医疗信息，甚至涉及一些亲密场景的内容．

基于云服务器的SLU功能存在多方面的潜在隐私泄露风险．其中，生物特征方面的相关隐私泄露问题已经被广泛研究，并已经可以在不影响SLU任务准确性的前提下，有效去除与说话人身份特征相关的信息[9,10]；然而，语音信号中的具体文本内容信息的保护，尤其是敏感词的保护，由于与语音意图理解任务本身深度耦合，尚未得到很好的解决．

近年来，基于解耦表示学习的隐私保护编码器[11~13]由于不受限于复杂的密码学运算[14~16]和外部物理设备的干扰[17~19]，成为了现在主流的语音隐私保护编码方式，其示意图如图1（b）所示．这些编码器通过诸如wav2vec[1,12]、Conformer[13,20]以及Preformer[11,21]等预训练好的声学模型提取语音表示，并通过对抗训练的方式[22]进一步促进语音信息的解耦．例如，隐私保护SLU（Privacy-Preserving Spoken Language Understanding，PPSLU）方法[11]采用了一个基于12层Transformer的Preformer作为其编码器，将语音信号分为语音意图、说话人身份信息等不同的特征层．

虽然基于解耦的编码器能够有效保护用户隐私，但它们需要大量的计算资源，通常需要超过数十GFLOPs才能实现有效的信息解耦[23]．而且，这些编码器对内存有较高要求，通常包含数千万模型参数．因此，它们并不适合直接部署在内存有限的嵌入式微处理单元上．此外，为了针对每个特定的SLU任务，需要生成不同的解耦模型，这往往需要耗时的对抗训练．这一特点限制了其对新出现SLU任务的灵活性和可扩展性．更多相关细节将在第2节中进行详细的说明．

本文的目标是在资源受限设备（如仅配备1 MB内存的STM32H7微控制单元）上实现实时、安全的SLU任务卸载．为实现这一目标，我们提出一种新型的编码器设计方案，该设计既要轻量化，又能够有效过滤敏感信息，其设计示意图如图1（c）所示．

我们提出了一个面向语音意图理解的高效编码器（SImpLe ENCodEr designed for efficient privacy-preserving SLU offloading，SILENCE）系统．该方法基于我们的一项独特实验观察：语音信息具有非对称依赖关系性．具体来说，在SLU任务中，意图提取任务（例如场景识别）通常需要整个语句的全局依赖关系知识，而语音识别任务（例如识别某个具体单词或短语）则更多地依赖于短期依赖关系知识，我们在第3节中对这一观察进行了详细的原理分析与实验验证．基于此，SILENCE创新性地提出在处理过程中将语句合理划分为多个片段，并选择性地掩蔽大部分片段，以通过模糊短期细节来达到隐私保护的目的，同时不会显著损害全局依赖关系，即不会影响后续语音意图提取任务的处理．处理后的音频波形随后被传输到云端进行准确的SLU意图分析．此外，我们结合了可解释性学习方法，基于文献[24]设计了一种差分掩码生成器，通过自动识别需要遮蔽的片段数量及具体位置来优化系统性能．

（a）资源受限终端上进行语意理解任务计算卸载时的隐私问题

（b）现有语音隐私保护编码方式

（c）本文方案：一种基于不对称信息依赖的新型编码器

图1 资源受限设备上SLU任务时的隐私问题及其隐私保护方法的示意图

2 相关工作及研究背景

2.1 隐私保护的语义理解任务

SLU是现代语音辅助系统的关键组成部分，负责理解人类语音并将其转换为结构化、可执行的指令．例如，当用户说“Set a meeting for tomorrow at 10 AM”时，SLU系统可能会将其映射为一个结构化意图，以便后续任务执行，如{scenario：Calendar，action：Create entry}．目前，意图、场景分类任务是SLU语音理解研究的主要研究目标，并具有广泛的应用场景[25~31]．

2.1.1 SLU系统的演进

SLU系统的演进经历了从传统的双组件系统，包括自动语音识别（Automatic Speech Recognition，ASR）和自然语言理解（Natural Language Understanding，NLU）两个模块，到现代端到端一体式神经网络的转变[32,33]．这些先进系统绕过了中间的文本表示，直接将语音信号

映射到其语义意义,从而提高了效率并减少了误差传播.

一个典型的端到端SLU模型包含一个编码器(通常结合了卷积和基于注意力的元素)和一个解码器(包括Transformer解码器和连接时序分类解码器). 许多SLU模型会借鉴预训练的ASR模型(如HuBERT(Hidden unit Bidirectional Encoder Representations from Transformers)[30])中的编码器部分,并将原始ASR文本解码器替换为适用于SLU意图理解任务的解码器.

2.1.2　攻击模型

本文的攻击模型与先前的研究保持一致[11,12]:假设用户(服务使用者)主动将音频数据传输到云服务器(潜在攻击者)以完成预定的SLU任务. 在接收到数据后,攻击者可能利用ASR技术转录音频,并识别其中的隐私敏感词[8,21,31]. 需要注意的是,这些转录结果往往极为详细,包含的信息远超用户原本希望披露的内容. 本文的目标是确保用户能够从潜在恶意的云服务提供商处可靠获取准确的SLU意图识别结果,同时防止服务提供商通过交互过程中的语音转录文本,推断出用户的敏感信息或关键词.

2.2　当前方法的局限性

2.2.1　传统隐私保护方法

当前语音隐私保护编码技术主要可分为三类:基于密码学的方法、基于物理设备的方法以及基于信息解耦的编码方法.

(1)基于密码学的隐私保护方法

此类方法通过数学加密机制实现隐私保护,主要包括同态加密(Homomorphic Encryption,HE)[14]和多方安全计算(secure Multi-Party Computation,MPC)[15]等方案. HE允许在加密数据上直接进行计算,而MPC通过分布式协议,确保各参与方无法获取其他方的原始数据. 尽管这些方法在理论上具备严格的安全性保证,但其计算复杂度呈现指数级增长特征. 以基于MPC的私有统一模型架构(Private Unified Model Architecture,PUMA)方案[16]为例,完成单次语音推理需耗时5 min,难以满足实时性需求. 此外,可信执行环境(Trusted Execution Environment,TEE)作为硬件级加密方案,虽能通过隔离内存空间保障数据安全,但其安全区域的有限存储容量(通常远小于512 MB)无法承载现代语音语义理解模型的推理开销,导致该方法在边缘计算场景中适用性受限.

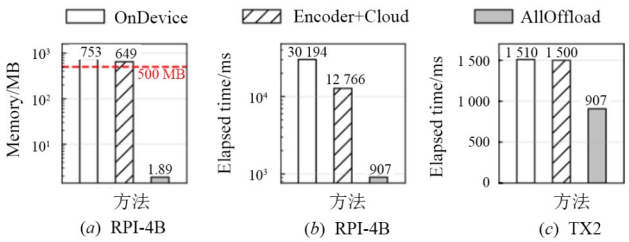
(2)基于物理设备的干扰方法

该方法通过物理手段对语音信号进行干扰,典型实现包括超声波麦克风干扰器(Ultrasonic Microphone Jammer,UMJ)[17,19]和语音转换技术. UMJ通过注入非线性噪声干扰窃听设备,但会不可逆地破坏语音的语义信息,导致后续语义理解任务失效. Preech方案[18]尝试结合语音转换与生成式预训练变换器(Generative Pre-trained Transformer,GPT)生成噪声,虽能保留部分语义特征,但其复杂的生成对抗网络在轻量级设备上存在严重的计算延迟. 这类方法的共同缺陷在于无法在隐私保护与语义保留之间建立有效平衡.

