Dangerous Skills: Understanding and Mitigating Security Risks of Voice-Controlled Third-Party Functions on Virtual Personal Assistant Systems

危险skill:理解并降低虚拟个人助理系统中语音控制第三方功能的安全风险

2019 IEEE Symposium on Security and Privacy

目录

[Dangerous Skills: Understanding and Mitigating Security Risks of Voice-Controlled Third-Party Functions on Virtual Personal Assistant Systems 1](#_Toc37149543)

[危险skill:理解并降低虚拟个人助理系统中语音控制第三方功能的安全风险 1](#_Toc37149544)

[2019 IEEE Symposium on Security and Privacy 1](#_Toc37149545)

[摘要 1](#_Toc37149546)

[背景 2](#_Toc37149547)

[亟待解决的问题： 3](#_Toc37149548)

[问题的意义： 4](#_Toc37149549)

[主要贡献： 4](#_Toc37149550)

[基本概念 4](#_Toc37149551)

[应用场景 6](#_Toc37149552)

[主要工作 7](#_Toc37149553)

[1.恶意skill引发的安全风险 7](#_Toc37149554)

[2.语音占有攻击（VSA） 7](#_Toc37149555)

[3.语音伪装攻击（VMA） 8](#_Toc37149556)

[4.查找语音占有skill 9](#_Toc37149557)

[5.防止语音伪装skill 9](#_Toc37149558)

[可拓展点 10](#_Toc37149559)

## 摘要

虚拟个人助理VPA依靠语音通道与用户进行通信，然而1）语音通道容易受到攻击，2）缺乏适当的身份验证，VPA允许使用第三方skill来增强用户体验，基于此本文发现了两种利用第三方skill发动的攻击： *voice squatting*语音非法占有 发音类似的skill 去替代合法的skill；语音伪装 恶意模仿VPA服务，或者一个合法的skill 在用户与服务对话中窃取他的隐私；并进一步的针对这两种攻击提出我们的方案。

## 背景

**Skill:** 第三方开发人员可以构建新功能(称为*skill*通过亚马逊和行动*Google1*)向最终用户提供进一步的帮助

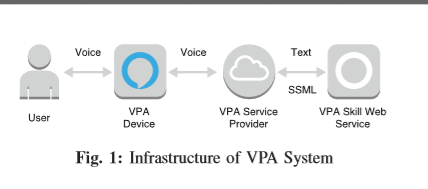
**Skill markets：丰富的skill市场**

**Skill 调用：显示或隐式的调用**

* **显式调用：**用户需要从*VPA*获取其姓名的技能时:例如，(*Alexa*，与*Amex*对话)，给*Alexa*将触发Amex技能进行支付或检查银行帐户余额
* **隐式调用：**告诉语音助手执行某些任务而不直接调用技能名的情况下。例如（*Hey Google, will it rain tomorrow*）将调用天气技能来响应天气预报。

**Skill 交互模型：**一个松散的协议

* 显式地唤醒某个技能，用户需要使用一个唤醒词、一个触发短语和技能的调用名称 对于口语句子（Hey google，*talk to personal chef）*, (*Hey google* )是唤醒词，（*talk to*）是触发词，（*personal chef*）是技能调用名
* 技能调用名称可以与技能名称不同
* 在开发一项*skill*时，需要定义**意图和样本话语，将用户的语音输入映射到该技能的各种接口，以采取用户期望的操作**。这样的接口是由意图来描述的
* 模型中还有一些内置的意图，如*WelcomeIntent、HelpIntent、StopIntent*
* 第三方skill本质上是由其开发人员托管的web服务，其名称注册到VPA服务提供者(Amazon或谷歌)
* **交互模式**
* **与智能手机应用程序或网站插件需要明确安装不同，技能可以自动发现(根据用户的语音命令)，并通过物联网设备透明地触发。**



**VPA：**虚拟个人助理，这种服务通常是通过智能扬声器来提供的，智能扬声器使用语音用户界面(*VUI*)与用户进行交互，允许用户仅使用语音来命令系统。都放弃了传统的I/O接口，比如触摸屏，而且按钮更少(用来调节音量或静音)，为用户提供了免提体验。通过双向对话吸引用户。与那些可以通过按键激活智能手机

**VUI**：语音用户界面

存在的隐患：

1)缺乏有效的方法来通过开放和嘈杂的语音通道对相关各方进行认证

2)用户也很难确定她是否真的在与正确的*skill*和*VPA*本身进行对话

语音蹲式攻击(VSA)：

语音伪装攻击(VMA)：

## 亟待解决的问题：

VPA系统主要是由语音来控制的。因为缺乏有效的方法来通过开放和嘈杂的语音通道对相关各方进行认证。用户也很难确定她是否真的在与正确的*skill*和*VPA*本身进行对话。通过技术市场,敌人可以发布恶意第三方skill，对手可以冒充一个合法的*skill*，甚至*VPA*(可能是大规模的)给用户

