(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 102903362 B
(45)授权公告日 2015.08.19

(21)申请号 201210320230.7
(22)申请日 2012.08.31
(30)优先权数据
13/224,778 2011.09.02 US
(73)专利权人 微软技术许可有限责任公司
地址 美国华盛顿州
(72)发明人 T·M·苏摩 L·宋 M·H·金
C·R·海涅曼 D·H·霍金斯
(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 顾嘉运
(51)Int.Cl.
G10L 15/34(2013.01)
H04L 29/08(2006.01)
(56)对比文件
CN 1540625 A,2004.10.27,
US 2007/0050191 A1,2007.03.01,
US 2007/0265850 A1,2007.11.15,
US 7873349 B1,2011.01.18,
审查员 张飞弦

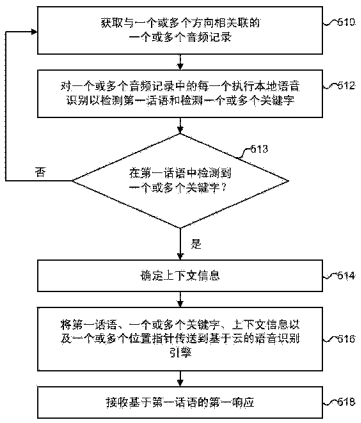
权利要求书2页 说明书18页 附图13页

(54)发明名称

集成的本地和基于云的语音识别

(57)摘要

本发明描述了集成的本地和基于云的语音识别。描述了一种用于将本地语音识别与基于云的语音识别集成以提供高效的自然用户界面的系统。在一些实施例中,计算设备确定与环境中的特定人相关联的方向,并生成与该方向相关联的音频记录。该计算设备然后对该音频记录执行本地语音识别以检测该特定人说出的第一话语并检测该第一话语中的一个或多个关键字。该第一话语可通过对音频记录应用话音活动检测技术来检测。该第一话语以及一个或多个关键字随后被传送至服务器,该服务器可标识第一话语中与该一个或多个关键字相关联的语音,并基于所标识的语音来使一种或多种语音识别技术进行适应。



CN 102903362 B

1. 一种用于执行语音识别的方法，包括：

从多个话筒获取 (502) 多个音频信号，所述多个音频信号中的每一个都与所述多个话筒中的不同话筒相关联，所述多个音频信号与第一环境相关联；

确定 (507) 所述第一环境中的一个或多个方向，所述第一环境包括一个或多个人，所述一个或多个方向中的每一个都与所述一个或多个人中的不同人相关联；

基于所述多个音频信号来生成 (508) 一个或多个音频记录，所述一个或多个音频记录中的第一音频记录通过以下操作生成：对所述多个音频信号应用音频信号处理技术以使得源自所述一个或多个方向中的第一方向的声音被放大，同时源自一个或多个其他方向的其他声音被削弱；

对所述多个音频记录中的每一个执行 (512) 本地语音识别，所述执行本地语音识别包括检测第一话语以及检测所述第一话语中的一个或多个关键字，所述第一话语通过对所述一个或多个音频记录中的第一音频记录应用一种或多种语音检测技术来检测；

将所述第一话语以及所述一个或多个关键字传送 (516) 到第二计算设备，所述第二计算设备对所述第一话语执行语音识别技术，所述语音识别技术检测所述第一话语中的一个或多个单词，所述第二计算设备是联网计算环境中的应用服务器；以及

从所述第二计算设备接收 (518) 基于所述第一话语的第一响应。

2. 如权利要求 1 所述的方法，其特征在于：

所述第二计算设备标识与在所述第一话语中发音的一个或多个关键字相关联的一个或多个语音，所述第二计算设备基于所述一个或多个语音来使所述语音识别技术进行适应；以及

所述第一响应包括与所述第一话语中由所述第二计算设备检测到的一个或多个单词相关联的文本消息。

3. 如权利要求 1 所述的方法，其特征在于：

所述传送所述第一话语以及所述一个或多个关键字包括传送与所述第一话语相关联的音频文件以及将与所述一个或多个关键字相关联的文本信息传送到所述第二计算设备；以及

所述第一话语通过对所述第一音频记录应用一个或多个话音活动检测技术来检测。

4. 如权利要求 1 所述的方法，其特征在于，还包括：

将与所述一个或多个关键字相关联的一个或多个位置指针传送到所述第二计算设备，所述检测所述第一话语中的一个或多个关键字包括确定所述第一话语中的所述一个或多个位置指针。

5. 如权利要求 1 所述的方法，其特征在于，还包括：

在所述确定一个或多个方向之前执行对所述多个音频信号的回声抵消。

6. 如权利要求 1 所述的方法，其特征在于：

所述确定一个或多个方向包括执行声源定位，所述执行声源定位包括确定与所述一个或多个人中的每一个相关联的角度和置信度。

7. 如权利要求 1 所述的方法，其特征在于，还包括：

获取所述第一环境中的一个或多个图像，所述多个音频信号在第一时间段期间与所述第一环境相关联，所述一个或多个图像在所述第一时间段期间与所述第一环境相关联，所

述一个或多个图像包括一个或多个深度图像，所述确定一个或多个方向包括基于所述一个或多个人中的每一个的一个或多个图像来执行骨架跟踪。

8. 一种用于集成本地和基于云的语音识别的电子设备，包括：

包括多个话筒的捕捉设备 (20)，所述捕捉设备从所述多个话筒获取一个或多个声音，所述一个或多个声音与第一环境相关联；以及

一个或多个处理器 (42)，所述一个或多个处理器确定所述第一环境中的一个或多个方向，所述第一环境包括一个或多个人，所述一个或多个方向中的每一个都与所述一个或多个人中的不同人相关联，所述一个或多个处理器基于所述一个或多个声音来生成一个或多个音频记录，所述一个或多个音频记录中的每一个都与所述一个或多个方向中的不同方向相关联，所述一个或多个处理器通过对第一音频记录应用一种或多种语音检测技术来检测所述一个或多个音频记录中的第一音频记录中的第一话语，所述一个或多个处理器检测所述第一话语中的一个或多个关键字，所述一个或多个处理器将所述第一话语以及所述一个或多个关键字传送到第二计算设备，所述第二计算设备基于所述一个或多个关键字来对所述第一话语执行语音识别技术，所述语音识别技术检测所述第一话语中的一个或多个单词，所述一个或多个处理器从所述第二计算设备接收基于所述第一话语的第一响应，所述第二计算设备是联网计算环境中的应用服务器。

9. 如权利要求 8 所述的电子设备，其特征在于：

所述第二计算设备标识与在所述第一话语中发音的一个或多个关键字相关联的一个或多个语音，所述第二计算设备基于所述一个或多个语音来使所述语音识别技术进行适应。

10. 如权利要求 8-9 中的任一项所述的电子设备，其特征在于：

所述一个或多个处理器标识正在所述电子设备上执行的特定应用，所述一个或多个处理器将与所述特定应用相关联的标识信息传送到所述第二计算设备，所述第二计算设备基于所述标识信息以及在所述第一话语中检测到的所述一个或多个单词来执行因特网搜索；以及

所述第一响应包括基于所述标识信息和所述第一话语的因特网搜索结果。

集成的本地和基于云的语音识别

技术领域

[0001] 本发明涉及语音识别，尤其涉及本地和基于云的语音识别的集成。

背景技术

[0002] 语音识别技术可用于将说出的单词或词组转换成文本。基于统计数据的语音识别技术通常利用声学建模和 / 或语言建模。声学模型可通过以下操作来创建 :取得各种语言音频记录(例如，各种单词或词组)，将各种音频记录与文本转录相关联，然后创建构成各种单词或词组中的每一个的声音的统计表示。语言模型尝试捕捉特定语言的特性并预测语音序列中的下一个单词或词组。语音模型可包括特定语言中的单词频率和 / 或单词序列的概率。

发明内容

[0003] 描述了用于将本地语音识别与基于云的语音识别集成以提供高效的自然用户界面的技术。在一些实施例中，计算设备确定与环境中的特定人相关联的方向并生成与该方向相关联的音频记录，其中源自该方向的声音被放大，而源自其他方向的声音被抑制。该计算设备然后对该音频记录执行本地语音识别以检测该特定人说出的第一话语并检测该第一话语中的一个或多个关键字。该第一话语可通过对音频记录应用话音活动检测技术来检测。该第一话语以及一个或多个关键字随后被传送至服务器，该服务器可标识第一话语中与该一个或多个关键字相关联的语音，并基于所标识的语音来使一种或多种语音识别技术进行适应。

[0004] 一个实施例包括从与第一环境相关联的多个话筒获取一个或多个声音，确定该第一环境中与一个或多个人相关联的一个或多个方向，以及基于该一个或多个声音来生成一个或多个音频记录，其中该一个或多个音频记录中的每一个都与该一个或多个方向中的不同方向相关联。该方法还包括对该一个或多个音频记录中的每一个执行本地语音识别，包括检测第一话语和检测该第一话语中的一个或多个关键字。该方法还包括将该第一话语以及该一个或多个关键字传送到第二计算设备以及从该第二计算设备接收基于该第一话语的第一响应。

[0005] 一个实施例包括捕捉设备以及一个或多个处理器。该捕捉设备包括多个话筒。该捕捉设备从与第一环境相关联的多个话筒获取一个或多个声音。该一个或多个处理器确定该第一环境中与一个或多个人相关联的一个或多个方向。该一个或多个处理器基于该一个或多个声音来生成一个或多个音频记录，其中该一个或多个音频记录中的每一个都与该一个或多个方向中的不同方向相关联。该一个或多个处理器检测该一个或多个音频记录中的第一音频记录中的第一话语，并检测该第一话语中的一个或多个关键字。该一个或多个处理器将该第一话语以及该一个或多个关键字传送到第二计算设备，该第二计算设备基于该一个或多个关键字来检测该第一话语中的一个或多个单词。该一个或多个处理器从该第二计算设备接收基于该第一话语的第一响应。

[0006] 提供本发明内容以便以简化形式介绍将在以下具体实施例中进一步描述的一些概念。本发明内容并非旨在标识所要求保护的主题的关键特征或必要特征，也不旨在用于帮助确定所要求保护的主题的范围。

[0007] 附图简述

[0008] 图 1 是可在其中实现所公开的技术的联网计算环境的一个实施例的框图。

[0009] 图 2 描绘了目标检测和跟踪系统的一个实施例，用户正在玩拳击游戏。

[0010] 图 3 描绘了目标检测和跟踪系统以及与该目标检测和跟踪系统相关联的捕捉设备的视野内的环境的一个实施例。

[0011] 图 4 示出了包括捕捉设备和计算环境的计算系统的一个实施例。

[0012] 图 5A 是描述集成本地和基于云的语音识别的过程的一个实施例的流程图。

[0013] 图 5B 是描述获取一个或多个音频记录的过程的一个实施例的流程图。

[0014] 图 5C 是描述获取一个或多个音频记录的过程的一个实施例的流程图。

[0015] 图 5D 是描述获取一个或多个音频记录的过程的一个实施例的流程图。

[0016] 图 5E 是描述生成一个或多个音频记录的过程的一个实施例的流程图。

[0017] 图 6 是描述执行基于云的语音识别的过程的一个实施例的流程图。

[0018] 图 7 是描述执行本地语音识别的过程的一个实施例的流程图。

[0019] 图 8 是游戏和媒体系统的实施例的框图。

[0020] 图 9 是计算系统环境的实施例的框图。

具体实施例

[0021] 描述了用于将本地语音识别与基于云的语音识别集成以提供高效的自然用户界面的技术。在一些实施例中，计算设备确定与环境中的特定人相关联的方向并生成与该方向相关联的音频记录，其中源自该方向的声音被放大，而源自其他方向的声音被抑制。该计算设备然后对该音频记录执行本地语音识别以检测该特定人说出的第一话语并检测该第一话语中的一个或多个关键字。该第一话语可通过对音频记录应用语音活动检测技术来检测。该第一话语以及一个或多个关键字随后被传送至服务器，该服务器可标识第一话语中与该一个或多个关键字相关联的语音，并基于所标识的语音来使一种或多种语音识别技术进行适应。

[0022] 将本地语音识别与基于云的语音识别集成有若干好处。一个好处是利用云计算的更强大的处理能力和存储容量。例如，与通常受限于能力和 / 或形状因子约束的本地语音识别相比，基于云的语音识别可利用大规模机器学习和更大的声学模型。在基于云的语音识别之前执行本地语音识别还节省网络带宽，因为本地语音识别器可过滤发送到云以供处理的不必要或非预期请求。例如，本地语音识别器可以只在检测到包括在本地关键字文件（即，预先确定的词典）中的一个或多个关键字时才启动基于云的语音识别器。基于云的语音识别的另一个好处是能够利用具有最新语法（例如，与特定电视节目的最新一集相关联的已更新的关键字）的大型动态目录。

