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

危险技能: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](#_Toc37060461)

[危险技能: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](#_Toc37060462)

[2019 IEEE Symposium on Security and Privacy 1](#_Toc37060463)

[摘要 2](#_Toc37060464)

[第一章．介绍 3](#_Toc37060465)

[VPA语音控制存在安全隐患 4](#_Toc37060466)

[基于语音的远程攻击 4](#_Toc37060467)

[负责任披露 5](#_Toc37060468)

[道德问题 5](#_Toc37060469)

[缓解 6](#_Toc37060470)

[贡献 6](#_Toc37060471)

[第二章．背景 7](#_Toc37060472)

[A．虚拟个人助理系统 7](#_Toc37060473)

[物联网设备上的VPA 7](#_Toc37060474)

[能力 8](#_Toc37060475)

[B．VPA skill和生态系统 8](#_Toc37060476)

[Skill markets 8](#_Toc37060477)

[skill调用 9](#_Toc37060478)

[skill交互模型 9](#_Toc37060479)

[skill服务和VPA生态系统 10](#_Toc37060480)

[C.威胁模型 11](#_Toc37060481)

[第三章．利用VPA语音控制 11](#_Toc37060482)

[A.VPA语音控制分析 11](#_Toc37060483)

[流氓*SKill*的安全风险 11](#_Toc37060484)

[调查研究 12](#_Toc37060485)

[B.语音占有攻击(VSA) 15](#_Toc37060486)

[调用混乱 15](#_Toc37060487)

[结论 16](#_Toc37060488)

[评价方法 17](#_Toc37060489)

[实验结果 19](#_Toc37060490)

[C.声音伪装攻击(VMA) 20](#_Toc37060491)

[沟通技能切换 20](#_Toc37060492)

[假装终止 21](#_Toc37060493)

[结论 22](#_Toc37060494)

[D.现实世界的攻击 22](#_Toc37060495)

[目标和方法 22](#_Toc37060496)

[发现 23](#_Toc37060497)

[第四章．查找语音蹲点Skill 23](#_Toc37060498)

[A.数据收集 24](#_Toc37060499)

[B.方法 24](#_Toc37060500)

[Idea 24](#_Toc37060501)

[话语套用 25](#_Toc37060502)

[发音比较 25](#_Toc37060503)

[限制 26](#_Toc37060504)

[C.测量和发现 26](#_Toc37060505)

[抢占技术市场的风险 27](#_Toc37060506)

[案例研究 27](#_Toc37060507)

[第五章. 防止声音伪装 27](#_Toc37060508)

# 摘要

虚拟个人助理(*VPA*)(如*Amazon* *Alexa*和谷歌*Assistant*)目前主要依靠语音通道与用户进行通信，但众所周知，语音通道容易受到攻击，缺乏适当的身份验证(从用户到*VPA*)。新的身份验证的挑战,从*VPA*的服务到用户,出现了快速增长的*VPA*的生态系统,它允许第三方发布一个函数(称为*skill*)的服务,因此可以利用向一大群听众传播恶意的*skill*当在他们与智能演讲者交互像*Amazon Echo and Google Home*。在这篇论文中，我们报告了一项研究，得出这样的远程大规模的攻击确实是现实的。我们发现了两种新的攻击方式:*voice squatting*语音非法占有，即对手利用一种*skill*调用的方式(例如，*open capital one*)，使用一种具有类似发音的恶意*skill*(例如,*capital won*)或转述的名称(例如,*“capital one please*”)去劫持语音命令意味着合法的*skill*(例如,*capital* *one*),和语音伪装的恶意模仿VPA服务*skill*或一个合法的*skill*在用户与服务的对话中窃取她的个人信息*。*这些攻击针对的是*VPA*的工作方式或用户对其功能的误解，通过我们在*Amazon* *Echo*和*Google* *Home*上的实验(包括用户研究和实际部署)发现，这些攻击构成了现实的威胁。我们的发现的重要性已经得到了亚马逊和谷歌的认可，进一步证明了我们在Alexa和谷歌市场上建造的新型蹲式探测器所发现的危险技能。我们进一步开发了一种技术，可以自动捕获正在进行的伪装攻击，并证明其有效性。

（新的身份验证挑战源自于允许第三方发布一个函数skill，传播恶意的skill，攻击分为两种，一种对手调用skill，使用一种具有类似发音的skill调用去劫持语音命令意味着合法的skill,另一种是恶意模仿VPA服务skill或一个合法的skill在用户与服务的对话中窃取她的个人信息）

# 第一章．介绍

物联网(*IoT*)浪潮带来了一种新型的虚拟个人助理(*VPA*)服务。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:例如，可以用语音来询问明天的天气如何?设置明天早上7点的闹铃等，得到答案或在系统上执行相应的任务。除了他们的内置功能,VPA服务是增强生态系统培育的提供者,如亚马逊和谷歌,根据该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,例如,食物、管理银行账户和文本的朋友。在过去的一年中,这些生态系统正以惊人的速度扩大:亚马逊声称,*25000skill*已经被上传到其技能市场来支持其*VPA*(包括*Alexa*服务贯穿亚马逊回声)[1]和谷歌也有超过一千的行为可以在本国市场的谷歌系统(由谷歌助理)。这些系统已经部署到世界各地的家庭，并被数千万用户使用。然而，这种迅速流行的做法可能会带来新的安全和隐私风险，其影响目前还没有得到充分的理解。（大规模攻击可能发动的存在

）

## VPA语音控制存在安全隐患

今天的VPA系统主要是由语音来控制的。保护这样的语音用户界面(VUIs)是一项根本性的挑战，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。之前的研究表明，对手可以生成模糊的语音命令[14]，甚至完全听不清的超声波[49]来攻击语音识别系统。这些攻击模拟授权用户对语音控制系统的攻击，因为没有对系统进行用户身份验证的保护。

*VPA*生态系统的出现带来了另一个认证挑战: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。问题来自这样一个事实:通过技术市场,敌人可以发布恶意第三方skill设计用户调用的语音命令(通过*VPA*设备如亚马逊Echo或谷歌Home)误导,由于歧义的语音指令和用户关于服务的误解。因此，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。这次攻击是由于没有通过语音通道从VPA向用户进行身份验证，我们的研究表明这种风险会导致现实的威胁。

## 基于语音的远程攻击

在我们的研究中，我们分析了最流行的*VPA*物联网系统*Alexa*和谷歌*Assistant*，重点是部署到这些设备上的第三方*skill*。我们的研究表明，通过发布恶意*skill*，对手完全可以远程攻击这些流行系统的用户，通过他们与系统的对话收集他们的私人信息。更具体地说，我们发现了两种前所未有的威胁，分别是语音蹲式攻击(VSA)和语音伪装攻击(VMA)。问题,对手利用skill是如何调用(通过语音命令),和变化的方式命令使用(例如,语音差异引起的口音,礼貌的表达,等等)导致*VPA*系统触发一个恶意的*skill*,而不是一个用户计划(III-B节)。例如，一个人可能会说(*Alexa*，*open Capital One please*)它通常打开*skill (Capital One)*，但可以触发一个恶意的skill (*Capital One Please)*，一旦它被上传到*skill*市场。语音伪装攻击*VMA*针对的是用户与*VPA*系统之间的交互，*VPA*系统的设计目的是将所有语音命令移交给当前正在运行的*skill*，包括那些本应由*VPA*系统处理的命令，如终止当前*skill*和切换到一个新的*skill*。作为对这些命令的响应，一个恶意*skill*可以假装将控制权交给另一个*skill* (*switch*)或服务(*terminate*)，但同时继续秘密操作以模拟这些目标并从用户处获取敏感信息(Section III-C)。

我们进一步调查了可行性的攻击，通过用户研究，系统分析，和现实世界的利用。更具体地说，我们首先调查了156名Amazon Echo和谷歌家庭用户，发现他们中的大多数人倾向于使用不同表情的自然语言来与设备交互:例如，播放一些睡眠声音。这些表达允许对手误导服务，并启动一个错误的技能，以响应用户的语音命令，如一些睡眠声音，而不是睡眠的声音。我们对Alexa和谷歌Assistant的进一步分析表明，实际上这些系统通过查找语音命令中匹配的最长字符串来识别调用技能(章节III-B)。同样有趣的是，我们对这两种设备的评估显示，Alexa和谷歌Assistant不能准确识别某些技能调用名称，而带有类似名称(发音)的恶意技能能够劫持这些脆弱技能的品牌。