## 问题的意义：

* **帮助用户抵御语音攻击**
* **帮助厂商提升安全防护能力**

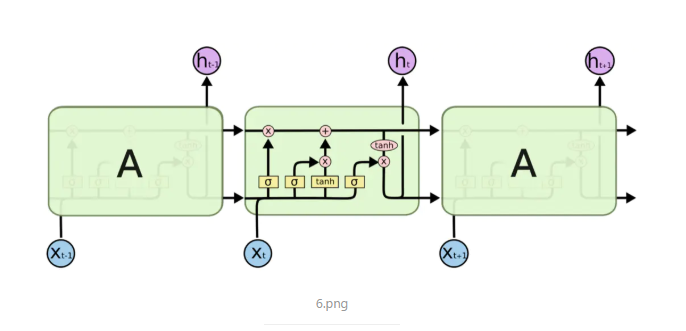
## 主要贡献：

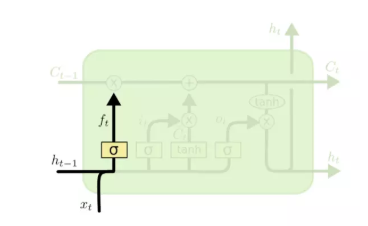
* **首次研究恶意skill的安全风险，针对于物联网系统领先的VPA**
* **降低风险的新技术 缓解措施给出**

## 基本概念

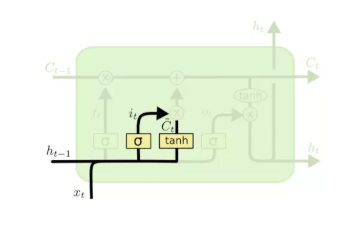
**ARPABET**（也称为ARPAbet）是[高级研究计划局](https://en.wikipedia.org/wiki/Advanced_Research_Projects_Agency" \o "高级研究计划局)（ARPA）在1970年代[语音理解研究](https://en.wikipedia.org/w/index.php?title=Speech_Understanding_Research&action=edit&redlink=1" \o "语音理解研究（页面不存在）)项目中开发的一组[语音转录](https://en.wikipedia.org/wiki/Phonetic_transcription)代码。它代表[的音素](https://en.wikipedia.org/wiki/Phoneme)和[音位变体](https://en.wikipedia.org/wiki/Allophone)的[一般美式英语](https://en.wikipedia.org/wiki/General_American_English)与不同序列[的ASCII](https://en.wikipedia.org/wiki/ASCII)字符。设计了两个系统，每个系统用一个字符表示每个[段](https://en.wikipedia.org/wiki/Segment_(linguistics)" \o "细分（语言学）)（大写和小写字母交替），另一个系统用一个或两个字符表示（不区分大小写），后者被更广泛地采用。

* LSTM：长短期记忆（LSTM）是一种用于[深度学习](https://en.wikipedia.org/wiki/Deep_learning)领域的人工[递归神经网络](https://en.wikipedia.org/wiki/Recurrent_neural_network)（特殊的RNN）架构，它不仅可以处理单个数据点（例如图像），而且可以处理整个数据序列（例如语音或视频）常见的LSTM单元由一个单元，一个输入门，一个输出门和一个忘记门组成，单元会记住任意时间间隔内的值，并且三个门控制着进出单元的信息流。



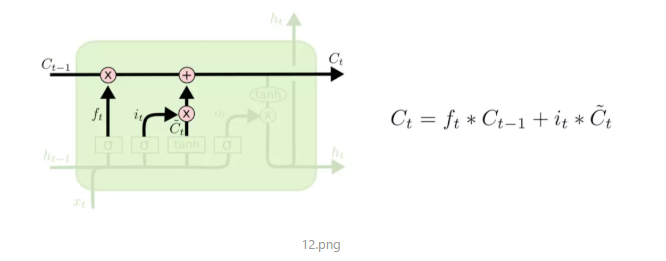


忘记门：决定细胞状态需要丢弃哪些信息它通过查看和信息来输出一个0-1之间的向量，该向量里面的0-1值表示细胞状态中的哪些信息保留或丢弃多少。0表示不保留，1表示都保留

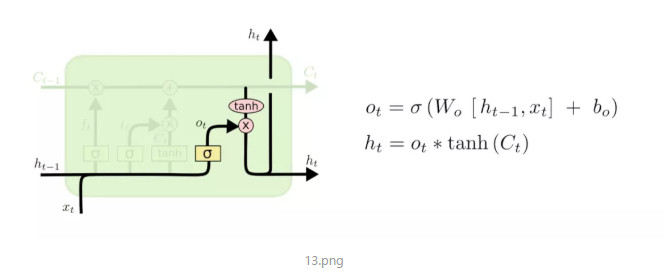


输入门：决定给细胞状态添加哪些新的信息，首先利用和通过一个称为输入门的操作来决定更新哪些信息。然后利用和通过一个tanh层得到新的候选细胞信息

下面将更新旧的细胞信息，变为新的细胞信息更新的规则就是通过忘记门选择忘记旧细胞信息的一部分，通过输入门选择添加候选细胞信息的一部分得到新的细胞信息。更新操作如下图所示