(3)基于解耦的编码器方法[11,12]

该技术路线通过构建分层表征实现敏感信息剥离,可分为基础解耦方法与增强解耦方法两类. 基础方法如变分自动编码器方案[12]通过对抗训练方法启发模型学习不同任务的信息表征,其信息分离层次表浅,主要用于分离说话人身份信息,难以应对复杂语义场景. 针对此缺陷,PPSLU[11]引入深度神经网络构建多级解耦编码器,通过注意力机制实现更细粒度的信息分离. 这些方法旨在建立分层的语音信号表示,从原始语音中解耦分离出敏感数据. 但是,其性能依然不足以支撑在资源受限的微处理单元上有效部署.

为了揭示传统方法的性能问题,我们进行了初步实验,测量基于信息解耦的隐私保护编码器在Raspberry PI 4B(RPI-4B)(https://www.raspberrypi.com/products/raspberry-pi-4-model-b)和Jetson TX2(TX2)(https://developer.nvidia.com/embedded/jetson-tx2)上的资源消耗. 本文的观察结论如下:基于解耦的隐私保护SLU系统资源消耗过高,难以实际部署. 如图2所示,解耦编码器在RPI-4B上完成一次推理需要消耗648.7 MB内存,并花费12.8 s. 即使是在配备图形处理单元(Graphics Processing Unit,GPU)的高性能TX2上,编码器完成一次信息加密仍需要593.0 ms. 考虑到网络延迟,基于解耦的SLU语意卸载系统的端到端延迟相比On-Device推理(即无卸载)仅节省了0.7%的整体时间,而内存占用依然超过500 MB.



注:编码对象为4 s音频.

图2　基于解耦编码器方法[11]的成本分析

2.2.2　启示与思考

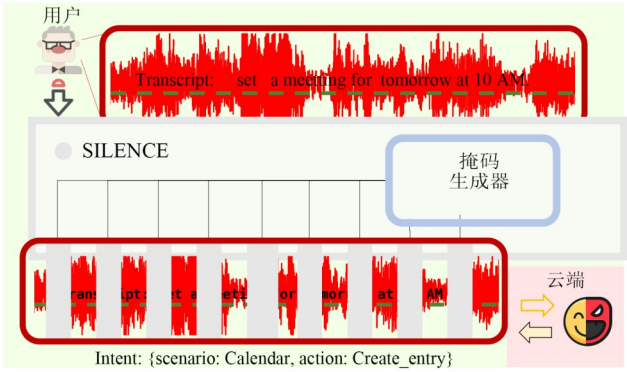
基于解耦的编码器由于需要将语音信号中的敏感信息完全分离,因此需要复杂的模型结构设计,这导致其运行速度缓慢且内存占用高. 鉴于微处理单元的资源限制,这种方法并不适用. 为了在微处理单元上实现

实用的语意理解任务的安全卸载，编码器结构和推理过程需要进一步简化.

3 SILENCE 系统

3.1 系统设计原理

本文提出的SILENCE编码器用于高效地清理原始音频中的隐私信息，以实现隐私保护的SLU任务的卸载，如图3所示. SILENCE的核心思想简单且新颖：在将音频发送到云端进行SLU任务之前，对部分音频片段进行掩码处理. 这一设计基于如图4(c)中所示的一个独特观察：当部分音频片段被掩码时，ASR模型无法识别掩码帧中的音素，而SLU模型仍能够识别意图.



注：实线表示长期依赖关系，虚线表示短期依赖关系.

图3 SILENCE系统概述

3.1.1 设计原理

SILENCE在保护敏感词隐私的同时能够保持SLU的准确性. 这一能力源于ASR任务和SLU任务之间的非对称依赖关系：ASR模型的输出具有短期依赖性，而SLU模型具有长期依赖性. 语音由许多元音素组成，单个元音素的生成依赖于其相邻的帧[8]. 依赖关系定义为模型输出依赖的帧长度.

形式化地说，模型输出某个音素y_t的时候，其概率仅依赖输入序列$\boldsymbol{x}_1^{\mathrm{T}}$中的少数关键帧，其可以表示为

$$p\left(y_t=\hat{s}\middle|\boldsymbol{x}_1^{\mathrm{T}}\right)\gg p\left(y_t=s'\middle|\boldsymbol{x}_1^{\mathrm{T}}\right),\quad \forall s'\neq\hat{s},\ \text{对多数}\ t\in[1,T] \tag{1}$$

其中，$\hat{s}$为时序上的空白标签；s'为音素标签. 模型在大部分时间步t上预测空白标签，仅在少数关键帧输出较大的目标音素预测概率. 其本质原因来源于传统ASR模型所使用的连接时序分类(Connectionist Temporal Classification，CTC)损失函数，为了对齐语音，允许在任意位置插入空白标签，导致空白标签的全局计数显著高于其他标签[34]. 在模型参数均匀初始化的前提下，梯度下降会偏向提升空白标签的概率，因为其在路径总数上占优. 从而在模型训练到局部最优解的情况下，模型会在ASR领域，这一现象被称为“峰值行为”[35]. 如图4(a)所示，每个音素主要依赖于少数帧，当缺少这些帧时，模型就无法对关键音素形成准确率的预测. 这表明ASR模型的输出具有短期依赖性.

相比之下，SLU模型使用基于注意力的解码器[33]来捕捉整句话语与意图之间的关系. 其模型输出的概率通过全局注意力权重来计算，依赖所有的输入帧：

$$p\left(y_t\middle|\boldsymbol{x}_1^{\mathrm{T}}\right)=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_t\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

其中，$\boldsymbol{Q}_t$是当前输出位置t的查询向量；$\boldsymbol{K}$、$\boldsymbol{V}$是所有输入帧的键值矩阵. 注意力权重矩阵显式地建模了输出y_t对输入序列$\boldsymbol{x}_1^{\mathrm{T}}$的全局依赖. 这表明意图对整句话语具有长期依赖性.

本文的SILENCE系统便是一个基于上述非对称信息依赖粒度性质的高效编码器. 该编码器定义为

$$\hat{\boldsymbol{x}}=\boldsymbol{x}\odot\boldsymbol{Z} \tag{3}$$

其中，$\hat{\boldsymbol{x}}$为掩蔽后的音频信号；$\boldsymbol{x}$为输入音频信号；$\odot$表示逐元素乘法；$\boldsymbol{Z}$为与$\boldsymbol{x}$维度相同的二值掩蔽向量. $\boldsymbol{Z}$由k个均匀分区组成，每个分区内全为0或1，分别用于完全掩蔽或保留相邻帧. 这一简单有效的编码器构成了SILENCE高效性和隐私保护能力的基础，从而使得在轻量微处理单元上安全卸载SLU任务成为可能.