[0023] 图 1 是可在其中实现所公开的技术的联网计算环境 100 的一个实施例的框图。联网计算环境 100 包括多个计算设备，所述计算设备通过一个或多个网络 180 互连。所述一个或多个网络 180 允许特定计算设备连接到另一计算设备以及与其通信。所描绘的计算设

备包括计算环境12、移动设备11、计算机13和应用服务器150。在一些实施例中，所述多个计算设备可以包括未示出的其他计算设备。在一些实施例中，所述多个计算设备可以包括比图1所示计算设备的数目更多或更少的计算设备。所述一个或多个网络180可以包括诸如企业专用网络之类的安全网络、诸如无线开放式网络之类的不安全网络、局域网(LAN)、广域网(WAN)、以及因特网。所述一个或多个网络180中的每个网络都可以包括集线器、网桥、路由器、交换机、以及有线传输介质，比如有线网络或直接有线连接。

[0024] 诸如应用服务器150之类的服务器可以允许客户机从该服务器下载信息(例如，文本、音频、图像和视频文件)或者执行与存储在该服务器上的特定信息相关的搜索查询。一般而言，“服务器”可以包括在客户机-服务器关系中充当主机的硬件设备、或者与一个或多个客户机共享资源或为所述客户机执行工作的软件进程。客户机-服务器关系下的计算设备之间的通信可以由客户机通过向服务器发送要求访问特定资源或执行特定工作的请求来发起。服务器随后可以执行所请求的动作并且将响应发送回客户机。

[0025] 计算环境12的一个实施例包括网络接口145、处理器146、以及存储器147，所有这些都彼此通信。网络接口145允许计算环境12连接到一个或多个网络180。网络接口145可以包括无线网络接口、调制解调器、和/或有线网络接口。处理器146允许计算环境12执行存储在存储器147中的计算机可读指令以执行在此讨论的过程。在一个示例中，计算环境12可包括游戏控制台。

[0026] 联网计算环境100可以为一个或多个计算设备提供云计算环境。云计算是指基于因特网的计算，其中共享的资源、软件和/或信息通过因特网(或其他全局网络)被按需提供给一个或多个计算设备。基于在计算机网络图中使用云图来将因特网描绘成对其所表示的底层基础设施的抽象，术语“云”被用作对因特网的比喻。

[0027] 在一个实施例中，应用服务器150可以从计算环境12接收音频文件以及一个或多个关键字。应用服务器150可标识音频文件中与该一个或多个关键字相关联的一个或多个语音。随后，应用服务器150可基于该一个或多个语音来使基于云的语音识别技术进行适应，对音频文件执行基于云的语音识别技术，并将音频文件中所标识的一个或多个关键字传送到计算环境12。

[0028] 图2描绘了包括计算环境12和捕捉设备20的目标检测和跟踪系统10的一个实施例。目标检测和跟踪系统10可用于检测、识别、分析、和/或跟踪人类目标(诸如用户18)和/或非人类目标(诸如用户18拿着的道具(未示出))。目标检测和跟踪系统10可包括用于生成用户18存在于其中的位置空间环境的深度图的深度检测系统。

[0029] 如图所示，目标检测和跟踪系统10可包括计算环境12。计算环境12可包括计算机、游戏系统或控制台等等。在一个实施例中，计算环境12可包括硬件组件和/或软件组件，使得计算环境12可用于执行操作系统和诸如游戏应用、非游戏应用等的应用。在一个实施例中，计算环境12可包括诸如标准化处理器、专用处理器、微处理器之类的处理器，该处理器可执行存储在处理器可读的存储设备上的用于执行此处所描述的过程的指令。

[0030] 目标检测和跟踪系统10还可包括捕捉设备20。捕捉设备20可包括用于捕捉或记录声音的一个或多个话筒以及用于捕捉或记录图像的一个或多个相机。在一个实施例中，捕捉设备20可包括可用于在视觉上监视包括诸如用户18等一个或多个用户的一个或多个目标。由一个或多个用户执行的姿势(包括姿态)可被捕捉、分析、和跟踪，以便执行对操作

系统或应用的用户界面的一个或多个控制或动作。在一些实施例中，捕捉设备 20 可包括深度传感相机。

[0031] 用户可通过移动他或她的身体来创建姿势。姿势可包括用户的运动或姿态，其可被捕捉为图像数据并解析其意义。姿势可以是动态的，包括运动，如模仿投球。姿势可以是静态姿势，诸如保持其前臂交叉。姿势也可结合道具，如挥动仿制的剑。

[0032] 捕捉设备 20 可捕捉与一个或多个用户和 / 或物体相关的图像和音频数据。例如，捕捉设备 20 可用于捕捉与一个或多个用户的部分或全部身体移动、姿势和语音相关的信息。由捕捉设备 20 捕捉的信息可通过计算环境 12 和 / 或捕捉设备 20 内的处理元件来接收，并用于对游戏或其他应用的各方面进行呈现、交互和控制。在一个示例中，捕捉设备 20 捕捉与特定用户有关的图像和音频数据，并且计算环境 12 处理所捕捉的信息以便通过执行面部和语音识别软件来标识该特定用户。

[0033] 在一些实施例中，目标检测和跟踪系统 10 可生成并利用深度图来检测和 / 或跟踪环境中的物体。深度图可包括环境中的包括与该环境相关联的深度信息的图像或帧。在一个示例中，深度图像可包括多个观测到的像素，其中每一观测到的像素具有相关联的深度值。例如，每一个像素都可包括深度值，诸如从捕捉设备的观点到环境中的物体的长度或距离。

[0034] 目标检测和跟踪系统 10 可被连接到向诸如用户 18 等用户提供游戏或应用视觉和 / 或音频的视听设备 16，如电视机、监视器、高清电视机 (HDTV)。例如，计算环境 12 可包括诸如图形卡等视频适配器和 / 或诸如声卡等音频适配器，这些适配器可提供与游戏应用、非游戏应用等相关联的视听信号。视听设备 16 可从计算环境 12 接收视听信号，然后可向用户 18 输出与视听信号相关联的游戏或应用视觉和 / 或音频。视听设备 16 可经由例如，S- 视频电缆、同轴电缆、HDMI 电缆、DVI 电缆、VGA 电缆等连接到计算环境 12。

[0035] 如图 2 所示，在计算环境 12 上执行的应用可以是用户 18 可能正在玩的拳击游戏。计算环境 12 可使用视听设备 16 来向用户 18 提供拳击对手 22 的视觉表示。计算环境 12 还可使用视听设备 16 来提供用户 18 可通过他的或她的移动来控制的玩家化身 24 的视觉表示。例如，用户 18 可在物理空间中挥拳猛击，这使得玩家化身 24 在游戏空间中挥拳猛击。在一个实施例中，目标检测和跟踪系统 10 的计算机环境 12 和捕捉设备 20 可用于识别和分析用户 18 在物理空间中的重拳，从而使得该重拳可被解释为对游戏空间中的玩家化身 24 的游戏控制。

[0036] 在一个实施例中，用户移动可被解释为可对应于除控制玩家化身 24 之外的动作的控制。例如，用户 18 可以使用特定移动来结束游戏、暂停游戏、保存游戏、选择级别、查看高分、或与朋友交流。在另一实施例中，目标检测和跟踪系统 10 将目标的移动解释为游戏领域之外的操作系统和 / 或应用控制。例如，事实上操作系统和 / 或应用程序的任何可控方面都可由诸如用户 18 等目标的移动来控制。在另一实施例中，用户 18 可使用移动来从主用户界面选择游戏或其他应用。由此，用户 18 的全范围运动可以用任何合适的方式来获得、使用并分析以与应用或操作系统进行交互。

[0037] 目标检测和跟踪系统 10 及其组件的合适的示例在以下共同待审的专利申请中找到，所有这些专利申请的全部内容都通过引用结合于此 ：于 2009 年 5 月 29 日提交的名称为“Environment And/Or Target Segmentation (环境和 / 或目标分割)”的美国专利申请

序列号 No.12/475,094 ;于 2009 年 7 月 29 日提交的名称为“Auto Generating a Visual Representation(自动生成视觉表示)”的美国专利申请序列号 No.12/511,850 ;于 2009 年 5 月 29 日提交的名称为“Gesture Tool(姿势工具)”的美国专利申请序列号 No.12/474,655 ;于 2009 年 10 月 21 日提交的名称为“Pose Tracking Pipeline (姿态跟踪流水线)”的美国专利申请序列号 No.12/603,437 ;于 2009 年 5 月 29 日提交的名称为“Device for Identifying and Tracking Multiple Humans Over Time (用于随时间标识和跟踪多个人类的设备)”的美国专利申请序列号 No.12/475,308 ;于 2009 年 10 月 7 日提交的名称为“Human Tracking System(人类跟踪系统)”的美国专利申请序列号 No.12/575,388 ;于 2009 年 4 月 13 日提交的名称为“Gesture Recognizer System Architecture(姿势识别器系统架构)”的美国专利申请序列号 No.12/422,661 ;于 2009 年 2 月 23 日提交的名称为“Standard Gestures (标准姿势)”的美国专利申请序列号 No.12/391,150 ;以及于 2009 年 5 月 29 日提交的名称为“Gesture Tool (姿势工具)”的美国专利申请序列号 No.12/474,655。

[0038] 图 3 描绘了目标检测和跟踪系统 10 和捕捉设备 20 的视野内的环境 300 的一个实施例。环境 300 包括人 28 和 29 以及非人类物体(椅子 16)。如图所示,人 28 比人 29 更靠近捕捉设备 20。从捕捉设备 20 的观点来看,人 28 还与不同于人 29 的方向(或角度)相关联。在一些实施例中,目标检测和跟踪系统 10 确定环境 300 中与人 28 相关联的第一方向(例如,经由声源定位)以及环境 300 中与人 29 相关联的第二方向。第一方向和第二方向各自可由相对于同捕捉设备 20 相关联的基准坐标的特定方向(或角度)来标识。特定方向还可由与环境 300 中的特定方向相关联的一组坐标来标识。

[0039] 一旦第一方向和第二方向被确定,目标检测和跟踪系统 10 就可生成与第一方向相关联的第一音频记录以及与第二方向相关联的第二音频记录。第一音频记录和第二音频记录各自可使用波束形成(beamforming)技术来生成。波束形成技术可应用于同多个话筒相关联的声音,以使得该多个话筒充当单个高度方向性话筒(即,源自特定方向范围内的声音被放大,而源自特定方向范围外的声音被削弱)。这些技术允许聚焦于源自人 28 的声音,同时抑制源自人 29 的声音。波束形成技术可以用硬件或软件来实现,并且可以并行执行(即,可执行对第一音频记录和第二音频记录两者的并行处理)。

[0040] 一旦生成特定音频记录(例如,第一音频记录),目标检测和跟踪系统 10 就可对特定音频记录执行本地语音识别。本地语音识别可包括检测特定人(例如,人 28)说出的第一话语以及检测该第一话语中的一个或多个关键字。该第一话语可包括与特定人相关联的完整语音单元。例如,第一话语可包括说出的句子。

[0041] 在一个实施例中,第一话语的开始可通过检测人 28 说出的一个或多个关键字来标识。在标识出第一话语以及一个或多个关键字后,目标检测和跟踪系统 10 可经由因特网或其他全球网络将第一话语以及该一个或多个关键字传送到一个或多个服务器以便进行语音识别处理。第一话语可作为音频文件来传送。该一个或多个服务器可以用人 28 作出的第一话语中检测到的一个或多个单词来响应。在一个实施例中,该一个或多个服务器可返回与该一个或多个单词相关联的文本。在另一实施例中,该一个或多个服务器可基于该一个或多个单词来执行因特网搜索并返回结果。

[0042] 在一些实施例中,基于云的语音识别引擎可基于同如在第一话语中发音的一个或多个关键字相关联的语音来适应该基于云的语音识别引擎执行的语音识别技术。

[0043] 图 4 示出了包括捕捉设备 20 和计算环境 12 的目标检测和跟踪系统 10 的一个实施例。在一些实施例中，捕捉设备 20 和计算环境 12 可以集成在单个计算设备中。该单个计算设备可以是移动设备，诸如图 1 中的移动设备 11。

[0044] 在一个实施例中，捕捉设备 20 可以包括用于捕捉图像和视频的一个或多个图像传感器。图像传感器可以包括 CCD 传感器或 CMOS 传感器。在一些实施例中，捕捉设备 20 可包括 IR CMOS 图像传感器。捕捉设备 20 还可以包括深度相机(或深度感测相机)，该相机被配置为经由包括例如飞行时间、结构化光、立体图像等在内的任何合适的技术来捕捉包括深度图像的带有深度信息的视频，该深度图像可包括深度值。

[0045] 捕捉设备 20 可包括图像相机组件 32。在一个实施例中，图像相机组件 32 可以包括可捕捉场景的深度图像的深度相机。深度图像可包括所捕捉的场景的二维(2-D)像素区域，其中 2-D 像素区域中的每个像素都可以表示深度值，比如所捕捉的场景中的物体与图像相机组件 32 相距的例如以厘米、毫米等为单位的距离。