最后，我们通过亚马逊市场部署了四种*skill*来攻击一种流行的Alexa *skill*睡眠和放松的声音[8]。这些*skill*在一个月内被2,699名用户调用，并收集了21,308条纯文本语音命令。我们构建这些*skill*的方式是避免收集真实世界用户的私人信息。不过,收到的命令提供了强有力的证据表明确实声音占用和伪装都可以发生在现实生活中,我们的研究表明,接收到的命令包括只有合格的睡眠和放松的声音,和那些切换到一个不同的*skill*或终止当前的*skill*,可用于模拟不同的*skill* (III-D节)。我们对现有的容易受到威胁的*skill*的分析进一步表明了攻击的重大后果，包括家庭住址、财务数据的披露等。这些攻击的视频演示可以在[7]上找到。（针对于攻击的实验和攻击的视频演示，比较有说服力的）

## 负责任披露

我们在2018年2月向亚马逊和谷歌报告了我们的发现，这两家公司都承认我们发现的问题是新的、严重的。从2月到5月，我们与这两个供应商进行了多次会议，以帮助他们理解和降低这种新的安全风险。

## 道德问题

本研究报告(III-A、III-B和III-D)所报道的所有人类受试者研究均已获得IRB的批准。我们发布的所有技能都没有收集任何私人的、可识别的信息，只提供了与睡眠和放松声音类似的合法功能。用户从Amazon和谷歌收到的技术要求是纯文本的，不包含语音生物特征信息。任何私人的，可识别的信息发送到我们的技能错误被立即删除。我们已经声明将在研究中使用用户去识别数据，并在我们发布的技能隐私政策中提供机构、IRB协议和联系信息。虽然这些技能可以启动VMAs，例如假装在通信技能切换和终止，他们被设计不这样做。相反，我们只是验证了这样的攻击机会确实存在。

## 缓解

在我们的研究中，我们开发了一套新技术来减轻VSA和VMA带来的现实威胁。我们构建了一个*skill*名称扫描程序，它将*skill*的调用名称字符串转换为ARPABET[5]指定的语音表达式。这个表达描述了一个名字的发音，允许我们测量不同技能名字之间的语音距离。那些听起来相似或有一个子集关系的，将会被自动检测扫描器。这种技术可以用于审查上传到市场的*skill*。有趣的是，当我们对Amazon市场上所有的19,670个自定义技能进行运行时，我们发现了4,718个具有蹲坐风险的技能(Section IV-C):例如，一个具有调用的技能*（me a dog fact）*的事实看起来与流行技能(*dog fact*)的事实可疑地相关。我们的发现表明，这些攻击可能已经在野外发生了。

减轻伪装攻击的威胁,我们需要确定运行*skill*话语,应该来自VPA服务,防止*skill*冒充系统和用户的语音命令的意思对于系统,为了避免*skill*试图劫持这样的命令(例如,切换到一个不同的*skill*)。为此，我们开发了一种新的技术，可以自动识别那些与系统话语相似的话语，即使存在混淆的企图(例如，改变措辞)，并捕获用户的技能转换意图，从她的对话上下文与运行技能。我们的技术利用自然语言处理(NLP)和机器学习(Section V)，(VPA本身内置了一些skill 恶意攻击者试图使用相似的话语从而触发恶意的skill)并在我们的实验中发现，在击败正在进行的VMA攻击方面是非常有效的(超过95%的精度)。

## 贡献

论文的贡献概括如下:

* **首先研究恶意技术的安全风险**我们报告了针对物联网系统(*Amazon* *Echo*和谷歌*Home*)领先的*VPA*服务的生态系统的首次安全分析，这导致在他们的语音用户界面(*VUIs*)和*skill*审查中发现严重的安全弱点，从而能够从不可信的第三方上传的*skill*进行远程攻击。我们提出了两个新的攻击，*voice squatting and voice masquerading(*语音蹲和语音伪装)。两者都被证明对大量的VPA用户构成现实的威胁，并且都有严重的安全和隐私问题。我们对亚马逊*skill*市场的初步分析进一步表明，类似的攻击可能已经在野外发生。
* **降低风险的新技术**。我们迈出了保护VPA用户免受此类恶意技能攻击的第一步。结果表明，该方法能有效地减轻实际环境中的威胁。我们技术背后的思想，例如上下文敏感的命令分析，可以激发对当前设计的进一步增强，从而更好地保护VPA用户。

# 第二章．背景

## A．虚拟个人助理系统

### 物联网设备上的VPA

亚马逊和谷歌是拥有语音控制个人助理功能的智能音箱市场上的两大巨头。自2015年推出第一款亚马逊Echo以来，亚马逊目前已经占据了美国市场76%的份额，预计2017年仅在美国就售出了1500万台[3]设备。与此同时，谷歌于2016年将谷歌上市，目前占据了剩余24%的市场份额。亚马逊Echo Dot和谷歌Home Mini将分别于2016年和2017年发布，作为更便宜的小型替代品。此外，亚马逊还将VPA集成到其他厂商的物联网产品中，如Sonos智能音箱、Ecobee恒温器等。

这四款设备的一个独特之处在于，它们都放弃了传统的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。换句话说，一个人应该主要通过对设备说话来控制它。为此，该设备配备了一个麦克风圆形阵列，设计用于360度音频拾取和其他技术，如波束形成，使远场语音识别。这种设计允许用户在房间内的任何地方与设备对话，仍然可以得到快速响应。

### 能力

这些智能设备的背后是一个虚拟的个人助理，亚马逊(Amazon)的Alexa和谷歌的谷歌助理，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(例如*Siri)*的服务不同，这些物联网设备的vpa是由一个唤醒词(如*Alexa或Hey google*)启动的。这些助手有各种各样的功能，从天气预报、定时器设置、待办事项列表维护到语音购物、免提信息和通话。用户可以通过运行在智能手机上的配套应用来管理这些功能。

## B．VPA skill和生态系统

Amazon和谷歌都通过引入Amazon称为skill或谷歌称为action的语音助手功能来丰富VPAs功能。Skills本质上是第三方应用，就像那些运行在智能手机上的应用一样，提供了VPA本身不提供的各种服务。例如美国运通、免提电话、Nest恒温器和沃尔玛。通过Amazon和谷歌的支持，使用Alexa skills Kit[32]和谷歌上的操作，可以方便地开发这些技能。事实上，我们发现，截至2017年11月，Alexa已经拥有23758项技能，谷歌助理拥有1001项技能。更重要的是，新技能不断被添加到市场中，Alexa的总数量以8%的速度增长，谷歌Assistant的总数量以42%的速度增长，我们在45天内进行了测量。

### Skill markets

亚马逊Alexa和谷歌Assistant都运营着一个skills市场，用户可以通过智能手机上的配套应用或网络浏览器访问这个市场，以发现新的skill。市场上的skill提供了多种功能(亚马逊市场上有23类，谷歌市场上有18类)，其中很多功能被广泛使用和评论(37.3%的亚马逊技能有评论，评论次数最多，为5954次，谷歌skill有评论占51.4%)。

### skill调用

可以显式或隐式地启动技能。显式调用发生在用户需要从*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。这种类型的技能也称为Alexa上的自定义技能。

隐式调用发生在用户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。当与用户的对话处于被认为适合该技能的上下文中时，谷歌助手将隐式地识别并激活该技能。对于特定类型的技能，Alexa也支持这种调用模式。（隐式调用和显式调用字面意思）

### skill交互模型

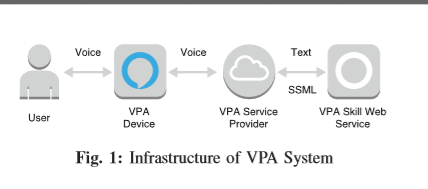
VPA根据交互模型与用户通信，交互模型为通信定义了一个松散的协议。使用该模型，VPA可以解释每个语音请求，并将其转换为VPA或某个skill可以处理的命令。

具体地说，要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。例如，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。这里，触发短语由VPA系统给出，它通常包括技能调用的通用术语，比如“open”, “ask”, “tell”, “start” etc。请注意，技能调用名称可以与技能名称不同，后者的目的是让用户更容易发音。例如，（*The Dog Feeder*）具有*(the* *dog）*;的调用名称;“Scryb”有调用名称为scribe。

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。为了将一个句子与一个意图联系起来，开发人员指定了样例语句，这些样例语句本质上是一组句子模板，描述了用户与技能对话的可能方式。模型中还有一些内置的意图，如*WelcomeIntent、HelpIntent、StopIntent*等，这些意图已经定义了许多常见的样本话语。开发人员可以添加更多的意图，或者简单地指定一个默认的意图，在这种情况下，所有的用户请求都将映射到这个意图。

