(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116783650 A
(43)申请公布日 2023.09.19

(21)申请号 202180089319.4
(22)申请日 2021.12.14
(30)优先权数据
17/321,994 2021.05.17 US
(85)PCT国际申请进入国家阶段日
2023.07.05
(86)PCT国际申请的申请数据
PCT/US2021/063359 2021.12.14
(87)PCT国际申请的公布数据
WO2022/245397 EN 2022.11.24
(71)申请人 谷歌有限责任公司
地址 美国加利福尼亚州
(72)发明人 维克托·克尔布内 马修·谢里菲 翁德热·斯科佩克 贾斯汀·路 丹尼尔·巴尔卡塞 凯文·基尔戈 穆罕默德·哈桑·罗姆 尼科洛·德尔科莱 迈克尔·戈利科夫
(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
专利代理师 邓聪惠 周亚荣
(51)Int.Cl.
*G10L 15/08*(2006.01)

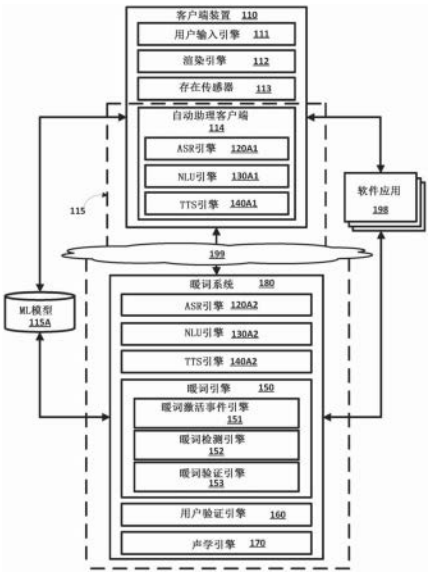
权利要求书5页 说明书29页 附图11页

(54)发明名称

助理命令的场境抑制

(57)摘要

一些实施方式使用暖词模型处理音频数据的流以确定对应于与助理命令相关联的特定单词和/或短语(例如,暖词)的音频数据的部分,使用自动语音辨识(ASR)模型处理音频数据的前序部分(例如,在暖词之前)和/或音频数据的后序部分(例如,在暖词之后)以生成ASR输出,并且基于处理ASR输出来确定用户是否打算执行助理命令。




CN 116783650 A

1.一种由一个或多个处理器实施的方法,所述方法包括:

使用暖词模型处理音频数据的流以监测一个或多个特定单词或短语的出现,所述音频数据的流由用户的客户端装置的一个或多个麦克风生成,并且所述一个或多个特定单词或短语中的每一个与助理命令相关联;

响应于确定所述音频数据的部分对应于所述特定单词或短语中的一个或多个:

使用自动语音辨识ASR模型处理所述音频数据的前序部分和/或所述音频数据的后序部分以生成ASR输出,

其中,所述音频数据的所述前序部分在对应于所述一个或多个特定单词或短语的所述音频数据的所述部分之前,并且

其中,所述音频数据的所述后序部分在对应于所述一个或多个特定单词或短语的所述音频数据的所述部分之后;以及

基于处理所述ASR输出来确定所述用户是否打算所述一个或多个特定单词或短语使所述助理命令执行;

响应于确定所述用户不打算所述一个或多个特定单词或短语使与所述特定单词或短语中的一个或多个相关联的所述助理命令执行:

避免使自动助理执行与所述特定单词或短语中的一个或多个相关联的所述助理命令;以及

响应于确定所述用户打算所述一个或多个特定单词或短语使与所述特定单词或短语中的一个或多个相关联的所述助理命令执行:

使所述自动助理执行与所述特定单词或短语中的一个或多个相关联的所述助理命令。

2.根据权利要求1所述的方法,进一步包括:

检测暖词激活事件的出现;以及

响应于检测到所述暖词激活事件的所述出现,激活利用所述暖词模型的一个或多个当前休眠的自动助理功能,

其中,使用所述暖词模型处理所述音频数据的流以监测所述一个或多个特定单词或短语的所述出现是响应于激活利用所述暖词模型的所述一个或多个当前休眠的自动助理功能。

3.根据权利要求2所述的方法,其中,所述暖词激活事件包括以下中的一个或多个:在所述客户端装置处接收到的电话呼叫、在所述客户端装置处接收到的文本消息、在所述客户端装置处接收到的电子邮件、在所述客户端装置处响起的警报、在所述客户端装置处响起的定时器、在所述客户端装置或所述客户端装置的环境中的附加客户端装置处播放的媒体、在所述客户端装置处接收到的通知、所述客户端装置的位置或在所述客户端装置处可访问的软件应用。

4.根据前述权利要求中的任一项所述的方法,其中,基于处理所述ASR输出来确定所述用户是否打算所述一个或多个特定单词或短语使与所述一个或多个特定单词或短语相关联的所述助理命令执行包括:

使用自然语言理解NLU模型处理所述ASR输出以生成NLU输出,其中,所述ASR输出是基于所述音频数据的所述前序部分而不是所述音频数据的所述后序部分而生成的;并且

基于所述NLU输出来确定所述用户是否打算所述一个或多个特定单词或短语使所述助

理命令执行。

5.根据权利要求4所述的方法,进一步包括:

响应于确定所述NLU输出不足以确定所述用户是否打算所述一个或多个特定单词或短语使与所述特定单词或短语中的一个或多个相关联的所述助理命令执行:

使用所述ASR模型处理所述音频数据的所述后序部分以生成附加ASR输出;并且

基于处理所述附加ASR输出,确定所述用户是否打算所述一个或多个特定单词或短语使与所述特定单词或短语中的一个或多个相关联的所述助理命令执行。

6.根据权利要求1至3中的任一项所述的方法,其中,基于处理所述ASR输出来确定所述用户是否打算所述一个或多个特定单词或短语使与所述特定单词或短语中的一个或多个相关联的所述助理命令执行包括:

使用自然语言理解NLU模型处理所述ASR输出以生成NLU输出,其中,所述ASR输出是基于所述音频数据的所述前序部分和所述音频数据的所述后序部分两者而生成的;并且

基于所述NLU输出来确定所述用户是否打算所述一个或多个特定单词或短语使所述助理命令执行。

7.根据权利要求6所述的方法,进一步包括:

响应于确定所述NLU输出不足以确定所述用户是否打算所述一个或多个特定单词或短语使与所述特定单词或短语中的一个或多个相关联的所述助理命令执行:

使用所述ASR模型处理所述音频数据的附加后序部分以生成附加ASR输出,其中,所述音频数据的所述附加后序部分在所述音频数据的所述后序部分之后;以及

基于处理所述附加ASR输出,确定所述用户是否打算所述一个或多个特定单词或短语使与所述特定单词或短语中的一个或多个相关联的所述助理命令执行。

8.根据前述权利要求中的任一项所述的方法,进一步包括:

使用端点模型处理所述音频数据的流,以生成在所述音频数据的流中捕获的并且包括所述一个或多个特定单词或短语的口头话语的多个时间戳。

9.根据权利要求8所述的方法,其中,所述多个时间戳至少包括:与当所述用户开始提供所述口头话语时的第一时间相关联的第一时间戳、与所述第一时间之后的当所述用户开始提供所述口头话语中包括的所述一个或多个特定单词或短语时的第二时间相关联的第二时间戳、与所述第二时间之后的当所述用户完成提供所述口头话语中包括的所述一个或多个特定单词或短语时的第三时间相关联的第三时间戳,以及与所述第三时间之后的当所述用户完成提供所述口头话语时的第四时间相关联的第四时间戳。

10.根据权利要求9所述的方法,其中,所述音频数据的所述前序部分包括与所述第一时间戳和所述第二时间戳之间的所述口头话语相对应的任何音频数据。

11.根据权利要求9或权利要求10所述的方法,其中,所述音频数据的所述后序部分包括与所述第三时间戳和所述第四时间戳之间的所述口头话语相对应的任何音频数据。

12.根据前述权利要求中的任一项所述的方法,进一步包括:

响应于确定所述口头话语包括所述特定单词或短语中的一个或多个,激活利用所述ASR模型的一个或多个当前休眠的自动助理功能。

13.根据前述权利要求中的任一项所述的方法,进一步包括:

结合所述音频数据的所述前序部分和/或所述音频数据的所述后序部分,使用所述ASR

模型处理与所述一个或多个特定单词或短语相对应的所述音频数据的部分,以生成所述ASR输出。

14.根据前述权利要求中的任一项所述的方法,进一步包括:

使用话音活动检测VAD模型处理所述音频数据的流,以在所述音频数据的所述部分对应于所述特定单词或短语中的一个或多个之前和/或在所述音频数据的所述部分对应于所述特定单词或短语中的一个或多个之后监测附加话音活动的出现;以及

响应于检测到所述附加话音活动的出现:

激活利用所述ASR模型的一个或多个当前休眠的自动助理功能。

15.根据权利要求14所述的方法,进一步包括:

响应于未检测到所述附加话音活动的出现:

使所述自动助理执行与所述特定单词或短语中的一个或多个相关联的所述助理命令。

16.根据前述权利要求中的任一项所述的方法,进一步包括:

使用声学模型处理所述音频数据以确定所述口头话语的一个或多个韵律特性,

其中,确定所述用户是否打算所述一个或多个特定单词或短语使与所述特定单词或短语中的一个或多个相关联的所述助理命令执行是进一步基于所述口头话语的所述一个或多个韵律特性。

17.根据前述权利要求中的任一项所述的方法,进一步包括:

从所述客户端装置的一个或多个传感器接收传感器数据的实例,

其中,确定所述用户是否打算所述一个或多个特定单词或短语使与所述特定单词或短语中的一个或多个相关联的所述助理命令执行是进一步基于所述传感器数据的所述实例。

18.根据权利要求17所述的方法,其中,所述传感器数据的所述实例包括以下中的一个或多个:由所述客户端装置的加速度计生成的加速度计数据、由所述客户端装置的一个或多个相机组件生成的图像数据、由所述客户端装置的一个或多个运动传感器生成的运动数据。

19.根据前述权利要求中的任一项所述的方法,进一步包括:

从所述客户端装置的音频缓冲器获得所述音频数据的所述前序部分;和/或

从所述音频数据的流获得所述音频数据的所述后序部分。

20.一种由一个或多个处理器实施的方法,所述方法包括:

使用暖词模型处理音频数据的流以监测一个或多个特定单词或短语的出现,所述音频数据的流由用户的客户端装置的一个或多个麦克风生成,并且所述一个或多个特定单词或短语中的每一个与助理命令相关联;

响应于确定所述音频数据的部分对应于所述特定单词或短语中的一个或多个,并且响应于确定验证提供对应于所述音频数据的口头话语的所述用户的身份:

使用说话者识别SID模型处理所述音频数据,以确定所述音频数据是否足以识别提供所述口头话语的所述用户;

响应于确定所述音频数据足以识别提供所述口头话语的所述用户,并且响应于确定提供所述口头话语的所识别的用户被授权使与所述特定单词或短语中的一个或多个相关联的所述助理命令执行:

使自动助理执行与所述特定单词或短语中的一个或多个相关联的所述助理命令;以及

响应于确定所述音频数据不足以识别提供所述口头话语的所述用户：

生成请求对提供所述口头话语的所述用户的身份进行附加验证的提示;以及

使所述提示被提供用于在所述客户端装置处呈现。

21.根据权利要求20所述的方法,进一步包括：

检测暖词激活事件的出现;以及

响应于检测到所述暖词激活事件的所述出现,激活利用所述暖词模型的一个或多个当前休眠的自动助理功能,

其中,使用所述暖词模型处理所述音频数据的流以监测所述一个或多个特定单词或短语的所述出现是响应于激活利用所述暖词模型的所述一个或多个当前休眠的自动助理功能。

22.根据权利要求21所述的方法,其中,所述暖词激活事件包括以下中的一个或多个：在所述客户端装置处接收到的电话呼叫、在所述客户端装置处接收到的文本消息、在所述客户端装置处接收到的电子邮件、在所述客户端装置处响起的警报、在所述客户端装置处响起的定时器、在所述客户端装置或所述客户端装置的环境中的附加客户端装置处播放的媒体、在所述客户端装置处接收到的通知、所述客户端装置的位置或在所述客户端装置处可访问的软件应用。

23.根据权利要求20至22中的任一项所述的方法,进一步包括：

响应于确定所述音频数据的所述部分对应于所述特定单词或短语中的一个或多个,并且响应于确定未验证提供对应于所述音频数据的所述口头话语的所述用户的身份：

使自动助理执行与所述特定单词或短语中的一个或多个相关联的所述助理命令。

24.根据权利要求20至23中的任一项所述的方法,其中,处理所述音频数据以确定所述音频数据是否足以验证提供所述口头话语的所述用户的所述身份包括：

使用文本相关TD SID模型作为所述SID模型,处理与所述特定单词或短语中的一个或多个相对应的所述音频数据的所述部分,以生成TD说话者嵌入,并且

其中,确定所述音频数据是否足以识别提供所述口头话语的所述用户包括：

在嵌入空间中,将所述TD说话者嵌入与本地存储在所述客户端装置处的一个或多个所存储的TD说话者嵌入进行比较;以及

基于将所述TD说话者嵌入与所述一个或多个所存储的TD说话者嵌入进行比较,确定所述音频数据的所述部分是否足以验证提供所述口头话语的所述用户的所述身份。

25.根据权利要求24所述的方法,其中,使用所述TD SID模型处理与所述特定单词或短语中的一个或多个相对应的所述音频数据的所述部分以生成所述TD说话者嵌入包括：

确定所存储的TD说话者嵌入中的一个或多个是否对应于所述特定单词或短语;并且

其中,使用所述TD SID模型处理与所述特定单词或短语中的一个或多个相对应的所述音频数据的所述部分以生成所述TD说话者嵌入是响应于确定所存储的TD说话者嵌入中的一个或多个与所述特定单词或短语相对应。

26.根据权利要求24所述的方法,其中,生成请求对提供所述口头话语的所述用户的所述身份进行所述附加验证的所述提示包括：

生成请求所述用户提供附加口头话语的所述提示,所述附加口头话语具有足够的长度以执行TI SID以用于验证提供所述口头话语的所述用户的所述身份。

27.根据权利要求24所述的方法,进一步包括:

响应于确定所述口头话语具有足够的长度以执行TISID:

使用TISID模型作为所述SID模型来处理所述音频数据的流以生成TI说话者嵌入,并且

其中,确定所述音频数据是否足以识别提供所述口头话语的所述用户包括:

在嵌入空间中,将所述TI说话者嵌入与本地存储在所述客户端装置处的一个或多个所存储的TI说话者嵌入进行比较;以及

基于将所述TI说话者嵌入与所述一个或多个所存储的TI说话者嵌入进行比较,确定所述音频数据的所述流是否足以验证提供所述口头话语的所述用户的所述身份。

28.根据权利要求27所述的方法,其中,生成请求对提供所述口头话语的所述用户的所述身份进行所述附加验证的所述提示包括:

生成请求附加生物特征信息以验证所述用户的所述身份的所述提示,其中,请求所述附加生物特征信息以验证所述用户的所述身份包括请求以下中的一个或多个:指纹验证或面纹验证。

29.根据权利要求20至28中的任一项所述的方法,进一步包括:

响应于所述提示被提供用于在所述客户端装置处呈现而接收附加用户输入;以及

响应于接收到所述附加用户输入:

确定所述附加用户输入是否足以识别提供所述口头话语的所述用户。

30.根据权利要求29所述的方法,进一步包括:

响应于确定所述附加用户输入足以识别提供所述口头话语的所述用户:

使所述自动助理执行对应的助理命令。

31.一种系统,包括:

至少一个处理器;以及

存储指令的存储器,所述指令在被实行时使所述至少一个处理器执行对应于根据权利要求1至30中的任一项所述的操作。

32.一种存储指令的非暂时性计算机可读存储介质,所述指令在被实行时使至少一个处理器执行对应于根据权利要求1至30中的任一项所述的操作。

# 助理命令的场境抑制

## 背景技术

[0001] 人类可以利用交互式软件应用参与人与计算机的对话，交互式软件应用在本文中称为“自动助理”(也称为“聊天机器人”、“交互式个人助理”、“智能个人助理”、“个人话音助理”、“会话代理”等)。例如，人类(当他们与自动助理交互时，可以称为“用户”)可以向自动助理提供口头自然语言输入(即，口头话语)——在一些情况下，说出的自然语言输入可以转换成文本并且然后被处理，和/或通过提供文本(例如，键入的)自然语言输入。自动助理通常通过提供响应用户界面输出(例如，听觉和/或视觉用户界面输出)、控制智能装置和/或执行其他动作来对口头话语作出响应。

[0002] 自动助理通常依赖于一系列组件来解释和响应口头话语。例如，自动语音辨识(ASR)引擎可以处理对应于用户的口头话语的音频数据以生成ASR输出，诸如口头话语的转录(即，术语序列和/或其他口令)。此外，自然语言理解(NLU)引擎可以处理ASR输出以生成NLU输出，诸如用户在提供口头话语时的打算以及可选地与打算相关联的参数的时隙值。此外，履行引擎可以用于处理NLU输出，并且生成履行输出，诸如获得对口头话语的响应内容的结构化请求。

[0003] 在一些情况下，可以绕过这一系列组件。例如，一些机器学习(ML)模型(也称为“暖词模型”)可以被训练成检测直接映射到履行输出的特定单词和/或短语(也称为“暖词”)。例如，如果用户正在听音乐并且提供“调大音量”的口头话语，这些ML模型中的一个或多个可以处理音频数据，并且生成使播放音乐的装置调大音量的结构化请求，而不生成任何ASR输出和/或NLU输出。然而，这些暖词通常是日常语音中可能出现的常见单词和/或短语。因此，与这些暖词相关联的助理命令可能被用户无意地触发，从而浪费计算资源。此外，在这些情况中的一些情况下，用户可能不得不撤消这些助理命令中的一些，从而浪费更多的计算资源。

## 发明内容

[0004] 本文公开的实施方式旨在基于场境音频数据和/或提供包括一个或多个特定单词和/或短语(例如，暖词)的口头话语的用户的身份，在场境上抑制与特定单词和/或短语中的一个或多个相关联的助理命令的执行(或实行)。一些实施方式使用一个或多个暖词模型处理音频数据的流，以确定对应于与助理命令相关联的特定单词和/或短语的音频数据的部分。这些实施方式中的一些使用自动语音辨识(ASR)模型处理音频数据的前序部分(例如，在与特定单词和/或短语中的一个或多个相对应的音频数据的部分之前的音频数据的部分)和/或音频数据的后序部分(例如，在与特定单词和/或短语中的一个或多个相对应的音频数据的部分之后的音频数据的部分)，以生成ASR输出。此外，这些实施方式中的一些基于处理ASR输出来确定提供音频数据中捕获的口头话语的用户是否打算执行与特定单词或短语中的一个或多个相关联的助理命令。附加的或可替代的实施方式可以使用说话者识别(SID)模型处理音频数据的流，以确定音频数据是否足以识别提供在音频数据的流中捕获的口头话语的用户，并且确定提供口头话语的用户是否被授权使得执行助理命令。

[0005] 例如，假设在用户的客户端装置处接收到电话呼叫。进一步假设客户端装置的用户响应于在客户端装置处接收到的电话呼叫而提供“接听”的口头话语。在此示例中，并且假设“接听”是与使至少部分在客户端装置处实施的自动助理接听电话呼叫的助理命令相关联的暖词，则自动助理可以基于使用一个或多个暖词模型检测到暖词“接听”的出现来代表用户接听电话呼叫。相反，假设客户端装置的用户响应于在客户端装置处接收到的电话呼叫而替代地提供“我不想接听它”的口头话语。在此示例中，尽管使用暖词模型中的一个或多个检测到暖词“接听”的出现，但自动助理可以处理场境音频数据以确定自动助理不应代表用户接听电话。

[0006] 在一些实施方式中，可以响应于检测到暖词激活事件的出现而激活利用暖词模型中的一个或多个的一个或多个自动助理组件。暖词激活事件可以包括例如在客户端装置处接收到的电话呼叫、在客户端装置处接收到的文本消息、在客户端装置处接收到的电子邮件、在客户端装置处响起的警报或定时器、在客户端装置或客户端装置的环境中的附加客户端装置处播放的媒体、在客户端装置处接收到的通知、客户端装置的位置、在客户端装置处可访问的软件应用，和/或与客户端装置相关联的其他事件，在这些事件中，用户可以提供口头话语以使客户端装置或与客户端装置通信的附加客户端装置受到控制。值得注意的是，这些暖词激活事件中的一些是离散事件(例如，在客户端装置处接收到的电话呼叫、在客户端装置处接收到的文本消息、在客户端装置处接收到的电子邮件、在客户端装置处响起的警报或定时器等)，而这些暖词激活事件中的一些其他事件是连续事件(例如，在客户端装置处可访问的软件应用)。此外，应注意，这些暖词激活事件并不是相互排斥的。换言之，可以在给定的时刻检测多个暖词激活事件，使得一个或多个自动助理组件基于多个不同的暖词激活事件来主动监测暖词。