3.1.2 配置挑战

图4(c)显示，掩蔽部分的比例在平衡隐私(敏感词识破成功率)和实用性(语意意图理解准确率)方面起着至关重要的作用. 目前，SILENCE采用一种相对基础的掩蔽机制，需要客户端花费大量时间进行超参数调整，以确定掩蔽的最优范围和位置. 错误的掩蔽配置可能导致全局依赖信息的大量丢失，从而显著影响SLU语意识别的准确性；或者敏感信息掩蔽不足，从而损害隐私保护性能. 因此，我们面临两个关键问题：应掩蔽量和应掩蔽位置.

3.2 自适应掩码生成器

为了解决这些挑战，本文受可解释学习方法[24]启发，设计了一种差分掩蔽生成器，作为SILENCE的掩码配置器. 该生成器能够自动生成掩码向量$\boldsymbol{Z}$. 掩码生成器通过训练来确定需要掩蔽的部分数量和具体位置，从而优化隐私与实用性的平衡.

3.2.1 可微掩码生成器

生成器模型的目标是通过生成掩码向量$\boldsymbol{Z}$，最小化掩码后输出与原始输出之间的信息差异. 形式化地来说，我们将未掩蔽部分的数量定义为$\mathcal{L}_0$损失：

$$\mathcal{L}_0(\phi,\boldsymbol{x})=\sum_{i=1}^{n}\mathbf{1}_{[\mathbb{R}_{>0}]}(Z_i) \tag{4}$$

其中，ϕ为掩码生成器；$\mathbf{1}(\cdot)$为计数函数. 我们在数据

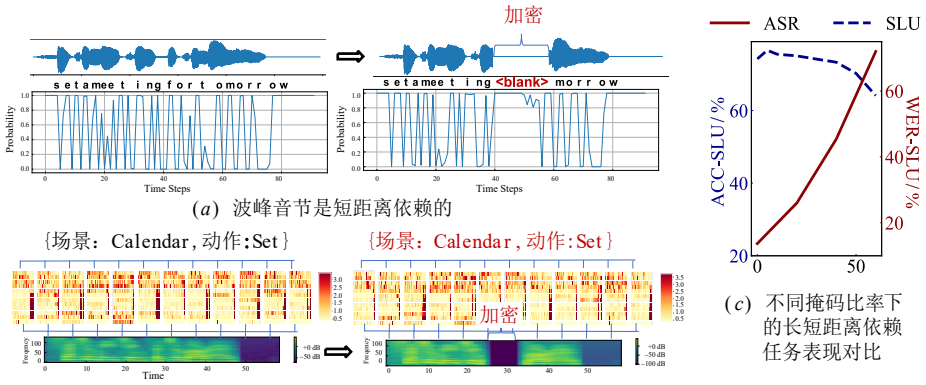


图4　SILENCE的原理示意图

集 $\mathcal{D}$ 上最小化 $\mathcal{L}_0$,确保掩蔽输入的预测结果与原始模型的预测结果相似:

$$\min_{\phi} \sum_{X \in \mathcal{D}} \mathcal{L}_0(\phi, \boldsymbol{x})$$
$$\text{s.t. } D_{\star}[y][\hat{y}] \leqslant \gamma \quad \forall x \in \mathcal{D} \tag{5}$$

其中,$\hat{y}=f(\hat{\boldsymbol{x}})$,$y$ 为语意意图标签;$D_{\star}[y][\hat{y}]$ 为 KL(Kullback-Leibler)散度,$\gamma \in \mathbb{R}_{>0}$ 为超参数. 由于 $\mathcal{L}_0$ 几乎处处不连续且导数为0,同时掩码生成器 ϕ 需要不连续的输出激活(如二值掩码的阶跃函数),本文在训练过程中使用稀疏松弛掩码法来替代二值掩码法[36-38].

3.2.2　整体工作流程

如图5所示,从系统服务角度来看,SILENCE包含两个阶段:离线训练阶段和在线推理阶段.

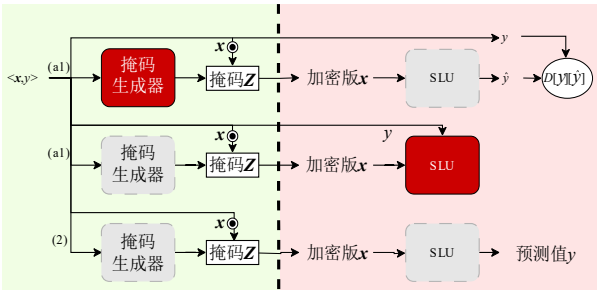


图5　SILENCE工作流程

(1)离线训练阶段. 首先,SILENCE训练一个可微的掩码生成器. 用户自行选择一个掩码生成器模型,该模型可能是预训练ASR模型的一个子模块. 例如,HuBERT的CNN特征提取器. 随后,一个小型门控模型将被集成至该子模块之后. 组合模型对输入音频进行处理并生成掩码. 这种掩码选择性地隐藏输入的一部分,确保输入语音仅保留关键的SLU意图信息,同时掩盖敏感数据信息. 掩蔽后的输入随后被发送至可信的云服务以获取对应的输出 $\hat{y}$. 掩码生成器通过微调,可以最小化掩码输出的logits与原始意图之间的差异,从而最大程度保证意图的正确识别.

其次,SILENCE会对云端模型进行微调. 客户端将编码后的输入和特定的SLU意图(例如"设置闹钟")发送到云端SLU模型. 云端SLU模型通过微调适应掩码后的加密输入. 这一过程会涉及云端SLU模型参数的更新调整,以便基于掩码输入更加准确识别和响应SLU命令.

(2)在线推理阶段. 在在线SLU阶段,客户端将编码后的输入发送至云端SLU模型. 通过更新后的云端SLU模型,云端服务提供商能够准确识别并执行预期的SLU操作或响应.

3.2.3　生成器成本分析

训练可微掩蔽生成器的开销对于用户端设备而言是可承受的. 实验表明:仅需约200个音频样本(相当于600 s音频)即可实现收敛. 在A40 GPU上,这一过程最多耗时30 s. 此外,将SLU模型适配至每个掩蔽生成器是一项一次性工作. 这种适配过程相对简单,可以基于开源SLU模型进行微调,而无需从头开始构建. 相较于云端SLU模型的预训练,这一过程的成本微乎其微. 从长远来看,这些成本可以在大量边缘用户之间分摊,从而成为一种经济上可行的解决方案.

3.2.4　讨论

需要注意的是,掩码生成器并非用于在语义层面对序列进行标注. 相反,其设计重点在于识别与SLU任务更相关的片段. 这一任务本质上是一个相对简单的二分类问题,其有效性已在先前的可解释学习文献中得到验证[24,38],且足够轻量化以支持实时推理.

4　系统实现和测试方法

4.1　隐私保护的语义理解任务

本文在基于PyTorch的统一语音工具包SpeechBrain[38]之上,完整实现了SILENCE的原型系统. 与文献[33]类似,本文使用SpeechBrain训练可微掩码生成

器并模拟云端训练过程. 随后,将训练好的掩码生成器部署到嵌入式设备上,并评估端到端性能.