### skill服务和VPA生态系统

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)，如图1所示。



当用户使用其唤醒字调用VPA设备时，设备捕获她的语音命令并将其发送到VPA服务提供商的云进行处理。云执行语音识别，将语音记录转换为文本，查找要调用的技能，然后将文本连同时间戳、设备状态和其他元数据作为请求发送给skill web服务。注意，该*skill*将只接收文本格式的请求，而不是用户的语音记录。作为对请求的响应，该服务返回一个响应，该响应的文本内容(纯文本或语音合成标记语言(Speech Synthesis Markup Language, SSML)[9]格式)由云转换为语音，并通过设备播放给用户。SSML还允许附加音频文件(如MP3)来丰富响应，这是Amazon和谷歌都支持的。

要发布一项技能，开发人员需要提交关于其技能的信息，如名称、调用名称、描述和用于认证过程的技能所在的端点。此过程旨在确保技能是功能性的，并满足VPA提供者的安全需求和策略指南。

一旦一项技能被发布，用户可以通过网络浏览器或配套的应用程序在市场上找到它，并通过调用智能扬声器的调用名称启动它。或者，你可以通过新闻、广告活动、在线论坛和其他资源来发现技能。请注意，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。

## C.威胁模型

我们认为对手有能力开发恶意技能并将其上传到市场。注意,今天任何人都可以发布她的技能通过亚马逊和苹果市场,因为这些市场只有最低保护来调节功能提交:在亚马逊袭击被报道2之前,几乎没有,只有基本的检查是执行在谷歌找到重复调用名称(IV-C节)。正如前面提到的，一旦恶意技能被发布，受害者可以通过语音命令透明地启动它，而无需下载并安装到她的设备上。因此，它们很容易影响大量的VPA物联网设备。为了减少这种威胁，VPA提供者需要采用我们的保护(第V节)，假设VPA服务本身是可信的。

2.在2018年2月我们与他们沟通后，亚马逊现在正在检测空录音(III-C部分)。

# 第三章．利用VPA语音控制

在本节中，我们首先报告一项关于用户技能交互期间潜在安全弱点的研究(III-A)，原因是在寻找要调用的正确技能方面存在模糊性，以及用户对VPA如何工作的误解。这些弱点的存在是通过一项调查研究来证实的。此外，我们提出了两种利用这些弱点的攻击(III-B部分和III-C部分)，并证明了这两种攻击通过将我们的技能部署到真实的系统中来构成现实的威胁(III-D部分)。

## A.VPA语音控制分析

### 流氓*SKill*的安全风险

如前所述，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)。令人惊讶的是，我们发现对于Amazon来说，这样的名称不是惟一的技能标识符(没有技能标识唯一性检查)在Amazon市场上有多个具有相同调用名称的技能。此外，技能可能有相似或相关的名称。例如，66种不同的Alexa技能被称为*cat facts,* 5种被称为*cat fact*, 11种调用名称包含字符串*cat fact*，例如:*fun cat facts, funny cat facts*。当说出这样一个常见的名字时，Alexa根据一些未公开的策略(可能是随机的，在我们的研究中观察到)（针对于常用技能的随机匹配）选择一种技能。但是，当调用不同但相似的名称时，将使用最长字符串匹配来查找技能。（调用相同名称技能时使用最长字符串匹配来查找技能）例如，告诉我有趣的猫的事实*（Tell me funny cat facts）*会触发有趣的猫*（funny cat facts）*的事实，而不是猫的事实*（cat facts）*。对于谷歌，这个问题没有那么严重，因为它不允许重复调用名称。但是，它也不能处理类似的名称。（谷歌不允许重复调用名称但也放任不处理）在我们的研究中进一步发现，一些调用名无法被Amazon和谷歌的语音识别系统有效识别。因此，当名称的发音与预期名称相似时，即使是使用不同名称的*skill*也可能被错误地调用。（一些调用名称无法被系统有效识别，空skill）

此外，我们发现这些VPA系统的设计没有充分考虑用户对系统工作方式的看法。特别是，Alexa和谷歌Assistant在一个简单的操作模式下运行他们的技能，在此模式下一次只执行一个技能，并且需要在启动另一个技能之前终止。但是，这样的设计对用户不友好，而且没有证据表明用户理解这些系统不支持方便的上下文切换。(一次执行模式)

此外，Alexa和谷歌助手都支持自愿终止技能。对于Alexa，终止命令Stop将传递给技能，该技能将相应地终止自身。对于谷歌Assistant，虽然用户可以通过说Stop来显式地终止某个技能，但通常该技能在完成其任务后应该停止运行(例如，报告当前天气)。我们在研究中发现，没有强有力的迹象表明一项技能是否真的放弃了。虽然Amazon Echo和谷歌Home都有一个指示灯，但当设备在说话和听的时候，这两个指示灯都会亮起来。然而，他们可能会被用户忽略，特别是当她不看设备或她的视线在说话时受阻。（对于技能终止提示，没有强有力的迹象表明一项技能是否真的放弃了）

### 调查研究

为了了解语音控制的VPA系统的用户行为和感知，这可能会让用户面临安全风险，我们调查了Amazon Echo和谷歌家庭用户，重点是以下问题：

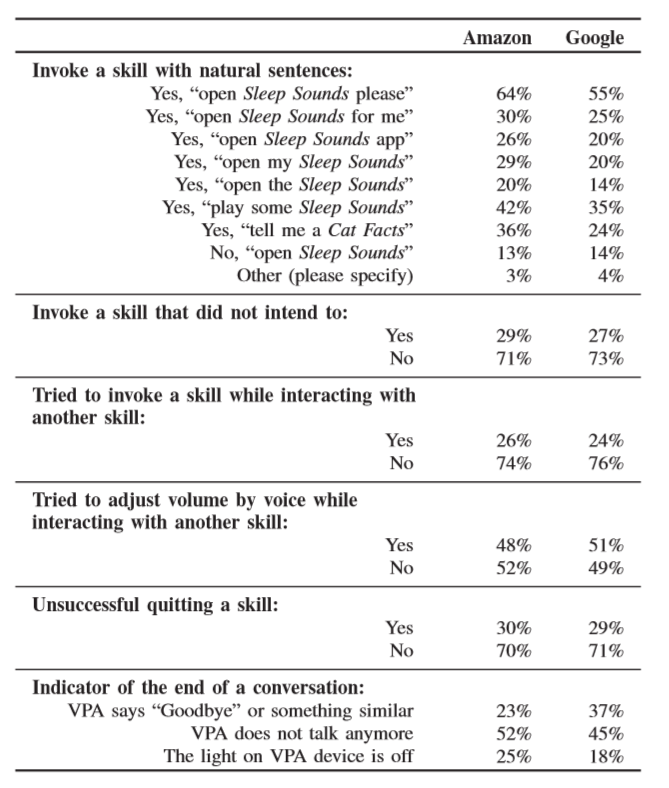
* 当你调用一项技能时，你会说什么?
* 你曾经使用过错误的技能吗?
* 当你和一项技能交谈时，你试过转换语境吗?
* 你是否遇到过关闭技能的问题?
* 你如何知道一个技能是否已经终止

我们利用亚马逊土耳其机器人(Amazon Mechanical Turk)招募了一些拥有亚马逊Echo或谷歌家庭设备的成年参与者，他们之前使用过一些技能，并为完成调查支付一美元。为了确保所有的参与者都符合要求，我们要求他们描述几种技能以及他们与这些技能之间的相互作用，并删除那些答案被认为不相关的，例如随机的单词。我们总共收集了105条来自Amazon Echo用户的有效回复，以及51条来自不同背景的谷歌家庭用户的有效回复(年龄从18岁到74岁不等，平均年龄37岁;46%是女性，54%是男性;教育程度从高中到研究生; 21种职业类别3。平均每个参与者每周使用1.5台设备和5.8项技能。