[0007] 本文描述的暖词模型可以包括多个不同的暖词模型(例如，音频关键词分类模型)，其被训练成检测一个或多个特定单词和/或短语(例如，暖词)，一个或多个特定单词或短语当被检测到时使自动助理执行与特定单词和/或短语中的一个或多个相关联的助理命令。在一些实施方式中，给定暖词模型可以被训练成检测与给定暖词激活事件相关联的单词和/或短语的特定子集。例如，假设正在客户端装置或与客户端装置通信的另一客户端装置(例如，智能扬声器)处播放音乐。在此示例中，给定暖词模型可以被训练成检测使得执行与控制音乐相关联的助理命令的特定单词和/或短语的子集，诸如可能使得音乐暂停的“暂停”暖词、可能使得音乐在暂停后继续的“继续”暖词、可能使得音乐调大音量的“调大音量”暖词、可能使得音乐调小音量的“调小音量”暖词、可能使得音乐跳到下一个歌曲的“下一个”暖词，以及针对与音乐相关联的其他特定单词和/或短语。因此，在此示例中，可以响应于确定正在客户端装置或与客户端装置通信的另一客户端装置处播放音乐，激活利用给定暖词模型的当前休眠的助理功能中的一个或多个。

[0008] 在附加的或可替代的实施方式中，多个暖词模型可以被训练成检测与给定暖词激活事件相关联的单词和/或短语的子集。继续以上示例，第一暖词模型可以被训练成检测一个或多个第一特定单词和/或短语，诸如“暂停”暖词和“继续”暖词，第二暖词模型可以被训练成检测一个或多个第二特定单词和/或短语，诸如“调大音量”暖词或“调小音量”暖词，第三暖词模型可以被训练成检测一个或多个第三特定单词和/或短语，诸如“下一个”暖词，以及针对与音乐暖词激活事件相关联的其他特定单词和/或短语。因此，在此示例中，可以响

应于确定正在客户端装置或与客户端装置通信的另一客户端装置处播放音乐，激活至少利用第一暖词模型、第二暖词模型和第三暖词模型的当前休眠的助理功能中的一个或多个。

[0009] 在一些实施方式中，并且响应于在音频数据的流中检测到特定单词和/或短语中的一个或多个，可以处理音频数据的前序部分和/或音频数据的后序部分，以确定提供口头话语的用户事实上是否打算执行与检测到的特定单词和/或短语中的一个或多个相关联的助理命令。例如，再次假设在用户的客户端装置处接收到电话呼叫，并且假设客户端装置的用户响应于在客户端装置处接收到的电话呼叫而提供“我不想接听它”的口头话语。在此示例中，音频数据的前序部分(例如，对应于“我不想”)可以从客户端装置的音频缓冲器获得，并且可以使用ASR模型进行处理以生成ASR输出。此外，可以使用NLU模型处理ASR输出以生成NLU输出。在此示例中，ASR输出和/或NLU输出指示用户不打算“接听”以使自动助理代表用户接听电话。在这些实施方式的一些版本中，可以结合音频数据的前序部分，附加地或可替代地处理与一个或多个特定单词和/或短语相对应的音频数据的部分(例如，与“接听”相对应)，以生成ASR输出和NLU输出。在这些实施方式的一些版本中，可以结合音频数据的前序部分和/或与特定单词和/或短语中的一个或多个相对应的音频数据的部分，附加地或可替代地处理音频数据的后序部分(例如，对应于“它”)，以生成ASR输出和NLU输出。

[0010] 在这些实施方式的一些版本中，可以响应于检测到除了特定单词和/或短语中的一个或多个之外的附加话音活动而激活利用ASR模型和/或NLU模型的一个或多个当前休眠的自动助理组件。例如，可以使用话音活动检测(VAD)模型来处理音频数据的流，以确定是否存在任何附加话音活动。在检测到附加话音活动的实施方式中，可以激活利用ASR模型和/或NLU模型的当前休眠的自动助理组件中的一个或多个以处理音频数据的前序部分、与特定单词和/或短语中的一个或多个相对应的音频数据的部分、和/或音频数据的后序部分。在没有检测到附加话音活动的实施方式中，可以使利用ASR模型和/或NLU模型的当前休眠的自动助理组件中的一个或多个保持休眠，并且可以执行与特定单词和/或短语中的一个或多个相关联的助理命令。

[0011] 在各种实施方式中，必须在执行助理命令之前验证提供口头话语的用户的身份。例如，与特定暖词激活事件、特定装置和/或特定软件应用相关联的助理命令可能要求验证用户的身份。例如，在助理命令使自动助理响应电子通信(例如，电话呼叫、文本消息、电子邮件、社交媒体消息和/或其他电子通信)的实施方式中，可能需要在自动助理响应电子通信之前验证提供口头话语的用户。此外，例如，在助理命令使自动助理控制特定智能装置(例如，智能锁、智能烤箱、智能车库门等)的实施方式中，可能需要在自动助理控制特定智能装置之前验证提供口头话语的用户。然而，一些助理命令可能不要求在执行助理命令之前验证用户的身份。例如，在助理命令使自动助理控制媒体(例如，音乐、视频、电视节目和/或其他媒体)的实施方式中，可能不需要在自动助理响应电子通信之前验证提供口头话语的用户。此外，例如，在助理命令使自动助理控制其他特定智能装置(例如，智能锁、智能烤箱、智能车库门等)的实施方式中，可能不需要在自动助理控制特定智能装置之前验证提供口头话语的用户。

[0012] 如上所述，可以使用SID模型处理音频数据的流，以验证提供口头话语的用户的身份。在这些实施方式的一些版本中，并且假设一个或多个先前生成的文本相关(TD)说话者嵌入是用户针对特定单词和/或短语中的一个或多个在客户端装置处可访问的，可以使用

TD SID模型处理与特定单词和/或短语中的一个或多个相对应的音频数据的部分，以生成TD说话者嵌入。所生成的TD说话者嵌入可以有效地与用户在客户端装置处可访问的先前生成的一个或多个TD说话者嵌入进行比较，以确定口头话语是否来自授权用户。值得注意的是，TD SID中的一个或多个特定单词和/或短语可以被限制为是被配置成调用自动助理的一个或多个调用短语(例如，热词和/或触发词，诸如例如“嘿助理”、“OK助理”和/或“助理”)或本文描述的一个或多个暖词。

[0013] 在这些实施方式的一些附加的或可替代的版本中，并且假设一个或多个先前生成的与文本无关(TI)说话者嵌入可供用户在客户端装置处访问，则可以使用TI SID模型处理与音频数据的流中的口头话语相对应的音频数据，以生成TI说话者嵌入。所生成的TI说话者嵌入可以有效地与用户在客户端装置处可访问的先前生成的一个或多个TI说话者嵌入进行比较，以确定口头话语是否来自授权用户。值得注意的是，在TD SID中，TI一个或多个先前生成的TI说话者嵌入可以不被限制为是被配置成调用自动助理的一个或多个调用短语(例如，热词和/或触发词，诸如例如“嘿助理”、“OK助理”和/或“助理”)或本文描述的一个或多个暖词。

[0014] 因此，在这些实施方式的一些版本中，可以将音频数据的流中捕获的口头话语的长度与长度阈值进行比较，以确定口头话语的长度是否足以执行TI SID。如果口头话语的长度足以执行TI SID，则可以使用TI SID模型和/或TD SID模型处理与口头话语相对应的音频数据，以尝试验证用户的身份。然而，即使口头话语的长度不足以执行TI SID，仍然可以使用TD SID模型处理与口头话语相对应的音频数据，以尝试验证用户的身份。假设口头话语的长度不足以执行TI SID(和/或TI SID不足以验证用户的身份)，并且假设TD SID不足以验证用户的身份，则可以提供包括示例话语的提示以呈现给用户，其中示例话语是执行TI SID的足够长度(或足够长)，或对应于一个或多个先前生成的TD说话者嵌入可用于的特定术语和/或短语。附加地或可替代地，提示可以请求用户提供其他形式的生物特征识别(例如，面纹验证和/或指纹识别)，以使得执行助理命令。

[0015] 通过使用本文描述的技术，可以实现一个或多个技术优点。作为一个非限制性示例，本文描述的技术使得自动助理能够基于处理暖词的场境音频数据来确定用户实际上是否打算暖词出现以使助理命令执行。因此，可以减少暖词的假阳性数量，从而在自动助理利用一个或多个网络来使得执行助理命令的情况下节省客户端装置处的计算资源和/或网络资源。此外，由于客户端装置的用户不需要采取行动来纠正或撤消任何助理命令，因此可以减少用户输入的数量，从而节省客户端装置处的计算资源。此外，本文描述的各种自动助理功能可以保持休眠直到如本文描述需要为止，从而节省客户端装置处的计算资源。作为另一非限制性示例，本文描述的技术使得自动助理能够使用SID和/或其他技术来验证提供包括暖词的口头话语的用户的身份。因此，只有授权用户才能使某些助理命令执行，从而提高与客户端装置的用户相关联的数据的数据安全性。

[0016] 以上描述仅作为本文公开的一些实施方式的概述而提供。本文另外更详细地描述这些实施方式和其他实施方式。

[0017] 应理解，本文更详细地描述的上述概念和附加概念的所有组合视为本文所公开的主题的一部分。例如，在本公开结尾处出现的所要求保护的主题的所有组合视为本文公开的主题的一部分。

## 附图说明

[0018] 图1描绘展示本公开的各个方面并且可以在其中实施本文公开的实施方式的示例环境的框图。

[0019] 图2描绘根据各种实施方式的使用图1的各种组件展示本公开的各个方面的示例过程流程。

[0020] 图3描绘示出根据各种实施方式的基于场境音频数据来确定是否使助理命令执行的示例方法的流程图。

[0021] 图4A、图4B和图4C描绘根据各种实施方式的基于场境音频数据来确定是否使助理命令执行的各种非限制性示例。

[0022] 图5描绘示出根据各种实施方式的基于说话者识别来确定是否使助理命令执行的示例方法的流程图。

[0023] 图6A、图6B和图6C描绘根据各种实施方式的基于说话者识别来确定是否使助理命令执行的各种非限制性示例。

[0024] 图7描绘根据各种实施方式的计算装置的示例架构。

## 具体实施方式

[0025] 现在转向图1，描绘了展示本公开的各个方面并且可以在其中实施本文公开的实施方式的示例环境的框图。示例环境包括客户端装置110和暖词系统180。在一些实施方式中，暖词系统180可以在客户端装置110本地实施。在附加的或可替代的实施方式中，暖词系统180可以从客户端装置110远程实施，如图1所示。在这些实施方式中，客户端装置110和暖词系统180可以经由诸如一个或多个有线或无线局域网(“LAN”，包括Wi-Fi LAN、网状网络、蓝牙、近场通信等)或广域网(“WAN”，包括互联网)的一个或多个网络199彼此通信耦合。

[0026] 客户端装置110可以是例如以下中的一个或多个：台式计算机、膝上型计算机、平板电脑、手机、车辆的计算装置(例如，车载通信系统、车载娱乐系统、车载导航系统)、独立交互式扬声器(可选地具有显示器)、诸如智能电视的智能设备，和/或包括计算装置的用户可穿戴设备(例如，具有计算装置的用户手表，具有计算装置、虚拟或增强现实计算装置的用户眼镜)。可以提供附加的和/或可替代的客户端装置。

[0027] 客户端装置110可以实行自动助理客户端114。自动助理客户端114的实例可以是与客户端装置110的操作系统分离的应用(例如，安装在操作系统的“顶部”)，或可以可替代地由客户端装置110中的操作系统直接实施。自动助理客户端114可以与在客户端装置110本地实施的暖词系统180交互，或经由如图1所示的一个或多个网络199与暖词系统180交互。自动助理客户端114(以及可选地通过其与其他远程系统(例如，服务器)的交互)可以形成从用户的角度看来似乎是自动助理115的逻辑实例的东西，用户可以与该自动助理115参与人机对话。自动助理115的实例在图1中描绘，并且由包括客户端装置110的自动助理客户端114和暖词系统180的虚线包围。因此，应理解，与在客户端装置110上实行的自动助理客户端114参与的用户实际上可以与他或她自己的自动助理115的逻辑实例(或在家庭或其他用户组当中共享的自动助理115的逻辑实例)参与。为简洁起见，本文使用的自动助理115将指在客户端装置110和/或可以实施暖词系统180的一个或多个服务器上实行的自动助理客户端114。

[0028] 在各种实施方式中，客户端装置110可以包括用户输入引擎111，用户输入引擎111被配置成检测由客户端装置110的用户使用一个或多个用户界面输入装置提供的用户输入。例如，客户端装置110可以配备有一个或多个麦克风，该一个或多个麦克风捕获音频数据，诸如与用户的口头话语或客户端装置110的环境中的其他声音相对应的音频数据。附加地或可替代地，客户端装置110可以配备有一个或多个视觉组件，一个或多个视觉组件被配置成捕获与在一个或多个视觉组件的视场中检测到的图像和/或运动(例如，姿势)相对应的视觉数据。附加地或可替代地，客户端装置110可以配备有一个或多个触敏组件(例如，键盘和鼠标、手写笔、触摸屏、触摸面板、一个或多个硬件按钮等)，一个或多个触敏组件被配置成捕获与指向客户端装置110的触摸输入相对应的信号。

[0029] 在各种实施方式中，客户端装置110可以包括渲染引擎112，渲染引擎112被配置成使用一个或多个用户接口输出装置向客户端装置110的用户提供用于听觉和/或视觉呈现的内容。例如，客户端装置110可以配备有一个或多个扬声器，一个或多个扬声器使得能够经由客户端装置110向用户提供用于听觉呈现的内容。附加地或可替代地，客户端装置110可以配备有显示器或投影仪，显示器或投影仪使得能够经由客户端装置110向用户提供用于视觉呈现的内容。

[0030] 在各种实施方式中，客户端装置110可以包括一个或多个存在传感器113，一个或多个存在传感器被配置成在得到对应用户的批准的情况下提供指示检测到的存在(特别是人类存在)的信号。在这些实施方式中的一些实施方式中，自动助理115可以至少部分地基于用户在客户端装置110处(或在与客户端装置110的用户相关联的另一计算装置处)的存在来识别客户端装置110(或与客户端装置110的用户相关联的另一计算装置)以满足口头话语。通过在客户端装置110和/或与客户端装置110的用户相关联的其他计算装置处渲染响应内容(例如，经由渲染引擎112)、通过使得客户端装置110或与客户端装置110的用户相关联的其他计算装置受到控制、和/或通过使客户端装置110和/或与客户端装置110的用户相关联的其他计算装置执行任何其他动作来满足口头话语，可以满足口头话语。如本文所描述，自动助理115可以利用基于存在传感器113确定的数据来基于用户在哪里附近或最近在哪里附近来确定客户端装置110(或其他计算装置)，并且仅向客户端装置110(或那些其他计算装置)提供对应命令。在一些附加或替代实施方式中，自动助理115可以利用基于存在传感器113确定的数据来确定是否有任何用户(任何用户或特定用户)当前接近客户端装置110(或其他计算装置)，并且可以可选地基于接近客户端装置110(或其他计算装置)的用户来抑制向客户端装置110(或其他计算装置)提供数据和/或从客户端装置110(或其他计算装置)提供数据。

[0031] 存在传感器113可以具有各种形式。例如，客户端装置110可以利用上文关于用户输入引擎111描述的用户界面输入组件中的一个或多个来检测用户的存在。附加地或可替代地，客户端装置110可以配备有其他类型的基于光的存在传感器113，诸如测量从其视场内的物体辐射的红外(“IR”)光的无源红外(“PIR”)传感器。

[0032] 附加地或可替代地，在一些实施方式中，存在传感器113可以被配置成检测与人类存在或装置存在相关联的其他现象。例如，在一些实施例中，客户端装置110可以配备有存在传感器113，存在传感器113检测由例如用户携带/操作的其他计算装置(例如，移动装置、可穿戴计算装置等)和/或其他计算装置发射的各种类型的无线信号(例如，诸如无线电、超

声波、电磁波等的波)。例如,客户端装置110可以被配置成发射人类无法察觉的波,诸如超声波或红外波,这些波可以被其他计算装置(例如,经由诸如具有超声波功能的麦克风之类的超声波/红外接收器)检测到。

[0033] 附加地或可替代地,客户端装置110可以发射其他类型的人类无法感知的波,诸如无线电波(例如,Wi-Fi、蓝牙、蜂窝等),这些波可以被用户携带/操作的其他计算装置(例如,移动装置、可穿戴计算装置等)检测到并用于确定用户的特定位置。在一些实施方式中,可以使用GPS和/或Wi-Fi三角测量,例如基于GPS和/或去往/来自客户端装置110的Wi-Fi信号来检测人的位置。在其他实施方式中,客户端装置110可以单独或共同使用诸如飞行时间、信号强度等其他无线信号特性,以基于由用户携带/操作的其他计算装置发射的信号来确定特定人的位置。

[0034] 附加地或可替代地,在一些实施方式中,客户端装置110可以执行说话者识别(SID),以根据用户的话音来辨识用户(例如,关于用户验证引擎160所描述的)。在一些实施方式中,然后可以例如由客户端装置110的存在传感器113(以及可选地客户端装置110的GPS传感器、Soli芯片和/或加速度计)来确定说话者的移动。在一些实施方式中,基于这种检测到的移动,可以预测用户的位置,并且当至少部分地基于客户端装置110和/或其他计算装置与用户位置的接近度使得在客户端装置110和/或其他计算装置处渲染任何内容时,可以假设此位置是用户的位置。在一些实施方式中,可以简单地假设用户处于他或她与自动助理115参与的最后位置,特别是在自上次参与以来没有经过太多时间的情况下。

[0035] 此外,客户端装置110和/或暖词系统180可以包括用于存储数据和/或软件应用198的一个或多个存储器、用于访问数据和实行软件应用198的一个或多个处理器、和/或便于通过网络199中的一个或多个进行通信的其他组件。在一些实施方式中,软件应用198中的一个或多个可以被本地安装在客户端装置110处,而在其他实施方式中,软件应用198中的一个或多个可以被远程托管(例如,由一个或多个服务器)并且可以由客户端装置110通过网络199中的一个或多个访问。客户端装置110、其他计算装置和/或自动助理115执行的操作可以分布在多个计算机系统上。自动助理115可以被实施为例如在客户端装置110和/或在通过网络(例如,图1的网络199)彼此耦合的一个或多个位置中的一个或多个计算机上运行的计算机程序。

[0036] 在一些实施方式中,由自动助理115执行的操作可以经由自动助理客户端114在客户端装置110本地实施。如图1所示,自动助理客户端114可以包括自动语音辨识(ASR)引擎120A1、自然语言理解(NLU)引擎130A1和文本转语音(TTS)引擎140A1。在一些实施方式中,由自动助理115执行的操作可以分布在多个计算机系统上,诸如当如图1所描绘从客户端装置110远程实施暖词系统180时。在这些实施方式中,自动助理115可以附加地或可替代地利用暖词系统180的ASR引擎120A2、NLU引擎130A2和TTS引擎140A2。

[0037] 这些引擎中的每一个可以被配置成执行一个或多个功能。例如,ASR引擎120A1和/或120A2可以使用存储在机器学习(ML)模型数据库115A中的ASR模型(例如,递归神经网络(RNN)模型、变换器模型和/或能够执行ASR的任何其他类型的ML模型)来处理捕获口头话语并且由客户端装置110的麦克风生成的音频数据,以生成ASR输出。此外,NLU引擎130A1和/或130A2可以使用存储在ML模型数据库115A中的NLU模型(例如,长短期存储器(LSTM)、门控递归单元(GRU)和/或能够执行NLU的任何其他类型的RNN或其他ML模型)和/或基于语法的

规则来处理ASR输出，以生成NLU输出。另外，TTS引擎140A1和/或140A2可以使用存储在ML模型数据库115A中的TTS模型来处理文本数据(例如，由自动助理115制定的文本)，以生成包括计算机生成的合成语音的合成语音音频数据。