[0046] 图像相机组件 32 可包括可用来捕捉捕捉区域的深度图像的 IR 光组件 34、三维(3-D)相机 36、以及 RGB 相机 38。例如，在飞行时间分析中，捕捉设备 20 的 IR 光组件 34 可以将红外光发射到捕捉区域上，然后可以使用传感器，用例如 3-D 相机 36 和 / 或 RGB 相机 38 来检测从捕捉区域中的一个或多个物体的表面反向散射的光。在某些实施例中，可以使用脉冲式红外光从而可以测量出射光脉冲和相应的入射光脉冲之间的时间差并将其用于确定从捕捉设备 20 到捕捉区域中的一个或多个物体上的特定位置的物理距离。此外，可将出射光波的相位与入射光波的相位进行比较来确定相移。然后可以使用该相移来确定从捕捉设备到与一个或多个物体相关联的特定位置的物理距离。

[0047] 在另一示例中，捕捉设备 20 可使用结构化光来捕捉深度信息。在该分析中，图案化光(即，被显示为诸如网格图案或条纹图案等已知图案的光)可经由例如 IR 光组件 34 被投影到捕捉区域上。在撞击到捕捉区域中的一个或多个物体(或目标)的表面时，作为响应，图案可变形。图案的这种变形可由例如 3-D 相机 36 和 / 或 RGB 相机 38 来捕捉并被分析以确定从捕捉设备到一个或多个物体上的特定位置的物理距离。捕捉设备 20 可包括用于产生准直光的光学器件。在一些实施例中，可使用激光投影仪来创建结构化光图案。光投影仪可包括激光器、激光二极管和 / 或 LED。

[0048] 在某些实施例中，可将两个或更多个相机整合到一个集成捕捉设备中。例如，深度相机和视频相机(例如 RGB 视频相机)可以被合并到共同的捕捉设备中。在一些实施例中，可协同使用相同或不同类型的两个或更多分开的捕捉设备。例如，可以使用深度相机和分开的视频相机，可以使用两个视频相机，可以使用两个深度相机，可以使用两个 RGB 相机，或者可以使用任何组合和数目的相机。在一个实施例中，捕捉设备 20 可包括可以从不同的角度观察捕捉区域的两个或更多个在物理上分离的相机，以获取可以被解析以生成深度信息的视觉立体数据。还可通过使用多个检测器(可以是单色、红外、RGB)或任意其它类型的检测器捕捉图像、以及执行视差计算，来确定深度。也可使用其他类型的深度图像传感器来创建深度图像。

[0049] 如图 4 所示，捕捉设备 20 可以包括一个或多个话筒 40。该一个或多个话筒 40 中的每一个都可以包括可以接收声音并将其转换成电信号的换能器或传感器。该一个或多个话筒可包括话筒阵列，其中该一个或多个话筒可以按预定布局排列。

[0050] 捕捉设备 20 可以包括可以与图像相机组件 32 可操作地通信的处理器 42。处理器 42 可包括标准处理器，专用处理器、微处理器等。处理器 42 可以执行指令，所述指令可以包括用于存储过滤器或简档、接收和分析图像、确定是否已经发生特定情况的指令或任何其他合适的指令。应当理解，至少一些图像分析和 / 或目标分析和跟踪操作可以由诸如捕捉设备 20 之类的一个或多个捕捉设备内所包含的处理器来执行。

[0051] 捕捉设备 20 可包括存储器 44，该存储器可存储可由处理器 42 执行的指令、由 3-D 相机或 RGB 相机捕捉的图像或图像帧、过滤器或简档、或任何其他合适的信息、图像等等。在一个示例中，存储器 44 可包括随机存取存储器（RAM）、只读存储器（ROM）、高速缓存、闪存、硬盘或任何其他合适的存储组件。如图所示，存储器 44 可以是与图像捕捉组件 32 和处理器 42 进行通信的单独的组件。在另一实施例中，存储器 44 可被集成到处理器 42 和 / 或图像捕捉组件 32 中。在其他实施例中，图 4 所示的捕捉设备 20 的组件 32、34、36、38、40、42 和 44 中的部分或全部被容纳在单个外壳中。

[0052] 捕捉设备 20 可以经由通信链路 46 与计算环境 12 进行通信。通信链路 46 可以是包括例如 USB 连接、火线连接、以太网电缆连接等有线连接和 / 或诸如无线 802.11b、802.11g、802.11a 或 802.11n 连接等无线连接。计算环境 12 可以向捕捉设备 20 提供时钟，可以使用该时钟来通过通信链路 46 确定何时捕捉例如场景。在一个实施例中，捕捉设备 20 可将由例如 3D 相机 36 和 / 或 RGB 相机 38 捕捉的图像经由通信链路 46 提供给计算环境 12。

[0053] 如图 4 中所示，计算环境 12 包括与操作系统 196 通信的图像和音频处理引擎 194。图像和音频处理引擎 194 包括专用音频处理单元 197、物体和姿势识别引擎 190、结构数据 198、处理单元 191 和存储器单元 192，所有这些都彼此通信。图像和音频处理引擎 194 处理从捕捉设备 20 接收的视频、图像和音频数据。为了辅助物体的检测和 / 或跟踪，图像和音频处理引擎 194 可以利用结构数据 198 以及物体和姿势识别引擎 190。专用音频处理单元 197 处理捕捉设备 20 获取的音频信号。音频信号处理技术可包括声学回声抵消、音源定位和波束形成技术。

[0054] 处理单元 191 可以包括用于执行物体、面部和语音识别算法的一个或多个处理器。在一个实施例中，图像和音频处理引擎 194 可以将物体识别和面部识别技术应用于图像或视频数据。例如，物体识别可以用于检测特定物体（例如足球、汽车或陆标），并且面部识别可以用于检测特定人的面部。图像和音频处理引擎 194 可以将音频和语音识别技术应用于音频数据。例如，音频识别可以用于检测特定声音。要检测的特定面部、语音、声音和物体可以存储在存储器单元 192 中所包含的一个或多个存储器中。处理单元 191 可执行存储在存储器单元 192 中的计算机可读指令以执行此处讨论的过程。

[0055] 在一些实施例中，可以用诸如 IR 回射标记之类的一个或多个标记来扩充所跟踪的一个或多个物体，以便改进物体检测和 / 或跟踪。也可以使用平面基准图像、已编码 AR 标记、QR 码和 / 或条形码来改进物体检测和 / 或跟踪。在检测到一个或多个物体以后，图像和音频处理引擎 194 可以向操作系统 196 报告所检测的每个物体的标识以及相应的位置和 / 或定向。

[0056] 图像和音频处理引擎 194 可以在执行物体识别时利用结构数据 198。结构数据 198 可以包括关于要跟踪的目标和 / 或物体的结构信息。例如，可以存储人类的骨架模型以帮助识别身体部位。在另一示例中，结构数据 198 可以包括关于一个或多个无生命物体的结

构信息以便帮助识别所述一个或多个无生命物体。

[0057] 图像和音频处理引擎 194 还可以在执行物体识别时利用物体和姿势识别引擎 190。在一个示例中，物体和姿势识别引擎 190 可以包括姿势过滤器的集合，每个姿势过滤器都包括关于骨架模型可执行的姿势的信息。物体和姿势识别引擎 190 可将由捕捉设备 20 所捕捉的数据（其形式为骨架模型以及与其相关联的移动）与姿势库中的姿势过滤器进行比较来标识用户（其由骨架模型来表示）何时执行了一个或多个姿势。在一个示例中，图像和音频处理引擎 194 可以使用物体和姿势识别引擎 190 来帮助解释骨架模型的移动以及检测特定姿势的执行。

[0058] 关于物体检测和跟踪的更多信息可在 2009 年 12 月 18 日提交的美国专利申请 12/641,788 “Motion Detection Using Depth Images（使用深度图像的运动检测）”，以及美国专利申请 12/475,308 “Device for Identifying and Tracking Multiple Humans over Time（用于随时间标识和跟踪多个人类的设备）”中找到，这两个申请的全部内容通过引用并入本申请。关于物体和姿势识别引擎 190 的更多信息参见 2009 年 4 月 13 日提交的美国专利申请 12/422,661 “Gesture Recognition System Architecture（姿势识别系统架构）”，该申请通过整体引用合并于此。关于识别姿势的更多信息可在 2009 年 2 月 23 日提交的美国专利申请 12/391,150 “Standard Gestures（标准姿势）”；以及 2009 年 5 月 29 日提交的美国专利申请 12/474,655 “Gesture Tool（姿势工具）”中找到，这两个申请的全部内容通过引用并入本申请。

[0059] 图 5A 是描述集成本地和基于云的语音识别的过程的一个实施例的流程图。图 5A 的过程可由一个或多个计算设备来连续执行。图 5A 的过程中每一步骤都可由与在其他步骤中所使用的那些计算设备相同或不同的计算设备来执行，且每一步骤不必由单个计算设备来执行。在一个实施例中，图 5A 的过程可由诸如图 2 ～ 3 中的目标检测和跟踪系统 10 之类的目标检测和跟踪系统来执行。

[0060] 在步骤 510，获取一个或多个音频记录。该一个或多个音频记录中的每一个都可与游戏空间或其他环境中的不同收听方向相关联。在一个实施例中，一个或多个音频记录中的特定音频记录只包括源自特定方向的声音。在步骤 512，对一个或多个音频记录中的每一个执行本地语音识别。本地语音识别可包括检测第一话语以及检测该第一话语中的一个或多个关键字。话语可包括与特定人相关联的完整语音单元，并且通常但并非始终可以通过持续预定时长的静音来划界。例如，话语可包括说出的词组，其中该说出的词组之前和之后存在一秒钟的静音。在一个实施例中，本地语音识别可由移动设备来执行，并且可检测与特定人的语音相关联的第一话语。本地语音识别还可包括确定标识第一话语中的一个或多个关键字的位置的一个或多个位置指针。

[0061] 在一些实施例中，步骤 512 中检测到的一个或多个关键字可用于发起计算设备上的本地动作。例如，如果电影应用正在特定计算设备上运行并且检测到关键字“pause（暂停）”，则该电影应用可以在该特定计算设备上暂停。在另一实施例中，步骤 512 中检测到的一个或多个关键字可用于基于该一个或多个关键字来发起非本地动作，诸如基于云的语音识别。

[0062] 在步骤 513，确定是否已经在步骤 512 中的第一话语中检测到一个或多个关键字。如果尚未检测到一个或多个关键字，则执行步骤 510。在这种情况下，如果没有在话语中检

测到关键字，则不执行后续基于云的语音处理。否则，如果检测到一个或多个关键字，则执行步骤 514。

[0063] 在步骤 514，确定上下文信息。在一个实施例中，上下文信息可以与特定计算设备相关联。例如，特定计算设备可包括执行本地语音识别的移动或非移动计算设备。与特定计算设备相关联的上下文信息可包括在特定计算设备上运行的特定应用的标识。在一个示例中，特定应用可包括游戏应用或在线市场应用。与特定计算设备相关联的上下文信息还可包括与特定计算设备相关联的地理位置信息（例如，GPS 坐标）。

[0064] 在另一实施例中，上下文信息可以与第一环境中的特定人相关联。例如，上下文信息可包括与特定人相关联的简档信息。简档信息可包括特定人的兴趣、特定人的朋友或联系人以及其他个人信息。上下文信息还可包括与特定人相关联的日历信息。

[0065] 在步骤 516，将第一话语以及一个或多个关键字传送到第二计算设备。图 5A 的步骤 514 中确定的上下文信息以及图 5A 的步骤 512 中确定的一个或多个位置指针也可被传送到第二计算设备。在一个实施例中，第二计算设备可包括基于云的语音识别引擎或应用服务器，诸如图 1 中的应用服务器 150。第一话语可经由与该第一话语相关联的音频文件（例如，WAV 文件或 MP3 文件）来传送。音频文件可使用音频编解码器来创建，音频编解码器将接收到的模拟音频信号转换成压缩或不压缩的数字表示。一个或多个关键字可经由文本消息来传送。

[0066] 在步骤 518，接收基于第一话语的第一响应。在一个示例中，第一响应可包括与第一话语中由第二计算设备检测到的一个或多个单词相关联的文本消息。在另一示例中，第一响应可包括基于第一话语的因特网搜索结果。第一响应还可取决于传送到第二计算设备的上下文信息。

[0067] 图 5B 是描述获取一个或多个音频记录的过程的一个实施例的流程图。图 5B 中描述的过程是用于实现图 5A 中的步骤 510 的过程的一个示例。图 5B 的过程可由一个或多个计算设备来连续执行。图 5B 的过程中每一步骤都可由与在其他步骤中所使用的那些计算设备相同或不同的计算设备来执行，且每一步骤不必由单个计算设备来执行。在一个实施例中，图 5B 的过程可由诸如图 2 ～ 3 中的目标检测和跟踪系统 10 之类的目标检测和跟踪系统来执行。