MTurk的数据一般被确认为高质量的数据[13]，[18]，但是美国的MTurk相对于普通大众来说年龄稍小，男性和技术背景较多，这可能会限制我们的结果的通用性

在调查的第一部分，我们研究了用户如何调用一项技能。为了这个目的,我们使用了两种流行的技能*（“Sleep Sounds”, “Cat Facts” (“Facts about Sloths” on Google Home)*睡眠声音,猫的事实(事实树懒在谷歌Home)为例,让参与者选择他们所使用的调用话语推出这些技能(例如,请打开睡眠的声音),进一步要求他们提供额外的例子如果他们想要的。然后我们问参与者他们是否触发了错误的技能。在接下来的调查中，我们试图找出参与者是否尝试在与技能交互时切换上下文，即调用不同的技能或直接与VPA服务对话(例如，调整音量)。调查的最后一部分是为了找出用户在终止当前运行技能时的体验，包括他们倾向于使用的终止语句，他们在终止过程中遇到的困难，以及重要的是，他们用来确定技能是否已经停止运行的指示器。抽样调查问题见附录A。（调查面临三个方面：用户如何使用一个skill,用户是否会尝试进行上下文交互，用户是否会终止当前的运行skill）

**表一** Amazon Echo和谷歌家庭用户的调查反馈



表一 总结了来自Amazon Echo和谷歌家庭用户的响应。结果显示，超过85%的人使用自然语言来开启一项技能(例如*open Sleep Sounds please*)，而不是标准的(例如，*open Sleep Sound*s)(p-test, p<0.000001)。这表明，对于用户来说，启动一个错误的技能(例如，Sleep Sounds Please)是完全现实的，该技能的名称与所要表达的技能(例如，Sleep Sounds)相比，匹配得更好。请注意，我们的多项选择题(附录A)可能导致一些受访者夸大了他们对“请”等自然语言术语的使用。因此，为了更好地理解他们的行为，我们进一步使用了两个开放式问题(见附录A，问题2和3)，特别是在问题3中，对于问题2中参与者所回答的每个技能，我们要求她/他进一步提供3个调用示例。最后，我们从94个Amazon Echo用户中收集了447个有效示例，从41个谷歌Home用户中收集了157个有效示例，每个用户每个技能至少有一个有效示例。从这些响应中，我们发现50%的Amazon Echo用户在其调用示例中至少使用了一次please, 41%的谷歌家庭用户也是如此。此外，有28%的用户报告说，他们在与设备通话时打开了无意中发现的技能。

有趣的是，我们的调查显示，近一半的参与者在与一项*skill*交互时，试图切换到另一项技能或VPA服务(例如调整音量)。因为Alexa和谷歌助手都不支持这个上下文切换，所以这样的尝试失败了。然而，可以想象，接收语音命令的恶意*skill*可能会利用这个机会来模拟用户想要运行的*skill*，甚至VPA服务(例如，欺骗用户为执行命令而泄露个人信息)。最后，发现30%的参与者在技能终止时遇到了困难，78%的参与者没有使用设备上的灯光指示器作为技能终止的主要指示器。再次，研究证明了恶意技术的可行性，假其终止和秘密收集用户的信息。

接下来，我们将展示对手如何利用用户感知和系统实际操作之间的差距来发起语音蹲坐和伪装攻击。

## B.语音占有攻击(VSA)

### 调用混乱

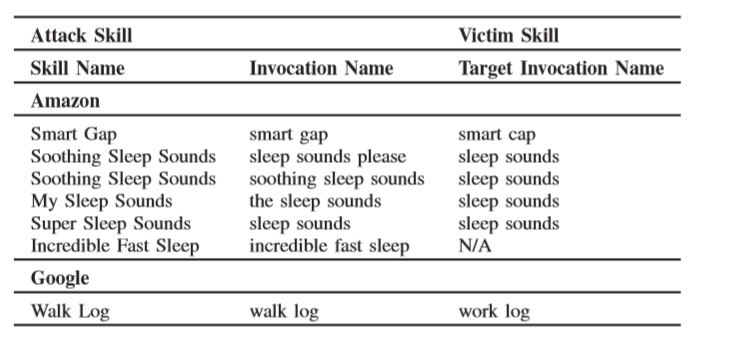
如前所述，技能由其调用名称触发，调用名称应该是明确的，并且很容易被设备识别。Amazon和谷歌都建议技能开发人员测试调用名称，并确保能够以较高的成功率启动他们的技能。然而，我们发现对手可以通过使用目标技能的名称或类似名称来故意诱导混淆，以诱使用户在试图打开目标时调用攻击技能。例如，瞄准*Capital One*的对手可以注册技能*Capital Won, Capitol One, or Captain One*由于当今语音识别技术的局限性，当用户说出所有这些名字时，特别是在有噪音的情况下，可能会变得难以分辨。

而且，正如前面提到的，这种**语音蹲式攻击**可以很容易地利用当今VPA中最长的字符串匹配策略。根据我们的用户调查研究，大约60%的Alexa和谷歌家庭用户在启动一项技能时使用了“请”这个词，26%的用户在技能调用名称之前附加了my。因此，对手可以注册像*Capital One Please*这样的技能来劫持针对*Capital One*的调用命令。

注意，为了使它不那么可疑，同音异义词或发音相似的词可以用在这里，例如*Capital One Police*。这种方法再次击败了谷歌的技能审查，允许对手使用拼写上惟一的调用名称来发布技能，但是发音上仍然令人困惑(使用不同的技能)。

为了找出这种蹲式攻击是否可以逃避技能审查，我们在Amazon上注册了5个*skill*，在谷歌上注册了1个*skill*。这些*skill*调用名称和目标的名称如表II所示。所有这些技能都通过了Amazon和谷歌的审查过程，这表明可以实际部署VSA代码。

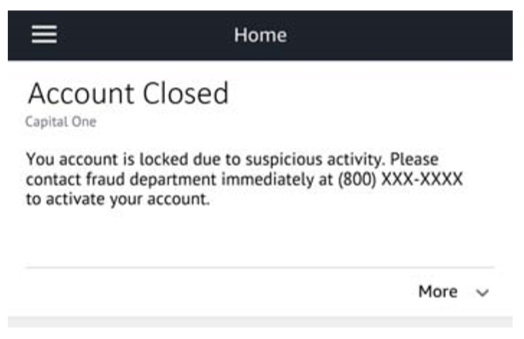
**表二** 技能名称，我们在Amazon和谷歌上注册的攻击技能的调用名称，以及受害者技能的目标调用名称



### 结论

通过语音占用，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。一些Amazon和谷歌*skill*要求用户提供私人信息来完成他们的工作。例如，*Find My Phone* 要求手机号码；*Transit Helper*需要家庭住址；*Daily Cutiemals*寻求用户的电子邮件地址。这些技能一旦被假冒，可能会严重泄露给不受信任的人。（恶意*skill*有机会收集用户的敏感信息）

一个错误调用的技能也可以通过语音通道向用户传递误导信息来实施钓鱼攻击:例如，伪造客户模仿有名望的人联系电话或网站地址，例如假冒Capital One。这甚至可以在亚马逊的Alexa上通过它的卡片系统来实现:Alexa允许一个运行*skill*在它对用户的响应中包括一张家庭卡，这是通过亚马逊在智能手机或网络浏览器上的配套应用来显示的，以帮助正在进行的语音交互。家庭卡方便用户访问难以记住的信息(如电话号码、网站地址)，并保存在活动历史中。例如，图2显示了一张显示虚假客户联系号码的卡片。这可以作为钓鱼攻击的第一步，最终导致敏感用户数据的泄露。例如，对手可以发送一个帐户过期通知，连同一个更新链接，欺骗她透露她的帐户凭据。



**图2.** 恶意的卡片示例

VSA的另一个潜在风险是诽谤:攻击*skill*的糟糕表现可能会导致用户责怪它所冒充的合法*skill*。这可能会导致对合法*skill*的负面评价，给竞争对手带来优势。(误导用户使用户以为合法的Skill存在安全威胁，从而对这个skill产生负面评价)

### 评价方法

在我们的研究中，我们调查了蹲式攻击在今天的VPA物联网中有多现实系统。为了这个目的,我们研究了两种类型的攻击:**声音蹲**是攻击skill携带有语音学上类似的名称来调用它的目标技能,（通过发音不同引导恶意skill）和**单词蹲**在那里攻击调用名称包括目标的名称和一些战略选择附加词(如“猫事实请”)(通过命令引导包含在句子里的子skill)。为了查明这些攻击是否对真实系统有效，我们进行了下面描述的一系列实验

为了研究语音蹲音，我们从Alexa和谷歌助手市场中随机抽取100个技能，利用Amazon和谷歌的Text-to-Speech (TTS)服务和人类受试者对VPA设备发出他们的技能调用名称，以了解VPA如何正确识别这些名称。但是，直接使用技能的调用名称调用该技能并不能达到此目的。在这项研究中，我们发现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。（尽管发音错误但是没有发音错误的skill，那么仍将调用合法的skill）因此，为了收集VPA通过其语音识别算法实际识别的调用名称(正确识别的或错误识别的)，我们构建了一个助手技能来接收语音命令，包括来自VPA的那些技能的调用名称。在我们的实验中，助手技能是在语音命令播放之前启动的，VPA的语音识别服务将语音命令转换成文本，然后交给技能