4.1.1 硬件与环境

离线训练在配备8张NVIDIA A40 GPU的服务器上完成. 训练好的掩码生成器被部署到STM32H7或Raspberry PI 4 (RPI-4B)上. STM32H7是一款内存仅为1 MB的资源受限微控制单元,而RPI-4B是一款具有4 GB内存的常用开发板. 本文将无法适配STM32H7的方案嵌入到RPI-4B上进行测试.

4.1.2 模型

本文设计了四种掩码生成器结构:(1)Random,生成50%掩蔽比例的随机二值向量;(2)SILENCE-S,仅包含一个MLP门控的可学习掩码生成器;(3)SILENCE-M,包含一个HuBERT编码器层和门控的可学习掩码生成器;(4)SILENCE-L,包含三个HuBERT编码器层和门控的可学习掩码生成器. 云端SLU模型使用当前最先进的端到端SLU模型进行模拟[33],该模型将预训练HuBERT中的ASR解码器替换为SLU注意力解码器.

4.1.3 数据集与指标

本文在SLU资源包(Spoken Language Understanding Resource Package,SLURP)[25]和联邦语音分类(Federated Speech Classification,FSC)[39]数据集上运行了实验. FSC是SLU研究中广泛使用的一个数据集,而SLURP中的语音话语更为复杂,更接近日常人类语音. 因此,本文选择意图分类准确率作为衡量SLU理解性能的指标.

此外,按照先前研究[11],本文在大规模英语阅读语料库LibriSpeech[40]上进行多任务保护性能测试实验. 在多任务保护场景中,上传到云端的不仅包括SLU命令话语(SLURP/FSC),还包括背景语音或后续话语(LibriSpeech). 攻击性能将通过词错误率(Word Error Rate)衡量,即敏感词保护率:测量攻击者识别上传SLU音频或伴随音频中单词信息的能力. 同时,本文还测量了关键敏感词的识别错误率(关键敏感词保护率),以确保云端模型无法识别语音信号中的具体私人信息. 例如,在语音控制信号"Siri,请帮我设置好明天的闹钟,我要去银行取钱(背景音:去帮我从xxx银行卡里取1万元)"中,模型需要准确识别出设置闹钟的意图,同时需要防止云端服务器识别出冗余的隐私信息. 隐私保护的关键在于提高敏感词保护率,该指标越高,隐私性越好. 在数据集中,"取钱"被预设为关键敏感词,需要特别关注其保护率.

4.1.4 基线方法

本文将SILENCE与以下方法进行对比:(1)OnDevice,云端SLU模型被下载到客户端设备上运行;(2)AllOffload,将原始音频上传至云端进行SLU推理;(3)变分自编码器(Variational AutoEncoder,VAE)[12],基于变分自编码器的标准方法,通过对抗训练解耦语音信号中的私人信息;(4)PPSLU[11],当前最先进的基于解耦的SLU隐私保护系统,使用12层Transformer分离SLU信息,仅上传隐藏层至云端进行SLU推理.

4.1.5 攻击场景

如图6所示,我们定义了五种攻击场景,包括主动攻击和被动攻击. (1)Azure,代表被动的黑盒攻击场景,其中掩蔽后的音频被传输至微软Azure语音识别服务中心(https://azure.microsoft.com/en-us/products/ai-services/speech-to-text/)进行ASR. (2)Whisper,模拟当前最先进的云端ASR模型. 这种被动的黑盒攻击者将使用从HuggingFace[41]直接下载的预训练Whisper. medium.en模型[42]. (3)Whisper(白盒攻击),构成主动的白盒攻击. 在该场景中,假设某些用户为恶意用户. 这些恶意用户将掩码生成器的结构和权重,以及他们自己的音频数据泄露给Whisper攻击模型. 随后,Whisper利用这些恶意用户收集的数据,尽力将预训练的Whisper.medium.en模型适配到特定的掩码模式. (4)U型网络(U-Net),一种基于卷积U-Net结构的传统修复模型,常用于主动重建丢失的音频信号[42~44]. 本文使用SLURP训练集及其掩码后的语音,从头训练该修复模型以重建丢失的音频. (5)常数Q变换扩散(Constant-Q Transform Diffusion,CQT-Diff),一种神经扩散模型,结合了可逆的Constant-Q变换以利用音高等变对称性[45],使其能够在无需重新训练的情况下有效重建音频.

4.1.6 超参数

在图5所示的离线阶段,我们使用Adam优化器,学习率设为1×10^{-5},批量大小为4. 在推理阶段,我们将批量大小设置为1,以模拟真实的流式音频输入场景. 端到端云端SLU推理延迟参考了先前研究方法[46],通过调用Azure API测量得出. 除非特别声明,所有掩码生成器的KL阈值λ均设为0.15,攻击模型默认为Whisper.

4.1.7 掩蔽生成器的可视化

本文在图7中可视化了一些由训练好的掩蔽生成器生成的掩蔽. 可以看出,掩蔽生成器能够在一定程度上将适当的掩蔽粒度分配到合适的语音粒度. 当周围语义更丰富时,掩蔽变得更加细致,掩蔽片段也相应地改变分布.

5 测试结果与讨论

本文将SILENCE部署在STM32H7微控制单元上,并在黑盒和白盒攻击环境下使用SLURP数据集[25]评估其性能. 在SLURP数据集上,SILENCE实现了81.2%的意图分类准确率,比现有隐私保护的SLU意图识别系统的识别准确率高出至多8.3%. 在隐私保护方面,SILENCE提供了与其他隐私保护系统相当的安全性.

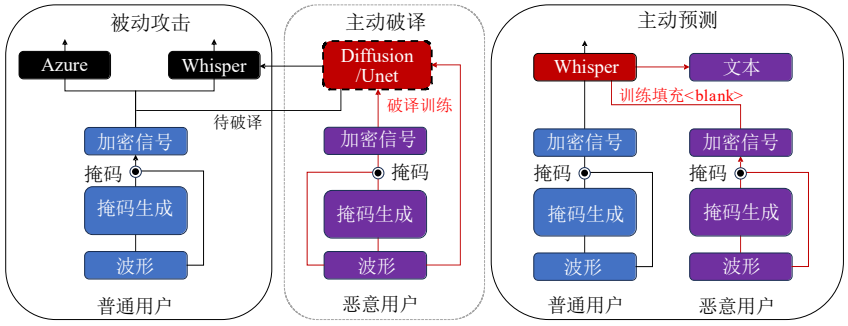
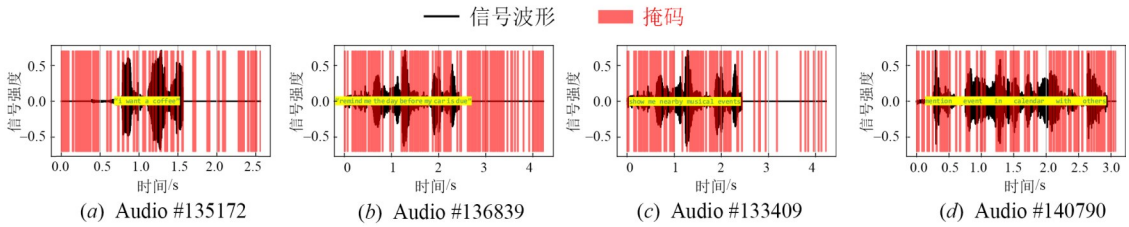


图6　掩码生成器与不同攻击场景(包括被动攻击和主动攻击)



注:根据不同的信息分布,具体音节被高效地切断了.