[0068] 在步骤 502，获取来自第一环境的一个或多个声音。该一个或多个声音可使用一个或多个话筒或话筒阵列来获取。一个或多个话筒可以按预定布局来排列，并且用于捕捉来自各个方向和 / 或源自环境中的不同点的声音（例如，与房间中的活动说话者相关联的语音）。一个或多个声音可作为模拟信号来捕捉并且通过使用模数转换器来数字化。

[0069] 在步骤 504，获取第一环境中的一个或多个图像。该一个或多个图像可包括一个或多个深度图像。在步骤 506，可执行对该一个或多个声音的回声抵消。在一个实施例中，声学回声抵消用于移除源自一个或多个音频说话者的声音。回声抵消可用于抑制源自第一环境之外的、通过一个或多个音频说话者投影到第一环境中的语音和 / 或其他声音。噪声抑制技术还可应用于一个或多个声音以移除背景噪声。在一个示例中，可以对一个或多个声音应用带通滤波器以移除背景噪声。

[0070] 在步骤 507，确定一个或多个方向。该一个或多个方向中的每一个都与特定声源相关联。在一个实施例中，该一个或多个方向中的每一个都与第一环境中的一个或多个人中

的不同人相关联。该一个或多个方向中的每一个可由特定方向(或角度)和关于特定声源的置信度来标识。

[0071] 在一个实施例中,该一个或多个方向可使用声源定位来确定。声源定位技术可用于通过检测不同声音的到达时间的时差(因为多个话筒捕捉到的声音的速度)来定位声音的方向。声源定位技术还可包括对多个话筒中的每一个接收到的每一个音频信号执行模式匹配。一个或多个方向可由以下各项来表示:一维定位(例如,表示特定声源位于其中的平面的角度)、二维定位(例如,表示角度和仰角的向量)、或三维定位(例如,在三维空间中定位与特定声源相关联的点)。

[0072] 在一些实施例中,通过利用步骤 504 中获取的一个或多个图像来执行骨架跟踪,可确定一个或多个方向。例如,骨架跟踪可用于检测第一环境中的一个或多个活动骨架。该一个或多个方向中的每一个都可与一个或多个活动骨架中的不同骨架相关联。该一个或多个方向还可通过对一个或多个声音应用话音识别技术以标识特定人的话音来确定。在一个实施例中,与特定人相关联的方向可通过应用骨架跟踪和话音识别两者来实时跟踪。关于骨架跟踪的更多信息可以在美国专利申请 12/475,308,“Device for Identifying and Tracking Multiple Humans over Time (用于随时间标识和跟踪多个人的设备)”中找到,该申请通过引用整体结合于此。

[0073] 在步骤508,生成基于一个或多个声音的一个或多个音频记录。该一个或多个音频记录中的每一个都可与一个或多个方向中的不同方向相关联。在一个实施例中,该一个或多个音频记录中的每一个都包括与第一环境中的不同人相关联的音频信息。

[0074] 在一些实施例中,可执行波束形成技术以放大源自特定方向(或位置)的声音和抑制源自其他方向的声音。波束形成允许多个话筒(例如,包括话筒阵列)用作可操纵的方向性话筒。在某些情况下,波束形成技术可包括对与多个话筒相关联的音频信号进行时移和组合(例如,延迟和求和波束形成)。时移度可基于同特定声源相关联的特定方向(或位置)。当不止一个人在环境中同时说话时,声音聚焦技术是特别有用的。

[0075] 在一个实施例中,音频记录将仅在特定人被确定为在特定方向说话(例如,朝着捕捉设备说话)的情况下生成。特定人的头部的方向可经由对在步骤 504 中获取的一个或多个图像进行图像处理来检测。

[0076] 在步骤 509,输出步骤 508 中生成的一个或多个音频记录。在一个实施例中,该一个或多个音频记录被输出到本地语音识别引擎以便进行语音到文本处理。

[0077] 图 5C 是描述获取一个或多个音频记录的过程的一个实施例的流程图。图 5C 中描述的过程是用于实现图 5A 中的步骤 510 的过程的另一个示例。图 5C 的过程可由一个或多个计算设备来连续执行。图 5C 的过程中每一步骤都可由与在其他步骤中所使用的那些计算设备相同或不同的计算设备来执行,且每一步骤不必由单个计算设备来执行。在一个实施例中,图 5C 的过程可由诸如图 2 ～ 3 中的目标检测和跟踪系统 10 之类的目标检测和跟踪系统来执行。

[0078] 在步骤540,获取来自第一环境的一个或多个声音。该一个或多个声音可使用一个或多个话筒或话筒阵列来获取。在步骤542,获取第一环境中的一个或多个图像。该一个或多个图像可以从诸如图 2 ～ 3 中的捕捉设备 20 之类的捕捉设备获取。该一个或多个图像可包括一个或多个彩色和 / 或深度图像。

[0079] 在步骤550，检测第一环境中的特定人。该特定人可通过对步骤542中获取的一个或多个图像应用物体识别技术来检测。在一些实施例中，特定人可通过对步骤542中获取的一个或多个图像应用面部识别技术并且对步骤540中获取的一个或多个声音应用话音识别技术来检测(或标识)。在步骤552，确定第一环境中与特定人相关联的位置。该位置可由特定人存在于第一环境中的空间中的单个点来表示。在一个实施例中，可使用来自与特定人相关联的一个或多个图像的各部分的深度信息来确定与特定人相关联的位置。

[0080] 在步骤554，确定特定人是否面向特定方向。特定人的面部或身体的方向可经由对在步骤542中获取的一个或多个图像进行图像处理来检测。在一个示例中，步骤550中检测到的特定人面向捕捉设备，诸如图2～3中的捕捉设备20。如果特定人未面向特定方向，则执行步骤540并且不生成音频记录。否则，如果特定人面向特定方向，则执行步骤556。

[0081] 在步骤556，生成包括源自步骤552中确定的位置的声音的音频记录。该音频记录可包括步骤540中获取的一个或多个声音的子集。在一个实施例中，使用波束形成技术来生成聚焦于源自特定人所存在的特定方向的声音的音频记录。音频记录还可使用参考图5B的步骤508描述的过程来生成。

[0082] 在步骤558，输出步骤556中生成的音频记录。在一个实施例中，音频记录被输出到本地语音识别引擎，以便在被传送到基于云的语音识别引擎之前进行语音到文本预处理。

[0083] 图5D是描述获取一个或多个音频记录的过程的一个实施例的流程图。图5C中描述的过程是用于实现图5A中的步骤510的过程的另一个示例。图5D的过程可由一个或多个计算设备来连续执行。图5D的过程中每一步骤都可由与在其他步骤中所使用的那些计算设备相同或不同的计算设备来执行，且每一步骤不必由单个计算设备来执行。在一个实施例中，图5D的过程可由诸如图2～3中的目标检测和跟踪系统10之类的目标检测和跟踪系统来执行。

[0084] 在步骤572，获取来自第一环境的一个或多个声音。该一个或多个声音可使用一个或多个话筒或话筒阵列来获取。一个或多个话筒可以按预定布局来排列，并且用于捕捉来自各个方向和/或源自环境中的不同点的声音(例如，与房间中的活动说话者相关联的语音)。一个或多个声音可作为模拟信号来捕捉并且通过使用模数转换器来数字化。

[0085] 在步骤576，过滤一个或多个声音。在一个实施例中，声学回声抵消用于移除源自一个或多个音频说话者的声音。回声抵消可用于抑制源自第一环境之外的、通过一个或多个音频说话者投影到第一环境中的语音和/或其他声音。噪声抑制技术还可被应用于一个或多个声音以移除背景噪声。在一个示例中，可以对一个或多个声音应用带通滤波器以移除背景噪声。

[0086] 在步骤577，确定一个或多个方向。该一个或多个方向中的每一个都与特定声源相关联。在一个实施例中，该一个或多个方向中的每一个都与第一环境中的一个或多个人中的不同人相关联。该一个或多个方向中的每一个可由特定方向(或角度)和关于特定声源的置信度来标识。

[0087] 在一个实施例中，该一个或多个方向可使用声源定位来确定。声源定位技术可用于通过检测不同声音的到达时间的时差(因为多个话筒捕捉到的声音的速度)来定位声音的方向。声音定位技术还可包括对多个话筒中的每一个接收到的每一个音频信号执行模式

匹配。一个或多个方向可由以下各项来表示：一维定位(例如，表示特定声源位于其中的平面的角度)、二维定位(例如，表示角度和仰角的向量)、或三维定位(例如，三维空间中与特定声源相关联的点)。

[0088] 在步骤578，生成基于一个或多个声音的一个或多个音频记录。该一个或多个音频记录中的每一个都可与一个或多个方向中的不同方向相关联。在一个实施例中，该一个或多个音频记录中的每一个都包括与第一环境中的不同人相关联的音频信息。

[0089] 在一些实施例中，可执行波束形成技术以放大源自特定方向(或位置)的声音和抑制源自其他方向的声音。波束形成允许多个话筒(例如，包括话筒阵列)用作可操纵的方向性话筒。在某些情况下，波束形成技术可包括对与多个话筒相关联的音频信号进行时移和组合。时移度可基于同特定声源相关联的特定位置。当不止一个人在环境中同时说话时，声音聚焦技术是特别有用的。

[0090] 在步骤579，输出步骤578中生成的一个或多个音频记录。在一个实施例中，该一个或多个音频记录被输出到本地语音识别引擎以便进行语音到文本处理。

[0091] 图5E是描述获取一个或多个音频记录的过程的一个实施例的流程图。图5E中描述的过程是用于实现图5B中的步骤508、用于实现图5C中的步骤556或用于实现图5D中的步骤578的过程的一个示例。图5E的过程可由一个或多个计算设备来连续执行。图5E的过程中每一步骤都可由与在其他步骤中所使用的那些计算设备相同或不同的计算设备来执行，且每一步骤不必由单个计算设备来执行。在一个实施例中，图5E的过程可由诸如图2～3中的目标检测和跟踪系统10之类的目标检测和跟踪系统来执行。

[0092] 在步骤592中，接收特定方向。在步骤596，接收多个音频信号。该多个音频信号中的每一个都可以与多个话筒中的不同话筒相关联。在步骤597，确定基于步骤592中接收到的特定方向的一组时延和一组加权因子。这组时延和这组加权因子可基于多个话筒的空间排列以及源自特定方向的声音到多个话筒中的每一个的到达时差来确定。

[0093] 在步骤598，通过对多个音频信号应用该组时延和该组加权因子来生成新音频信号。在一些实施例中，通过对步骤596中接收到的多个音频信号应用音频信号处理技术以使得源自步骤592中接收到的特定方向的声音被放大(例如，经由相长干涉)而源自另一方向的其他声音被削弱(例如，经由相消干涉)，生成新音频信号。音频信号处理技术可组合多个音频信号中的每一个的时移版本以便将源自特定方向的声音与环境中的其他声音隔开。在一个示例中，新音频信号可通过应用延迟和求和波束形成技术来生成，通过该技术多个音频信号中的每一个都可以在执行对多个音频信号的加权求和之前相对于彼此延迟。延迟量可确定波束形成器“收听”的波束角。

[0094] 在步骤599，输出新音频信号。该新音频信号可作为新音频记录来存储和输出。在一个实施例中，生成多个新音频信号并将其作为多个音频记录来并行输出。

[0095] 图6是描述执行基于云的语音识别的过程的一个实施例的流程图。图6的过程可由一个或多个计算设备来连续执行。图6的过程中的每个步骤都可由与在其他步骤中所使用的那些计算设备相同或不同的计算设备来执行，且每个步骤不必由单个计算设备来执行。在一个实施例中，图6的过程由诸如图1中的应用服务器150之类的可访问因特网的应用服务器来执行。

[0096] 在步骤602，接收第一话语以及一个或多个关键字。步骤602中接收到的一个或多

个关键字可包括诸如“Xbox Bing”或“search the Internet for（在因特网中搜索）”等词组。图 5A 的步骤 514 中确定的上下文信息以及图 5A 的步骤 512 中确定的一个或多个位置指针也可被接收。在步骤604，标识与如在第一话语中发音的一个或多个关键字相关联的一个或多个语音。在一个示例中，对特定人唯一且与一个或多个关键字相关联的语音可通过对第一话语应用传统的语音到文本处理技术并然后将检测到的单词与该一个或多个关键字进行比较来在第一话语中标识。

[0097] 在一些实施例中，与一个或多个关键字相关联的位置信息可用于帮助在第一话语中定位一个或多个关键字。在一个示例中，可接收一个或多个位置指针并使用该一个或多个位置指针来帮助标识一个或多个关键字在第一话语中的位置。该一个或多个位置指针可包括以第一话语的开头为基准的时间偏移量。