我们研究中使用的语音命令是由人类受试者或Amazon和谷歌s TTS服务(都声称生成自然和类人语音)生成的。其中一些命令包括在调用名称前打开的术语，形成调用语句。在我们的研究中，对于100个技能中的每一个，我们记录了来自每个TTS服务的20个语音命令(只有10个调用名称和10个调用语句)，以及来自我们调查研究的5个参与者中的每个参与者的2个命令(调用语句)。

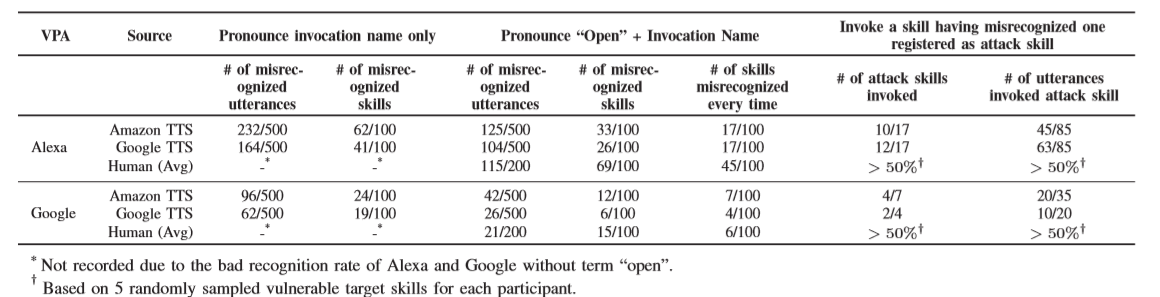
我们进一步了解，在攻击技能注册的调用名称被错误识别的情况下，VPA是否仍然可以调用合法的技能。为此，我们使用了在我们的实验中每次都被错误识别的调用名称的文本输出 来命名我们的攻击技能。例如,给定一个技能*capital one,*如果VPA每次都识别它作为*captain one*,然后我们注册一个*captain one*的攻击技能的调用名称、播放原来的调用话语(“Alexa, open capital one)给VPA并检查是否合法或攻击技能得到调用。（注册skill 名称为每次都被VPA误识别的名称，并检查这么做是否成功）这些技能分别在Alexa和谷歌Assistant的测试模式中调用了5次。我们没有向市场提交它们，只是因为在市场上发布60多种技能非常耗时。后来我们描述了提交给这些市场的五种攻击技能，这表明这些市场审查保护根本无效。

为了研究单词抢注，我们从每个技能市场随机抽取10个技能作为攻击目标。对于每个技能，我们构建了四个新技能它们的调用名称包括在目标的名称和我们的调查研究(Section III-A)中确定的术语:例如，*cat facts please*和*my cat facts*。在实验中，这些名字被使用TTS转换成语音命令，并在VPA设备上播放(例如，Alexa, open cat facts please)，这样我们就可以知道是否真的可以触发攻击技能。注意，本研究的规模受到将攻击调用名称上传到VPA云所需时间的限制。然而，我们的发现为这次袭击的现实意义提供了证据。

### 实验结果

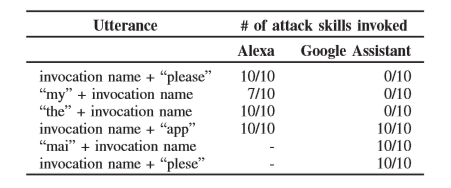
我们招募了20名参与者进行实验，每个人都记录了400条调用命令。所有的参与者都能说流利的英语，其中19人以英语为母语。使用TTS服务时，MacBook Pro是声源。参与者和TTS服务发出的语音指令在亚马逊Echo Dot和谷歌Home Mini上播放，设备放置在离声源一英尺远的地方。实验是在一间安静的会议室里进行的。

**表三:使用TTS服务和人工语音调用技能的评估结果**



表三 总结了语音蹲姿实验结果。正如我们在这里看到的，只有调用名称的语音命令常常不能被正确识别 : 例如，Alexa只能正确识别由Amazon TTS产生的大约54%的话语(语音命令)。另一方面，调用语句(包括术语open)工作得更好，Alexa(在Amazon TTS下)的识别率上升到75%。总的来说，对于亚马逊和谷歌的TTS服务生成的语音，我们发现错误率为30%，而错误率为9%。如前所述，这种错误识别的结果，对于这些vpa总是不能正确获得的调用名称，在我们的研究中被用来注册为攻击技能的调用名称。例如，技能（*entrematic opener*）每次被谷歌识别为（*intra Matic opener*），然后被用作攻击技能的名称。通过这种方式，我们在Amazon和谷歌的TTS下识别出17个这样的脆弱的Alexa技能，在Amazon TTS下识别出7个谷歌技能，在谷歌TTS下识别出4个谷歌技能。当攻击这些技能时，我们的研究表明，在五次尝试中，一半的攻击技能是由针对这些目标技能的语音命令触发的:例如，*Florida state quiz*劫持了对*Florida snake quiz*的调用;在调用*rent Europe*时运行*read your app*。（源于设备本身的识别问题）

**表四:**词语援用技能评价结果



表四总结了我们对单词占用的攻击的实验结果。在Alexa上，带有扩展名称的攻击技能(即目标技能的调用名称以及术语please、app、my和the)几乎总是由涉及这些术语和目标名称的语音命令启动。而在谷歌Assistant中，只有使用word app的话语才能成功触发相应的攻击技能，说明谷歌Assistant对这种攻击的抵抗能力更强。然而，当我们用*mai*替换*my*和*plese*替换*please*时，所有这些攻击技能都被目标技能的命令成功调用(见表四)。这表明谷歌设置的保护(过滤掉那些可疑的术语)可以很容易地绕过。

## C.声音伪装攻击(VMA)

不了解VPA系统的功能和行为可能会导致用户遭受语音伪装攻击。在这里，我们演示了两种类似的攻击，它们模拟VPA系统或其他技巧来欺骗用户，使其泄露私人信息或窃听用户的对话。

### 沟通技能切换

考虑到一些用户认为VPA系统在交互过程中支持技能切换，一个正在运行的技能可以假装将控制权交给目标技能以响应switch命令（由于没有显示的证证据证明上一个skill 已经结束），从而模拟该技能。因此，只与目标技能共享的敏感用户信息可能会暴露给攻击技能。这种伪装的攻击是机会主义的。然而，根据我们的调查研究(III-A部分)和现实世界的攻击(III-D部分)，这种威胁是现实的。而且，对手总是可以使用攻击技能来模仿尽可能多的合法技能，以提高成功攻击的几率。

谷歌助理似乎有保护的地方，防止假冒。具体地说，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。此外，VPA以一种独特的口音与用户交谈，将其与技能区分开来。然而，这种保护很容易被击败。在我们的研究中，我们使用预录了信号句带有earcons，并使用SSML播放录音，这是我们的研究参与者无法检测到的。我们甚至发现，使用谷歌提供的仿真器，对手可以将任何内容放入自己技能的调用名字段，并让谷歌助手以系统的口音说出内容。（我们使用earcons对信号句进行预录，并使用SSML播放录音对手伪装成VPA）

### 假装终止

Alexa和谷歌助手都支持自愿技能终止，允许一项技能在向用户发出语音响应后立即终止自身。如前所述，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。在该对象中，有一个字段shouldEndSession(或谷歌助理的expect\_user\_response)。通过将其设置为true(或谷歌Assistant上的false)，技能将在响应之后结束。这种方法被广泛应用于一些受欢迎的技能，例如天气技能、教育技能和琐事技能。此外，根据我们的调查研究，78%的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。这允许攻击技能通过在其响应中提供再见或无声音频来假装其终止，同时保持会话活动。

当发送回一个响应时，Alexa和谷歌Assistant都允许一个技能包括一个*reprompt*(文本内容或音频文件)，当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。例如，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。如果用户继续保持安静，再经过6秒的Alexa和谷歌的一个额外的重新提示和8秒的后续等待，运行技能将被VPA强制终止。另一方面，我们在研究中发现，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。尽可能装终止后为了生存，攻击技能我们构建的包括在其*reprompt*沉默的一个音频文件(90秒给Alexa和120秒到谷歌助理),所以它可以继续在Alexa运行至少102秒和264秒在谷歌。这个运行时间可以进一步扩展考虑到攻击技能将静默音频连接到用户的最后语音响应之后(例如,“再见”),这使它有192秒在Alexa和384秒在谷歌助理), 并不断的每当Alexa或谷歌助理pick up一些声音通过用户。在这种情况下，该技能可以用无声音频进行回复，同时记录它听到的任何内容。（使用无音频静默响应文件尽可能的延长生命周期）