图7　从SLURP数据集中随机选取音频所对应的掩码效果示意图

在恶意ASR敏感词识别攻击下,SILENCE词错误率达到81.6%,关键词错误率达到90.7%. 即使在白盒攻击环境下,即攻击者被假设已知与SILENCE相同的编码器结构和权重,并可从恶意客户端获取部分数据的情况下,SILENCE仍能保持67.3%的词错误率和64.3%的关键词错误率. 此外,SILENCE证明了其在资源受限设备上的高效性和可行性,其仅需要使用394.9 KB内存,便可在912.0 ms内完成对一段4 s语音信号的隐私保护编码. 为了进一步的公平比较验证,本文还在RPI-4B上对常见的语音隐私保护方法进行了集成测试. 研究发现,与之前的语音保护编码器相比,SILENCE的内存使用量最多减少134.1倍,运行速度最多提升至其他系统的53.3倍. 同时,SILENCE的语意理解准确率仅比未受保护的明文SLU任务卸载系统平均低3.8%,最高仍可达到99%以上的场景识别准确率,处于商业化高度可用的状态.

5.1　综合性能与隐私评估

SILENCE在准确性和隐私保护能力上可与现有编码器媲美. 如图8所示,本文在SLURP数据集上将SILENCE的准确性与所有基线方法进行了对比. OnDevice未将任何信号卸载到云端,因此具有最佳隐私保护性(敏感词保护率为100%). 观察发现,SILENCE的准确性最高可达到81.1%,相较于未保护的AllOffload和本地OnDevice SLU模型,其准确性损失不足7%. 其原理在于,SILENCE主要掩蔽了局部依赖帧,这些帧对SLU性能影响不大. 本文还将SILENCE的性能与最先进的PPSLU[11]进行了对比. SILENCE的准确性比PPSLU高7.2%. 这是因为PPSLU尝试对隐藏层应用复杂的非线性变换以防止恶意重建,但这可能也会损害部分SLU信息. 在隐私保护方面,我们的可学习掩码生成器在SILENCE-L配置下可达到78.6%的敏感词保护率,表明其隐私保护能力与PPSLU相当. 如表1所示,同样的效果也在FSC数据集中体现. SILENCE在意图理解准确性方面超过99%,与所有基线方法相似,同时在防御敏感词识别攻击方面表现优异,敏感词保护率超过80%,优于所有基于解耦的保护方法. SILENCE对不同攻击模型具有抵抗能力. 如图9所示,在攻击模型Whisper下,SILENCE将语音意图提取准确率从14.7%提升至78.6%. 对于在线攻击模型Azure,SILENCE将语音意图提取准确率从14.7%提升至81.6%. 根据返回的服务细节发现,超过50%的发送音频被标记为ResultReason.NoMatch,这意味着这些音频被Azure ASR模型识别为无效话语.

Whisper是一个白盒攻击模型,意味着攻击者拥有与SILENCE相同的掩蔽生成器结构和权重. 在该攻击模型下,SILENCE仍然实现了超过50%的语音意图提取准确率. 这是因为,即使Whisper经过微调能够填补部分缺失帧,仍然无法恢复缺失的隐私信息帧. 原因在于,掩蔽短期依赖帧从根本上破坏了原始音频信号. 在缺乏语音信息的情况下,不可能重建相应的敏感词. 在最后的子图中,本文展示了较高的实体错误率(Entity Error Rate,EER),以证明私人隐私信息未被泄露.

SILENCE同样可以防御主动修复攻击. 重建后的音频被发送到Whisper进行自动识别. 重建波形的可视

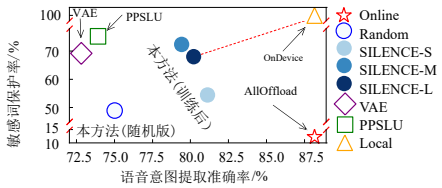


图8 不同隐私保护SLU方法的性能比较

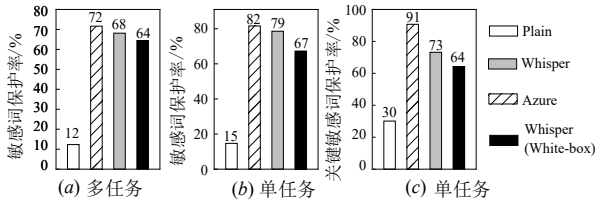


图9 SILENCE在不同攻击模型下的隐私保护能力

表1 FSC数据集上的隐私保护与SLU性能评估 单位:%

方法	语音意图提取准确率	敏感词保护率
AllOffloaded	99.7	1.2
VAE	98.3	65.5
PPSLU	99.2	78.5
OnDevice	99.7	100.0
Random	76.6	86.4
本方法	99.1	81.4

化结果如图10和表2所示. U-Net几乎无法重建掩蔽的音频,甚至引入了错误的噪声信号,降低了攻击成功率. CQT-Diff修复可以填补缺失的波形,但无法成功重建内容,因为它的设计目的是重建背景音乐(例如钢琴协奏曲). 而包含人类意图和对话的SLU音频难以被重建.

掩码生成器在不同语音粒度任务下均有效果. 我们引入了两个更细粒度的SLU任务:动作识别(Action)和组合意图(Intent:scenario_action)识别. 这些任务包含18种不同的场景和46种定义的动作,共有828种可能的意图组合. 如表3所示,本文方法能够在不同粒度下正确识别语音意图. 例如,可以正确识别76.8%的组合意图. 因此,本文系统可以较好地识别不同类型的SLU意图识别任务,适用范围广泛.

相比之下,基于解耦的方法需要为不同的语义粒度重新对表示进行信息解耦操作. 因此,用于场景分类的分类器无法直接应用于其他意图识别任务. 此外,这些方法也未设计为在命令音频中保护敏感信息. 这突出了本文方法的显著优势,即无需为不同的意图粒度重新训练模型.

SILENCE可以通过更大的掩蔽生成器实现更优的隐私-准确性权衡. 我们研究了SILENCE在不同掩蔽生成器结构下的阈值γ对隐私与实用性权衡的影响. 如图11所示,阈值γ控制了隐私与实用性之间的权衡. 当γ较小时,掩蔽生成器更为保守,掩蔽比例较低,从而提高了实用性. 然而,如第3节所讨论的,较低的掩蔽比例会增加隐私实体泄露的可能性. 当γ较大时,掩蔽生成器更为激进,从而增强隐私保护能力. 另一种实现更优的隐私-实用性平衡的方法是使用更复杂的掩蔽生成器结构,例如SILENCE-L. 与SILENCE-S相比,它在相同的隐私水平下实现了更高的实用性,但效率层面会有部分牺牲.