[0098] 在步骤 606，基于步骤 604 中标识的一个或多个语音来配置基于云的语音识别引擎。在一个实施例中，基于云的语音识别引擎可以在执行步骤 608 中的基于云的语音识别步骤之前针对与第一话语中的一个或多个关键字的发音相关联的话音模式进行适应。在某些情况下，基于云的语音识别引擎可以在执行步骤 608 之前通过更新一个或多个声学模型来对与特定说话者相关联的声音进行适应。在步骤 608，对第一话语执行基于云的语音识别。在一个示例中，可生成在第一话语中检测到的一个或多个单词的列表。在步骤 609，执行基于步骤 608 中检测到的一个或多个单词的第一任务。该第一任务可包括基于在第一话语中检测到的一个或多个单词来执行因特网搜索。在步骤610，输出基于第一话语的第一响应。第一响应可包括第一话语中检测到的一个或多个单词或与该一个或多个单词相关联的搜索结果。

[0099] 在一些实施例中，从第一话语中检测到的一个或多个单词中生成的搜索结果可基于步骤 602 中接收到的上下文信息来细化。例如，特定人可通过说出话语“search the Internet for Harry Potter（在因特网中搜索哈里波特）”来请求使用术语“Harry Potter（哈里波特）”来执行因特网搜索。如果一普通应用正在运行本地语音识别器（或引擎）的计算设备上执行，则可以只使用术语“Harry Potter”来执行普通因特网搜索。然而，如果一特殊应用（例如，游戏市场应用）正在运行本地语音识别器的计算设备上执行，则可例如通过使用术语“Harry Potter games（哈里波特游戏）”执行因特网搜索来执行经修改的搜索。在某些情况下，普通搜索结果可使用步骤 602 中接收到的上下文信息来过滤或加权。

[0100] 图 7 是描述执行本地语音识别的过程的一个实施例的流程图。图 7 中描述的过程是用于实现图 5A 中的步骤 512 的过程的一个示例。图 7 的过程可由一个或多个计算设备来连续执行。图 7 的过程中的每个步骤都可由与在其他步骤中所使用的那些计算设备相同或不同的计算设备来执行，且每个步骤不必由单个计算设备来执行。在一个实施例中，图 7 的过程可由诸如图 2 ～ 3 中的目标检测和跟踪系统 10 之类的目标检测和跟踪系统来执行。

[0101] 在步骤 702，获取关键字的列表。该关键字的列表可包括一个或多个关键字或词组，诸如“call（呼叫）”、“purchase（购买）”、“search the Internet for（在因特网中搜索）”或“search my email for（在我的电子邮件中搜索）”。该关键字的列表可被包括在本地语法文件（或关键字约束文件）中并且可以与特定应用相关联。例如，第一关键字列表可以在第一应用活动时激活并在第一应用暂停或关闭时停用。在步骤 704，接收音频记录。在某些情况下，该音频记录可通过利用图 5A 中的步骤 510 来获取。

[0102] 在步骤706，从音频记录中检测第一话语。可经由衰减的波束形成置信度(例如通过检测音频记录中包括的语音信号的置信水平的阈值改变以确定一个人何时停止说话)，或经由对其间音频记录中包括的语音信号被认为是静音的时间段的检测(例如，通过检测说出的词组之前和之后的持续长于预定时长的静音)来检测(或标识)第一话语。还可对音频记录应用其他语音检测技术以确定第一话语的起始点和停止点。

[0103] 在一个实施例中，使用话音活动检测器(VAD)通过处理与特定收听方向相关联的音频记录来检测第一话语。在这种情况下，由于源自其他方向的声音被抑制，因此只对源自特定收听方向的声音应用语音检测技术。这与其中对源自游戏空间中的任何地方的声音执行语音处理的典型VAD形成对比。对与特定收听方向相关联的音频记录应用语音检测技术的一个优点是即使多个人在游戏空间中同时说话，也可检测第一话语。例如，如果游戏空间包括同时说话的第一游戏玩家和第二游戏玩家，则可生成与第一游戏玩家的第一方向相关联的第一音频记录，并且可生成与第二玩家的第二方向相关联的第二音频记录。通过对第一音频记录应用语音检测技术，即使第二游戏玩家与第一游戏玩家同时说话，也可检测到与第一游戏玩家相关联的第一话语。类似地，通过对第二音频记录应用语音检测技术，即使第一游戏玩家与第二游戏玩家同时说话，也可检测到与第二游戏玩家相关联的第二话语。

[0104] 在步骤708，在第一话语中检测来自步骤702中获取的关键字列表的一个或多个关键字。该一个或多个关键字可通过对第一话语应用传统的语音到文本处理技术并然后将检测到的单词与该一个或多个关键字进行比较来在第一话语中检测到。该一个或多个关键字还可通过对与第一话语以及该一个或多个关键字相关联的音频信号应用模式匹配技术来在第一话语中检测到。在步骤710，可确定与第一话语中的一个或多个关键字中的每一个的位置相关联的一个或多个位置指针。该一个或多个位置指针中的每一个都由从第一话语的开头开始的时间偏移量来表示。在某些情况下，该一个或多个位置指针可指向在第一话语中间的位置(即，感兴趣的关键字可能在第一话语的中间)。在步骤712，输出步骤706中确定的第一话语以及步骤708中检测到的一个或多个关键字。还可输出步骤710中确定的一个或多个位置指针。在一个示例中，第一话语、一个或多个关键字以及一个或多个位置指针可被传送到第二计算设备。

[0105] 所公开的技术可以与各种计算系统一起使用。图8-9提供了可用于实现所公开的技术的实施例的各种计算系统的示例。

[0106] 图8是作为图3中的计算环境12的一个示例的游戏和媒体系统7201的一个实施例的框图。控制台7203具有中央处理单元(CPU)7200以及便于处理器访问各种存储器的存储器控制器7202，这些存储器包括闪速只读存储器(ROM)7204、随机存取存储器(RAM)7206、硬盘驱动器7208，以及便携式媒体驱动器7107。在一种实现中，CPU 7200包括1级高速缓存7210和2级高速缓存7212，这些高速缓存用于临时存储数据并因此减少对硬盘驱动器7208进行的存储器访问周期的数量，从而提高了处理速度和吞吐量。

[0107] CPU 7200、存储器控制器7202、以及各种存储器设备经由一个或多个总线(未示出)互连在一起。所述一个或多个总线可以包括下列各项中一个或多个：串行和并行总线、存储器总线、外围总线、使用各种总线体系结构中的任何一种的处理器或局部总线。作为示例，这样的体系结构可以包括工业标准体系结构(ISA)总线、微通道体系结构(MCA)总线、增强型ISA(EISA)总线、视频电子标准协会(VESA)局部总线、以及外围部件互连(PCI)总

线。

[0108] 在一个实施方式中，CPU 7200、存储器控制器 7202、ROM 7204、以及 RAM 7206 被集成到公用模块 7214 上。在此实施方式中，ROM 7204 被配置为通过 PCI 总线和 ROM 总线(两者都没有示出)连接到存储器控制器 7202 的闪速 ROM。RAM 7206 被配置为多个双倍数据速率同步动态 RAM (DDR SDRAM)模块，它们被存储器控制器 7202 通过分开的总线(未示出)独立地进行控制。硬盘驱动器 7208 和便携式媒体驱动器 7107 被示为通过 PCI 总线和 AT 附加(ATA)总线 7216 连接到存储器控制器 7202。然而，在其他实施方式中，也可以在替代方案中应用不同类型的专用数据总线结构。

[0109] 三维图形处理单元 7220 和视频编码器 7222 构成了视频处理流水线，用于进行高速度和高分辨率(例如，高清晰度)图形处理。数据通过数字视频总线(未示出)从图形处理单元 7220 传输到视频编码器 7222。音频处理单元 7224 和音频编解码器(编码器 / 解码器) 7226 构成了对应的音频处理流水线，用于对各种数字音频格式进行多通道音频处理。通过通信链路(未示出)在音频处理单元 7224 和音频编解码器 7226 之间传输音频数据。视频和音频处理流水线向 A/V (音频 / 视频)端口 7228 输出数据，以便传输到电视机或其他显示器。在所示出的实现中，视频和音频处理组件 7220-7228 安装在模块 7214 上。

[0110] 图 8 示出了包括 USB 主控制器 7230 和网络接口 7232 的模块 7214。USB 主控制器 7230 通过总线(未示出)与 CPU 7200 和存储器控制器 7202 通信，并用作外围控制器 7205(1)-7205(4) 的主机。网络接口 7232 提供对网络(例如，因特网、家庭网络等等)的访问，并可以是各种有线或无线接口组件中的任何一种，包括以太网网卡、调制解调器、无线接入卡、蓝牙模块、电缆调制解调器等等。

[0111] 在图 8 中描述的实现中，控制台 7203 包括用于支持四个控制器 7205(1)-7205(4) 的控制器支持子部件 7240。控制器支持子部件 7240 包括支持与诸如，例如，媒体和游戏控制器之类的外部控制设备的有线和无线操作所需的任何硬件和软件组件。前面板 I/O 子部件 7242 支持电源按钮 7213、弹出按钮 7215，以及任何 LED (发光二极管)或暴露在控制台 7203 的外表面上的其他指示器等多个功能。子部件 7240 和 7242 通过一个或多个电缆部件 7244 与模块 7214 进行通信。在其他实现中，控制台 7203 可以包括另外的控制器子部件。所示出的实施方式还示出了被配置为发送和接收可传递给模块 7214 的信号(例如从遥控器 7290)的光学 I/O 接口 7235。

[0112] MU 7241(1) 和 7241(2) 被示为可以分别连接到 MU 端口“A”7231(1) 和“B”7231(2)。附加 MU (例如，MU 7241(3)-7241(6))被示为可连接到控制器 7205(1) 和 7205(3)，即每一个控制器两个 MU。控制器 7205(2) 和 7205(4) 也可以被配置成接纳 MU(未示出)。每一个 MU 7241 都提供附加存储，在其上面可以存储游戏、游戏参数、及其他数据。诸如便携式 USB 设备之类的附加存储器设备可用来代替 MU。在一些实现中，其他数据可以包括数字游戏组件、可执行的游戏应用，用于扩展游戏应用的指令集、以及媒体文件中的任何一种。当被插入到控制台 7203 或控制器中时，MU 7241 可以被存储器控制器 7202 访问。系统供电模块 7250 向游戏系统 7201 的组件供电。风扇 7252 冷却控制台 7203 内的电路。

[0113] 包括机器指令的应用 7260 被存储在硬盘驱动器 7208 上。当控制台 7203 被上电时，应用 7260 的各个部分被加载到 RAM 7206 和 / 或缓存 7210 和 7212 中以供在 CPU 7200 上执行。其他应用也可以存储在硬盘驱动器 7208 上以供在 CPU 7200 上执行。

[0114] 可以通过简单地将系统连接到监视器、电视机、视频投影仪、或其他显示设备来将游戏和媒体系统 7201 用作独立系统。在此独立模式下，游戏和媒体系统 7201 允许一个或多个玩家玩游戏或欣赏数字媒体(例如观看电影或欣赏音乐)。然而，随着宽带连接的集成通过网络接口 7232 而成为可能，游戏和媒体系统 7201 还可以作为较大的网络游戏社区的参与者来操作。

[0115] 图 9 是作为图 3 中的计算环境 12 的一个示例的计算系统环境 2200 的一个实施例的框图。计算系统环境 2200 包括计算机 2210 形式的通用计算设备。计算机 2210 的组件可以包括、但不限于处理单元 2220、系统存储器 2230、以及将包括系统存储器 2230 在内的各种系统组件耦合到处理单元 2220 的系统总线 2221。系统总线 2221 可以是若干类型的总线结构中的任一种，包括使用各种总线体系结构中的任一种的存储器总线、外围总线、以及局部总线。作为示例，而非限制，这样的体系结构包括工业标准体系结构(ISA)总线、微通道体系结构(MCA)总线、增强型 ISA (EISA)总线、视频电子技术标准协会(VESA)局部总线和外围部件互连(PCI)总线。

[0116] 计算机 2210 通常包括各种计算机可读介质。计算机可读介质可以是能被计算机 2210 访问的任何可用介质，而且包含易失性和非易失性介质、可移动和不可移动介质。作为示例而非局限，计算机可读介质可以包括计算机存储介质。计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其它数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括，但不限于，RAM、ROM、EEPROM、闪存或其他存储器技术，CD-ROM、数字多功能盘(DVD)或其他光盘存储设备，磁带盒、磁带、磁盘存储设备或其他磁存储设备，或者能用于存储所需信息且可以由计算机 2210 访问的任何其他介质。上述中任一组合也应包括在计算机可读介质的范围之内。

[0117] 系统存储器 2230 包括易失性和 / 或非易失性存储器形式的计算机存储介质，如只读存储器(ROM) 2231 和随机存取存储器(RAM) 2232。包含诸如在启动期间帮助在计算机 2210 内的元件之间传输信息的基本例程的基本输入 / 输出系统 2233 (BIOS)通常储存储在 ROM 2231 中。RAM 2232 通常包含处理单元 2220 可立即访问和 / 或当前正在操作的数据和 / 或程序模块。作为示例而非限制，图 9 示出了操作系统 2234、应用程序 2235、其它程序模块 2236 和程序数据 2237。