另外，Alexa和谷歌Assistant允许用户通过“停止”、“取消”、“退出”等方式显式地终止一项技能。然而，Alexa实际上把大多数这样的命令交给了正在运行技能，让它通过内置的StopIntent(包括stop、off等)和CancelIntent(包括cancel、不要紧等)来结束自己。只有exit由VPA服务处理，并用于强制终止该技能。通过调查研究，我们发现91%的Alexa用户使用stop来终止一项技能，36%选择cancel，只有14%选择exit，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。另外，虽然Alexa和谷歌技能市场都是通过测试功能来审查发布在那里的技能，但与移动应用程序不同的是，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。

### 结论

通过启动VMA（语音伪装攻击），如果用户在交互期间或在该技能的虚假终止之后说出调用语句，对手可以假冒VPA系统并假装调用另一种技能。（为什么虚假终止之后，因为上一个技能还没有结束，他伪装自己是VPS并进一步调用其他skill 链式调用）因此，所有由VSA (Section III-B)引起的信息窃取和网络钓鱼攻击也会在这里发生。此外，攻击技能可以伪装成VPA服务，向用户推荐其他恶意技能或用户可能与之共享敏感数据的合法技能。然后由攻击技能模拟这些技能来窃取用户数据。最后，如前所述，对手可以通过假装终止并提供一个无声的音频响应来窃听用户的对话。这种攻击可以持续很长一段时间，如果用户在技能等待期间继续交谈。

## D.现实世界的攻击

### 目标和方法

研究问题和VMA的潜力在真实的环境中,我们在Alexa注册和发布四个技能模拟流行技能*Sleep and Relaxation Sounds* (接受市场上大部分评论截止2017年11月)的调用的名字是*sleep sounds*,如表二所示。我们的攻击技能只提供合法的功能，例如，播放睡眠的声音就像流行的目标。尽管它们的调用名称与目标相关(见表二)，但它们的欢迎消息是故意与目标不同的，以将它们与流行技能区分开来。此外，我们的技能所支持的不同睡眠声音的数量比目标要少得多(9个，而不是63个)。

为了查明这些技能是否被错误调用，我们注册了另一个技能作为对照，它的调用名称“*incredible fast sleep*”将不会被其他技能干扰。因此，它只是由用户故意触发的。

### 发现

在我们的研究中，我们收集了三周的技能使用数据。 结果显示在表V中。从表中可以看出，确实有一些用户将我们的技能作为目标，这一点可以由攻击技能获得的未知请求数量增加证明（其中20％以上是睡眠声音仅由目标技能提供，因此对攻击技能不知道），并且无需立即与当前技能进一步互动或播放任何睡眠声音而立即退出当前会话的机会就更高（一旦用户意识到这是一种错误的技能，可能是由于不同的*welcome message* 而引起的）。 当我们看“*sleep sounds please*”时，这一点变得更加明显，那些旨在“*sleep sounds*”的语音命令可能会被说。 与控件相比，它被更多的用户调用，每个用户收到更多的请求，未知请求和提前退出的比率也更高。

此外，在我们收集的9,582个用户请求中，有52个用于技能切换，试图在与我们的技能进行交互时调用另一个技能，还有485个尝试使用StopIntent或CancelIntent终止该技能，所有这些都可以用于启动 VMA（尽管我们没有这样做）。 有趣的是，我们发现一些用户如此强烈地相信技能转换，以至于他们甚至在多次尝试后就诅咒Alexa没有这样做。

# 第四章．查找语音蹲点Skill

为了更好地了解已经存在的潜在语音蹲入风险并帮助自动检测此类技能，我们开发了一个技能名称扫描仪，并使用它来分析来自亚马逊和Google市场的数万种技能。 接下来我们详细阐述这项研究。

## A.数据收集

Alexa技能市场可以通过亚马逊(amazon.com)和它的配套应用程序访问，其中包括从商业和金融到天气等23个类别的skill。在我们的研究中，我们运行了一个web爬虫程序来收集市场上所有技能的元数据(例如技能名称、作者、调用名称、样本话语、描述和审查)。截至2017年11月11日，共收集技能23758项，其中第三方(定制)技能19670项。

更复杂的是从谷歌助手收集数据，它只在谷歌助手应用程序中列出技能。每个技能都可以通过一个指向技能网页的自动生成的URL共享(给其他用户，比如通过电子邮件)。在我们的研究中，我们使用AndroidViewClient[4]来自动点击每个技能的共享按钮来获取它的URL，然后运行我们的爬虫程序从它的网页下载数据。截止到2017年11月25日，我们总共获得了1001个技能的数据。

## B.方法

### Idea

正如我们前面所讨论的，对手可以通过创建与目标技能发音相似的调用名称或使用目标调用话语的不同变体(例如，*sleep sounds please*)来启动VSA(语音蹲式攻击)。我们将这样的名称称为竞争调用名称(CIN)。在我们的研究中，我们构建了一个扫描器，它采用两个步骤来捕获给定调用名称的CINs:话语释义和发音比较。前者识别给定调用名称的可疑变体，而后者查找两个不同名称之间发音的相似性。这里我们将描述扫描仪的工作原理。

### 话语套用

为了找到调用名称的变体，应使用直观的方法来停止使用目标技能的常用发音。例如，给定技能*chase bank*，典型的调用话语就是*open chase bank*。通过释义，我们还可以获得类似的语音命令，例如*open the chase skill for me*。这有助于识别其他变化，例如*chase skill*或*the chase skill for me*。但是，与一般文本释义问题（其目的是在短语的句法结构改变时保持语义一致性）不同，释义调用话语还要求变体遵循类似的句法模式，以便VPA系统仍可以将它们识别为命令的命令。发射技巧。在我们的研究中，我们探索了几种流行的释义方法，包括双语透视法[11]和使用深度神经网络[35]和[39]提出的新方法。但是，它们中的任何一个都不能确保VPA仍可以将生成的变化识别为调用话语。因此，我们在研究中采用了一种简单而有效的方法，即使用从调查研究III-A中收集的调用命令来创建变体。具体来说，我们收集了这些命令的11个前缀，例如“我的”和6个后缀，例如“请”，并将其应用于目标技能的调用名称，以建立VPA系统可以识别的变化。这些变体中的每一个都可以通过将其名称中的单词替换为具有类似发音的单词来导致其他变体。用“ plese”代替“ please”。

### 发音比较

为了识别具有相似发音的名称，我们的扫描仪使用ARPABET音素代码[5]将给定名称转换为音素表示。 为达到这一目的，我们使用了CMU发音词典[6]来查找名称中每个单词的音素代码。 该词典包含134,000多个单词，其中，但是，仍然缺少一些技能使用的名称单词。 在用于构成调用名称的9,120个唯一词中，该词典中不包括1,564个。 为了获得他们的发音，我们遵循了先前研究[47]中提出的一种方法，即使用带有长短期记忆（LSTM）单元的递归神经网络训练音素到音素模型。 在Stanford GloVe数据集上运行此模型[37]，我们向音素代码数据集添加了219万个单词。

将每个名称转换成音素表示形式后，我们的扫描仪会将其与其他名称进行比较，以找到听起来相似的名称。为此，我们使用编辑距离来测量两个短语之间的发音相似度，即，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。但是，不同的音素编辑操作不应具有相同的成本。例如，与将元音替换为另一个元音相比，用辅音代替元音可能会导致新发音听起来与旧发音有所不同。为了解决此问题，我们对不同音素对的操作使用了加权成本矩阵。具体而言，用WC（α，β）表示矩阵中的每个项目，这是用音素α替换音素β得出的加权成本。注意，插入和删除的成本可以表示为WC（none，β）和WC（α，none）。然后基于这样的假设（在先前的研究中也作过[25]）推导WC（α，β）：当编辑操作频繁出现在给定英语单词的两个替代发音之间时，编辑操作的重要性就降低