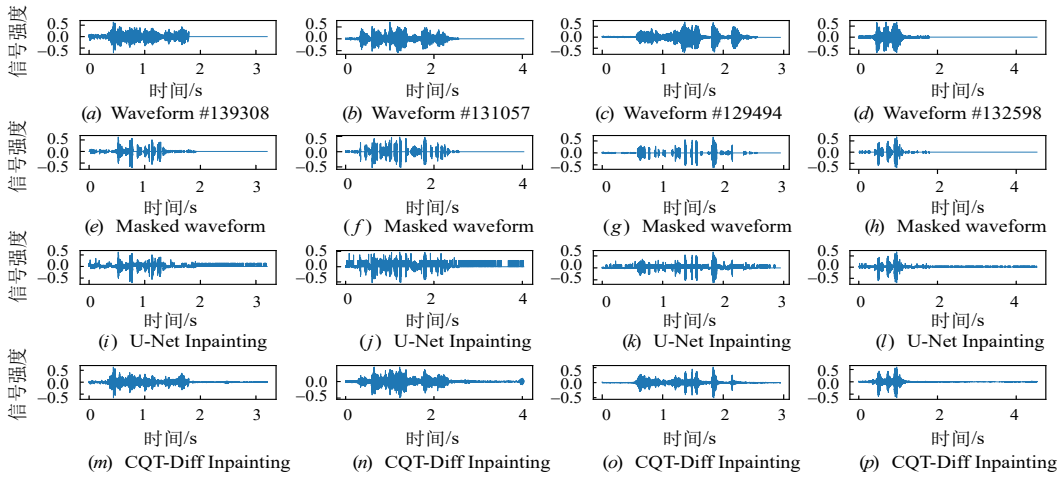


图10 在数据集SLURP上不同主动修复攻击对应的重建波形

具体来说,真实场景下的数据丢失或传输差错确实会对语音意图识别系统的性能指标产生影响,当丢失10%语音帧的情况下,对于SILENCE-S掩码来说,其意图预测和真实意图的KL散度会从0.14上升到0.22. 这意味着意图的预测准确率将会显著下降. 一种行之有效的方式是将SILENCE-S掩码器替换为更强的SILENCE-L掩码器,从而可以使得在相同的语音帧丢失情况下,保证KL散度仍然保持在0.14以内.

表2　不同生成攻击方式下的语音隐私保护性能　单位:%

生成攻击方式	单任务下的敏感词保护率	多任务下的敏感词保护率
PlainText	14.7	12.3
Azure	81.6	71.6
Whisper	78.6	68.1
Whisper (Whitebox)	67.3	64.4
U-Net	82.5	71.4
CQT-Diff	74.3	65.9

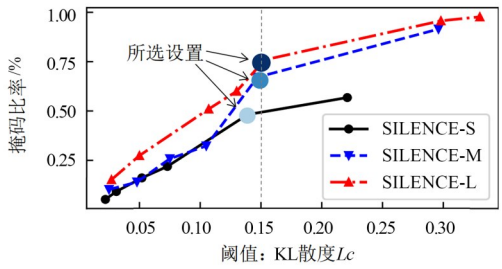
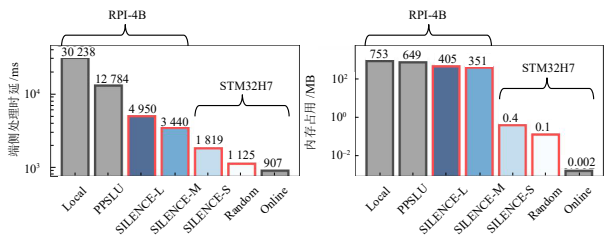


图11　不同掩码生成器下阈值的影响

5.2　系统开销

如图12所示,SILENCE能够高效保护用户敏感词信息. 不同于先前使用复杂解耦模型的编码器,SILENCE仅需一个轻量化的掩蔽生成器即可清理私人信息. 掩蔽生成器的大小根据不同的生成器结构而变化. 对于最小的掩蔽生成器SILENCE-S,其内存占用仅为394.9 KB,可成功嵌入到内存仅有1 MB的STM32H7中.



(a) 端侧时延　(b) 内存占用

图12　不同隐私保护方法的资源成本比较

SILENCE不仅在内存占用方面高效,在延迟方面也同样表现出色. SILENCE-S在资源受限的STM32H7上完成本地编码仅需912.2 ms.

为了公平比较,本文将SILENCE-S嵌入到RPI-4B中,发现其速度比PPSLU快18.1倍,内存占用减少134.1倍. 即使使用更强大的掩蔽生成器SILENCE-L,SILENCE的编码延迟仍然降低了7.5倍,内存消耗比OnDevice减少1.9倍.

5.3　讨论

SILENCE是一个高效且注重隐私保护的端到端SLU系统. 它基于ASR词语识别和SLU意图识别之间的非对称信息依赖关系. SILENCE通过选择性地掩蔽局部依赖的敏感词,同时保留全局依赖的SLU意图信息,实现了隐私保护和意图识别的平衡. 结合可微掩码生成器,SILENCE在不同攻击场景下表现出了卓越的端到端推理速度和隐私保护能力.

5.3.1　局限性讨论

SILENCE首次为资源受限的音频设备提供了可行的隐私保护解决方案,但其引入了一个巨大的掩码生成器结构设计空间. 掩码生成器类似于一把锁,优秀的锁设计能够在最小空间内保护隐私,而糟糕的设计则可能体积庞大且易被破解. 本文直接继承了SLU模型结构,并从中实例化了三个子模型,以便展示其相较于现有编码器的更高效率. 未来的研究可以探索其他结构,以实现更优的隐私、准确性和效率之间的平衡.

此外,关于隐私保护能力的可靠性、云端SLU模型微调的需求、威胁模型的范围及其对离线场景的扩展等潜在局限性,我们在下文中进行了更为详细的讨论.

5.3.2　系统的隐私保护能力

本文方法所达到的敏感词保护率达到了80%,这一水平被认为足够安全,与先前的编码器[11,12]达到了相同水平. 此外,部分SLU语音中被识别出来的文字内容可能包含非隐私的意图词. 例如,在一段测试音频中,“I want some jazz music to play”,其意图为{‘scenario’:‘play’,‘action’:‘music’}. 恶意云ASR对该语音的识别结果为“all subjects were used to play”是可以接受的,因为预测的短语“to play”不包含私人信息. 这一情况几乎体现在绝大多数被测音频上. 如图8所示,在预定义敏感词保护方面,本文方法成功保护了90%的敏感词信息. 本文方法在隐私保护能力上与最先进方法一致的同时,将系统延迟降低了30倍,内存消耗减

表3　不同语音粒度下隐私保护与SLU性能的比较　单位:%

方法	语音场景识别准确率	语音内动词识别准确率	语音意图识别准确率	意图相关敏感词保护率	整体敏感词保护率
AllOffloaded	88.2	77.1	83.3	14.7	12.3
VAE	72.8	—	—	—	69.3
PPSLU	73.9	—	—	—	75.3
OnDevice	88.2	77.1	83.3	100.0	100.0
本方法	80.2	76.4	76.8	68.6	68.1

注:由于本地方案OnDevice未上传任何数据,不存在单词泄露的风险.

少了100倍.