[0038] 在各种实施方式中，ASR输出可以包括例如被预测为与在音频数据中捕获的用户的口头话语相对应的一个或多个语音假设(例如，术语假设和/或转录假设)、针对一个或多个语音假设中的每一个的一个或多个对应的所预测值(例如，概率、对数似然性和/或其他值)、被预测为与在音频数据中捕获的用户的口头话语相对应的多个音素，和/或其他ASR输出。在这些实施方式的一些版本中，ASR引擎120A1和/或120A2可以选择语音假设中的一个或多个，作为对应于口头话语的所辨识文本(例如，基于对应的所预测值)。

[0039] 在各种实施方式中，NLU输出可以包括例如注释的所辨识文本，注释的所辨识文本包括所辨识文本的术语中的一个或多个(例如，所有)的所辨识文本的一个或多个注释。例如，NLU引擎130A1和/或130A2可以包括词性标记器(part of speech tagger)(未示出)，词性标记器被配置成用术语的语法角色来注释术语。附加地或可替代地，NLU引擎130A1和/或130A2可以包括实体标记器(未示出)，实体标记器被配置成注释所辨识文本的一个或多个片段中的实体引用，诸如对人(例如，包括文学人物、名人、公众人物等)、组织、地点(真实和想象的)等的引用。在一些实施方式中，关于实体的数据可以存储在一个或多个数据库中，诸如存储在知识图(未示出)中。在一些实施方式中，知识图可以包括表示已知实体(并且在一些情况下，实体属性)的节点，以及连接节点并表示实体之间关系的边。实体标记器可以高级别的粒度(例如，以使得能够识别对诸如人之类的实体类的所有引用)和/或较低级别的粒度(例如，以使得能够识别对诸如特定人之类的特定实体的所有引用)来注释对实体的引用。实体标记器可以依赖于自然语言输入的内容来解析特定实体，和/或可以可选地与知识图或其他实体数据库通信以解析特定实体。附加地或可替代地，NLU引擎130A1和/或130A2可以包括共指关系解析器(未示出)，共指关系解析器被配置成基于一个或多个场境线索来对针对同一实体的引用进行分组或“聚类”。例如，基于在接收到输入“购买它们”之前立即渲染的客户端装置通知中提到的“剧院门票”，可以使用共指关系解析器在自然语言输入“购买它们”中将术语“它们”解析为“购买剧院门票”。在一些实施方式中，NLU引擎130A1和/或130A2的一个或多个组件可以依赖于来自NLU引擎130A1和/或130A2的一个或多个其他组件的注释。例如，在一些实施方式中，实体标记器可以依赖于来自共指关系解析器的注释来注释对特定实体的所有提及。此外，例如，在一些实施方式中，共指关系解析器可以依赖于来自实体标记器的注释来对同一实体的引用进行聚类。

[0040] 尽管关于具有单个用户的单个客户端装置描述了图1，但是应理解，这是为了示例的目的而并不意味着限制。例如，用户的一个或多个附加客户端装置也可以实施本文描述的技术。例如，用户的客户端装置110、一个或多个附加客户端装置和/或任何其他计算装置可以形成可以采用本文描述的技术的装置生态系统。这些附加客户端装置和/或计算装置可以与客户端装置110通信(例如，通过网络199)。作为另一示例，给定客户端装置可以由共享设置中的多个用户(例如，一组用户、一个家庭)使用。

[0041] 如本文描述，自动助理115可以基于场境音频数据和/或基于提供在音频数据的流中捕获的口头话语的说话者的身份来确定是否响应于在音频数据的流中检测到暖词而使得执行助理命令。在进行此确定时，自动助理可以利用暖词引擎150、用户验证引擎160和/

或声学引擎170。在各种实施方式中并且如图1所描绘，暖词引擎150可以包括暖词激活事件引擎151、暖词检测引擎152和暖词验证引擎153。

[0042] 在一些实施方式中，暖词激活事件引擎151可以监测暖词激活事件的出现。暖词激活事件可以包括例如在客户端装置110处接收到的电话呼叫、在客户端装置110处接收到的文本消息、在客户端装置110处接收到的电子邮件、在客户端装置处110响起的警报、在客户端装置处110响起的定时器、在客户端装置110或客户端装置110的环境中的附加客户端装置处播放的媒体、在客户端装置110处接收到的通知、客户端装置110的位置(例如，用户在家、工作等)、一个或多个软件应用198中在客户端装置110处可访问的软件应用，和/或与客户端装置110相关联的其他事件，在这些事件中，用户可以提供口头话语以使客户端装置110或与客户端装置110通信的附加客户端装置基于口头话语而受到控制。在这些实施方式的一些版本中，并且响应于检测到暖词激活事件的出现，暖词激活事件引擎151可以使得利用存储在ML模型数据库115A中的暖词模型的一个或多个当前休眠的自动助理功能被激活(例如，低功率数字信号处理器(DSP)、中央处理单元(CPU)等)。暖词模型可以被训练成检测一个或多个特定单词或短语(本文中也称为“暖词”)，其每个与助理命令相关联。一个或多个特定单词或短语可以与助理命令相关联，使得当被检测到时，自动助理115可以使得执行特定动作，而无需经由调用单词或短语(例如“助理”、“嘿助理”等)、调用姿势、或用于调用自动助理115的其他手段明确地调用自动助理115。换言之，一个或多个特定单词或短语可以被映射到在客户端装置110处可访问的一个或多个软件应用198的特定打算、特定履行数据和/或特定软件应用，使得自动助理115可以绕过各种助理组件(例如，ASR引擎120A1和/或120A2、NLU引擎130A1和/或者130A1、和/或其他自动助理组件)的使用，以使得执行与一个或多个特定单词或短语相关联的助理命令。

[0043] 在这些实施方式的一些进一步版本中，利用暖词模型的当前休眠的自动助理功能中的一个或多个可以保持活动，直到满足一个或多个条件为止。一个或多个条件可以包括例如阈值持续时间结束，检测到特定单词或短语中的一个或多个，确定暖词激活事件已经结束，和/或本文描述的其他条件。例如，假设暖词激活事件包括在客户端装置110处响起的定时器。在此示例中，利用暖词模型的当前休眠的自动助理功能中的一个或多个可以保持活动，直到用户提供暖词以停止定时器的响起，和/或直到用户手动停止定时器。作为另一示例，假设暖词激活事件包括在客户端装置110处接收到的通知。在此示例中，利用暖词模型的当前休眠的自动助理功能中的一个或多个可以保持活动，直到用户提供暖词以与通知交互、直到用户手动解除通知，和/或直到阈值持续时间(例如，30秒、60秒等)结束。作为又一示例，假设暖词激活事件包括在客户端装置110处接收到的电话呼叫。在此示例中，利用暖词模型的当前休眠的自动助理功能中的一个或多个可以保持活动，直到用户提供暖词以与电话呼叫交互，和/或直到电话呼叫停止振铃。因此应理解，一旦被激活，可以基于检测到的暖词激活事件和/或客户端装置110的用户响应暖词激活事件的方式，在静态或动态持续时间激活利用本文描述的暖词模型的当前休眠的自动助理功能中的一个或多个。

[0044] 在附加的或可替代的实施方式中，利用暖词模型的一个或多个自动助理功能可以始终是活动的。换言之，无论暖词激活事件引擎151是否检测到暖词激活事件，利用暖词模型的自动助理功能中的一些始终是活动的。例如，利用暖词模型来检测以下的自动助理功能中的一个或多个：与特定查询相关联的特定单词或短语，诸如“天气怎么样?”、“我的日历

上有什么?”等;与特定助理例程相关联的特定单词或短语,诸如“早上好”例程、“晚上好”例程等;与特定装置相关联的特定单词或短语,诸如“开灯”、“打开[室内灯/指示灯]”等;与在客户端装置110处可访问的一个或多个软件应用198中的处于启动或未启动状态的特定软件应用相关联的特定单词或短语,诸如“打开第一应用”、“启动第一应用”、当导航应用处于启动或未启动状态时的“前往主街123号的指示”、当相机应用处于启动状态时的“起司”等;和/或其他特定单词或短语。

[0045] 本文描述的暖词模型可以包括多个不同的暖词模型(例如,音频关键词分类模型),该多个不同的暖词模型被训练成检测一个或多个特定单词或短语(例如,暖词),一个或多个特定单词或短语当被检测到时使自动助理115执行与特定单词或短语中的一个或多个相关联的助理命令。在一些实施方式中,给定暖词模型可以被训练成检测与给定暖词激活事件相关联的单词或短语的特定子集。例如,假设正在客户端装置110或与客户端装置110通信的另一客户端装置处播放音乐。在此示例中,给定暖词模型可以被训练成检测使得执行与控制音乐相关联的助理命令的特定单词或短语的子集,诸如可能使得音乐暂停的“暂停”暖词、可能使得音乐在暂停后继续的“继续”暖词、可能使得音乐调大音量的“调大音量”暖词、可能使得音乐调小音量的“调小音量”暖词、可能使得音乐跳到下一个歌曲的“下一个”暖词,以及针对与音乐相关联的其他特定单词或短语。因此,在此示例中,可以响应于确定正在客户端装置110或与客户端装置110通信的另一客户端装置处播放音乐,激活利用给定暖词模型的当前休眠的助理功能中的一个或多个。因此,在单个时刻,一个或多个暖词模型可以在客户端装置110处是活动的。

[0046] 在附加的或可替代的实施方式中,多个暖词模型可以被训练成检测与给定暖词激活事件相关联的单词或短语的子集。继续以上示例,第一暖词模型可以被训练成检测一个或多个第一特定单词或短语,诸如“暂停”暖词和“继续”暖词,第二暖词模型可以被训练成检测一个或多个第二特定单词或短语,诸如“调大音量”暖词或“调小音量”暖词,第三暖词模型可以被训练成检测一个或多个第三特定单词或短语,诸如“下一个”暖词,以及针对与音乐相关联的其他特定单词或短语。因此,在此示例中,可以响应于确定正在客户端装置110或与客户端装置110通信的另一客户端装置处播放音乐,激活至少利用第一暖词模型、第二暖词模型和第三暖词模型的当前休眠的助理功能中的一个或多个。

[0047] 本文描述的暖词模型可以基于多个训练实例来训练。多个训练实例中的每一个可以包括训练实例输入和训练实例输出。训练实例输入可以包括音频数据,其中音频数据的部分对应于给定暖词模型正被训练以检测的特定单词或短语中的一个或多个,并且训练实例输出可以包括基准真值输出,其中基准真值输出包括训练实例输入的音频数据是否包括给定暖词模型正被训练以检测的特定单词或短语中的一个或多个的指示。例如,在关于音乐播放描述的上述示例中训练给定暖词模型时,对于给定训练实例,训练实例输入可以包括对应于“调大音量”的音频数据,并且对于给定训练实例,训练实例输出可以包括指示音频数据包括特定单词或短语中的一个或多个(例如“调大音量”)的基准真值输出。训练实例输入可以被应用为给定暖词模型的输入,以生成指示训练实例输入是否包括特定单词或短语中的一个或多个的可能性的所预测输出(例如,值,诸如概率、对数似然、二进制值或其他值)。所预测输出可以与训练实例输入的基准真值输出(例如,指示训练实例输入的音频数据实际上包括特定单词或短语中的一个或多个的值,诸如此示例中的值“1”)进行比较,以

生成一个或多个损失，并且给定暖词模型可以基于损失中的一个或多个进行更新。在此示例中，给定训练实例可以被认为是正训练实例，因为训练实例输入包括给定暖词模型正被训练以检测的音频数据的部分。此外，在此示例中，可以可选地基于负训练实例，或基于具有与不包括给定暖词模型正被训练以检测的特定词或短语中的一个或多个的音频数据相对应的训练实例输入的训练实例，来附加地训练给定暖词模型。

[0048] 例如，在关于音乐播放描述的上述示例中训练给定暖词模型时，对于给定训练实例，训练实例输入可以包括对应于“开灯”的音频数据，并且对于给定训练实例，训练实例输出可以包括指示音频数据包括不包括特定单词或短语中的一个或多个(例如“调大音量”)的基准真值输出。训练实例输入可以被应用为给定暖词模型的输入，以生成指示训练实例输入是否包括特定单词或短语中的一个或多个的可能性的所预测输出(例如，值，诸如概率、对数似然、二进制值或其他值)。所预测输出可以与训练实例输入的基准真值输出(例如，指示训练实例输入的音频数据实际上包括特定单词或短语中的一个或多个的值，诸如此示例中的值“1”)进行比较，以生成一个或多个损失，并且给定暖词模型可以基于损失中的一个或几个进行更新。在此示例中，给定训练实例可以被认为是负训练实例，因为训练实例输入不包括给定暖词模型正被训练以检测的音频数据的任何部分(例如，“调大音量”或与控制音乐相对应的任何其他音频)。

[0049] 在一些实施方式中，并且假设利用暖词模型的自动助理功能中的一个或多个是活动的，暖词检测引擎152可以使用暖词模型处理由客户端装置的麦克风生成的音频数据的流，以监测一个或多个特定单词或短语的出现。例如，假设在客户端装置110处接收到文本消息。进一步假设暖词激活事件引擎151检测到在客户端装置110处正在接收文本消息，并且激活一个或多个自动助理组件，一个或多个自动助理组件利用一个或多个暖词模型来检测在客户端装置110处正在接收的文本消息的暖词激活事件的特定单词或短语(例如，“回复”、“不理会”、“给我看”等)，直到满足一个或多个条件为止。在此示例中，进一步假设客户端装置110的用户提供了“回复:我五分钟后到”的口头话语。因此，在处理音频数据流的时，暖词检测引擎152可以检测到音频数据的部分对应于暖词“回复”的出现。在一些实施方式中，暖词检测引擎152能够响应于检测到暖词的出现而使得利用ASR模型和/或NLU模型的一个或多个附加的当前活动的自动助理组件(例如，ASR引擎120A1和/或120A2、NLU引擎130A1和/或130A2，和/或其他自动助理组件)被激活。

[0050] 如下文更详细地描述的(例如，关于图2、3和4A至4C)，ASR引擎120A1和/或120A2可以使用存储在ML模型数据库115A中的ASR模型，处理与暖词的出现相对应的音频数据的部分、在与暖词的出现相对应的音频数据的部分之前的音频数据的前序部分(如果有的话)，和/或在与暖词的出现相对应的音频数据的部分之后的音频数据的后序部分(如果有的话)，以生成ASR输出。此外，NLU引擎130A1和/或130A2可以使用存储在ML模型数据库115A中的NLU模型和/或基于语法的规则来处理ASR输出，以生成NLU输出。在这些实施方式的一些版本中，ASR引擎120A1和/或120A2可以从客户端装置110的音频缓冲器中获得这些部分中的一个或多个(例如，音频数据的前序部分(如果有的话)和/或与暖词的出现相对应的音频数据的部分)。在各种实施方式中，ASR输出可以偏向于与暖词模型相关联的特定术语和/或短语。例如，如果暖词激活事件对应于媒体事件(例如，客户端装置110的用户正在客户端装置110处收听音乐)，则ASR输出可以偏向于与音乐相关的术语和/或短语，诸如艺术家、歌

曲、音乐流派等。

[0051] 例如，自动助理115可以使用存储在ML模型数据库115A中的端点模型来使得音频数据的流被处理，以生成在音频数据的流中捕获的口头话语的多个时间戳。多个时间戳可以至少包括：与当用户开始提供口头话语时的第一时间相关联的第一时间戳、与第一时间之后的当用户开始提供口头话语中包括的一个或多个特定单词或短语时的第二时间相关联的第二时间戳、与第二时间之后的当用户完成提供口头话语中包括的一个或多个特定单词或短语时的第三时间相关联的第三时间戳，以及与第三时间之后的当用户完成提供口头话语时的第四时间相关联的第四时间戳。在此示例中，音频数据的前序部分包括与第一时间戳和第二时间戳之间的口头话语相对应的任何音频数据，音频数据的部分包括与第二时间戳和第三时间戳之间的口头话语相对应的任何音频数据，并且音频数据的后序部分包括与第三时间戳和第四时间戳之间的口头话语相对应的任何音频数据。应注意，提供此示例是为了说明的目的，而并不意味着限制。例如，在一些情况下，诸如当口头话语仅包括特定单词或短语中的一个或多个时，多个时间戳可以至少包括与用户开始提供口头话语的第一时间相关联的第一时间戳，和与第一时间之后的当用户完成提供口头话语时的第二时间相关联的第二时间戳。在此示例中，可能不存在音频数据的前序部分或音频数据的后序部分。

[0052] 在一些实施方式中，暖词验证引擎153可以处理ASR输出，以确定用户是否打算一个或多个特定单词或短语使助理命令执行。例如，暖词验证引擎153可以使得NLU引擎130A1和/或130A2使用存储在ML模型数据库115A中的NLU模型或基于语法的规则来处理ASR输出，以生成NLU输出。暖词验证引擎153可以基于ASR输出和/或NLU输出来确定用户是否打算用户是否打算一个或多个特定单词或短语使助理命令执行。继续以上示例，响应于暖词(例如，捕获到口头话语“回复：我五分钟后到”)。继续以上示例，响应于在音频数据的流中检测到暖词(例如“回复”)的出现，ASR引擎120A1和/或120A2和/或NLU引擎130A1和/或130A2可以被激活并用于处理音频数据的流，以生成音频数据的流的ASR输出(例如，捕获到口头话语“回复：我五分钟后到”)。此外，NLU引擎130A1和/或130A2可以使用存储在ML模型数据库115A中的NLU模型来处理ASR输出，以生成NLU输出。在此示例中，暖词验证引擎153可以确定“回复”是由用户打算用“我五分钟后到”的回复消息来回复在客户端装置110处接收到的文本消息。

[0053] 与以上示例相反，假设当在客户端装置110处接收到文本消息时，客户端装置110的用户替代地提供了“我稍后回复”的口头话语。在此示例中，暖词激活事件引擎151可以检测暖词激活事件(例如，在客户端装置处接收到的文本消息)的出现，并且使得利用暖词模型的当前休眠的自动助理功能中的一个或多个被激活。此外，暖词检测引擎152在使用针对暖词激活事件被激活的暖词模型中的一个或多个来处理音频数据的流时，可以与上述相同或类似的方式来检测音频数据的部分中一个或多个特定单词或短语(例如，“回复”)的出现。然而，在此示例中，自动助理115可以基于使用ASR引擎120A1和/或120A2和/或NLU引擎130A1和/或130A2处理音频数据的后序部分(例如，对应于“稍后”的后序部分)来确定用户不打算“回复”是使自动助理回复所接收的文本消息的特定单词或短语。因此，在此示例中，即使用户提供的口头话语包括通常使得执行助理命令的一个或多个特定单词或短语(例如，“回复”)，自动助理115也可以避免执行对文本消息进行回复的助理命令。

[0054] 在附加的或可替代的实施方式中，仅当用户验证引擎160确定提供口头话语的用

户的身份被验证并且提供口头话语的用户被授权使助理命令执行时，自动助理115才可以使助理命令执行。在这些实施方式的一些版本中，只有助理命令的子集可能要求验证提供口头话语的用户的身份，并且授权提供口头话语的用户使助理命令执行。例如，与电话呼叫暖词激活事件、文本消息暖词激活事件和电子邮件暖词激活事件相关联的助理命令可能要求验证提供口头话语的用户的身份，并且授权提供口头话语的用户使助理命令执行。然而，与媒体暖词激活事件相关联的助理命令可能不要求验证提供口头话语的用户的身份，并且授权提供口头话语的用户使助理命令执行。可以使用各种技术来验证提供口头话语的用户的身份。

[0055] 在这些实施方式的一些版本中，用户验证引擎160可以使用存储在ML模型数据库115A中的说话者识别(SID)模型来处理音频数据的流，以生成提供口头话语的用户的说话者嵌入。所生成的说话者嵌入可以在嵌入空间中与客户端装置110的用户的一个或多个先前生成的说话者嵌入(例如，与用户的用户简档相关联的嵌入)进行比较。例如，如果基于嵌入空间中的所生成的说话者嵌入与针对用户的一个或多个先前生成的说话者嵌入之间的比较确定的距离度量满足阈值，则可以将用户验证为提供口头话语的用户。可以基于以对音频数据的一个或多个实例的处理为基础而生成的输出来生成用户的一个或多个先前生成的说话者嵌入，音频数据的一个或多个实例包括来自用户的口头话语(例如，当最初配置客户端装置110时)。例如，一个或多个先前生成的说话者嵌入可以基于多个不相同嵌入的平均值(或其他组合)，该多个不相同嵌入中的每一个都是基于对包括来自客户端装置的用户的对应口头话语的音频数据的对应实例进行处理而生成的。在这些实施方式的一些版本中，SID模型可以是文本无关(TI)SID模型，而在其他实施方式中，SID模型可以是文本相关(TD)SID模型。