[0118] 计算机 2210 也可以包括其他可移动 / 不可移动、易失性 / 非易失性计算机存储介质。仅作为示例，图 9 示出了从不可移动、非易失性磁介质中读取或向其写入的硬盘驱动器 2241，从可移动、非易失性磁盘 2251 中读取或向其写入的磁盘驱动器 2252，以及从诸如 CD ROM 或其它光学介质等可移动、非易失性光盘 2255 中读取或向其写入的光盘驱动器 2256。可在示例性操作环境中使用的其他可移动 / 不可移动、易失性 / 非易失性计算机存储介质包括但不限于，磁带盒、闪存卡、数字多功能盘、数字录像带、固态 RAM、固态 ROM 等。硬盘驱动器 2241 通常由例如接口 2240 等不可移动存储器接口连接至系统总线 2221，而磁盘驱动器 2251 和光盘驱动器 2255 通常由例如接口 2250 等可移动存储器接口连接至系统总线 2221。

[0119] 上文讨论并在图 9 中示出的驱动器及其相关联的计算机存储介质为计算机 2210 提供了对计算机可读指令、数据结构、程序模块和其他数据的存储。例如，在图 9 中，硬盘驱动器 2241 被示为存储操作系统 2244、应用程序 2245、其它程序模块 2246 和程序数据 2247。

注意，这些组件可与操作系统 2234、应用程序 2235、其他程序模块 2236 和程序数据 2237 相同，也可与它们不同。在此操作系统 2244、应用程序 2245、其他程序模块 2246 以及程序数据 2247 被给予了不同的编号，以说明至少它们是不同的副本。用户可以通过输入设备如键盘 2262 和定点设备 2261（通常指鼠标、跟踪球或触摸垫）向计算机 2210 输入命令和信息。其他输入设备（未示出）可包括话筒、操纵杆、游戏手柄、圆盘式卫星天线、扫描仪等。这些以及其他输入设备通常通过耦合到系统总线的用户输入接口 2260 连接到处理单元 2220，但也可通过诸如并行端口、游戏端口或通用串行总线（USB）之类的其他接口和总线结构来连接。监视器 2291 或其他类型的显示设备也通过诸如视频接口 2290 之类的接口连接至系统总线 2221。除了监视器以外，计算机还可包括诸如扬声器 2297 和打印机 2296 之类的其他外围输出设备，它们可通过输出外围接口 2295 来连接。

[0120] 计算机 2210 可使用到一个或多个远程计算机（诸如，远程计算机 2280）的逻辑连接而在联网环境中操作。远程计算机 2280 可以是个人计算机、服务器、路由器、网络 PC、对等设备或其它常见网络节点，且通常包括上文相对于计算机 2210 描述的许多或所有元件，但在图 9 中只示出存储器存储设备 2281。图 9 中所示的逻辑连接包括局域网（LAN）2271 和广域网（WAN）2273，但也可以包括其它网络。此类联网环境在办公室、企业范围的计算机网络、内联网和因特网中是常见的。

[0121] 当在 LAN 联网环境中使用时，计算机 2210 通过网络接口或适配器 2270 连接到 LAN 2271。当在 WAN 联网环境中使用时，计算机 2210 通常包括调制解调器 2272 或用于通过诸如因特网等 WAN 2273 建立通信的其他手段。调制解调器 2272 可以是内置的或外置的，可经由用户输入接口 2260 或其他适当的机制连接到系统总线 2221。在联网环境中，相对于计算机 2210 所示的程序模块或其部分可被存储在远程存储器存储设备中。作为示例而非限制，图 9 示出了远程应用程序 2285 驻留在存储器设备 2281 上。应当理解，所示的网络连接是示例性的，并且可使用在计算机之间建立通信链路的其他手段。

[0122] 所公开的技术可用各种其它通用或专用计算系统环境或配置来操作。适合在该技术中使用的公知的计算系统、环境和 / 或配置的示例包括，但不限于，个人计算机、服务器计算机、手持或膝上型设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费者电子产品、网络 PC、小型机、大型机、包含上述系统或设备中的任一个的分布式计算机环境等。

[0123] 所公开的技术可在诸如程序模块等由计算机执行的计算机可执行指令的一般上下文中描述。一般而言，如此处所述的软件和程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构和其他类型的结构。硬件或硬件和软件的组合可代替如此处所述的软件模块。

[0124] 所公开的技术也可以在任务由通过通信网络链接的远程处理设备执行的分布式计算环境中实现。在分布式计算环境中，程序模块可以位于包括存储器存储设备在内的本地和远程计算机存储介质中。

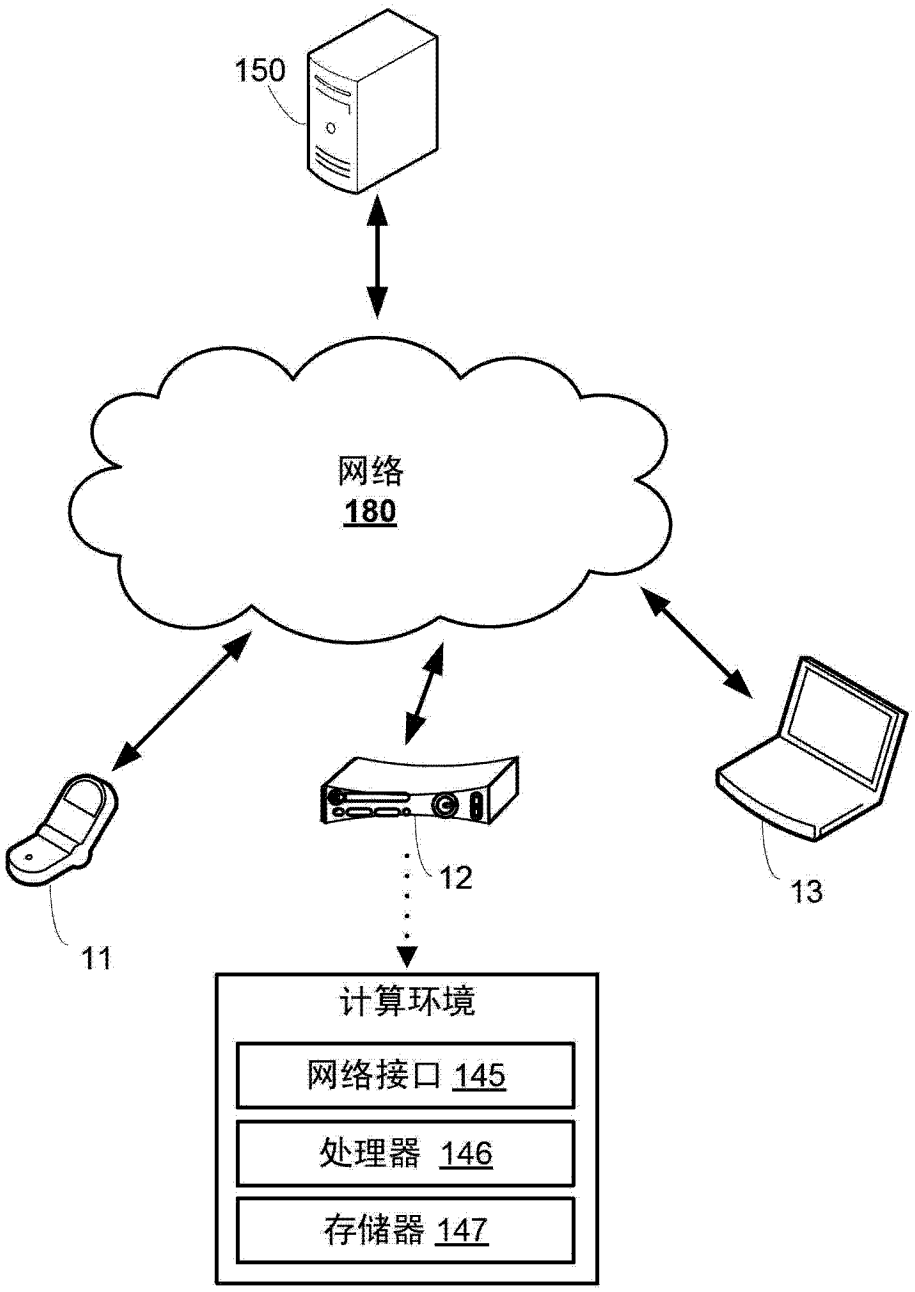
[0125] 出于本文的目的，说明书中引述的“一实施例”、“一个实施例”、“某些实施例”或“另一实施例”用于描述不同的实施例并且不必然指的是同一实施例。