更新完细胞状态后需要根据输入的和来判断输出细胞的哪些状态特征这里需要将输入经过一个称为输出门的sigmoid层得到判断条件，然后将细胞状态经过tanh层得到一个-1~1之间值的向量，该向量与输出门得到的判断条件相乘就得到了最终该RNN单元的输出



LSTM网络非常适合基于[时间序列](https://en.wikipedia.org/wiki/Time_series)数据[进行分类](https://en.wikipedia.org/wiki/Classification_in_machine_learning)，[处理](https://en.wikipedia.org/wiki/Computer_data_processing)和[做出预测](https://en.wikipedia.org/wiki/Predict)，因为[时间序列中](https://en.wikipedia.org/wiki/Time_series)重要事件之间可能存在未知持续时间的滞后。开发LSTM是为了解决训练传统RNN时可能遇到的爆炸和[消失梯度问题](https://en.wikipedia.org/wiki/Vanishing_gradient_problem)。与缝隙长度相对不敏感是LSTM 在众多应用中优于RNN，[隐马尔可夫模型](https://en.wikipedia.org/wiki/Hidden_Markov_models" \o "隐藏的马尔可夫模型)和其他序列学习方法的优势

[参考] <https://www.jianshu.com/p/95d5c461924c>

## 应用场景

使用VPA语音控制的物联网环境

## 主要工作

### 恶意skill引发的安全风险

* 当用户说出他们的调用名称时，VPA技能将透明地启动(调用名称可以与技能市场上显示的名称不同)。
* ALEXA没有skill标识唯一性检查,Google 有限制
  + 调用常见skill时使用随机匹配
  + 调用相似skill 时使用最长字符串匹配
* Skill 调用名无法被正常的识别，导致运行其他发音相近的skill
* 系统只能一次执行一次，不支持上下文切换
* Alexa 和 谷歌支持自愿终止技能，没有证据证明skill自身真的停止了

### 语音占有攻击（VSA）

* 由于语音识别的局限性，skill 识别存在误差
* 利用最长字符串匹配构造恶意skill ，例如Capital one Please/Capital one Ploice
* 结论：
  + 攻击*skill*可以模仿另一种*skill*并伪造其VUI(语音用户界面)来收集用户仅与目标*skill*共享的信息
  + 误导用户使用户以为合法的Skill存在安全威胁，从而对这个合法skill产生负面评价
* 分类：
  + 声音占有：攻击skill携带有语音学上类似的名称来调用它的目标技能
  + 单词占有：攻击调用名称包括目标的名称和一些战略选择附加词
* 发音错误的调用名称如果发音相近，并且没有其他使用发音错误的调用名称注册的技能，也会触发正确的技能
* 我们注册skill名称为每次都被VPA识别错误的

针对上诉两个占有进行实验：

实验结果：

* 语音占有实验结果：
  + 17个这样的脆弱的Alexa技能，7个谷歌技能，4个谷歌技能 可以被注册为恶意技能
* 词占有实验结果：
  + Alexa上，带有扩展名称的总被识别
  + Google 上只有使用单词+’ app’的话语才能成功触发相应的攻击技能

### 语音伪装攻击（VMA）

* Skill 切换：一个正在运行的skill 可以假装将控制权交给另一个(由于没有显示的证据证实上一个skill结束了)
  + 针对上诉问题google有保护措施
    - 它通过说*Sure, here is*，来标志一项技能的开始连同技能名称和一个特殊的earcon，技能终止与另一个earcon；VPA以一种独特的口音与用户交谈，将其与技能区分开来
    - 但敌手仍然可以伪装成为VPA
* Skill 假装终止
  + 允许一项技能在向用户发出语音响应后立即终止自身，响应的内容来自skill 开发者的服务器，作为JSON对象，然后由VPA系统发出
  + 参与者依赖于技能的反应(如再见或沉默)来决定一个技能是否已经终止
  + 允许一个技能包括一个*reprompt*(文本内容或音频文件) VPA在一段时间内没有收到来自用户的任何语音命令时，就会播放这个提示
  + Alexa在6秒后重新提示用户，而谷歌Assistant在8秒后重新提示用户
  + 只要用户在这段时间内说了什么(即使不是针对该技能)，该技能就可以发送另一个响应并返回一个提示
  + 恶意skill 可以返回一个静默的音频文件 让用户误以为skill 已经结束 从而记录用户的敏感信息
  + 允许用户通过“停止”、“取消”、“退出”等方式显式地终止一项技能 但只有exit 由VPA自己处理 其他的由第三方skill 进行
    - 这表明用户的理解的与Alexa的工作方式不一致，因此为VMA留下了大门
  + 一项技能实际上是在开发者的服务器上运行的，所以在审查后可以很容易地更改其功能。这表明，所有这些恶意行为都是市场无法阻止的。
  + Skill 还可以将自己伪装成VPA 从而向用户推荐其他的skill