[0056] 在TD SID中，用户的一个或多个先前生成的说话者嵌入是基于仅包括一个或多个特定单词或短语的口头话语而生成的。此外，在使用中，用户必须说出要使用TD SID模型生成的一个或多个TD SID说话者嵌入的一个或多个特定单词或短语，一个或多个TD SID说话者嵌入可以有效地与用户的一个或多个先前生成的TD说话者嵌入进行比较，以确定口头话语是否来自授权用户(例如，客户端装置110的用户或与客户端装置110相关联的另一用户)。例如，TD说话者辨识中的一个或多个特定单词或短语可以被限制为被配置成调用自动助理的一个或多个调用短语(例如，热词和/或触发词，诸如例如“嘿助理”、“OK助理”和/或“助理”)或本文描述的一个或多个暖词。相反，在TI SID中，使用TI SID模型处理的口头话语不限于一个或多个特定单词或短语。换言之，可以使用TI SID模型处理基于几乎任何口头话语的音频数据以生成TI SID说话者嵌入，TI SID说话者嵌入可以有效地与用户的一个或多个先前生成的TI说话者嵌入进行比较，以确定口头话语是否来自授权用户(例如，客户端装置110的用户或与客户端装置110相关联的另一用户)。此外，在各种实施方式中，在TI SID中使用的用户的一个或多个先前生成的TI说话者嵌入是基于包括不同单词和/或短语并且不限于调用单词和/或短语、暖词和/或任何其他特定口头话语的口头话语而生成的。

[0057] 在这些实施方式的附加的或可替代的版本中，可以利用其他生物特征信息来验证提供口头话语的用户的身份。用于验证用户身份的生物特征信息可以包括例如面纹识别、指纹识别和/或其他类型的生物特征信息。例如，客户端装置110的存在传感器113的视觉组件可以捕获包括提供口头话语的用户的面部的视觉数据。用户验证引擎160可以使用存储

在ML模型数据库115A中的面部识别(FID)模型来处理视觉数据,以生成提供口头话语的用户的面纹。提供口头话语的用户的面纹可以对应于嵌入,嵌入可以在嵌入空间中和与客户端装置的用户的面纹相对应的一个或多个先前生成的嵌入进行比较,以确定提供口头话语的用户是否被授权使助理命令执行。作为另一示例,存在传感器113中的指纹传感器可以接收提供口头话语的用户的触摸输入,并且将检测到的指纹与客户端装置110的用户的已知指纹进行比较,以确定提供口头话语的用户是否被授权使助理命令执行。

[0058] 在附加的或可替代的实施方式中,声学引擎170可以处理音频数据以生成与音频数据相关联的一个或多个声学特性。在这些实施方式中,自动助理115可以附加地或可替代地考虑与音频数据相关联的声学特性中的一个或多个,以确定是否响应于在音频数据的流中检测到暖词而使助理命令执行。声学引擎170可以利用各种技术来生成与音频数据相关联的一个或多个声学特性。

[0059] 在这些实施方式的一些版本中,声学引擎170可以使用存储在ML模型数据库115A中的话音活动检测(VAD)模型来处理音频数据,以确定在对应于一个或多个特定单词或短语的音频数据的部分之前是否存在音频数据的前序部分和/或在对应于一个或多个特定单词或短语的音频数据的部分之后是否存在音频数据的后序部分。在这些实施方式的一些进一步版本中,ASR引擎120A1和/或120A2和/或NLU引擎130A1和/或130A2可以仅响应于确定音频数据的前序部分和/或音频数据的后序部分存在而被激活。例如,假设当在客户端装置110处接收到通知时,客户端装置110的用户提供了“查看”的口头话语。在此示例中,暖词激活事件引擎151可以检测暖词激活事件(例如,在客户端装置处接收到的通知)的出现,并且使得利用暖词模型的当前休眠的自动助理功能中的一个或多个被激活。此外,暖词检测引擎152在使用针对暖词激活事件被激活的暖词模型中的一个或多个来处理音频数据的流时,可以与上述相同或类似的方式来检测音频数据的部分中一个或多个特定单词或短语(例如,“查看”)的出现。然而,声学引擎170可以基于使用VAD模型处理音频数据的流来确定不存在音频数据的前序部分(例如,在音频数据的流中在“查看”之前)或音频数据的后序部分(例如,在音频数据的流中在“查看”之后)。结果,ASR引擎120A1和/或120A2和/或NLU引擎130A1和/或130A2可以不被激活,并且自动助理115可以使通知被提供用于呈现给用户(并且可选地基于确定用户被授权查看通知)。

[0060] 在这些实施方式的附加的或可替代的版本中,声学引擎170可以使用存储在ML模型数据库115A中的语言识别模型来处理音频数据,以确定在音频数据中捕获的口头话语的语言。在这些实施方式的一些版本中,口头话语的语言可以提供音频数据的前序部分或音频数据的后序部分中的哪一个提供场境信息的指示,以通知自动助理115是否引起和与音频数据的部分相对应的特定单词或短语中的一个或多个相关联的助理命令,和/或避免使和与音频数据的部分相对应的特定单词或短语中的一个或多个相关联的助理命令执行。例如,在英语中,指示避免引起助理命令的某些术语或短语倾向于在与特定单词或短语中的一个或多个(例如,“不接听”、“不查看”等)相对应的音频数据的部分之前(例如,在音频数据的前序部分中)。然而,在其他语言(例如,德语)中,指示避免引起助理命令的某些术语或短语倾向于在与特定单词或短语中的一个或多个相对应的音频数据的部分之后(例如,在音频数据的后序部分中)。

[0061] 在附加的或可替代的实施方式中,口头话语的语言可以提供应利用哪个(哪些)暖

词模型和/或哪个(哪些)ASR模型来进一步处理音频数据的流的指示。例如,假设跨语言识别模型生成的输出指示提供在音频数据的流中捕获的口头话语的用户正在说英语,则自动助理组件随后使用的暖词模型和/或ASR模型可以是特定于英语的。然后,假设跨语言识别模型生成的输出指示提供在音频数据的流中捕获的口头话语的用户正在说德语(或其他语言),则自动助理组件随后使用的暖词模型和/或ASR模型可以是特定于德语(或其他语言)的。在附加的或可替代的实施方式中,可以由一个或多个主动的自动助理组件使用被训练成检测不相同语言的特定单词和/或短语中的一个或多个的暖词模型,并且可以基于检测到的特定单词和/或短语中的一个或多个的语言来确定哪个ASR模型应随后被一个或多个自动助理组件用来进一步处理音频数据的流。例如,假设跨暖词模型生成的输出指示检测到的暖词是以西班牙语提供的,则自动助理组件随后使用的ASR模型可以是特定于西班牙语的。然而,假设跨暖词模型生成的输出指示检测到的暖词是以英语(或另一种语言)提供的,则自动助理组件随后使用的ASR模型可以是特定于英语(或所述另一种语言)的。

[0062] 在这些实施方式的附加的或可替代的版本中,声学引擎170可以使用存储在ML模型数据库115A中的韵律特性分类模型来处理音频数据,以确定在音频数据的流中捕获的口头话语的一个或多个韵律特性。在音频数据的流中捕获的口头话语的韵律特性可以包括例如以下中的一个或多个:语调、音调、重音、旋律、节奏和停顿。这些韵律属性中的一个或多个可以通知自动助理115是否使和与音频数据的部分相对应的特定单词或短语中的一个或多个相关联的助理命令执行,和/或避免使和与音频数据的部分相对应的特定单词或短语中的一个或多个相关联的助理命令执行。例如,如果在音频数据中捕获的口头话语不包括任何停顿,并且在特定单词或短语中的一个或多个上施加了附加应力,则口头话语的韵律属性中的一个或多个可以通知自动助理115,用户更有可能打算执行助理命令。相反,如果在音频数据中捕获的口头话语确实包括停顿和不确定的音调,则口头话语的韵律属性中的一个或多个可以通知自动助理115,用户更有可能不打算执行助理命令。

[0063] 在附加的或可替代的实施方式中,自动助理115可以分析由存在传感器113中的一个或多个生成的传感器数据的实例。自动助理115在确定是否响应于在音频数据的流中检测到暖词而使助理命令执行时考虑由一个或多个存在传感器113生成的传感器数据的实例。例如,如果客户端装置110的一个或多个视觉组件指示提供口头话语的用户的凝视没有指向客户端装置110,则视觉数据可以通知自动助理115,用户更有可能不打算使助理命令执行。附加地或可替代地,如果客户端装置110的一个或多个加速度计指示客户端装置110正在移动(并且可选地与缺少用户的凝视相结合),则视觉数据可以通知自动助理115,用户更有可能不打算执行助理命令。

[0064] 现在转向图2,描绘了使用图1的各种组件展示本公开的各个方面的示例过程流程。暖词激活事件引擎151可以监测暖词激活事件251的出现。响应于检测到暖词事件的出现,暖词检测引擎152可以被激活,并且使用与存储在ML模型数据库115A中的暖词激活事件251相关联的暖词模型开始处理由客户端装置110的一个或多个麦克风201生成的音频数据201A的流。如框299A所示,响应于确定音频数据不对应于一个或多个暖词,暖词检测引擎152可以继续监测暖词中的一个或多个的出现,直到满足一个或多个条件为止。

[0065] 然而,也如框299A所示,响应于确定音频数据的部分确实对应于暖词中的一个或多个,暖词检测引擎152可以使得获得在与暖词中的一个或多个相对应的音频数据的部分

之前的音频数据的前序部分201B(如果有的话)和/或在与暖词中的一个或多个相对应的音频数据的部分之后的音频数据的后序部分201C(如果有的话)。此外,并且还响应于确定音频数据的部分确实对应于暖词中的一个或多个,暖词检测引擎152可以使ASR引擎120A1和/或120A2和/或NLU引擎130A1和/或130A2被激活。ASR引擎120A1和/或120A2可以使用存储在ML模型数据库115A中的ASR模型,处理音频数据的前序部分201B和/或音频数据的后序部分201C(以及可选地与暖词中的一个或多个相对应的音频数据的部分),以生成ASR输出220(例如,针对音频数据中包括的用户的任何语音辨识)。此外,NLU引擎130A1和/或130A2可以使用存储在ML模型数据库115A中的NLU模型来处理ASR输出220,以生成NLU输出230(例如,针对音频数据中包括的用户的任何语音辨识的注释)。暖词验证引擎153可以基于ASR输出220和/或NLU输出230来确定用户实际上是否打算让自动助理115使助理命令执行。

[0066] 例如,假设在听音乐时音频数据201A经由客户端装置110或与客户端装置110通信的附加客户端装置(例如,智能扬声器)捕获到“你能调大音量吗”的口头话语。在此示例中,暖词激活事件251可以对应于媒体事件,并且暖词“调大音量”可以使用与媒体事件相关联的一个或多个暖词模型在捕获口头话语的音频数据中检测到。此外,ASR引擎120A1和/或120A2可以使用ASR模型处理与“你能调”相对应的音频数据的前序部分201B和/或与“大音量”相对应的音频数据的部分,以生成与“你能调”和/或“大音量”相对应的所辨识文本(例如,作为ASR输出220)。此外,NLU引擎130A1和/或130A2可以使用NLU模型来处理与“你能调”和/或“大音量”相对应的所辨识文本,以生成注释的所辨识文本(例如,作为NLU输出230),注释的所辨识文本用各种注释来注释所辨识文本的一个或多个术语。在此示例中,并且如框299B所示,自动助理115可以基于所辨识文本和/或注释的所辨识文本来确定应执行调整正在播放的音乐的音量的助理命令215A,并且使助理命令215A执行(例如,通过在客户端装置110处调大音量和/或向与客户端装置110通信的附加客户端装置传输调大音量的请求)。

[0067] 相反,假设在听音乐时音频数据201A经由客户端装置110或与客户端装置110通信的附加客户端装置(例如,智能扬声器)捕获到“请不要调大音量”的口头话语。在此示例中,暖词激活事件251可以对应于媒体事件,并且暖词“调大音量”可以使用与媒体事件相关联的一个或多个暖词模型在捕获口头话语的音频数据中检测到。此外,ASR引擎120A1和/或120A2可以使用ASR模型处理与“请不要调”相对应的音频数据的前序部分201B和/或与“大音量”相对应的音频数据的部分,以生成与“请不要调”和/或“大音量”相对应的所辨识文本(例如,作为ASR输出220)。此外,NLU引擎130A1和/或130A2可以使用NLU模型来处理与“请不要调”和/或“大音量”相对应的所辨识文本,以生成注释的所辨识文本(例如,作为NLU输出230),注释的所辨识文本用各种注释来注释所辨识文本的一个或多个术语。在此示例中,并且如框299B所示,自动助理115可以基于所辨识文本和/或注释的所辨识文本来确定不应执行调整正在播放的音乐的音量的助理命令,并且避免使助理命令执行。在这些示例中的一些中,可以丢弃音频数据201A和/或基于音频数据201A生成的任何其他数据,如215B所示。

[0068] 在一些实施方式中,自动助理115是否确定使助理命令执行可以基于是否验证了提供口头话语的用户的身份,以及所识别用户是否被授权执行助理命令(例如,使用如关于图1、5和6A至6C描述的用户验证引擎160确定的)。例如,在暖词激活事件对应于在客户端装置110处接收的电子通信(例如,电话呼叫、文本消息、电子邮件、社交媒体消息等)的实施方式中,可能需要在自动助理115使助理命令执行之前执行用户的身份。作为另一示例,在暖

词激活事件对应于被控制的某些装置(例如,智能锁、诸如智能烤箱的智能电器、智能车库门开启器等)的实施方式中,可能需要在自动助理115使助理命令执行之前执行用户的身份。

[0069] 尽管关于图2的处理流程描述了特定实施方式,但是应理解,这是为了示例的目的而并不意味着限制。例如,假设音频数据201A捕获到“展示我回家的路”的口头话语。在此示例中,检测到的暖词激活事件251可以对应于在客户端装置110处可访问的导航应用。因此,与导航事件相关联的一个或多个暖词模型可以连续地监测一个或多个与导航相关的暖词的出现。此外,自动助理可以利用附加的或可替代的信号来确定是否使助理命令执行(例如,如关于声学引擎170所描述的)。

[0070] 通过使用本文描述的技术,可以实现一个或多个技术优点。作为一个非限制性示例,本文描述的技术使得自动助理能够基于处理暖词的场境音频数据来确定用户实际上是否打算暖词出现以使助理命令执行。因此,可以减少暖词的假阳性数量,从而在自动助理利用一个或多个网络来使得执行助理命令的情况下节省客户端装置处的计算资源和/或网络资源。此外,由于客户端装置的用户不需要采取行动来纠正或撤消任何助理命令,因此可以减少用户输入的数量,从而节省客户端装置处的计算资源。此外,本文描述的各种自动助理功能可以保持休眠直到如本文描述需要为止,从而节省客户端装置处的计算资源。作为另一非限制性示例,本文描述的技术使得自动助理能够使用SID和/或其他技术来验证提供包括暖词的口头话语的用户的身份。因此,只有授权用户才能使某些助理命令执行,从而提高与客户端装置的用户相关联的数据的数据安全性。

[0071] 现在转向图3,描绘了示出基于场境音频数据来确定是否使助理命令执行的示例方法300的流程图。为了方便,参考执行操作的系统来描述方法300的操作。方法300的此系统包括计算装置(例如,图1、4A至4C和6A至6C的客户端装置110和/或图7的计算装置710、一个或多个服务器和/或其他计算装置)的一个或多个处理器、存储器和/或其他组件。此外,虽然方法300的操作以特定顺序示出,但这并不意味着限制。可以重新排序、省略和/或添加一个或多个操作。

[0072] 在框352处,系统监测暖词激活事件的出现。暖词激活事件可以包括例如在客户端装置处接收到的电话呼叫、在客户端装置处接收到的文本消息、在客户端装置处接收到的电子邮件、在客户端装置处响起的警报或定时器、在客户端装置或客户端装置的环境中的附加客户端装置处播放的媒体、在客户端装置处接收到的通知、客户端装置的位置、在客户端装置处可访问的软件应用,和/或与客户端装置相关联的其他事件,在其中,用户可以提供口头话语以使客户端装置或与客户端装置通信的附加客户端装置受到控制。值得注意的是,这些暖词激活事件中的一些是离散事件(例如,在客户端装置处接收到的电话呼叫、在客户端装置处接收到的文本消息、在客户端装置处接收到的电子邮件、在客户端装置处响起的警报或定时器等),而这些暖词激活事件中的一些其他事件是连续事件(例如,在客户端装置处可访问的软件应用)。此外,应注意,这些暖词激活事件并不是相互排斥的。换言之,系统可以在给定时间实例检测多个暖词激活事件。如果在框352的迭代中,系统没有检测到暖词激活事件的出现,则系统可以在框352处继续监测暖词激活事件的出现。如果在框352的迭代中,系统检测到暖词激活事件的出现,则系统可以前进到框354。

[0073] 在框354处,系统使用一个或多个暖词模型处理音频数据的流,以监测一个或多个

特定单词和/或短语(例如,暖词)的出现。响应于在框352处检测到暖词激活事件的出现,系统可以使一个或多个自动助理组件利用一个或多个暖词模型。用于处理音频数据的流的一个或多个暖词模型可以与在框352处检测到的暖词激活事件相关联。例如,如果在框352处检测到的暖词激活事件对应于在客户端装置处接收到的电话呼叫,则可以利用与电话呼叫事件相关联的一个或多个暖词模型处理音频数据的流。作为另一示例,如果在框352处检测到的暖词激活事件对应于在客户端装置处播放的媒体,则可以利用与媒体事件相关联的一个或多个暖词模型处理音频数据的流。

[0074] 在框356处,系统确定在音频数据的流中是否检测到特定单词和/或短语中的一个或多个的出现。系统可以继续监测一个或多个特定单词和/或短语的出现,直到满足一个或多个条件为止。一个或多个条件可以包括例如阈值持续时间结束,检测到特定单词或短语中的一个或多个,确定暖词激活事件已经结束,和/或其他条件。如果在框356的迭代中,系统没有检测到特定单词和/或短语中的一个或多个的出现,则系统可以返回到框354,并且在框356处继续处理音频数据的流并监测特定单词和/或短语中的一个或多个的出现,直到满足一个或多个条件为止。如果在框356的迭代中,系统确实检测到特定单词和/或短语中的一个或多个的出现,则系统前进到框358。

[0075] 在框358处,系统确定是否存在除了暖词之外的附加话音活动(例如,在暖词之前或之后)。系统可以使用VAD模型处理音频数据的流,以确定音频数据的流是否除了与特定单词和/或短语中的一个或多个相对应的音频数据的部分之外还捕获了任何附加音频数据。换言之,系统可以利用VAD模型来确定用户是否除了特定单词和/或短语中的一个或多个之外还提供了任何附加语音,使得解释特定单词和/或短语中的一个或多个时可以将与附加语音相对应的音频数据用作场境音频数据(例如,音频数据的前序部分和/或音频数据的后序部分)。如果在框358的迭代中,系统确定除了暖词之外不存在任何附加话音活动,则系统前进到框366,并且使自动助理执行与特定单词和/或短语中的一个或多个相关联的助理命令。值得注意的是,在前进到框366时,系统可以绕过ASR和/或NLU。如果在框358的迭代中,系统确定除了暖词之外还存在附加话音活动,则系统前进到框360。在一些实施方式中,可以省略框358的操作,并且系统可以直接前进到框360。

[0076] 在框360处,系统使用ASR模型处理音频数据的前序部分和/或音频数据的后序部分,以生成ASR输出。附加地或可替代地,系统可以结合音频数据的前序部分和/或音频数据的后序部分(例如,作为音频数据的单个实例),使用ASR模型处理与暖词相对应的音频数据的部分以生成ASR输出。可以响应于在框356处检测到特定单词和/或短语中的一个或多个的出现和/或响应于在框358处确定存在附加话音活动,激活利用ASR模型的一个或多个自动助理组件。在各种实施方式中,ASR输出可以偏向于与暖词模型相关联的特定单词和/或短语中的一个或多个。在框362处,系统使用NLU模型处理ASR输出以生成NLU输出。类似于上文关于框360所描述的,可以响应于在框356处检测到特定单词和/或短语中的一个或多个的出现和/或响应于在框358处确定存在附加话音活动,激活利用NLU模型的一个或多个自动助理组件。