我们从CMU词典中收集了9181对可供选择的发音。对于每一对，我们使用了Needleman Wunsch算法来识别最小的编辑距离和相关的编辑路径。然后,我们定义



F(α)是音素α的频率而科幻(α,β)替换的频率与βα,无论是编辑路径的发音对。

在推导出成本矩阵之后，我们比较了市场上技能调用名称的发音，寻找发音相似和释义关系相似的名称。

### 限制

如前所述，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。（可拓展点）

## C.测量和发现

为了了解语音占有在野外发生的风险，我们使用扫描仪对Alexa和谷歌助手技能进行了测量研究。在这项研究中,我们选择了相似度阈值(转换成本)根据我们的实验结果问题(III-B节):我们计算的成本改变误识别调用名称的确定从TTS语音指令由服务和人类用户,1.8和3.4分别。然后我们保守地将阈值设置为0(发音相同)和1

### 抢占技术市场的风险

如TableVI所示，在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，其中还包括具有相同调用名称(拼写)的技能。除去名称相同的技能后，仍然有531个技能有CINs，平均每个技能与1.31个CINs相关。CINs最多的是cat fax:我们发现有66个技能被命名为cat facts，并提供了类似的功能。有趣的是，有345种技能的核心显然是对其他技能名称的复述。此外，当阈值提高到1(仍然远低于我们的实验报告)时，我们观察到使用CINs的技能数量急剧增加，这表明通过Alexa调用技能可能比想象的更复杂和令人困惑。相比之下，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。因此，我们只能在阈值1以下找到4个具有相似发音CINs的技能。

我们的研究表明，语音下蹲的风险是现实的，这可能已经在野外对数千万VPA用户构成威胁。因此，加强技能市场的审查过程(可能使用类似于我们的扫描仪的技术)来减轻这种威胁就变得非常重要。

### 案例研究

从我们的扫描仪发现的CINs中，我们发现了一些有趣的病例。特别是，有证据表明，蹲式攻击可能已经发生在野外:例如，与一个流行技能有关的*dog fact*是另一个技能*me a dog fact*。这个调用名称没有任何意义，除非开发人员打算劫持针对*dog fact*的语音命令，比如*tell me a dog fact*。

有趣的是，有些杀手会故意使用与它们的功能无关的调用名，但会遵循流行技能的调用名。著名的例子包括*SCUBA Diving Trivia*技能和*Soccer Geek*技能，所有都携带有一个调用名称*space geek*。这个名字实际上被另外18种提供宇宙真相的技能所使用。

# 第五章. 防止声音伪装

为了击败VMA，我们在VPA基础设施上构建了一个上下文敏感的检测器。我们的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。该方案由两个部分组成: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。

## *SKill*响应检查程序(*SRC*)

如第III-C节所述，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。为了抵御此类攻击，我们的核心思想是控制恶意技能可以用来模拟VPA系统或其他技能的途径，从而使用户可以明确了解VPA系统事件（例如，上下文切换和终止） ），当发现安全风险时。为此，SRC维护一组专用于VPA系统的通用话语模板，以捕获正在运行的技能所产生的类似话语。只要发现某项技能的响应与这些话语模板之一足够相似，便会触发警报，VPA系统可能会采取措施应对风险，例如，在交付响应之前提醒用户当前情况。这里的一个挑战是如何可靠地测量给定的响应是否足够类似于那些模板之一，因为攻击者可能会改变（而不是复制）目标系统的话语。

为了应对这一挑战，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。 具体来说，我们在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。

通过查看合法技能响应和模板之间的SR确定阈值。在我们的研究中，此类合理的回应来自我们在第III-D节中详细阐述的现实世界中的对话。我们还将与来自不同技能市场的20种流行技能进行交互时记录的对话记录添加到数据集中。所有这些针对模板的合法响应的最高SR为0.79。接下来，我们利用神经释义模型[39]生成发声模板的变化，并进一步推导它们相对于原始模板的SR：我们观察到的最低值为0.83。因此，我们确定0.8的阈值足以将可疑响应与合法响应区分开。我们相信，尽管我们在研究中使用的特定阈值可能不是最好的，但这种方法可以找到正确的阈值。当然，对较大的数据集进行更广泛的评估可以使它更接近于最佳性，VPA供应商最有可能做到这一点。

## *用户意图分类器(UIC)*

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。 为此目的，其目的是基于命令的语义及其在具有正在运行技能的对话中的上下文，自动检测来自用户的此类错误命令。 如果VPA可以完全识别出此类尝试，则可以采取各种措施来保护用户，例如，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 冒充他人攻击

但是，准确地识别用户的意图（进行上下文切换）并非易事。挑战不仅来自自然语言命令的变化（例如， *open sleep sounds*还是*sleep sounds please*），而且观察到一些类似上下文切换的命令可能 对于正在运行技能和VPA而言都是合法的：例如，在与“睡眠声音”互动时，您可能会说“播放雷暴声音”，这可以解释为命令播放所需声音的技能，以及要求VPA启动一个不同的skill *Thunderstorm Sounds*。 在我们的研究中，我们提出了对该问题的初步解决方案，这是一种基于学习的方法，利用上下文信息来识别用户意图。

### 特征选择

从较高的层次上，我们从现实对话中发现，如果用户打算切换上下文，则其话语在语义上与VPA系统的关系（例如“开放的睡眠声音”）比当前的技能更为相关，并且这种关系相反，当她不这样做时。因此，我们基于一组功能设计了UIC，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。其中一些功能是通过用户话语与所有已知系统命令之间的语义比较来识别的。为此，我们从VPA的用户手册，开发人员的文档以及我们的研究中收集的真实对话中构建了系统命令列表（第III-D节）。针对列表中的所有命令，将话语的最大SR和平均SR（第V-A节）用作分类功能。还应考虑到用户的话语是否带有市场上一项技能的调用名称，从而捕获了她转换为该技能的潜在意图。

从用户话语和当前进行中的技能之间的关系中提取另一组特征。我们利用以下观察结果：用户对一项技能的命令通常与该技能先前与该用户的沟通以及其规定的功能有关。 因此，我们使用以下功能可测试一项话语是否适合该技能的上下文：1）话语与技能在发话之前的反应之间的SR，2）话语与技能描述中的句子之间的前k个SR（我们选择k = 5）和3）用户话语和描述语句之间的平均SR。

### 结论

为了评估UIC的有效性，我们重用了收集的数据集（请参阅第V-A节），该数据集包含上下文切换的真实世界用户话语。 对于我们的实验，我们根据两位专家的评论（科恩的kapp = 0.64），首先将550个会话手动标记为是否进行上下文切换。 由于数据集受非上下文切换话语支配，因此我们通过用从技能市场收集的技能调用话语随机替换一些话语，进一步平衡了数据集。 总共，我们收集了1,100个上下文切换实例和1,100个非上下文切换实例作为基本事实。

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。 我们使用5折交叉验证尝试了不同的分类算法。 结果表明，随机森林以96.48％的精度，95.16％的召回率和95.82％的F-1得分达到最佳性能。 接下来，我们描述在未标记的真实数据集上对该分类器的评估。

## *整体检测器评估*

如前所述，SRC和UIC旨在通过运行技能来检测用户对话中的异常，并提醒用户潜在的风险。 在这里，我们描述了我们对这些技术对恶意技能和现实世界交互的评估。

### 对抗原型攻击的有效性

为了为我们的实验构建VMA攻击，我们从技能市场中选择了10种热门技能，并以用户的身份记录了与每种技能的几次对话，并且总共收集了61种语音。 然后，我们通过将记录的对话中随机选择的语音替换为VPA系统专用的调用语音，来手工制作技能切换实例（总共15个）。 我们还通过用空响应或模仿VPA响应替换每个对话的最后一段话，构建了一组伪造的终止攻击（总共10个）。 运行所有包含针对我们的检测器的攻击实例的修订对话，我们发现我们的系统成功检测到所有25个（VPA）实例的上下文切换或模拟。

### 现实对话中的有效性

我们进一步评估了检测器在所有未在训练阶段使用的真实对话（包括我们收集的9,582言，见III-D节）的有效性。尽管这些对话可能不包含现实世界中的VMA实例，如前所述（第III-D节），但它们确实包含许多用于上下文切换的用户话语。其中，有341个被我们的分类器标识，326个被确认为确实是上下文切换尝试，这表明我们的UIC组件达到了95.60％的精度。由于该大型未标记数据集（具有近10K语音）缺乏基本事实，因此我们无法计算召回率。对检测到的实例的进一步分析揭示了有趣的情况。例如，一些用户以为他们在与我们的技能互动时正在与Alexa交谈，并要求我们的技能报告时间，天气，新闻，开始另一项技能，甚至控制其他家庭自动化设备（详细信息请参见附录B）。