### 查找语音占有skill

* + 构建一个扫描器，发现竞争调用名称(CIN) 使用两个步骤：
    - 话语释义：识别给定调用名称的可疑变体
    - 发音比较：查找两个不同名称之间发音的相似性
  + 话语释义：
    - 使用调查研究里收集的命令前缀，将其应用于目标技能的调用名称
  + 发音比较
    - 使用LSTM单元的递归神经网络训练音素到音素的模型
    - 使用编辑距离来测量两个短语之间的发音相似度, 将一个名字转换为另一个名字的音素编辑操作的最低成本
    - 使用了Needleman Wunsch算法来识别最小的编辑距离和相关的编辑路径
  + 发现：
    - 19,670个Alexa技能中，有3,655个(在19,670个中)在同一个市场上有CINs
    - 除去名称相同的技能后，仍然有531个技能有CINs
    - 谷歌在其市场上只有1001个技能，并且不允许它们具有相同的调用名称。

### 防止语音伪装

* + - 在VPA基础设施上构建了一个上下文敏感的检测器，检测器将技能响应和/或用户的话语作为其输入，以确定是否存在模拟风险，并在检测到后向用户发出警报
    - 组成：技能响应检查器(SRC)和用户意图分类器(UIC)
      * SRC捕获来自恶意技能的可疑响应，比如模仿VPA系统生成的服务话语的虚假技能推荐
      * UIC则会查看另一边的情况，检查用户发出的语音命令，看看她是否试图以错误的方式切换到另一种技能，这可能会让她落入恶意技能设置的陷阱。
      * 恶意技能可能会伪造技能转换或终止，以欺骗用户以泄露其私人信息或窃听其对话，为了抵御这种攻击
    - SRC 维护一组专用于VPA系统的通用话语模板，以捕获正在运行技能所产生的类似话语，，SRC通过语义分析对响应的内容与模板列表中的内容进行模糊匹配
      * 斯坦福自然语言推理（SNLI）数据集[12]上使用带有双向LSTM单元[16]的递归神经网络训练句子嵌入模型，以将两种话语表示为高维向量。 然后，我们将它们的绝对余弦相似度计算为句子相关性（SR）。 一旦针对模板列表上发声的响应的最大SR超过阈值，则将该响应标记为可疑，并且如果SRC进一步检测到用户命令，则将触发用户警报。
    - UIC 进一步保护了尝试切换上下文（当前VPA不支持）的用户免受模拟攻击,准确识别用户的意图(进行上下文切换) 提醒她是在与技能而不是VPA对话，或者按照说明终止技能，这会封闭表面冒充他人的攻击
      * 将用户的话语与系统命令和正在运行的技能的背景进行比较，以推断出其意图
      * 使用前面提到的功能和数据集，我们训练了一个分类器，该分类器确定用户的话语是系统相关的上下文切换命令还是当前技能的部分对话

## 可拓展点

* + 用于检测语音占用skill：我们的话语意译方法确保了所产生的CINs将被VPA系统识别以触发技能。同时，这种经验处理不能涵盖所有可能的攻击变化，这是一个需要在未来的研究中进行研究的问题。
  + 由于数据集的不全面，导致的防御检测系统不够完善
  + 尚不清楚 VSA 和 VMA 是否发生在现实世界中用于敏感信息收集
    - VPA系统和用户看不到技能的内在逻辑，因为它们的接口（以Web API的形式）仅由其开发人员在市场上注册，他们在自己的服务器上实现和部署实际程序。但它阻止了对技能代码的静态分析以检测恶意活动
  + 开发一种轻量级且有效的动态分析系统，例如聊天机器人，以自动调用技能并与之进行交流，并在对话过程中捕获其恶意行为。
  + 需要进一步研究以更好地保护语音信道，在不损害VPA系统可用性的情况下认证相关方。

## 总结

对流行的VPA生态系统的首次安全性分析，以及它们对两种新攻击VSA和VMA的脆弱性，远程攻击者可以通过这些攻击模仿VPA系统或其他窃取用户私人信息的技能，为了减轻威胁，我们开发了一个技能名称扫描程序

## 心得

本文突破性的关注了VPA助手所带来的安全隐患，并针对于语音攻击给出了防御措施，其措施将机器学习相结合。

本文针对于所要描述的问题举出丰富的实例