[0077] 在框364处,系统确定提供包括暖词并且在音频数据的流中捕获的口头话语的用户是否打算由自动助理执行与在框356处检测到的暖词相关联的助理命令。系统可以基于在框360处生成的ASR输出和/或在框362处生成的NLU输出来确定用户是否打算由自动助理

执行助理命令。如果在框364的迭代中，系统确定用户打算由自动助理执行助理命令，则系统前进到框366。在框366处，系统使自动助理执行与在框356处检测到的暖词相关联的助理命令。如果在框364的迭代中，系统确定用户不打算由自动助理执行助理命令，则系统前进到框368。在框368处，系统避免使自动助理执行与在框356处检测到的暖词相关联的助理命令。

[0078] 现在转向图4A、4B和4C，描绘了利用图3的方法300基于场境音频数据来确定是否使助理命令执行的各种非限制性示例。自动助理可以至少部分地在客户端装置110处实施(例如，关于图1描述的自动助理115)。自动助理可以利用暖词系统180(例如，关于图1描述的暖词系统)来确定是否基于场境音频数据来执行助理命令(例如，如关于图3的方法300所描述)。

[0079] 图4A、4B和4C中所描绘的客户端装置110可以包括各种用户接口组件，包括例如用于基于口头话语和/或其他听觉输入生成音频数据的麦克风、用于听觉渲染合成语音和/或其他听觉输出的扬声器，以及用于接收触摸输入和/或视觉渲染转录和/或其他视觉输出的显示器190。此外，客户端装置110的显示器190包括各种系统接口元件(例如，硬件和/或软件接口元件)，客户端装置110的用户可以与这些元件交互以使客户端装置110执行一个或多个动作(例如，如关于图6B和6C所描述)。尽管图4A至4C中描绘的客户端装置110是移动装置，但是应理解，这是为了示例的目的而并不意味着限制。

[0080] 具体参考图4A，假设与客户端装置110相关联的用户101与附加用户102共同位于环境400中。进一步假设在客户端装置110处接收到电话呼叫。在此示例中，自动助理可以确定在客户端装置110处接收到的电话呼叫对应于暖词激活事件。因此，自动助理可以激活利用与电话呼叫事件相关联的一个或多个暖词模型的一个或多个当前休眠的自动助理功能。可以使用与电话呼叫事件相关联的一个或多个暖词模型处理由客户端装置110的麦克风生成的音频数据的流，以监测与电话呼叫事件相关联的特定单词和/或短语(例如，“接听”、“拒绝”、“屏幕呼叫”、“发送到话音邮件”等)的出现，当检测到特定单词和/或短语时，使自动助理执行与特定单词和/或短语相关联的特定动作。

[0081] 例如，并且如图4A所示，假设当客户端装置110正在振铃时，客户端装置110的用户101提供“接听”的口头话语452A。在此示例中，在使用与电话呼叫事件相关联的一个或多个暖词模型处理由客户端装置110的麦克风生成的音频数据的流时，自动助理可以确定音频数据对应于暖词“接听”，当检测到暖词时，使自动助理代表用户101接听电话呼叫。此外，自动助理可以使用VAD模型处理音频数据的流，以确定是否存在任何附加话音活动。在此示例中，使用VAD模型生成的输出将指示不存在附加话音活动，并且因此不存在待处理的场境音频数据，以确定用户101实际上是否打算口头话语452A使电话呼叫被接听。因此，自动助理将使得在客户端装置110处接听电话呼叫。

[0082] 相反，并且如图4B所示，假设当客户端装置110正在振铃时，客户端装置110的用户101提供“我现在不接听，稍后再给他们回电话”的口头话语452B。在此示例中，在使用与电话呼叫事件相关联的一个或多个暖词模型处理由客户端装置110的麦克风生成的音频数据的流时，自动助理可以确定音频数据的部分对应于暖词“接听”，当检测到该暖词时，使自动助理代表用户101接听电话呼叫。此外，自动助理可以使用VAD模型处理音频数据的流，以确定是否存在任何附加话音活动。在此示例中，使用VAD模型生成的输出将指示存在附加话音

活动，并且因此存在待处理的场境音频数据，以确定用户101实际上是否打算口头话语452B使电话呼叫被接听。因此，自动助理将使得获得音频数据的前序部分和/或音频数据的后序部分。

[0083] 在此示例中，音频数据的前序部分对应于口头话语452B的“我不”。自动助理可以使用ASR模型处理音频数据的前序部分，以生成音频数据的前序部分的ASR输出。此外，自动助理可以使用NLU模型处理ASR输出，以生成音频数据的前序部分的NLU输出。自动助理可以基于音频数据的前序部分的ASR输出和/或NLU输出来确定用户101实际上是否打算口头话语452B使电话呼叫被接听。例如，自动助理可以基于ASR输出和/或NLU输出来确定“不”否定了暖词“接听”。因此，在此示例中，自动助理将基于对音频数据的前序部分的处理而避免使得在客户端装置110处接听电话呼叫。在各种实施方式中，除了本文描述的暖词模型之外，还可以附加地或可替代地使用一个或多个否定暖词模型。在这些实施方式中，一个或多个否定暖词模型可以用于检测与否定暖词的暖词非常接近的其他单词和/或短语(例如，“不”、“不能”、“不要”等)。此外，如果这些否定暖词模型中的一个或多个检测到与暖词非常接近的其他单词和/或短语，则利用ASR模型和/或NLU模型的一个或多个自动助理组件可以保持休眠。

[0084] 在一些实施方式中，并且假设自动助理不能基于处理音频数据的前序部分来确定用户101实际上是否打算口头话语452B使电话呼叫被接听，自动助理可以进一步使得使用ASR模型处理音频数据的后序部分以生成音频数据的后序部分的ASR输出(例如对应于口头话语452B的“稍后再给他们回电话”)。此外，自动助理可以使用NLU模型处理ASR输出，以生成音频数据的后序部分的NLU输出。自动助理可以基于音频数据的后序部分的ASR输出和/或NLU输出来确定用户101实际上是否打算口头话语452B使电话呼叫被接听。例如，自动助理可以确定“稍后再给他们回电话”指示用户101不打算在这一时刻及时接听电话，使得音频数据的后序部分基于ASR输出和/或NLU输出来否定暖词“接听”。因此，在此示例中，自动助理将基于对音频数据的后序部分的进一步处理来避免使得在客户端装置110处接听电话呼叫。

[0085] 作为另一示例，并且如图4C所示，假设当客户端装置110正在振铃时，共同位于与客户端装置110相关联的用户101的环境中的附加用户102提供“你要接听电话吗?”的口头话语452C。在此示例中，在使用与电话呼叫事件相关联的一个或多个暖词模型处理由客户端装置110的麦克风生成的音频数据的流时，自动助理可以确定音频数据对应于暖词“接听”，当检测到暖词时，使自动助理代表用户101接听电话呼叫。此外，自动助理可以使用VAD模型处理音频数据的流，以确定是否存在任何附加话音活动。在此示例中，使用VAD模型生成的输出将指示存在附加话音活动，且因此存在要处理的场境音频数据。然而，与图4A和4B的示例相反，口头话语452C是由附加用户102而不是与客户端装置110相关联的用户101提供的。

[0086] 在附加的或可替代的实施方式中，并且如关于图5和图6A至6C所描述的，在自动助理使助理命令执行之前，可能需要验证提供口头话语452A、452B和/或452C的用户的身份。此外，自动助理可能附加地或可替代地需要确定所识别用户被授权使助理命令执行。自动助理需要验证提供口头话语的用户的身份和/或确定用户是否被授权的这些助理命令可以是与特定暖词模型相关联的助理命令、与在客户端装置110处可访问的特定软件应用相关

联的助理命令、与基于助理命令要控制的特定装置相关联的助理命令和/或助理命令的其他子集。

[0087] 尽管如此，并且关于图4C，自动助理仍然可以使得获得音频数据的前序部分和/或音频数据的后序部分。换言之，如果口头话语不打算使助理命令执行，则自动助理可以不尝试验证提供口头话语452C的附加用户102的身份。在此示例中，音频数据的前序部分对应于口头话语452C的“你要”。自动助理可以使音频数据的前序部分以与上述相同或类似的方式进行处理，以生成音频数据的前序部分的ASR输出和NLU输出(并且可选地结合与特定术语或短语中的一个或多个相对应的音频数据的部分和/或音频数据的后序部分(例如，作为音频数据的单个实例或音频数据的多个不同实例))。此外，自动助理可以基于音频数据的前序部分的ASR输出和/或NLU输出来确定附加用户102实际上是否打算口头话语452C使电话呼叫被接听。例如，自动助理可以确定基于ASR输出和/或NLU输出的“你要”对应于附加用户102向与客户端装置110相关联的用户101提出的问题，并且不意在使电话被接听。因此，在此示例中，自动助理将基于对音频数据的前序部分的处理(以及可选地对与口头话语452C的“电话”相对应的后序音频数据的进一步处理)来避免在客户端装置110处接听电话呼叫。

[0088] 尽管关于使特定助理命令执行的特定暖词描述了图4A至4C，但是应理解，这是为了示例的目的而并不意味着限制。相反，应理解，关于图4A至4C描述的技术可以用于检测任何暖词激活事件，使用与暖词激活事件相关联的暖词模型检测任何暖词，以及使任何助理命令执行。

[0089] 现在转向图5，描绘示出基于说话者识别来确定是否执行助理命令的示例方法500的流程图。为了方便，参考执行操作的系统来描述方法500的操作。方法500的此系统包括计算装置(例如，图1、4A至4C和6A至6C的客户端装置110和/或图7的计算装置710、一个或多个服务器和/或其他计算装置)的一个或多个处理器、存储器和/或其他组件。此外，虽然方法500的操作以特定顺序示出，但这并不意味着限制。可以重新排序、省略和/或添加一个或多个操作。

[0090] 在框552处，系统监测暖词激活事件的出现。如果在框552的迭代中，系统没有检测到暖词激活事件的出现，则系统可以在框552处继续监测暖词激活事件的出现。如果在框552的迭代中，系统检测到暖词激活事件的出现，则系统可以前进到框554。在框554处，系统使用一个或多个暖词模型处理音频数据的流，以监测一个或多个特定单词和/或短语(例如，暖词)的出现。在框356处，系统确定在音频数据的流中是否检测到特定单词和/或短语中的一个或多个的出现。如果在框556的迭代中，系统没有检测到特定单词和/或短语中的一个或多个的出现，则系统可以返回到框554，并且在框556处继续处理音频数据的流并监测特定单词和/或短语中的一个或多个的出现，直到满足一个或多个条件为止。如果在框556的迭代中，系统确实检测到特定单词和/或短语中的一个或多个的出现，则系统前进到框558。框552、554和556的操作可以分别以关于图3的方法300的框352、354和356所描述的相同或相似的方式来执行。

[0091] 在框558处，系统确定是否验证提供在音频数据的流中捕获的包括特定单词和/或短语中的一个或多个的口头话语的用户的身份。系统可以基于特定单词和/或短语中的一个或多个来确定是否验证提供口头话语的用户的身份，特定单词和/或短语中的一个或多个与特定暖词模型相关联、与在用户的客户端装置处可访问的特定软件应用相关联，与基

于助理命令要控制的特定装置相关联，和/或助理命令的其他子集。例如，在助理命令使媒体在客户端装置处播放或使智能灯被控制的实施方式中，则可能不需要验证用户的身份。然而，在助理命令使对传入的电子通信(例如，电话呼叫、文本消息、电子邮件、社交媒体消息等)回复或使智能锁被控制的实施方式中，则可能需要验证用户的身份。如果在框558的迭代中，系统确定不需要验证提供口头话语的用户的身份，则系统可以前进到图3的框358并且继续执行图2的方法300。如果在框558的迭代中，系统确定需要验证提供口头话语的用户的身份，则系统可以前进到框560。

[0092] 在框560处，系统确定音频数据的流中包括的音频数据是否足以执行TI SID。系统可以通过将在音频数据的流中包括的音频数据中捕获的口头话语的长度与阈值长度进行比较来确定音频数据是否足以执行TI SID。如果口头话语满足阈值长度，则系统可以确定音频数据足以执行TI SID。如果在框560的迭代中，系统确定音频数据足以执行TI SID，则系统可以前进到框562。在框562处，系统使用TI SID模型处理音频数据以验证用户的身份。在框566处，从框562开始，系统确定是否验证了用户的身份。例如，系统可以使用TI SID模型处理音频数据，以生成提供口头话语的用户的TI说话者嵌入。此外，系统可以在嵌入空间中，将提供口头话语的用户的TI说话者嵌入与客户端装置的一个或多个用户的一个或多个先前生成的TI说话者嵌入进行比较。此外，系统可以以基于比较而确定的距离度量为基础来确定TI说话者嵌入是否对应于先前生成的TI说话者嵌入中的一个或多个。如果在框566的迭代中，从框562开始，系统确定距离度量满足距离阈值，则可以将提供口头话语的用户的身份验证为与客户端装置相关联的用户，并且假设所识别的用户被授权使助理命令执行，在框568处，系统可以使自动助理执行助理命令。然而，如果在框566的迭代中，从框562开始，系统确定距离度量不满足距离阈值，则系统可以在框570处生成并提供用于对用户的身份进行附加验证的提示(例如，请求提供口头话语的用户提供面纹识别或指纹识别)以使自动助理执行助理命令。

[0093] 如果在框560的迭代中，系统确定音频数据不足以执行TI SID，则系统可以前进到框564。在框564处，系统使用TD SID模型处理音频数据以验证用户的身份。在框566处，从框562开始，系统确定是否验证了用户的身份。例如，系统可以使用TD SID模型处理音频数据，以生成提供口头话语的用户的TD说话者嵌入。此外，系统可以在嵌入空间中，将提供口头话语的用户的TD说话者嵌入与客户端装置的一个或多个用户的一个或多个先前生成的TI说话者嵌入进行比较。值得注意的是，TD说话者嵌入对应于一个或多个特定单词和/或短语的说话者嵌入(例如，与在框556处检测到的暖词相对应的音频数据的部分)。此外，系统可以以基于比较而确定的距离度量为基础来确定TD说话者嵌入是否对应于特定单词和/或短语中的一个或多个的先前生成的TD说话者嵌入中的一个或多个。如果在框566的迭代中，从框564开始，系统确定距离度量满足距离阈值，则可以将提供口头话语的用户的身份验证为与客户端装置相关联的用户，并且假设所识别的用户被授权使助理命令执行，在框568处，系统可以使自动助理执行助理命令。然而，如果在框566的迭代中，从框564开始，系统确定距离度量不满足距离阈值，则系统可以在框570处生成并提供用于对用户的身份进行附加验证的提示(例如，请求提供口头话语的用户提供TI SID验证或TD SID验证的示例话语，和/或请求提供口头话语的用户提供面纹识别或指纹识别)以使自动助理执行助理命令。

[0094] 值得注意的是，尽管关于分别执行TI SID和TD SID以验证提供口头话语的用户的

身份描述了图5,但是应理解,这是为了示例的目的而并不意味着限制。例如,在口头话语的长度足以执行TI SID的实施方式中,可以执行TI SID和TD SID两者来验证提供口头话语的用户的身份。作为另一示例,在各种实施方式中,只有在确定先前生成的TD说话者嵌入中的一个或多个对应于在口头话语中捕获的一个或多个特定单词和/或短语的情况下,才可以执行TD SID。此外,这些技术可以用于验证在由同一用户提供的音频数据的流中捕获的口头话语的部分。例如,假设音频数据的流捕获来自第一用户的第一口头话语“不”,以及来自第二用户的第二口头话语“接听”。进一步假设第二用户与客户端装置相关联并且被授权接听,并且针对第二用户的“接听”的口头话语的TD说话者嵌入先前生成并存储在客户端装置处。在这种情况下,使用TD SID模型生成的TD说话者嵌入可以指示第二用户是被授权执行助理命令以接听客户端装置处传入的电话呼叫的经验证的用户。此外,口头话语“不”和口头话语“接听”的多个TI说话者嵌入可以指示不相同用户提供在音频数据的流中捕获的口头话语。因此,在此示例中,系统可以使自动助理接听传入的电话呼叫。

[0095] 现在转向图6A、6B和6C,描绘了利用图5的方法500基于说话者识别来确定是否执行助理命令的各种非限制性示例。自动助理可以至少部分地在客户端装置110处实施(例如,关于图1描述的自动助理115)。自动助理可以利用暖词系统180(例如,关于图1描述的暖词系统)来确定是否基于场境音频数据来执行助理命令(例如,如关于图5的方法500所描述)。

[0096] 类似于图4A至4C中所描绘的客户端装置,图6A至6C中所描绘的客户端装置110可以包括各种用户接口组件,包括例如用于基于口头话语和/或其他听觉输入生成音频数据的麦克风、用于听觉渲染合成语音和/或其他听觉输出的扬声器,以及用于接收触摸输入和/或视觉渲染转录和/或其他视觉输出的显示器190。此外,并且如图6B和6C所示,客户端装置110的显示器190包括各种系统接口元件191、192和193(例如,硬件和/或软件接口元件),客户端装置的用户可以与这些元件交互以使客户端装置110执行一个或多个动作。客户端装置110的显示器190使得用户能够通过触摸输入(例如,通过将用户输入引导到显示器190或其部分(例如,引导到文本输入框194或显示器190的其他部分))和/或通过口头输入(例如,通过选择麦克风接口元件195——或仅通过说话而不必选择麦克风接口单元195(即,自动助理可以监测一个或多个特定术语或短语、凝视姿势、嘴部运动、嘴唇运动和/或其他条件,以在客户端装置110处激活口头输入))与在显示器190上渲染的内容交互。尽管图6A-6C中描绘的客户端装置110是移动装置,但是应理解,这是为了示例的目的而并不意味着限制。

[0097] 具体参考图6A,假设与客户端装置110相关联的用户101与附加用户102共同位于环境600中。进一步假设在客户端装置110处接收到电话呼叫。在此示例中,自动助理可以确定在客户端装置110处接收到的电话呼叫对应于暖词激活事件。因此,自动助理可以激活利用与电话呼叫事件相关联的一个或多个暖词模型的一个或多个当前休眠的自动助理功能。可以使用与电话呼叫事件相关联的一个或多个暖词模型处理由客户端装置110的麦克风生成的音频数据的流,以监测与电话呼叫事件相关联的特定单词和/或短语(例如,“接听”、“拒绝”、“屏幕呼叫”、“发送到话音邮件”等)的出现,当检测到特定单词和/或短语时,使自动助理执行与特定单词和/或短语相关联的特定动作。

[0098] 例如,如图6A所示,假设当客户端装置110正在振铃时,附加用户102提供“接听”的

口头话语652A。值得注意的是，与图4A相比，“接听”的口头话语652A是由附加用户102而不是与客户端装置110相关联的用户101提供的。在此示例中，在使用与电话呼叫事件相关联的一个或多个暖词模型处理由客户端装置110的麦克风生成的音频数据的流时，自动助理可以确定音频数据对应于暖词“接听”，当检测到暖词时，使自动助理接听电话呼叫。进一步假设需要验证提供口头话语652A的用户(例如，附加用户102)的身份，以使自动助理接听电话。

[0099] 在此示例中，自动助理可以确定口头话语652A的长度是否足以执行TI SID，以验证提供口头话语的用户的身份。在此示例中，进一步假设口头话语652A的长度不足以执行TI SID。然而，自动助理可以尝试使用TD SID来验证提供口头话语652A的用户的身份(并且假设存在针对“接听”的一个或多个先前生成的TD说话者嵌入，而不管是否执行TI SID)。例如，自动助理可以使用TD SID模型处理与一个或多个特定单词和/或短语相对应的音频数据的部分，以生成针对“接听”的TD说话者嵌入。此外，自动助理可以在嵌入空间中将针对“接听”的TD说话者嵌入与针对“接听”的一个或多个先前生成的说话者嵌入进行比较。另外，自动助理可以以基于比较而确定的距离度量为基础来确定提供口头话语652A的用户是否是客户端装置110的用户101。在此示例中，自动助理可以确定针对“接听”的TD说话者嵌入与针对“接听”的一个或多个先前生成的说话者嵌入不匹配。因此，由于自动助理没有验证提供口头话语652A的用户(例如，附加用户102)的身份，因此自动助理可以避免执行代表用户101接听电话呼叫的助理命令。