5.3.3 微调云端SLU模型的必要性及方法

起初,云端SLU是一个通用的预训练语音模型,缺乏准确理解个性化用户意图的能力.因此,微调云端SLU对于第三方应用提供商更好地理解个性化用户意图是至关重要的.这里需要进一步注意的是,在图5中,用于训练本地掩码生成器时,通用语音模型已经足够,因为此处的重点并非生成精确的意图,而是获得粗粒度的数值logits分布,以便于掩码生成器的训练.其次,虽然局部依赖掩码并未完全消除意图信息,但确实会影响注意力图中的某些具体细节,如图4(*b*)所示.通过微调云端SLU模型,可以减轻这种影响,并增强对端侧用户意图的理解.目前,云服务提供商已提供相关API(https://azure.microsoft.com/en-us/blog/improve-speechtotext-accuracy-with-azure-custom-speech/),允许客户微调其个性化的云端语音模型.具体来说,客户可以利用Azure Speech的自定义(Custom Speech)功能对云端语音模型进行微调,具体做法包括:先基于微软预训练的基础模型,再向其中添加各自领域的专有数据(如针对特定行业、口音、背景噪声等),从而改进识别准确率.常见的数据形式包括文本(如领域词汇或短语库)以及带标注的音频(可帮助模型适应独特口音、发音、环境噪声),也可以将多种数据整合为一个更精细的模型.最终用户训练好自定义模型后,会部署在专用的服务端点上,应用或设备即可实时访问该端点,并获得比通用模型更准确的语音转文字结果.

5.3.4 防止语义检测攻击

需要明确指出,检测短依赖关键短语或特定指令并非本文工作的重点.例如,窃听特定的金融词汇和政治框架属于非本文讨论范围.然而,本文系统也可以部分提供针对这些攻击的防御能力.用户控制的掩码生成器经过训练,可以用于剔除与公共意图无关的语音内容.用户未预定义的敏感词几乎不会包含在掩码后的音频中.因此,即使攻击者拥有明确定义的语意和掩码生成器,训练检测威胁模型依然是困难的,因为合成的掩码音频缺乏相应语意词的清晰表示.

5.3.5 离线场景

在资源受限设备上,网络离线条件的情况会周期性出现.本系统可以进一步轻松集成到小型设备端离线运行部分SLU任务与复杂任务卸载到强大的云端模型这一协同架构中.此类架构已被许多现成产品正式采用,例如iOS 18中的Apple Intelligence.在这些情况下,本文系统仍然是不可或缺的,因为小型设备端SLU模型由于模型规模受限,可能无法生成令人满意的意图理解结果.即使设备端SLU模型能够正确生成意图理解,由于设备能量受限,它们也无法始终运行.因此,在线过程仍然是当前SLU解决方案的主要组成部分.而设备端功能可以在离线条件下作为替代方案使用.在我们的系统的帮助下,基于云端的SLU功能组件可以变得既具备隐私保护特性,又高效可靠.

此外,现有的语音编码方案也可以通过将解耦模型部署在云端来降低内存开销.但是此种卸载方式需要于TEE下进行,亦或者是依赖可信的云端SLU第三方来保证用户数据的隐私.对于第一种方案,拥有TEE的硬件通常依然会受限于内存(远小于500 MB),无法有效承载现有基于解耦的编码器方法所使用的模型.对于第二种方案,其实现成本较高,且依赖于第三方的可靠性,大大提升了方法的部署难度.本文方法的优势在于所有的隐私数据均保留在本地,这不仅增强了用户的信心,还实现了在微处理器单元上的有效部署,从而成功降低了隐私语音推理的门槛.

6 结论

SILENCE是一个高效且注重隐私保护的端到端SLU系统.它基于ASR词语识别和SLU意图识别之间的非对称信息依赖关系.SILENCE通过选择性地掩蔽局部依赖的敏感词,同时保留全局依赖的SLU意图信息,实现了隐私保护和意图识别的平衡.结合可微掩码生成器,SILENCE在不同攻击场景下表现出了卓越的语音意图提取准确率、隐私保护能力和端到端推理速度.

致谢 感谢林小竹教授给本文提出的参考意见.

参考文献

[1] BAEVSKI A, ZHOU Y, MOHAMED A, AULI M. wav2vec 2.0: A framework for self-supervised learning of speech representations[C]//Advances in Neural Information Processing Systems 33. Red Hook: Curran Associates, 2020: 12449-12460.

[2] SENEVIRATNE S, HU Y N, NGUYEN T, et al. A survey of wearable devices and challenges[J]. IEEE Communications Surveys & Tutorials, 2017, 19(4): 2573-2620.

[3] CLARK L, DOYLE P, GARAIALDE D, et al. The state of speech in HCI: Trends, themes and challenges[J]. Interacting with Computers, 2019, 31(4): 349-371.

[4] NORUWANA N C, OWOLAWI P A, MAPAYI T. Interactive IoT-based speech-controlled home automation system[C]//Proceedings of the 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC). Piscataway: IEEE, 2020: 1-8.

[5] RAJA J M, ELSAKR C, ROMAN S, et al. Apple watch,

wearables, and heart rhythm: Where do we stand?[J]. Annals of Translational Medicine, 2019, 7(17): 417.

[6] EMOKPAE L E, EMOKPAE R N, LALOUANI W, et al. Smart multimodal telehealth-IoT system for COVID-19 patients[J]. IEEE Pervasive Computing, 2021, 20(2): 73-80.

[7] KUMAR N, LEE S C. Human-machine interface in smart factory: A systematic literature review[J]. Technological Forecasting and Social Change, 2022, 174: 121284.

[8] SUN K, CHEN C, ZHANG X Y. "Alexa, stop spying on me!": Speech privacy protection against voice assistants[C]// Proceedings of the 18th Conference on Embedded Networked Sensor Systems. New York: ACM, 2020: 298-311.

[9] DANG T, THAKKAR O, RAMASWAMY S, et al. A method to reveal speaker identity in distributed ASR training, and how to counter IT[C]//2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2022: 4338-4342.

[10] QIAN J W, DU H H, HOU J H, et al. Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity[C]//Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems. New York: ACM, 2018: 82-94.

[11] WANG Y G, HUANG W, YANG L. Privacy-preserving end-to-end spoken language understanding[C]//Proceedings of the 32nd International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2023: 5224-5232.

[12] ALOUFI R, HADDADI H, BOYLE D. Privacy-preserving voice analysis via disentangled representations[C]// Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop. New York: ACM, 2020: 1-14.

[13] PEYSER C, HUANG R W, ROSENBERG A, et al. Towards disentangled speech representations[C]//Proceedings of Interspeech 2022. Baixas: ISCA, 2022: 3603-3607.

[14] ZHANG C L, LI S Y, XIA J Z, et al. BatchCrypt: Efficient homomorphic encryption for cross-silo federated learning[C]//USENIX Annual Technical Conference 2020. Berkeley: USENIX Association, 2020: 493-506.

[15] GOLDREICH O. Secure multi-party computation[EB/OL]. (1998)[2024-12-12]. https://www.researchgate.net/profile/Oded-Goldreich/publication/2934115_Secure_Multi-Par - ty_Computation/links/00b7d52bb04f7027d4000000/Secure-Multi-Party-Computation.pdf.