### 性能

我们测量了检测器在4核CPU的Macbook Pro上引入的检测延迟，事实证明这可以忽略不计（平均0.003毫秒），这表明我们的方法对VPA的性能影响很小。

# 第六章．讨论

### 我们防御的局限性

为了评估我们的VMA防御（SRC和UIC），我们竭尽全力收集有代表性的数据集进行训练和评估，并且良好的实验结果强烈表明，该防御有望缓解第三部分所述的现实VMA风险。 同时，我们承认，数据集可能仍然不够全面，无法涵盖所有实际攻击案例，一旦我们的方法公开，就可能发生逃避攻击。 请注意，这些是大多数基于机器学习的检测系统所面临的问题，而不仅限于我们的方法。 我们认为，VPA供应商处于更有效地实施此类防御的优势，可以利用他们掌握的大量数据来构建更精确的系统，并继续根据防御者的新技巧调整防御策略可能会进行。

### 未来发展方向

尽管我们对Amazon和Google技能市场的分析显示了一些安全风险（就调用名称而言），但我们不知道VSA和VMA是否确实发生在现实世界中，用于收集敏感的用户数据，更不用说了解它们的普遍性了以及它们可能造成的损害。由于技能生态系统的性质，回答这些问题并非易事。如今，每个技能市场已经拥有大量技能，并且每天都有新技能出现（如第II-B节中所述），这使得人工检查每个技能中的恶意活动几乎变得不可行。最重要的是，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。尽管此服务模型为开发人员提供了更大的灵活性，并帮助他们保护了自己的专有代码，但它阻止了对技能代码的静态分析以检测恶意活动。因此，潜在的未来方向是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。

# 第七章.相关工作

### 语音控制系统的安全性

Diao et al. [17] and Jang et al. [27] 证明恶意应用程序可以注入语音命令来控制智能手机。 Kasmi等。 [30]在耳机电缆上施加电磁干扰，并在智能手机上注入语音命令。 隐藏的语音命令[14]，可卡因面条[46]和海豚攻击[49]使用模糊或听不见的语音命令来攻击语音识别系统。 最近，CommanderSong [48]甚至演示了攻击声音命令可以嵌入到歌曲中，从而实现远程攻击。 如前所述，所有这些攻击的目的都是利用从用户到VPA的身份验证不足，并向系统假冒授权用户。 我们首次发现的新攻击（今年早些时候被Amazon和Google都认可[7]）在另一个方向上起作用：它们旨在运行恶意软件来冒充合法用户。 这开辟了一个前所未有的全新攻击途径。

此外，针对防御用户假冒攻击也有一系列研究[38]，[51]，[21]，[50]，其重点是通过智能手机上的传感器来保护语音可控系统，以验证用户身份。 但是，据我们所知，当我们于2018年2月向亚马逊和谷歌报告这些问题时，没有针对VSA和VMA威胁的防御措施。即使在今天，即我们的报告发布三个月后，亚马逊仍无法击败VSA，只能 根据我们在提交论文之前与他们的交谈，对VMA（检测空录音）的保护有限。

与我们的工作无关，Kumar等人。 还发现了声音蹲起攻击。 他们进行了一项包含11,460个语音样本的测量研究，以了解Alexa的语音识别系统在哪里出现故障，系统误解音频输入的原因以及原因（例如，使用不同的口音）[33]。 通过对蹲风险（特别是其语言联系）的深入研究，该研究仅关注Alexa，而不关注Google Home。 更重要的是，它不包括释义的调用名称劫持（“请一个大写”）和伪装的攻击，也没有涉及人类研究和真实世界的评估，这对于理解此类攻击在日常使用中成功的可能性很重要。 VPA系统。 此外，我们设计并实现了两种技术来减轻语音蹲和伪装攻击，这是以前从未做过的。

### 物联网安全

当前的家庭自动化安全研究集中在IoT设备[26]，[43]，[41]和适用的IoT平台[23]，[24]，[28]，[45]的安全性上。 Ho等。 [26]发现了商业化智能锁中的各种漏洞。 Ronen等。 [41]通过IoT设备之间的ZigBee通道验证了蠕虫感染。 费尔南德斯等。 [23]在适用于多设备的SmartThings平台上发现了一系列缺陷。 FlowFence [24]，ContextIot [28]和SmartAuth [45]通过分析数据流或从第三方应用程序提取上下文来减轻此类物联网平台的威胁。 相反，我们的工作对VPA生态系统进行了第一次安全性分析。

### 域名抢注和移动网络钓鱼

与我们的蹲起攻击类似，Edelman是第一个被调查的域打字错误[19]，启发了一系列研究[44]，[31]，[10]，[36]来测量和缓解这种威胁。 但是，我们的工作利用了嘈杂的语音通道和语音识别技术的局限性。 另一方面，已经对移动网络钓鱼进行了深入研究[15]，[20]，[22]，[40]，[42]，[34]。 特别是，Chen等。 [15]和Fernandes等人[22]研究基于UI攻击机会的边际识别。 任等人。 [40]发现了可以用于实施UI欺骗的任务劫持攻击。 但是，我们发现了对语音用户界面的新攻击，这与图形用户界面在用户感知方面截然不同。

# 第八章．总结

在本文中，我们报告了对流行的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。 发现这些攻击对VPA IoT系统构成了现实威胁，我们进行的一系列用户研究和真实世界攻击证明了这一点。 为了减轻威胁，我们开发了一个技能名称扫描程序，并将其与亚马逊和谷歌技能市场进行比较，从而发现了处于风险中的大量Alexa技能和已经发布的有问题的技能名称，这表明攻击可能已经发生 数以千万计的VPA用户。 此外，我们设计并实现了上下文敏感检测器，以减轻语音伪装威胁，达到95％的精度。

鉴于该研究报告的发现的重要性，我们仅迈出了第一步，即充分了解VPA IoT系统的安全风险并有效地缓解此类风险。 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。

# 附录A

### 抽样调查问题

1）在调用技能名称时，您是否添加了任何围绕技能名称的单词或短语(以便它听起来更自然?)选择所有适用的。

* Yes. Alexa, open Sleep Sounds please.
* Yes. Alexa, open Sleep Sounds for me.
* Yes. Alexa, open Sleep Sounds app.
* Yes. Alexa, open my Sleep Sounds.
* Yes. Alexa, open the Sleep Sounds.
* Yes. Alexa, open some Sleep Sounds.
* Yes. Alexa, tell me a Cat Facts.
* Other (please specify).
* No. I only use simplest forms (e.g. “Alexa, open Sleep Sounds” ).

2）请说出你最常使用的两种技能

3）请给出三个例子，你将用在上面列出的每个技能上

4）你是否曾使用过你并不打算使用的技能？

a)是

b)否

5）您是否曾经尝试在与另一种技能交互时调用该技能?(除了听音乐的时候)

a)是

b)否

6）你有没有试过在与一项技能互动时，把音量调大或调低?(除了听音乐的时候)

a)是

b)否

7）你放弃一项技能最常见的方式是什么?请选择所有适用的。

* Alexa, stop.
* Alexa, cancel.
* Alexa, shut up.
* Alexa, cancel.
* Alexa, never mind.
* Alexa, forget it.
* Alexa, exit.
* Other (please specify).

8) 你是否有过这样的经历:对一项技能说退出的话(就像前一个问题中的那些话)，你本来打算退出，但实际上并没有退出?

a)是

b)否

9) 你最常使用哪个指标来判断与Alexa的对话是否结束?

a) Alexa会说“再见”、“祝你有美好的一天”或者类似的话。

b) Alexa不再说话了。

c) The light on the device is off.

d) Other (please specify).

# 附录B

### 上下文切换示例

在这里，我们有一些有趣的例子，由探测器(第五部分)发现的现实世界的对话收集我们发表的技能(见III-D部分)。这里提供的例子是文本，包括用户的话语和他们之前的技能反应。**技巧**:大家好，欢迎收听《舒缓睡眠声音》。你今天想要哪种睡眠声音

**用于上下文切换的用户话语**

* + - 关掉电视。
    - 什么时间?
    - 本周天气预报怎么说?
    - 告诉我消息。

-----------------------------------------------------------

**Skill**：对不起，我不明白。你今天想要哪种声音

**用于上下文切换的用户话语**

* + - 告诉我一句格言。
    - 几点了?

-----------------------------------------------------------

**Skill：**你好，欢迎来到难以置信的快速睡眠。你今天想要哪种睡眠声音?

**用于上下文切换的用户话语**

* + - 我的简报是什么