[0100] 然而，在各种实施方式中，自动助理可以生成将被提供用于向用户101呈现(例如，视觉和/或听觉)的提示，其请求用户提供附加验证以使助理命令执行。例如，具体参考图6B，自动助理可以生成“请说[示例话语]以验证您的身份并接听来电”的提示652B，并且使提示被提供用于经由在客户端装置110处可访问的自动助理应用198A向用户进行视觉呈现。因此，客户端装置110的用户101可以响应于提示来提供附加用户输入，以验证用户101实际上是否打算让自动助理接听电话呼叫。在此示例中，自动助理可以请求用户101提供示例话语，示例话语具有足够的长度以对示例话语执行TI SID，或对应于客户端装置110可以访问一个或多个先前生成的TD说话者嵌入的另一特定术语或短语。假设客户端装置110的用户101没有提供使自动助理接听电话呼叫的口头话语(例如，在图6A的示例中)，则用户101可以忽略或不理会提示652B。然而，假设客户端装置110的用户101确实提供了使自动助理接听电话呼叫的口头话语(例如，在图4A的示例中)，则用户101可以提供示例话语。

[0101] 作为另一示例，并且具体参考图6C，自动助理可以生成“请提供面部识别或指纹识别以接听来电”的提示652C，并且使提示被提供用于经由在客户端装置110处可访问的自动助理应用198A向用户进行视觉呈现。因此，客户端装置110的用户101可以响应于提示来提供附加用户输入，以验证用户101实际上是否打算让自动助理接听电话呼叫。在此示例中，在提供口头话语的用户的身份不能被识别的情况下，自动助理可以请求用户101提供此生物特征识别。在一些情况下，可以仅响应于确定TD SID和TD SID不足以识别提供口头话语的用户而请求此生物特征信息。假设客户端装置110的用户101没有提供使自动助理接听电话呼叫的口头话语(例如，在图6A的示例中)，则用户101可以忽略或不理会提示652C。然而，假设客户端装置110的用户101确实提供了使自动助理接听电话呼叫的口头话语(例如，在图4A的示例中)，则用户101可以提供生物特征信息。

[0102] 尽管图6B和6C中描绘的提示652B和652C被提供用于经由在客户端装置110处可访问的自动助理应用198A呈现给用户101，但是应理解，这是为了示例的目的而并不意味着限制。例如，提示652B和652C可以附加地或可替代地听觉渲染以经由客户端装置110的扬声器呈现给用户、视觉渲染为通知(例如，弹出通知、横幅通知等)以经由客户端装置110的显示器190呈现给用户。此外，尽管关于使特定助理命令执行的特定暖词描述了图6A至6C，但是应理解，这是为了示例的目的而并不意味着限制。相反，应理解，关于图6A至6C描述的技术可以用于检测任何暖词激活事件，使用与暖词激活事件相关联的暖词模型检测任何暖词，并且当提供口头话语的用户的身份被验证为经授权用户时使任何助理命令执行。

[0103] 现在转向图7，描绘示例计算装置710的框图，示例计算装置710可以可选地用于执行本文描述的技术的一个或多个方面。在一些实施方式中，客户端装置、基于云的自动助理组件和/或其他组件中的一个或多个可以包括示例计算装置710的一个或多个组件。

[0104] 计算装置710通常包括经由总线子系统712与多个外围装置通信的至少一个处理器714。这些外围装置可以包括存储子系统724，存储子系统724包括例如存储器子系统725和文件存储子系统726、用户接口输出装置720、用户接口输入装置722和网络接口子系统716。输入装置和输出装置允许用户与计算装置710交互。网络接口子系统716提供到外部网络的接口，并且耦合到其他计算装置中的对应接口装置。

[0105] 用户接口输入装置722可以包括键盘，诸如鼠标、轨迹球、触摸板或图形输入板的指向装置，扫描仪，并入显示器中的触摸屏，诸如话音辨识系统的音频输入装置，麦克风和/或其他类型的输入装置。通常，术语“输入装置”的使用旨在包括将信息输入到计算装置710或通信网络上的所有可能类型的装置和方式。

[0106] 用户接口输出装置720可以包括显示子系统、打印机、传真机或诸如音频输出装置的非可视显示器。显示子系统可以包括阴极射线管(CRT)、诸如液晶显示器(LCD)的平板装置、投影装置或用于创建可见图像的一些其他机构。显示子系统还可以诸如经由音频输出装置提供非视觉显示。通常，术语“输出装置”的使用旨在包括将信息从计算装置710输出到用户或另一机器或计算装置的所有可能类型的装置和方式。

[0107] 存储子系统724存储提供本文描述的一些或所有模块的功能的编程和数据结构。例如，存储子系统724可以包括执行本文公开的方法的选定方面以及实施图1和2中所示的各种组件的逻辑。

[0108] 这些软件模块通常由处理器714单独实行或与其他处理器组合执行。存储子系统724中使用的存储器725可以包括多个存储器，包括用于在程序实行期间存储指令和数据的主随机存取存储器(RAM)730和存储固定指令的只读存储器(ROM)732。文件存储子系统726可以为程序和数据文件提供持久性存储，并且可以包括硬盘驱动器、软盘驱动器以及相关联的可移动媒体、CD-ROM驱动器、光学驱动器或可移动媒体盒。实施某些实施方式的功能的模块可以由文件存储子系统726存储在存储子系统724中，和/或存储在处理器714可访问的其他机器中。

[0109] 总线子系统712提供用于使计算装置710的各种组件和子系统按预期彼此通信的机制。尽管总线子系统712被示意性地示出为单个总线，但是总线子系统712的替代实施方式可以使用多个总线。

[0110] 计算装置710可以具有各种类型，包括工作站、服务器、计算集群、刀片服务器、服

务器场或任何其他数据处理系统或计算装置。由于计算机和网络的不断变化的性质，图7中描绘的计算装置710的描述仅旨在作为用于示出一些实施方式的特定示例。计算装置710的许多其他配置可能具有比图7中所描绘的计算装置更多或更少的组件。

[0111] 在本文描述的系统收集或以其他方式监测关于用户的个人信息或可以使用个人信息和/或所监测信息的情况下，可以为用户提供机会来控制程序或特征是否收集用户信息(例如，关于用户的社交网络、社交行为或活动、职业、用户爱好或用户的当前地理位置的信息)，或者控制是否和/或如何从内容服务器接收可能与用户更相关的内容。此外，在存储或使用某些数据之前，可以一种或多种方式对数据进行处理，从而删除个人可识别信息。例如，可以对用户的身份进行处理，使得不能针对用户确定个人可识别信息，或可以在获得地理位置信息的地方概括用户的地理位置(诸如城市、邮政编码或州级别)，使得不能确定用户的特定地理位置。因此，用户可以控制如何收集关于用户的信息和/或使用信息。

[0112] 在一些实施方式中，提供一种由一个或多个处理器实施的方法，并且包括：使用暖词模型处理音频数据的流以监测一个或多个特定单词或短语的出现，音频数据的流由用户的客户端装置的一个或多个麦克风生成，并且一个或多个特定单词或短语中的每一个与助理命令相关联；响应于确定音频数据的部分对应于特定单词或短语中的一个或多个：使用自动语音辨识(ASR)模型处理音频数据的前序部分和/或音频数据的后序部分以生成ASR输出；以及基于处理ASR输出来确定用户是否打算一个或多个特定单词或短语使助理命令执行。音频数据的前序部分在对应于一个或多个特定单词或短语的音频数据的部分之前，并且音频数据的后序部分在对应于一个或多个特定单词或短语的音频数据的部分之后。方法进一步包括：响应于确定用户不打算一个或多个特定单词或短语使与特定单词或短语中的一个或多个相关联的助理命令执行：避免使自动助理执行与特定单词或短语中的一个或多个相关联的助理；以及响应于确定用户打算一个或多个特定单词或短语使与特定单词或短语中的一个或多个相关联的助理命令执行：使自动助理执行与特定单词或短语中的一个或多个相关联的助理命令。

[0113] 本文公开的技术的这些和其他实施方式可以可选地包括以下特征中的一个或多个。

[0114] 在一些实施方式中，方法可以进一步包括检测暖词激活事件的出现；以及响应于检测到暖词激活事件的出现，激活利用暖词模型的一个或多个当前休眠的自动助理功能。使用暖词模型处理音频数据的流以监测一个或多个特定单词或短语的出现可以是响应于激活利用暖词模型的一个或多个当前休眠的自动助理功能。在这些实施方式的一些版本中，暖词激活事件可以包括以下中的一个或多个：在客户端装置处接收到的电话呼叫、在客户端装置处接收到的文本消息、在客户端装置处接收到的电子邮件、在客户端装置处响起的警报、在客户端装置处响起的定时器、在客户端装置或客户端装置的环境中的附加客户端装置处播放的媒体、在客户端装置处接收到的通知、客户端装置的位置，或在客户端装置处可访问的软件应用。

[0115] 在一些实施方式中，基于处理ASR输出来确定用户是否打算一个或多个特定单词或短语使与一个或多个特定单词或短语相关联的助理命令执行包括：使用自然语言理解(NLU)模型处理ASR输出以生成NLU输出；并且基于NLU输出来确定用户是否打算一个或多个特定单词或短语使助理命令执行。ASR输出是基于音频数据的前序部分而不是音频数据的

后序部分生成的。在这些实施方式的一些版本中，方法可以进一步包括：响应于确定NLU输出不足以确定用户打算一个或多个特定单词或短语使与特定单词或短语中的一个或多个相关联的助理命令执行：使用ASR模型处理音频数据的后序部分以生成附加ASR输出；并且基于处理附加ASR输出，确定用户是否打算一个或多个特定单词或短语使与特定单词或短语中的一个或多个相关联的助理命令执行。

[0116] 在一些实施方式中，基于处理ASR输出来确定用户是否打算一个或多个特定单词或短语使与特定单词或短语中的一个或多个相关联的助理命令执行可以包括：使用自然语言理解(NLU)模型处理ASR输出以生成NLU输出；并且基于NLU输出来确定用户是否打算一个或多个特定单词或短语使助理命令执行。ASR输出是基于音频数据的前序部分和音频数据的后序部分两者生成的。在这些实施方式的一些版本中，方法可以进一步包括：响应于确定NLU输出不足以确定用户打算一个或多个特定单词或短语使与特定单词或短语中的一个或多个相关联的助理命令执行：使用ASR模型处理音频数据的附加后序部分以生成附加ASR输出；并且基于处理附加ASR输出，确定用户是否打算一个或多个特定单词或短语使与特定单词或短语中的一个或多个相关联的助理命令执行。音频数据的附加后序部分可以在音频数据的后序部分之后。

[0117] 在这些实施方式的一些版本中，方法可以进一步包括：使用端点模型处理音频数据的流，以生成在音频数据的流中捕获的并且包括一个或多个特定单词或短语的口头话语的多个时间戳。在这些实施方式的一些版本中，多个时间戳可以至少包括：与当用户开始提供口头话语时的第一时间相关联的第一时间戳、与第一时间之后的当用户开始提供口头话语中包括的一个或多个特定单词或短语时的第二时间相关联的第二时间戳、与第二时间之后的当用户完成提供口头话语中包括的一个或多个特定单词或短语时的第三时间相关联的第三时间戳，以及与第三时间之后的当用户完成提供口头话语时的第四时间相关联的第四时间戳。在这些实施方式的一些进一步版本中，音频数据的前序部分可以包括与第一时间戳和第二时间戳之间的口头话语相对应的任何音频数据。在这些实施方式的另外的或可替代的版本中，音频数据的后序部分可以包括与第三时间戳和第四时间戳之间的口头话语相对应的任何音频数据。

[0118] 在一些实施方式中，方法可以进一步包括：响应于确定口头话语包括特定单词或短语中的一个或多个，激活利用ASR模型的一个或多个当前休眠的自动助理功能。

[0119] 在一些实施方式中，方法可以进一步包括：结合音频数据的前序部分和/或后序部分音频数据，使用ASR模型处理与一个或多个特定单词或短语相对应的音频数据的部分，以生成ASR输出。

[0120] 在一些实施方式中，方法可以进一步包括：使用话音活动检测(VAD)模型处理音频数据的流，以在音频数据的部分对应于特定单词或短语中的一个或多个之前和/或在音频数据的部分对应于特定单词或短语中的一个或多个之后监测附加话音活动的出现；以及响应于检测到附加话音活动的出现：激活利用ASR模型的一个或多个当前休眠的自动助理功能。在这些实施方式的一些版本中，方法可以进一步包括：响应于未检测到附加话音活动的出现：使自动助理执行与特定单词或短语中的一个或多个相关联的助理命令。

[0121] 在一些实施方式中，方法可以进一步包括：使用声学模型处理音频数据以确定口头话语的一个或多个韵律。确定用户是否打算一个或多个特定单词或短语使与特定单词或

短语中的一个或多个相关联的助理命令执行可以是进一步基于口头话语的一个或多个韵律特性。

[0122] 在一些实施方式中，方法可以进一步包括：从客户端装置的一个或多个传感器接收传感器数据的实例。确定用户是否打算一个或多个特定单词或短语使与特定单词或短语中的一个或多个相关联的助理命令执行可以是进一步基于传感器数据的实例。在这些实施方式的一些版本中，传感器数据的实例可以包括以下中的一个或多个：由客户端装置的加速度计生成的加速度计数据、由客户端装置的一个或多个相机组件生成的图像数据、由客户端装置的一个或多个运动传感器生成的运动数据。

[0123] 在一些实施方式中，方法可以进一步包括：从客户端装置的音频缓冲器获得音频数据的前序部分；和/或从音频数据的流获得音频数据的后序部分。

[0124] 在一些实施方式中，提供一种由一个或多个处理器实施的方法，并且包括：使用暖词模型处理音频数据的流以监测一个或多个特定单词或短语的出现，音频数据的流由用户的客户端装置的一个或多个麦克风生成，并且一个或多个特定单词或短语中的每一个与助理命令相关联；响应于确定音频数据的部分对应于特定单词或短语中的一个或多个，并且响应于确定验证提供对应于音频数据的口头话语的用户的身份：使用说话者识别(SID)模型处理音频数据，以确定音频数据是否足以识别提供口头话语的用户；响应于确定音频数据足以识别提供口头话语的用户，并且响应于确定提供口头话语的所识别用户被授权执行与特定单词或短语中的一个或多个相关联的助理命令：使自动助理执行与特定单词或短语中的一个或多个相关联的助理命令；以及响应于确定音频数据不足以识别提供口头话语的用户：生成请求对提供口头话语的用户的身份进行附加验证的提示；以及使得提示被提供用于在客户端装置处呈现。

[0125] 本文公开的技术的这些和其他实施方式可以可选地包括以下特征中的一个或多个。

[0126] 在一些实施方式中，方法可以进一步包括：检测暖词激活事件的出现；以及响应于检测到暖词激活事件的出现，激活利用暖词模型的一个或多个当前休眠的自动助理功能。使用暖词模型处理音频数据的流以监测一个或多个特定单词或短语的出现可以是响应于激活利用暖词模型的一个或多个当前休眠的自动助理功能。在这些实施方式的一些版本中，暖词激活事件可以包括以下中的一个或多个：在客户端装置处接收到的电话呼叫、在客户端装置处接收到的文本消息、在客户端装置处接收到的电子邮件、在客户端装置处响起的警报、在客户端装置处响起的定时器、在客户端装置或客户端装置的环境中的附加客户端装置处播放的媒体、在客户端装置处接收到的通知、客户端装置的位置，或在客户端装置处可访问的软件应用。

[0127] 在一些实施方式中，方法可以进一步包括：响应于确定音频数据的部分对应于特定单词或短语中的一个或多个，并且响应于确定未验证提供对应于音频数据的口头话语的用户的身份：使自动助理执行与特定单词或短语中的一个或多个相关联的助理命令。

[0128] 在一些实施方式中，处理音频数据以确定音频数据是否足以验证提供口头话语的用户的身份包括：使用文本相关(TD)SID模型作为SID模型，处理与特定单词或短语中的一个或多个相对应的音频数据的部分，以生成TD说话者嵌入。确定音频数据是否足以识别提供口头话语的用户可以包括：在嵌入空间中，将TD说话者嵌入与本地存储在客户端装置处

的一个或多个所存储的TD说话者嵌入进行比较;以及基于将TD说话者嵌入与一个或多个所存储的TD说话者嵌入进行比较,确定音频数据的部分是否足以验证提供口头话语的用户的身份。

[0129] 在这些实施方式的一些版本中,使用TD SID模型处理与特定单词或短语中的一个或多个相对应的音频数据的部分以生成TD说话者嵌入可以包括:确定所存储的TD说话者嵌入中的一个或多个是否对应于特定单词或短语。使用TD SID模型处理与特定单词或短语中的一个或多个相对应的音频数据的部分以生成TD说话者嵌入是响应于确定所存储的TD说话者嵌入中的一个或多个与特定单词或短语相对应。

[0130] 在这些实施方式的另外的或可替代的版本中,生成请求对提供口头话语的用户的身份进行附加验证的提示可以包括:生成请求用户提供附加口头话语的提示,附加口头话语的长度足以执行TI SID以用于验证提供口头话语的用户的身份。

[0131] 在这些实施方式的另外的或可替代的版本中,方法可以进一步包括:响应于确定口头话语的长度足以执行TI SID:使用TI SID模型作为SID模型来处理音频数据的流以生成TI说话者嵌入。确定音频数据是否足以识别提供口头话语的用户可以包括:在嵌入空间中,将TI说话者嵌入与本地存储在客户端装置处的一个或多个所存储的TI说话者嵌入进行比较;以及基于将TI说话者嵌入与一个或多个所存储的TI说话者嵌入进行比较,确定音频数据的流是否足以验证提供口头话语的用户的身份。在这些实施方式的一些进一步版本中,生成请求对提供口头话语的用户的身份进行附加验证的提示可以包括:生成请求附加生物特征信息以验证用户的身份的提示。请求附加生物特征以验证用户的身份可以包括请求以下中的一个或多个:指纹验证或面纹验证。

[0132] 在一些实施方式中,方法可以进一步包括:接收响应于提示被提供用于在客户端装置处呈现的附加用户输入;以及响应于接收到附加用户输入:确定附加用户输入是否足以识别提供口头话语的用户。在这些实施方式的一些版本中,方法可以进一步包括:响应于确定附加用户输入足以识别提供口头话语的用户:使自动助理执行对应的助理命令。

[0133] 另外,一些实施方式包括一个或多个计算装置的一个或多个处理器(例如,中央处理单元(CPU)、图形处理单元(GPU)和/或张量处理单元(TPU)),其中,一个或多个处理器可操作以实行相关联存储器中存储的指令,并且其中,指令被配置成使得执行前述方法中的任一方法。一些实施方式还包括一个或多个非暂时性计算机可读存储介质,其存储可由一个或多个处理器实行以执行前述方法中的任一方法的计算机指令。一些实施方式还包括计算机程序产品,计算机程序产品包括可由一个或多个处理器实行以执行前述方法中的任一方法的指令。