[16] DONG Y, LU W J, ZHENG Y C, et al. PUMA: Secure inference of Llama-7B in five minutes[EB/OL]. (2023-07-24)[2024-12-12]. https://arxiv.org/abs/2307.12533.

[17] CHEN Y K, GAO M, LI Y M, et al. Big brother is listening: An evaluation framework on ultrasonic microphone jammers[C]//IEEE INFOCOM 2022 - IEEE Conference on Computer Communications. Piscataway: IEEE, 2022: 1119-1128.

[18] AHMED S, CHOWDHURY A R, FAWAZ K, RAMANATHAN P. Preech: A system for privacy-preserving speech transcription[C]//USENIX Security Symposium 2020. Berkeley: USENIX Association, 2020: 2703-2720.

[19] GAO M, CHEN Y K, LIU Y J, et al. Cancelling speech signals for speech privacy protection against microphone eavesdropping[C]//Proceedings of the 29th Annual International Conference on Mobile Computing and Networking. New York: ACM, 2023: 1-16.

[20] GULATI A, QIN J, CHIU C C, et al. Conformer: Convolution-augmented transformer for speech recognition[C]// Interspeech 2020. Baixas: ISCA, 2020: 5036-5040.

[21] DENG K Q, CAO S J, ZHANG Y K, et al. Improving hybrid CTC/attention end-to-end speech recognition with pretrained acoustic and language models[C]//2021 IEEE Automatic Speech Recognition and Understanding Workshop. Piscataway: IEEE, 2021: 76-82.

[22] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2014: 2672-2680.

[23] ARORA S, DALMIA S, CHANG X K, et al. Two-pass low latency end-to-end spoken language understanding[C]// Proceedings of the Annual Conference of the International Speech Communication Association. Incheon: ISCA, 2022: 3478-3482.

[24] DE CAO N, SCHLICHTKRULL M S, AZIZ W, et al. How do decisions emerge across layers in neural models? Interpretation with differentiable masking[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020: 3243-3255.

[25] BASTIANELLI E, VANZO A, SWIETOJANSKI P, et al. SLURP: A spoken language understanding resource package[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020: 7252-7262.

[26] DESHMUKH S, ELIZALDE B, SINGH R, et al. Pengi: An audio language model for audio tasks[J]. Advances in Neural Information Processing Systems, 2023, 36: 18090-18108.

[27] WANG J X, RADFAR M, WEI K, et al. End-to-end spoken language understanding using joint CTC loss and self-supervised, pretrained acoustic encoders[C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 1-5.

[28] AGRAWAL B, MÜLLER M, CHOUDHARY S, et al. Tie your embeddings down: Cross-modal latent spaces for end-to-end spoken language understanding[C]// ICASSP 2022. Piscataway: IEEE, 2022: 7157-7161.

[29] WATANABE S, HORI T, KIM S, et al. Hybrid CTC/attention architecture for end-to-end speech recognition[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1240-1253.

[30] CHOROWSKI J, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition[C]//Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1. New York: ACM, 2015: 577-585.

[31] DE MORI R. Spoken language understanding: A survey[C]//2007 IEEE Workshop on Automatic Speech Recognition & Understanding. Piscataway: IEEE, 2007: 365-376.

[32] HAGHANI P, NARAYANAN A, BACCHIANI M, et al. From audio to semantics: Approaches to end-to-end spoken language understanding[C]//2018 IEEE Spoken Language Technology Workshop. Piscataway: IEEE, 2018: 720-726.

[33] HSU W N, BOLTE B, TSAI Y H, et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[34] PRABHAVALKAR R, HORI T, SAINATH T N, et al. End-to-end speech recognition: A survey[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2023, 32: 325-351.

[35] HUANG R Z, ZHANG X H, NI Z H, et al. Less peaky and more accurate CTC forced alignment by label priors[C]//ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2024: 11831-11835.

[36] LOUIZOS C, WELLING M, KINGMA D P. Learning sparse neural networks through L0 regularization[C]//International Conference on Learning Representations, 2018.

[37] BASTINGS J, AZIZ W, TITOV I. Interpretable neural predictions with differentiable binary variables[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 2963-2977.

[38] RAVANELLI M, PARCOLLET T, MOUMEN A, et al. Open-source conversational AI with SpeechBrain 1.0[J]. Journal of Machine Learning Research, 2024, 25: 1-11.

[39] MEHRISH A, MAJUMDER N, BHARADWAJ R, et al. A review of deep learning techniques for speech processing[J]. Information Fusion, 2023, 99: 101869.

[40] PANAYOTOV V, CHEN G G, POVEY D, et al. Librispeech: An ASR corpus based on public domain audio books[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2015: 5206-5210.

[41] WOLF T, DEBUT L, SANH V, et al. Transformers: State-of-the-art natural language processing[C]//Proceedings of EMNLP 2020: System Demonstrations. Stroudsburg: ACL, 2020: 38-45.

[42] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[C]//Proceedings of the 40th International Conference on Machine Learning. New York: ACM, 2023: 28492-28518.

[43] HAO X, SU X D, WEN S X, et al. Masking and inpainting: A two-stage speech enhancement approach for low SNR and non-stationary noise[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2020: 6959-6963.

[44] KEGLER M, BECKMANN P, CERNAK M. Deep speech inpainting of time-frequency masks[C]//Proceedings of the Annual Conference of the International Speech Communication Association. Shanghai: ISCA, 2020: 3276-3280.

[45] MOLINER E, LEHTINEN J, VÄLIMÄKI V. Solving audio inverse problems with a diffusion model[C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 1-5.

[46] WANG R X, LIN F X. Turbocharge speech understanding with pilot inference[C]//Proceedings of the 30th Annual International Conference on Mobile Computing and Networking. New York: ACM, 2024: 1299-1313.

作者简介

蔡栋琪　男,1999年8月生,江苏盐城人. 现为北京邮电大学计算机学院直博四年级博士研究生. 现于剑桥大学进行联合培养访问研究. 主要研究方向为高效的终端侧机器学习系统. 中国电子学会会员编号:E190182924A.

E-mail: dc912@cam.ac.uk

王尚广　男,1982年2月生,河南周口人. 2011年毕业于北京邮电大学,获博士学位. 现为北京邮电大学计算机学院教授. 主要研究方向为服务计算、移动边缘计算与卫星计算. 已发表论文150余篇. 中国电子学会会员编号:E190027924S.

E-mail: sgwang@bupt.edu.cn

张泽凌　男,2000年8月生,四川成都人. 现为北京邮电大学计算机学院硕士研究生.

E-mail: marovlo@bupt.edu.cn

马　骁　女,1990年9月生,山东德州人. 博士,2018年毕业于清华大学计算机科学与技术系. 现为北京邮电大学网络与交换技术国家重点实验室副教授. 主要研究方向为移动云计算与移动边缘计算.

E-mail: maxiao18@bupt.edu.cn

徐梦炜　男,1992年6月生,浙江绍兴人. 现为北京邮电大学计算机学院副教授. 主要研究方向为移动计算、边缘计算、人工智能与系统软件等. 中国电子学会会员编号:E190024575M.

E-mail: mwx@bupt.edu.cn