[0126] 出于本文的目的，连接可以是直接连接或间接连接（例如，经由另一方）。

[0127] 出于本文的目的，术语对象的“集合”指的是一个或多个对象的“集合”。

[0128] 尽管用结构特征和 / 或方法动作专用的语言描述了本主题，但可以理解，所附权

利要求书中定义的主题不必限于上述具体特征或动作。更确切而言，上述具体特征和动作是作为实现权利要求的示例形式公开的。



100

图 1

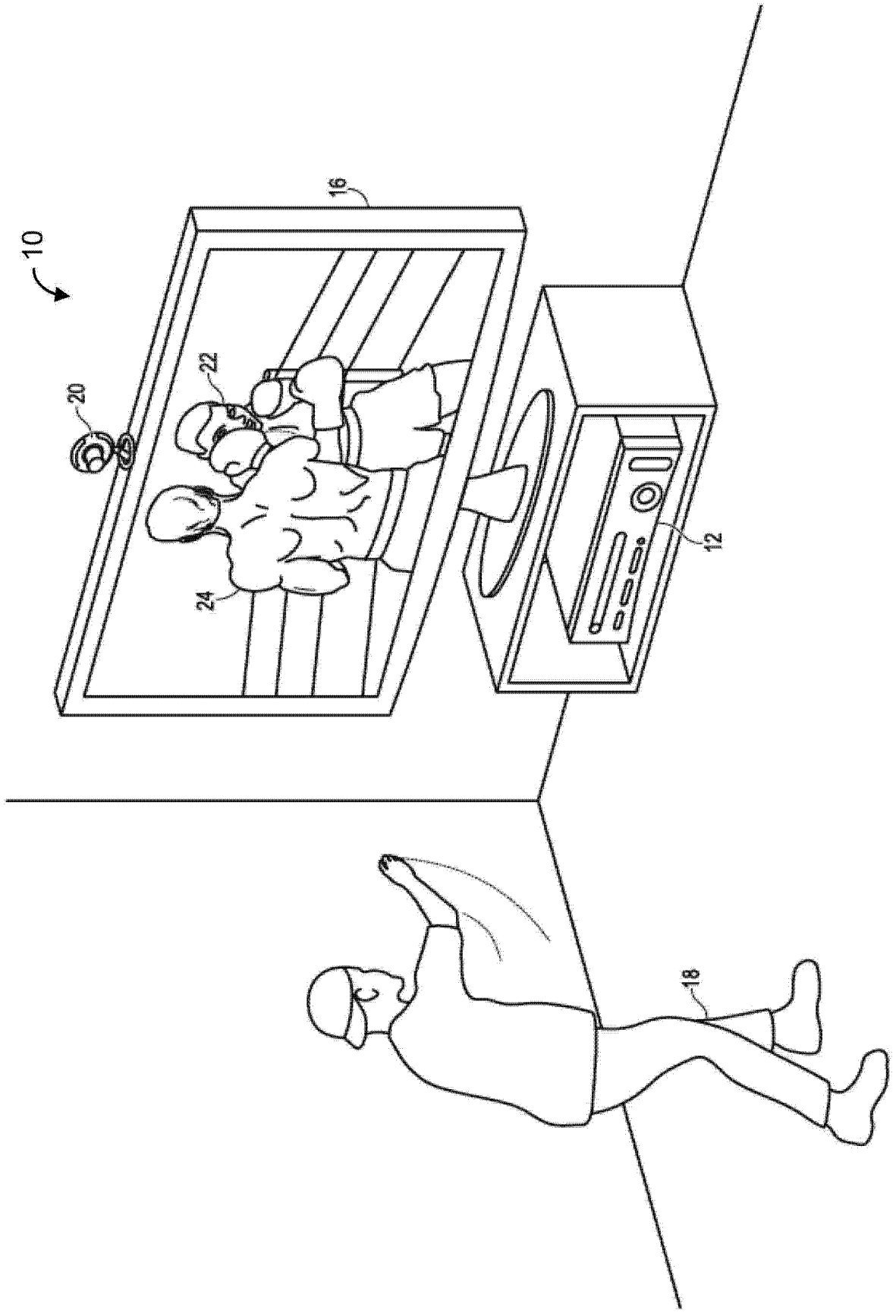


图 2

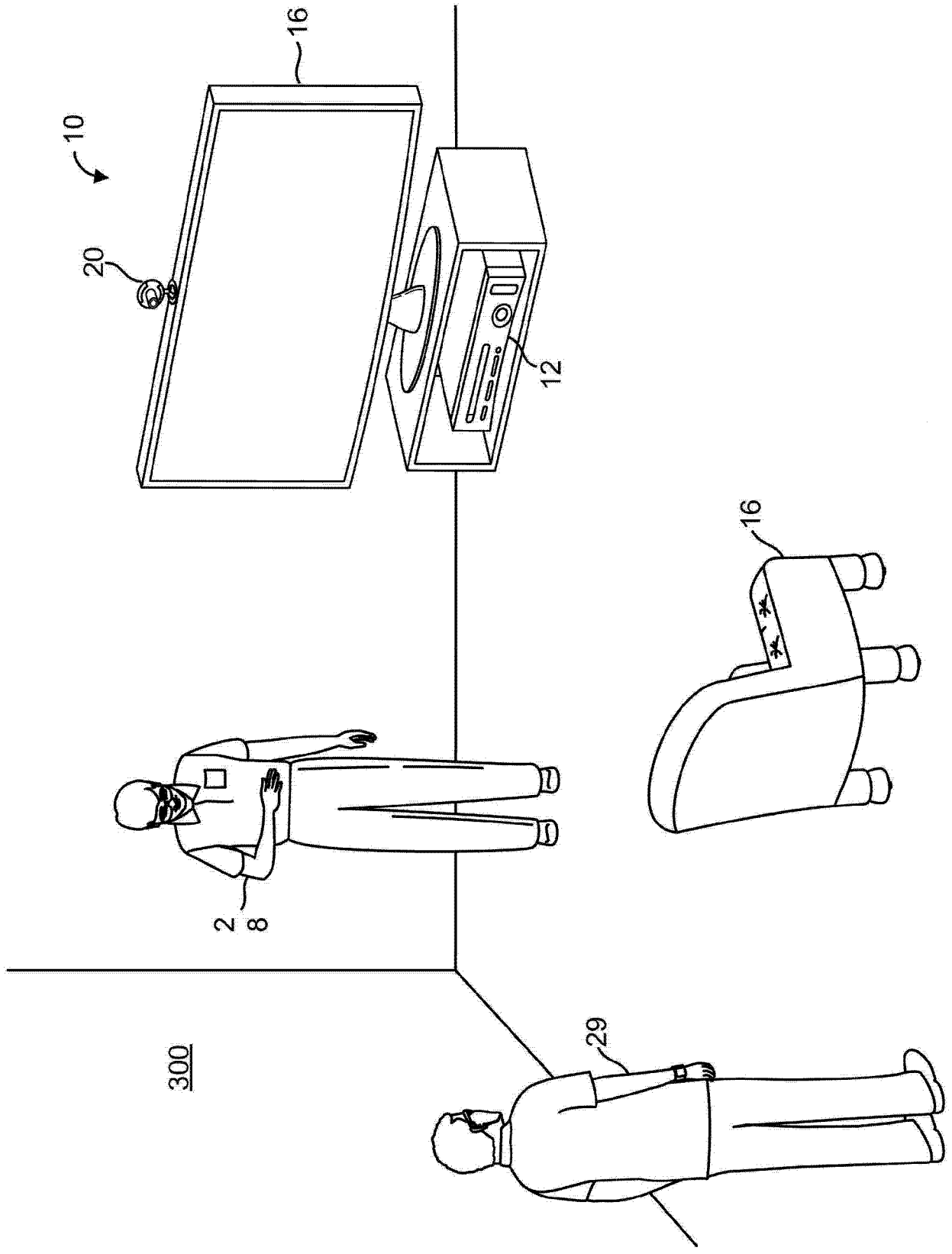


图 3

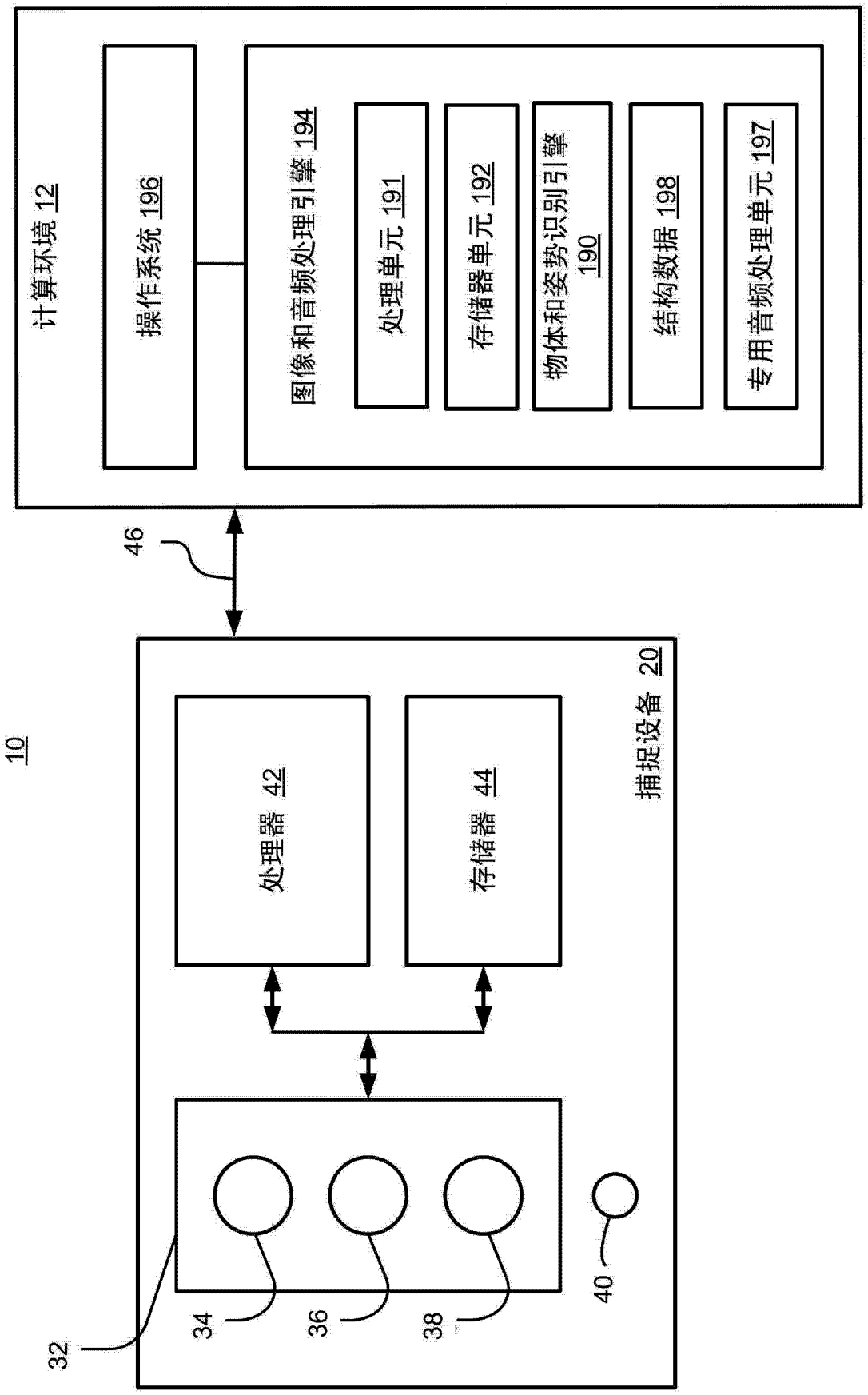

计算环境 12
操作系统 196
图像和音频处理引擎 194
处理单元 191
存储器单元 192
物体和姿势识别引擎 190
结构数据 198
专用音频处理单元 197
46
10
捕捉设备 20
处理器 42
存储器 44
32
34
36
38
40

图 4

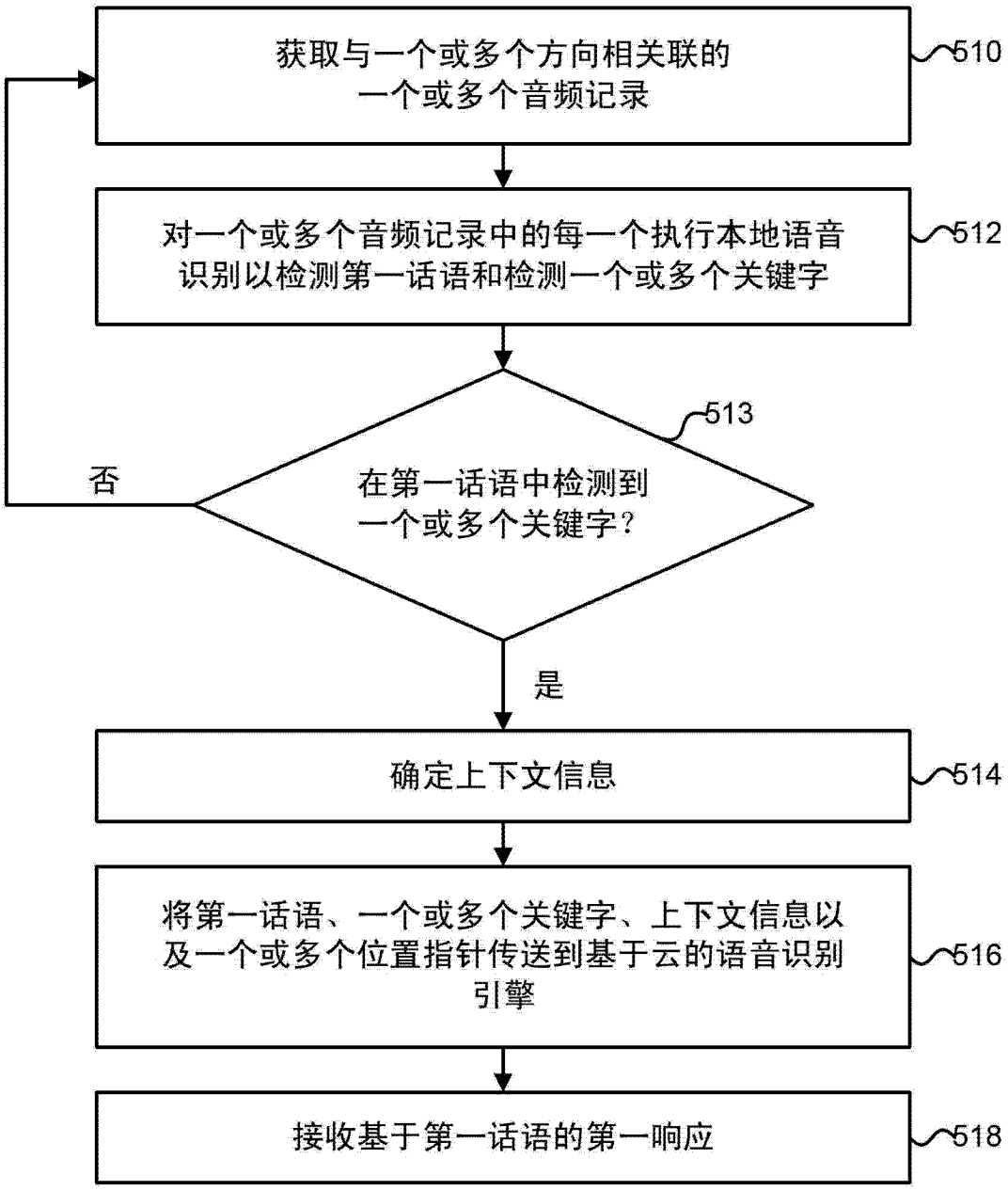

获取与一个或多个方向相关联的
一个或多个音频记录
510
对一个或多个音频记录中的每一个执行本地语音
识别以检测第一话语和检测一个或多个关键字
512
513
在第一话语中检测到
一个或多个关键字？
否
是
确定上下文信息
514
将第一话语、一个或多个关键字、上下文信息以
及一个或多个位置指针传送到基于云的语音识别
引擎
516
接收基于第一话语的第一响应
518