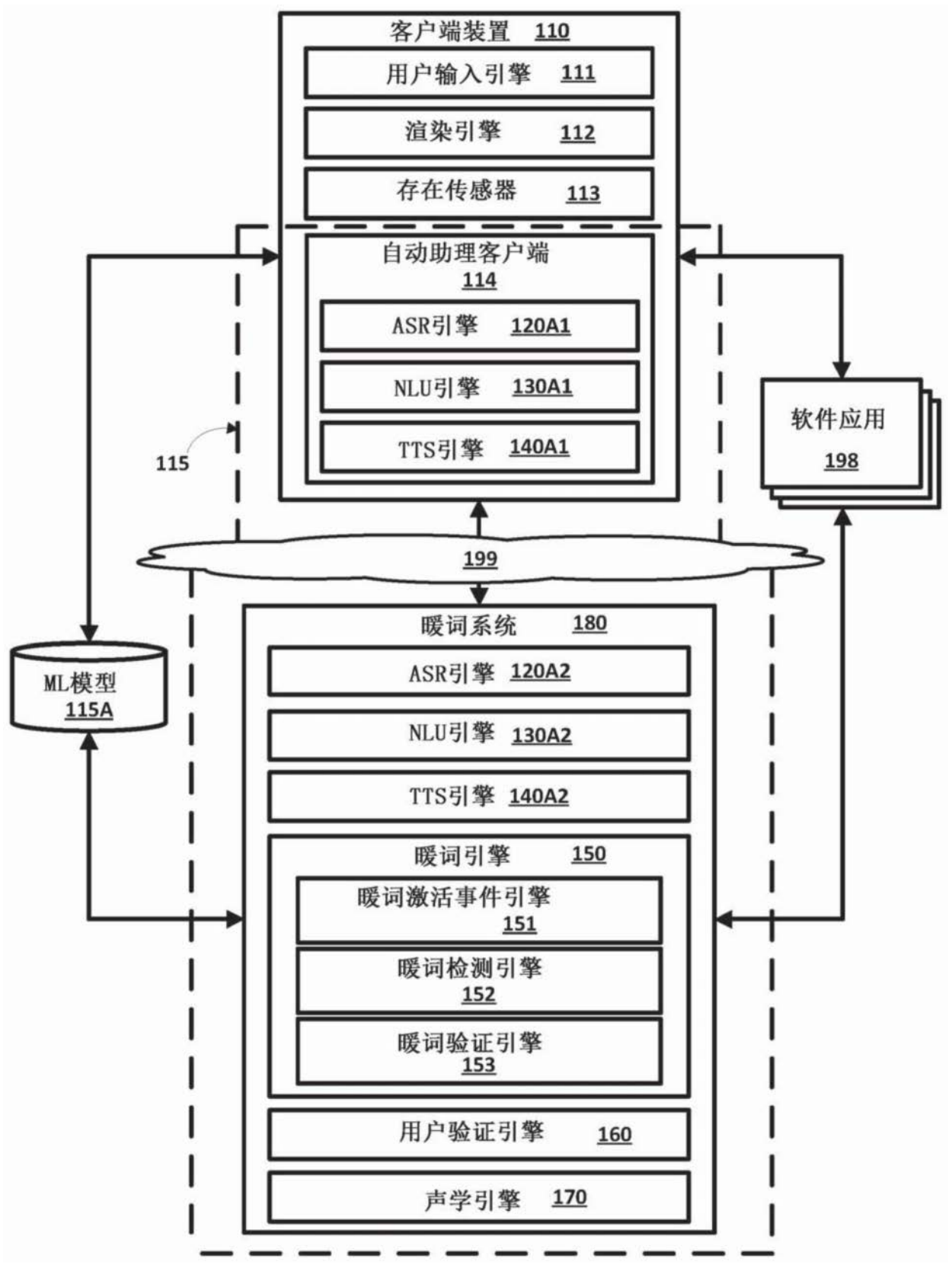

客户端装置 110
用户输入引擎 111
渲染引擎 112
存在传感器 113
自动助理客户端 114
ASR引擎 120A1
NLU引擎 130A1
TTS引擎 140A1
115
软件应用 198
199
暖词系统 180
ML模型 115A
ASR引擎 120A2
NLU引擎 130A2
TTS引擎 140A2
暖词引擎 150
暖词激活事件引擎 151
暖词检测引擎 152
暖词验证引擎 153
用户验证引擎 160
声学引擎 170

图1

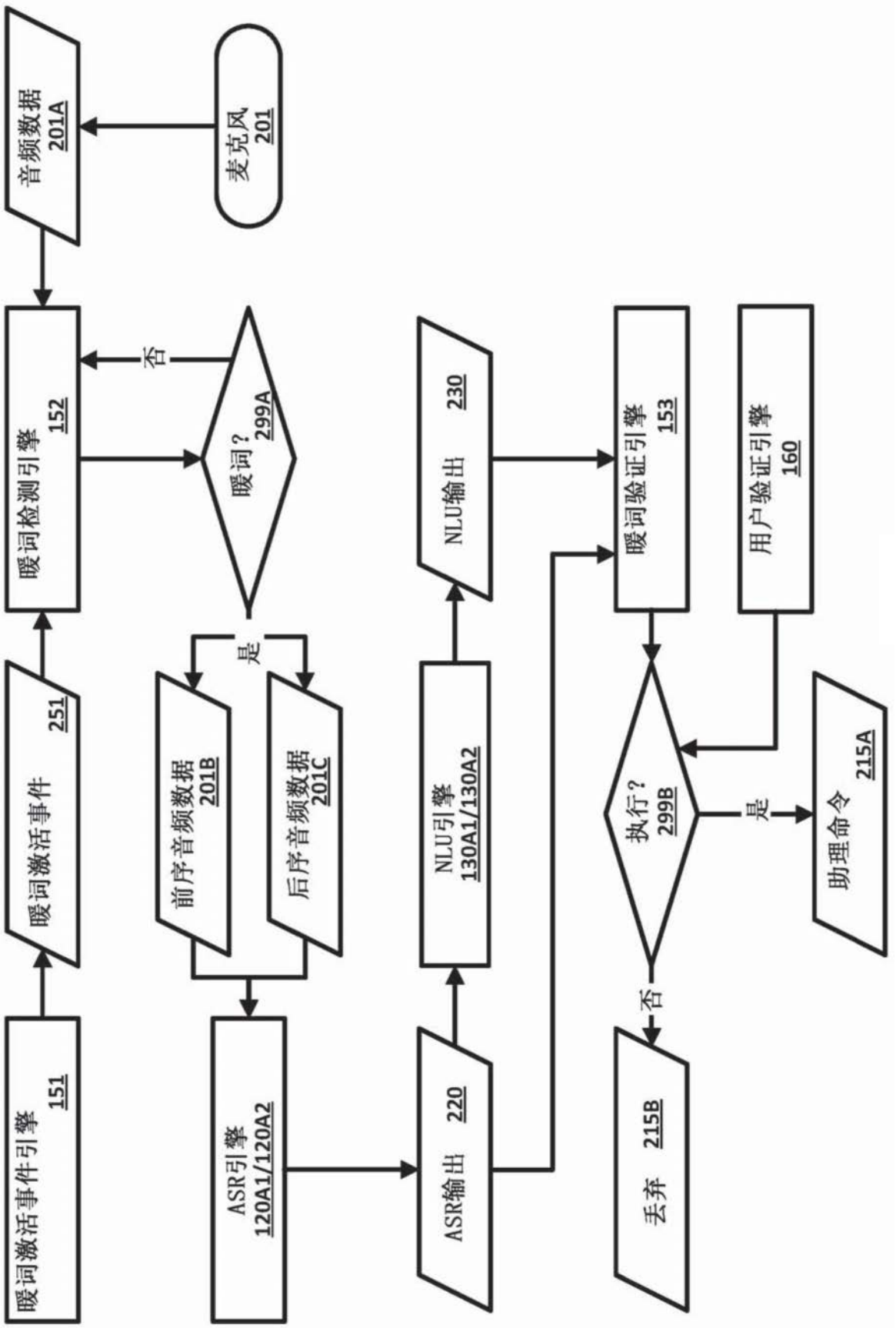

麦克风 201
音频数据 201A
暖词检测引擎 152
暖词？ 299A
否
是
暖词激活事件引擎 151
暖词激活事件 251
前序音频数据 201B
后序音频数据 201C
ASR引擎 120A1/120A2
ASR输出 220
NLU引擎 130A1/130A2
NLU输出 230
暖词验证引擎 153
用户验证引擎 160
执行？ 299B
是
否
助理命令 215A
丢弃 215B

图2

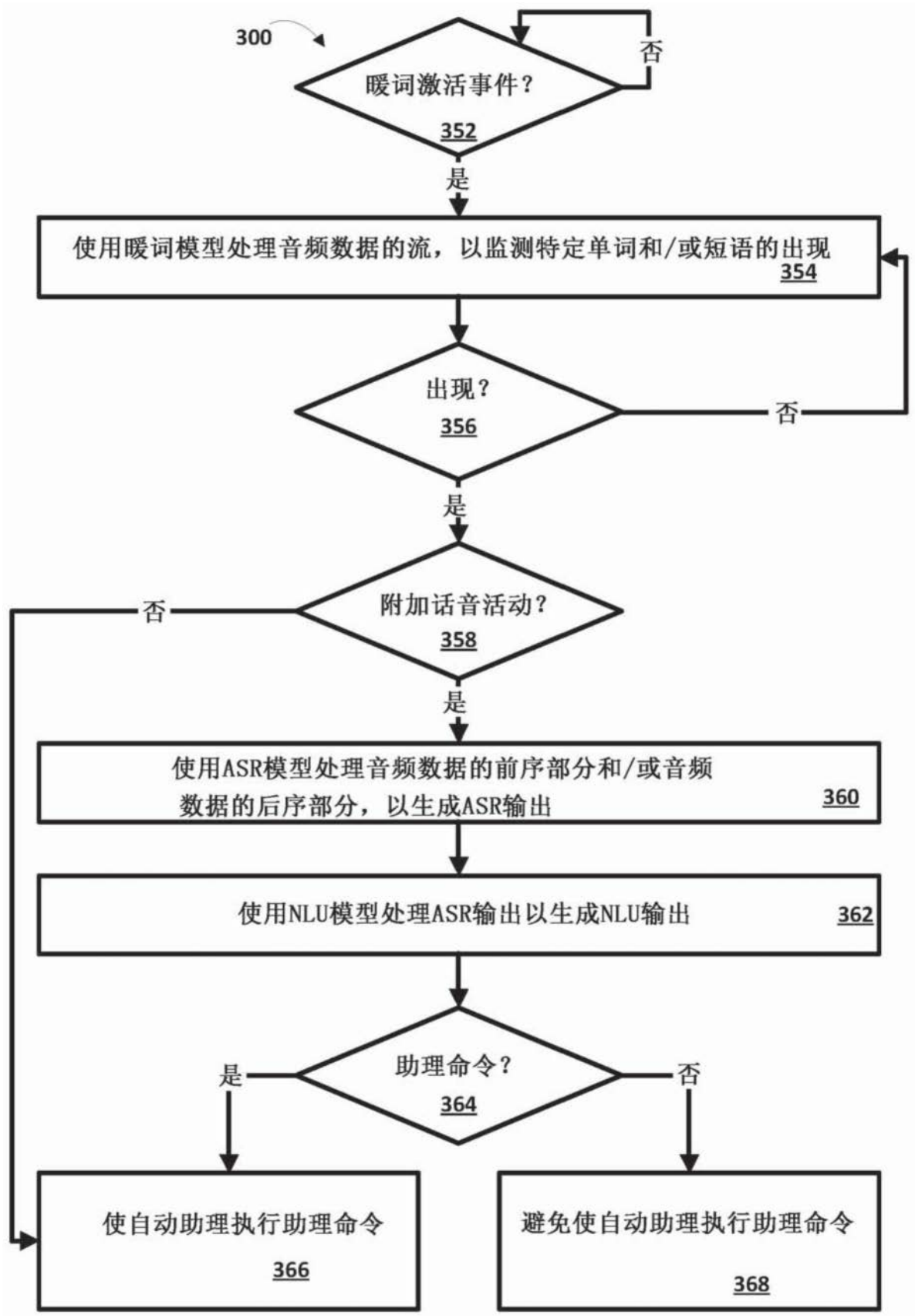

300
暖词激活事件？
352
否
是
使用暖词模型处理音频数据的流，以监测特定单词和/或短语的出现
354
出现？
356
否
是
附加话音活动？
358
否
是
使用ASR模型处理音频数据的前序部分和/或音频数据的后序部分，以生成ASR输出
360
使用NLU模型处理ASR输出以生成NLU输出
362
助理命令？
364
是
否
使自动助理执行助理命令
366
避免使自动助理执行助理命令
368

图3

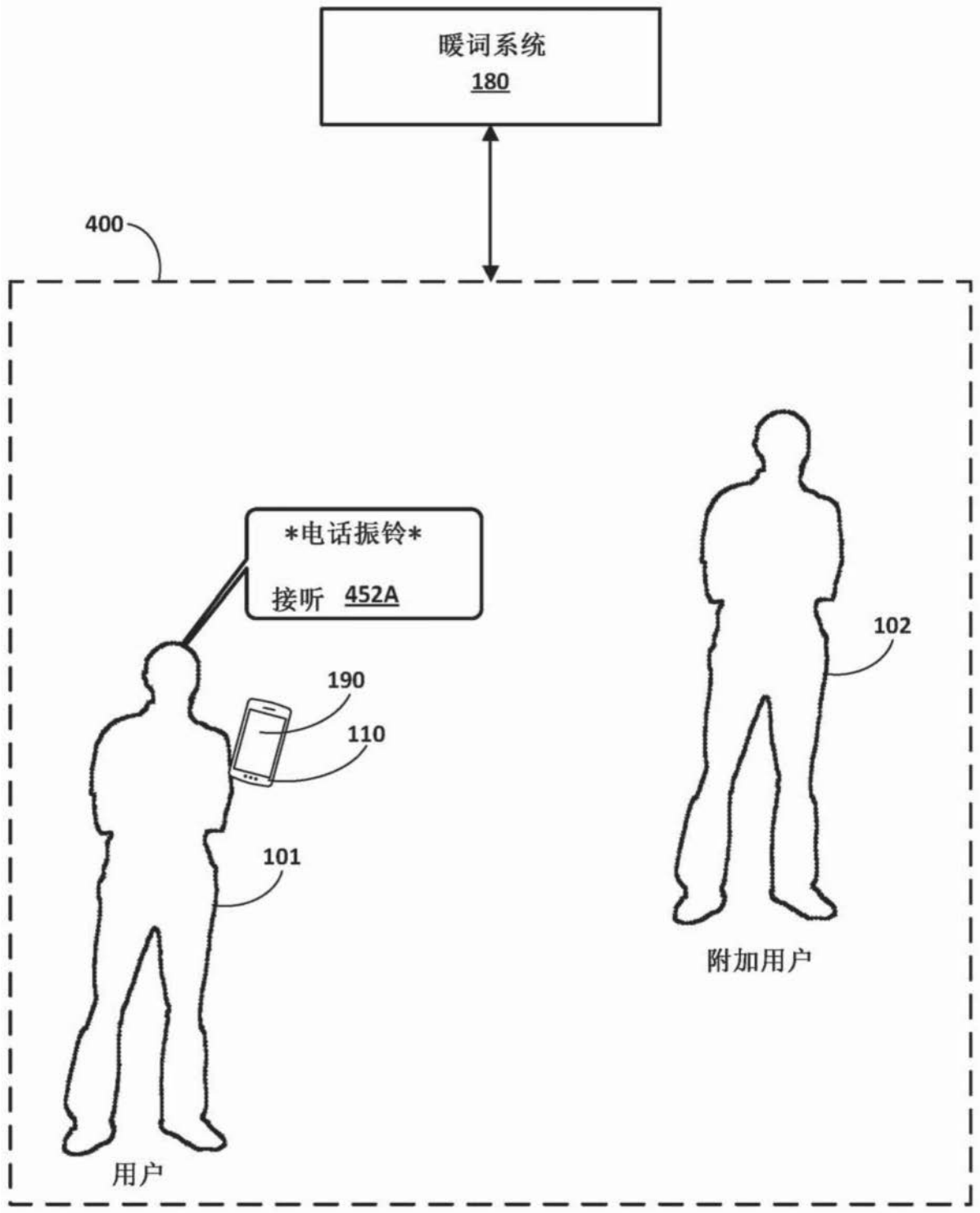



图4A

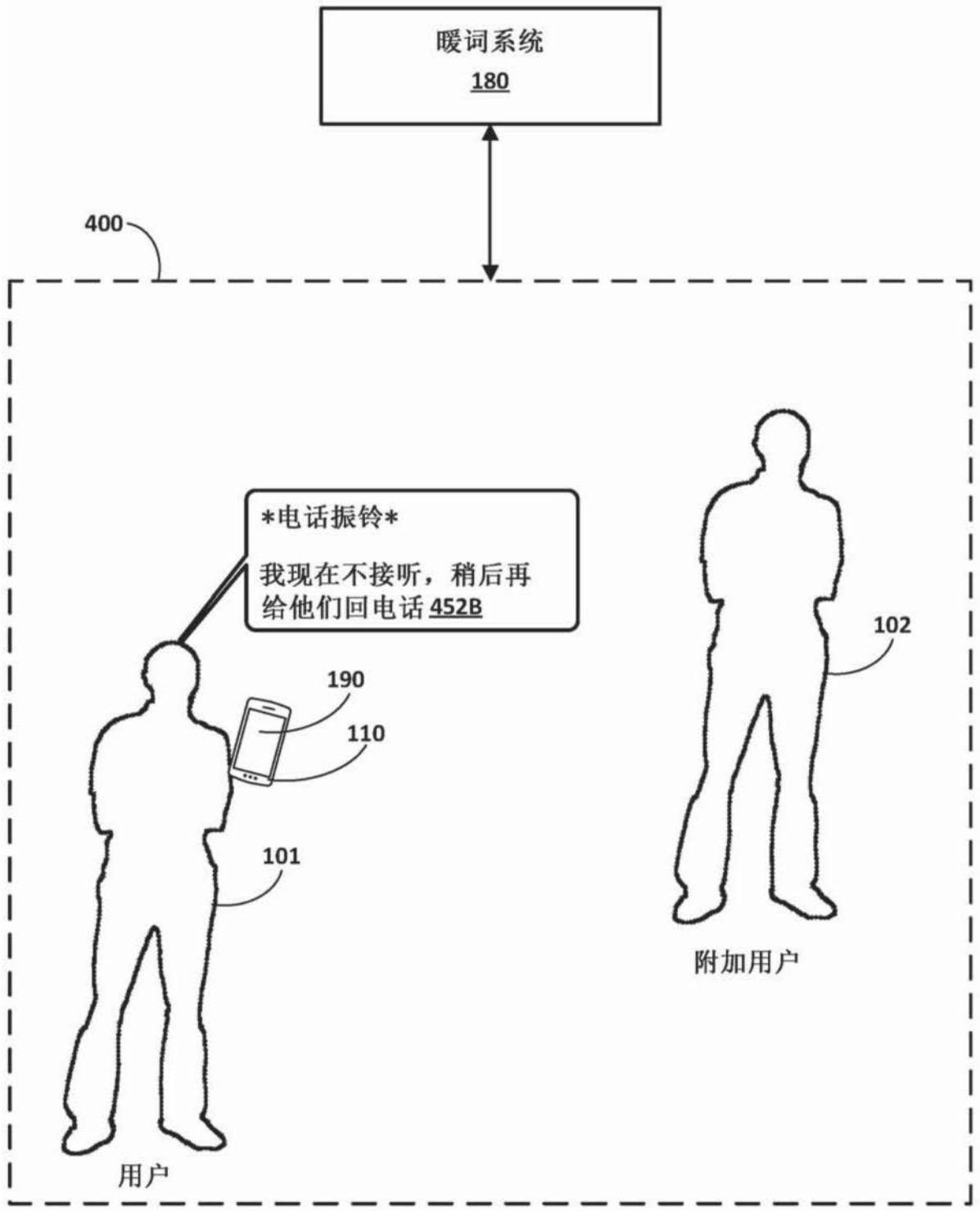

暖词系统
180
400
*电话振铃*
我现在不接听，稍后再给他们回电话 452B
190
110
101
用户
102
附加用户

图4B

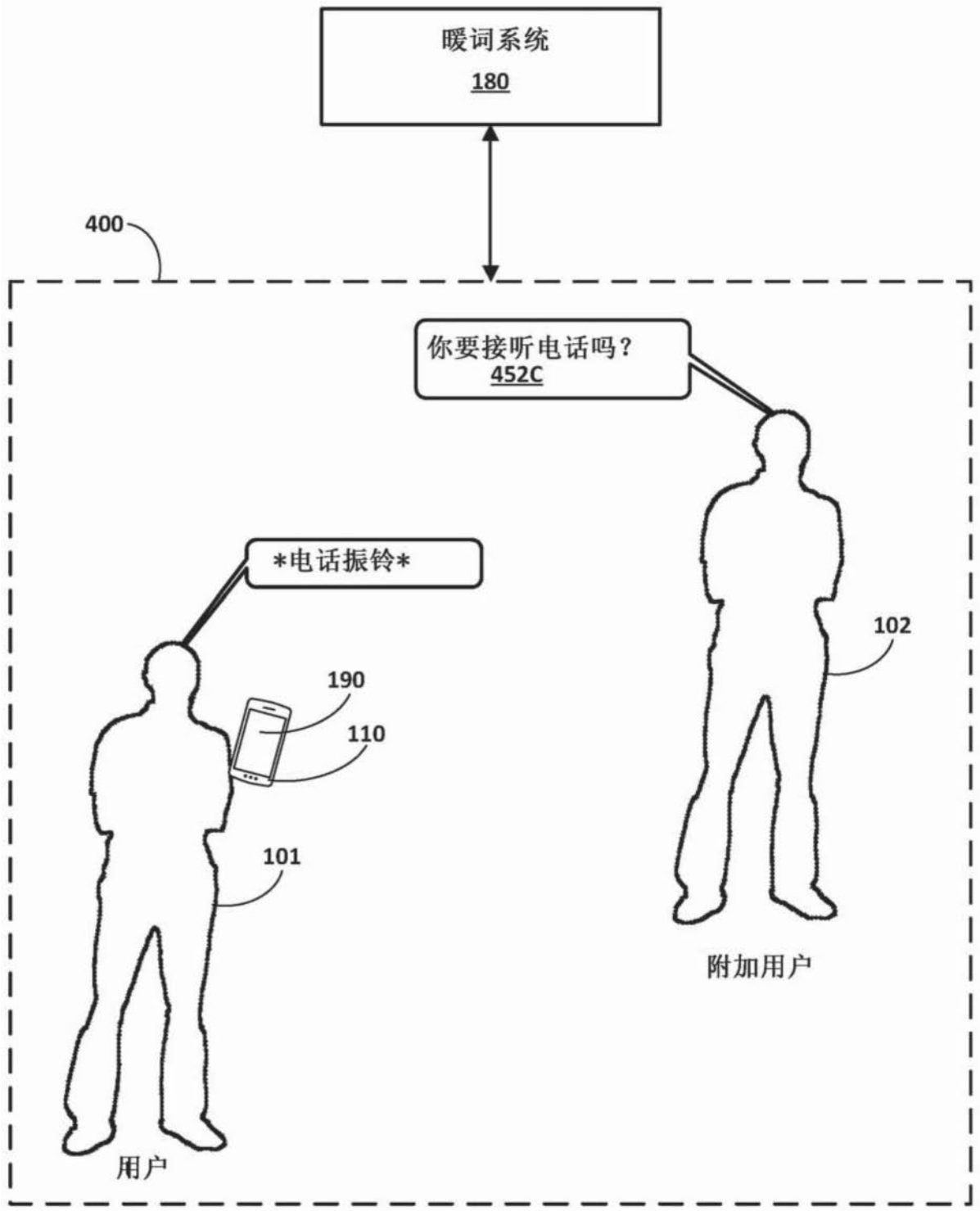



图4C

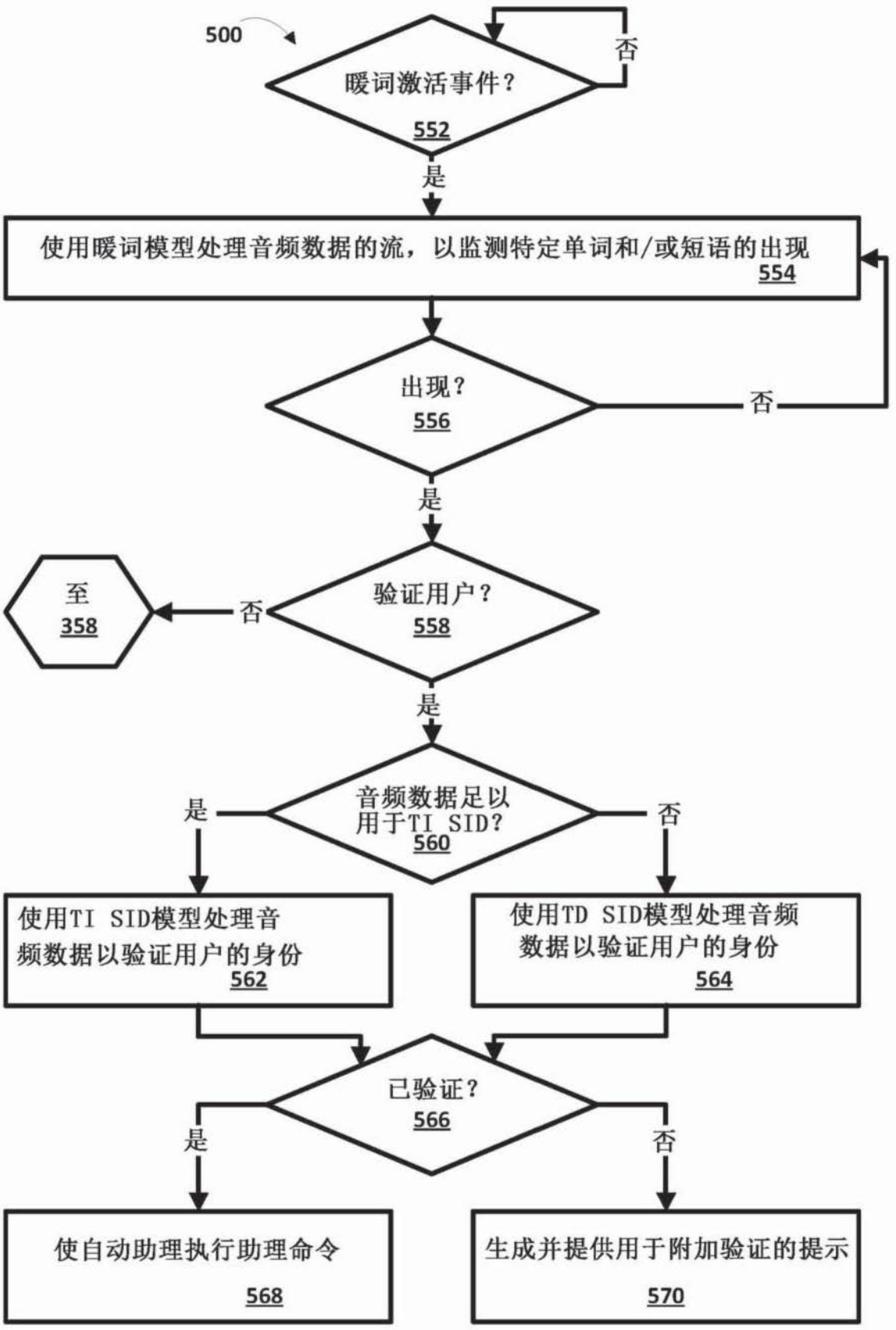

500
暖词激活事件？
552
否
是
使用暖词模型处理音频数据的流，以监测特定单词和/或短语的出现
554
出现？
556
否
是
验证用户？
558
否
至
358
是
音频数据足以用于TI SID？
560
是
否
使用TI SID模型处理音频数据以验证用户的身份
562
使用TD SID模型处理音频数据以验证用户的身份
564
已验证？
566
是
否
使自动助理执行助理命令
568
生成并提供用于附加验证的提示
570

图5

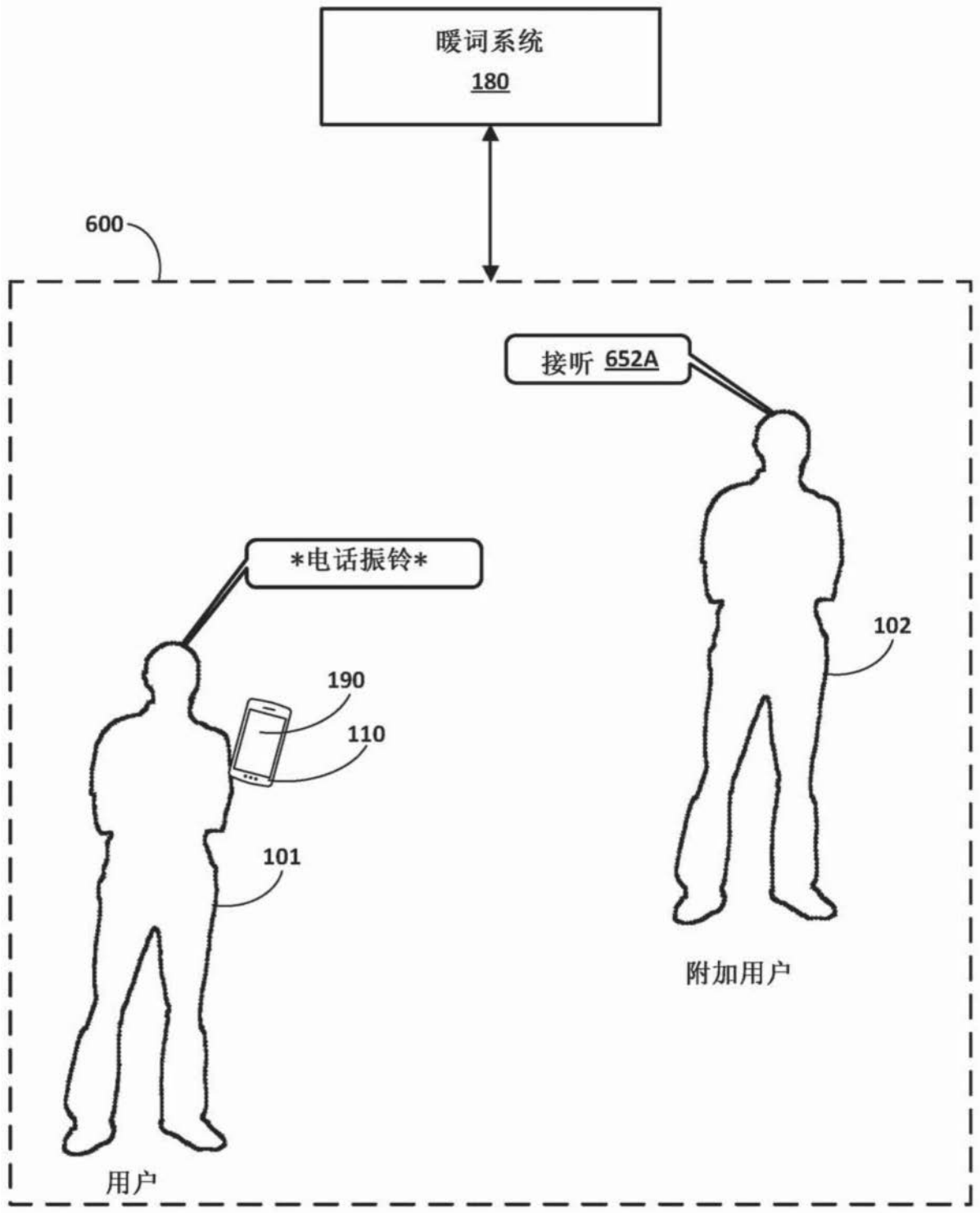



图6A

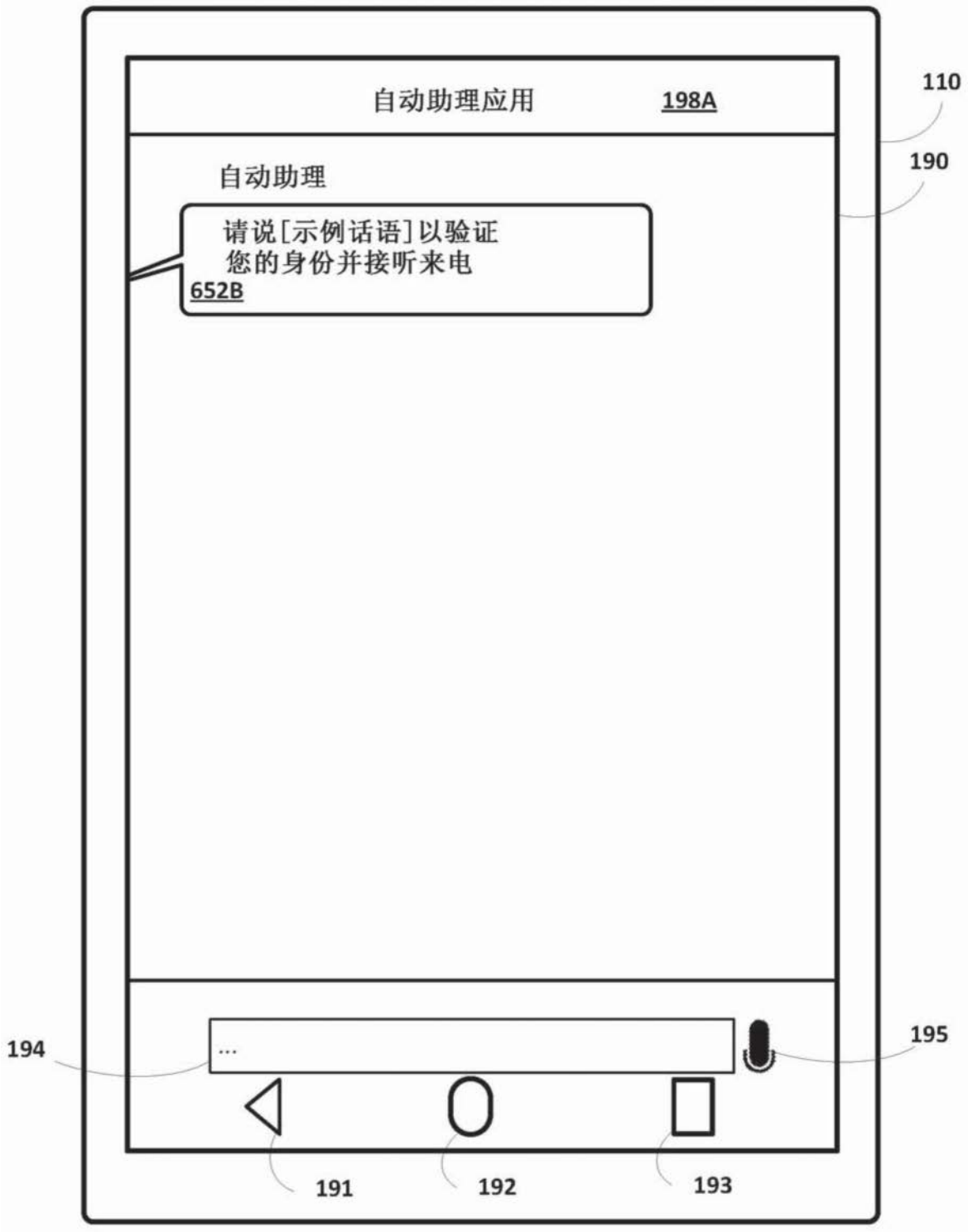

自动助理应用
198A
110
190
自动助理
请说[示例话语]以验证
您的身份并接听来电
652B
194
195
191
192
193

图6B

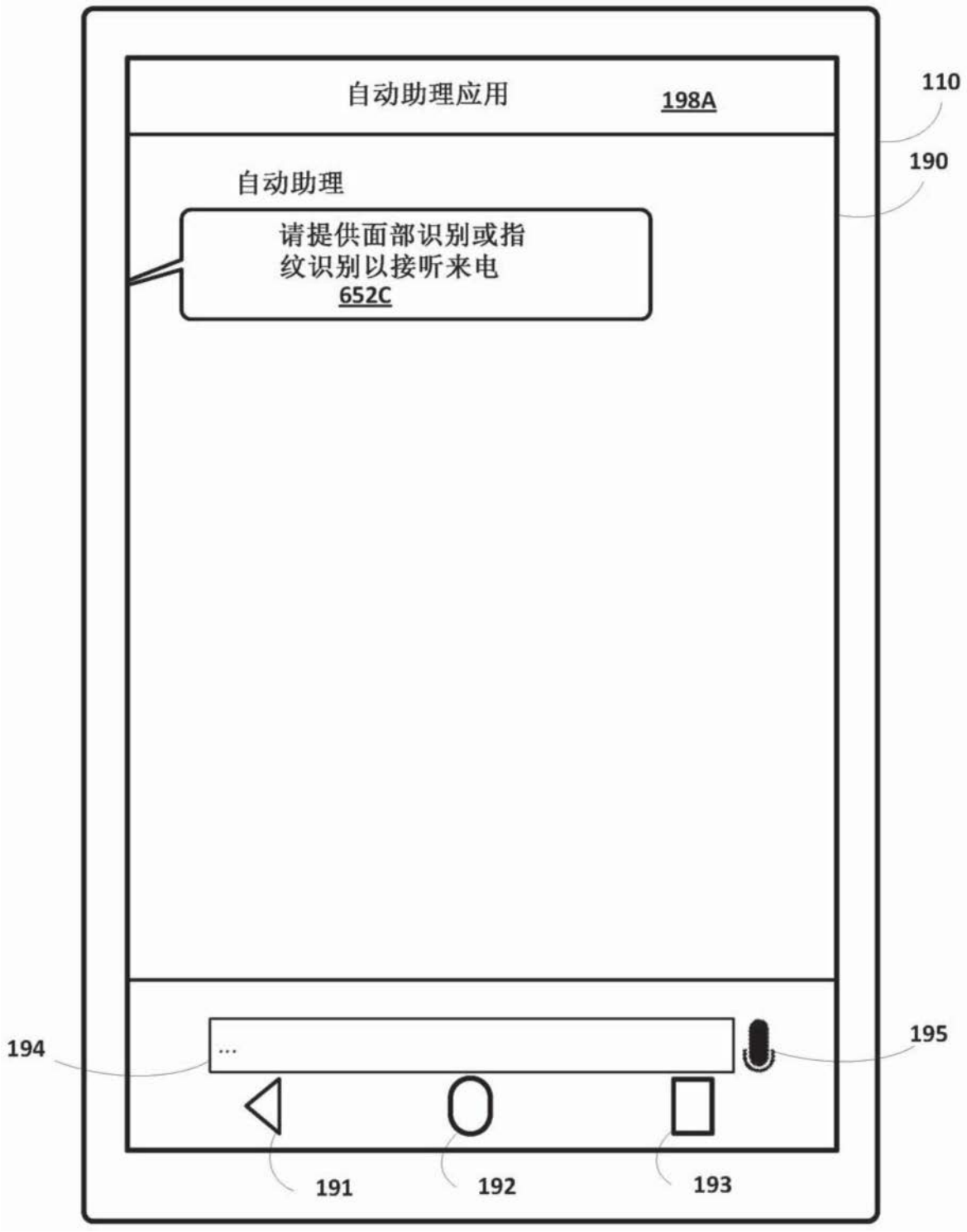

自动助理应用
198A
110
190
自动助理
请提供面部识别或指
纹识别以接听来电
652C
194
...
195
191
192
193

图6C

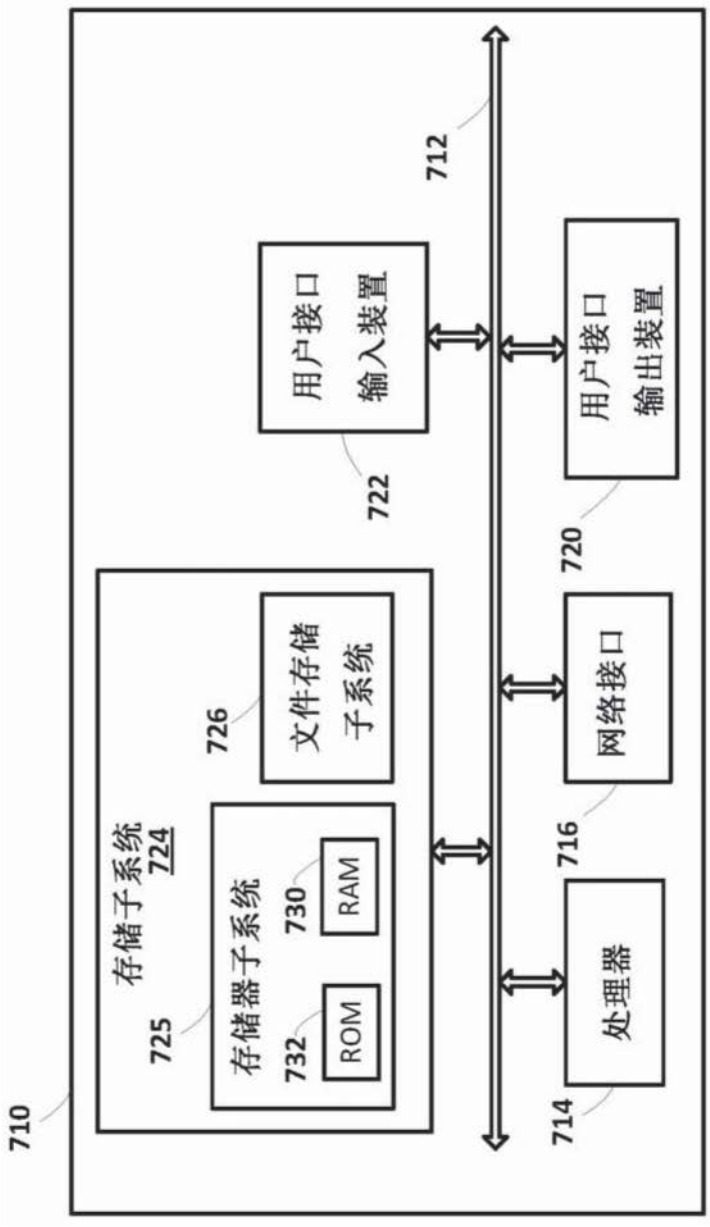

710
712
用户接口
输入装置
722
用户接口
输出装置
720
存储子系统
724
725
存储器子系统
730
RAM
732
ROM
726
文件存储
子系统
网络接口
716
处理器
714

图7