图 5A

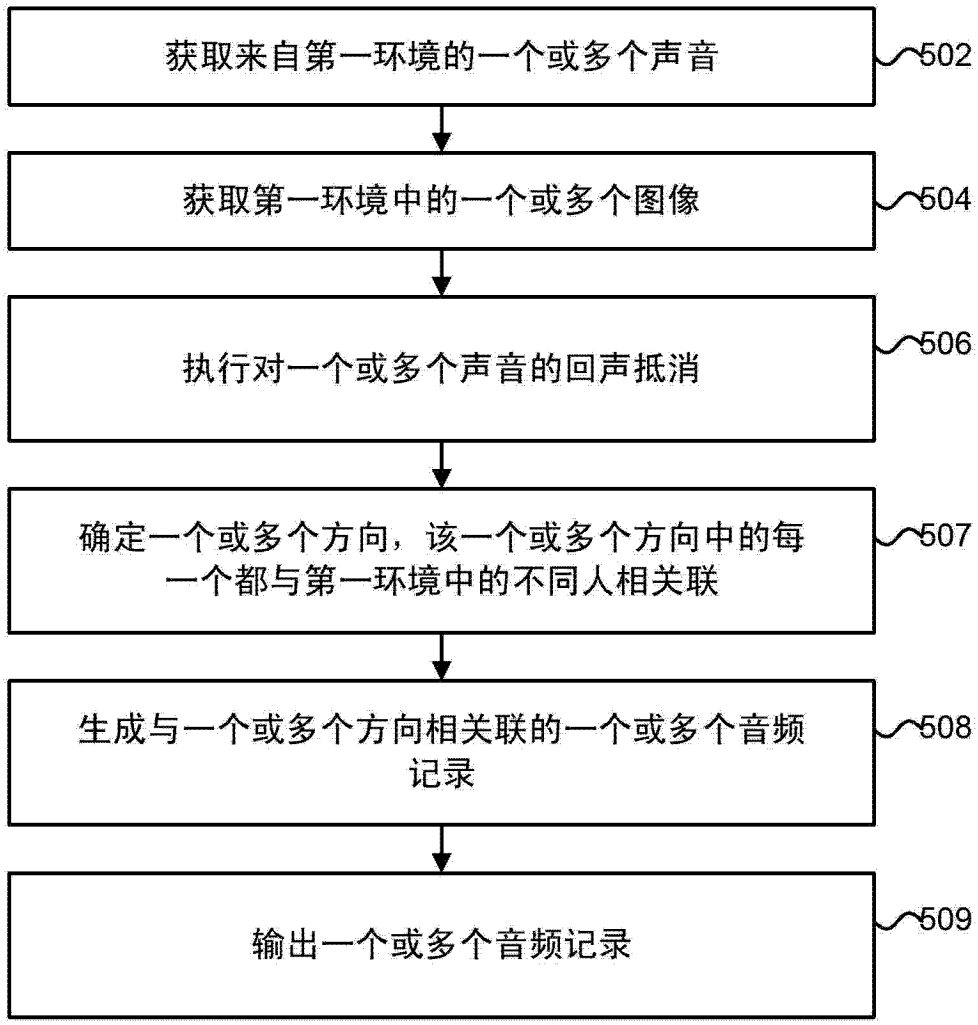

获取来自第一环境的一个或多个声音
502
获取第一环境中的一个或多个图像
504
执行对一个或多个声音的回声抵消
506
确定一个或多个方向，该一个或多个方向中的每一个都与第一环境中的不同人相关联
507
生成与一个或多个方向相关联的一个或多个音频记录
508
输出一个或多个音频记录
509

图 5B

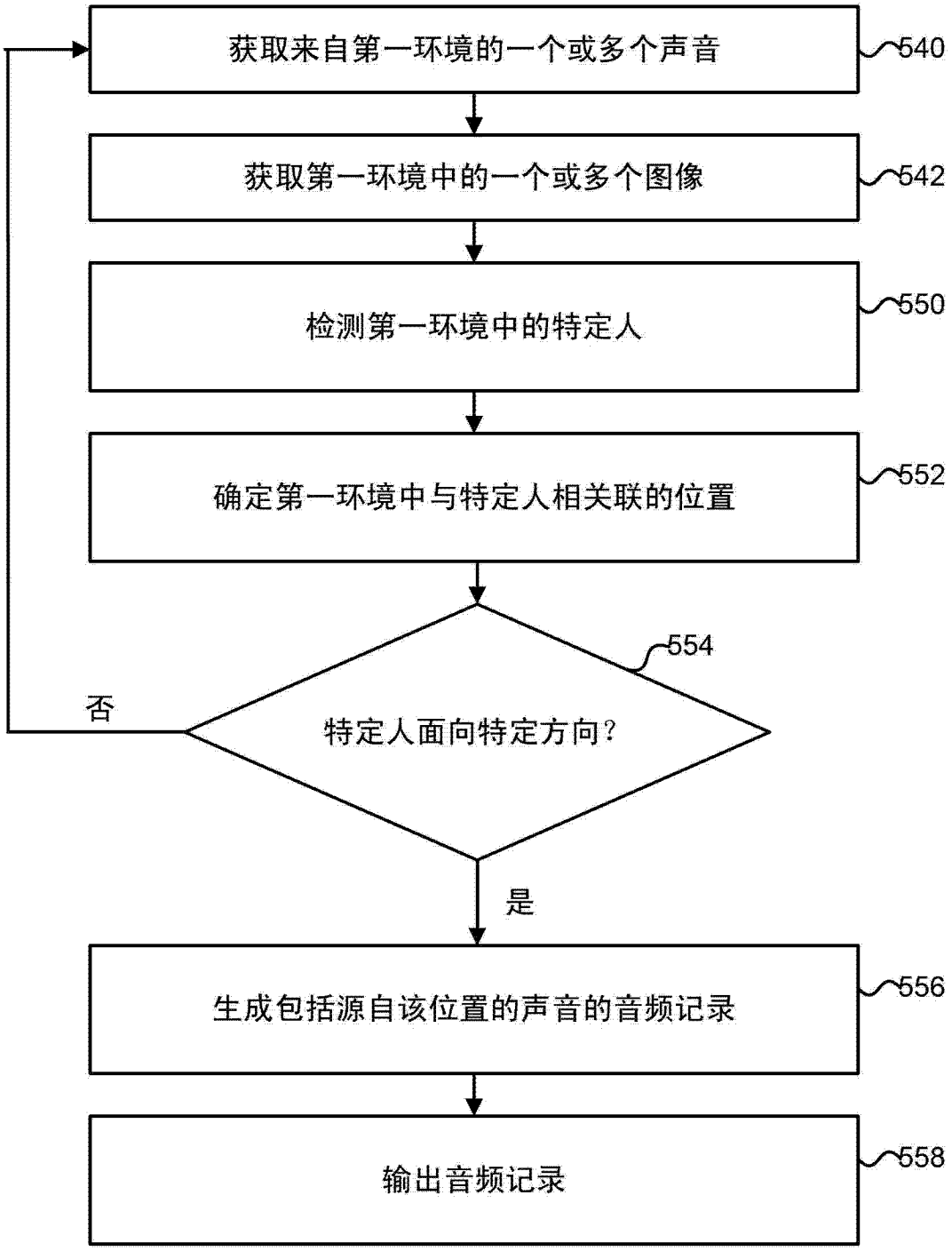

获取来自第一环境的一个或多个声音
540
获取第一环境中的一个或多个图像
542
检测第一环境中的特定人
550
确定第一环境中与特定人相关联的位置
552
554
特定人面向特定方向?
否
是
生成包括源自该位置的声音的音频记录
556
输出音频记录
558

图 5C

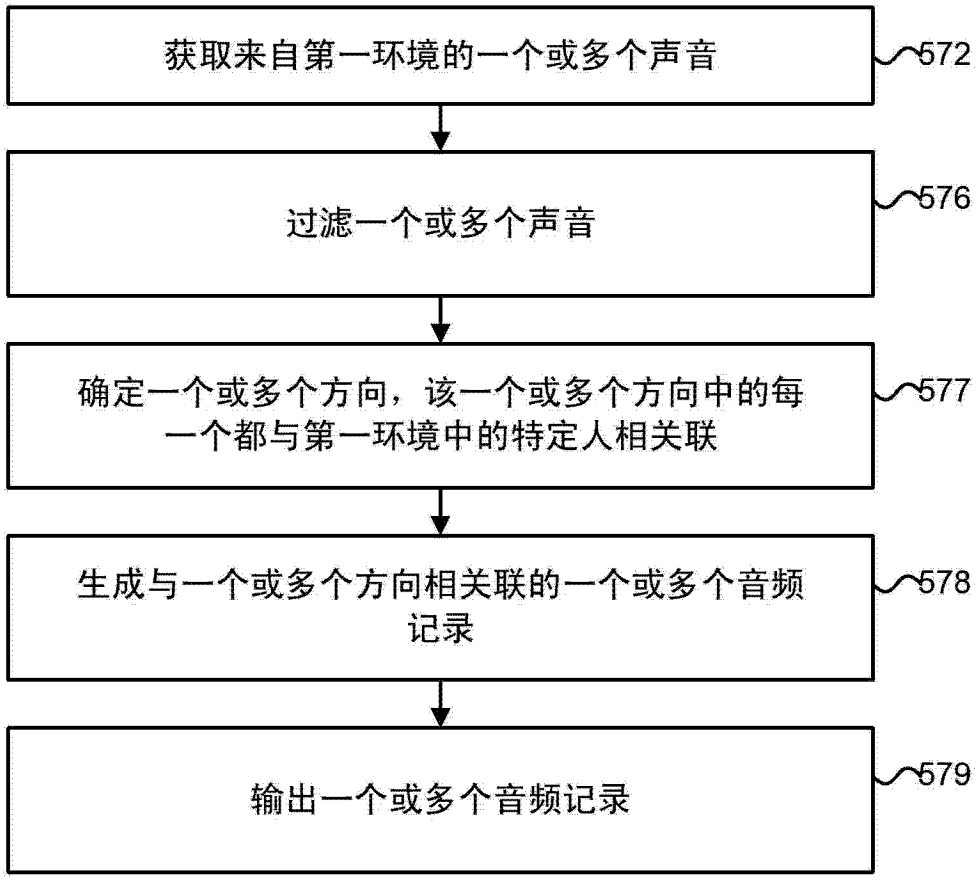

获取来自第一环境的一个或多个声音
572
过滤一个或多个声音
576
确定一个或多个方向，该一个或多个方向中的每一个都与第一环境中的特定人相关联
577
生成与一个或多个方向相关联的一个或多个音频记录
578
输出一个或多个音频记录
579

图 5D

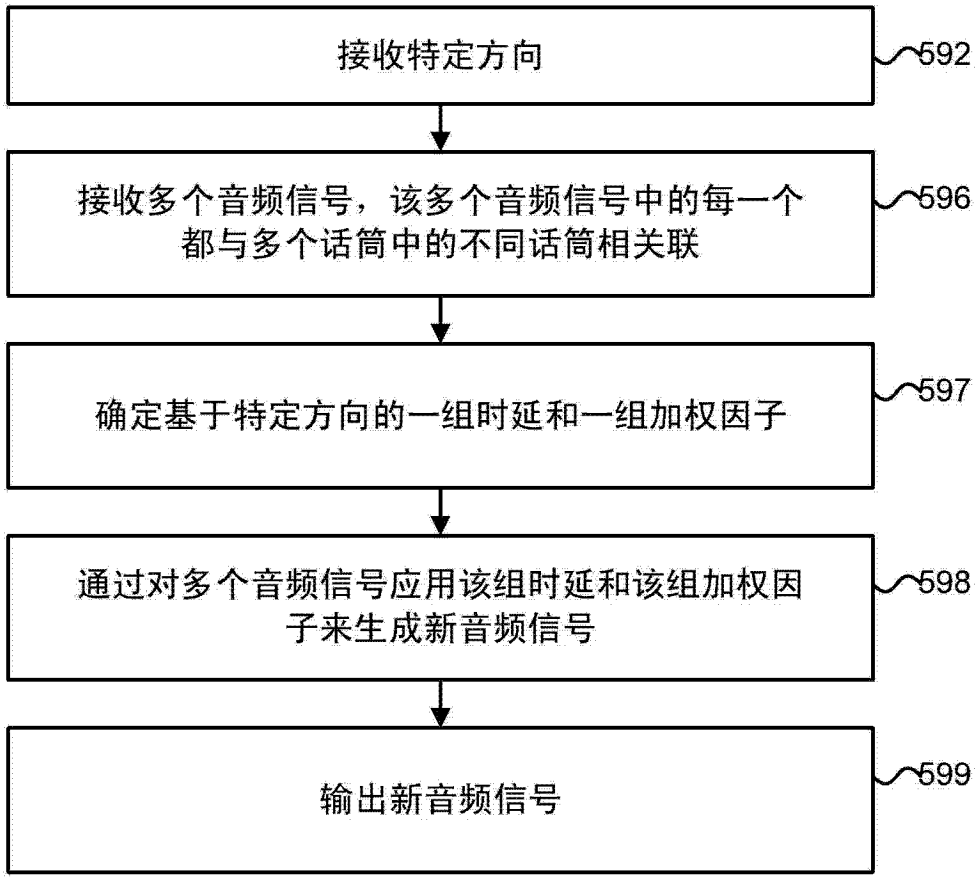

接收特定方向
592
接收多个音频信号，该多个音频信号中的每一个都与多个话筒中的不同话筒相关联
596
确定基于特定方向的一组时延和一组加权因子
597
通过对多个音频信号应用该组时延和该组加权因子来生成新音频信号
598
输出新音频信号
599

图 5E

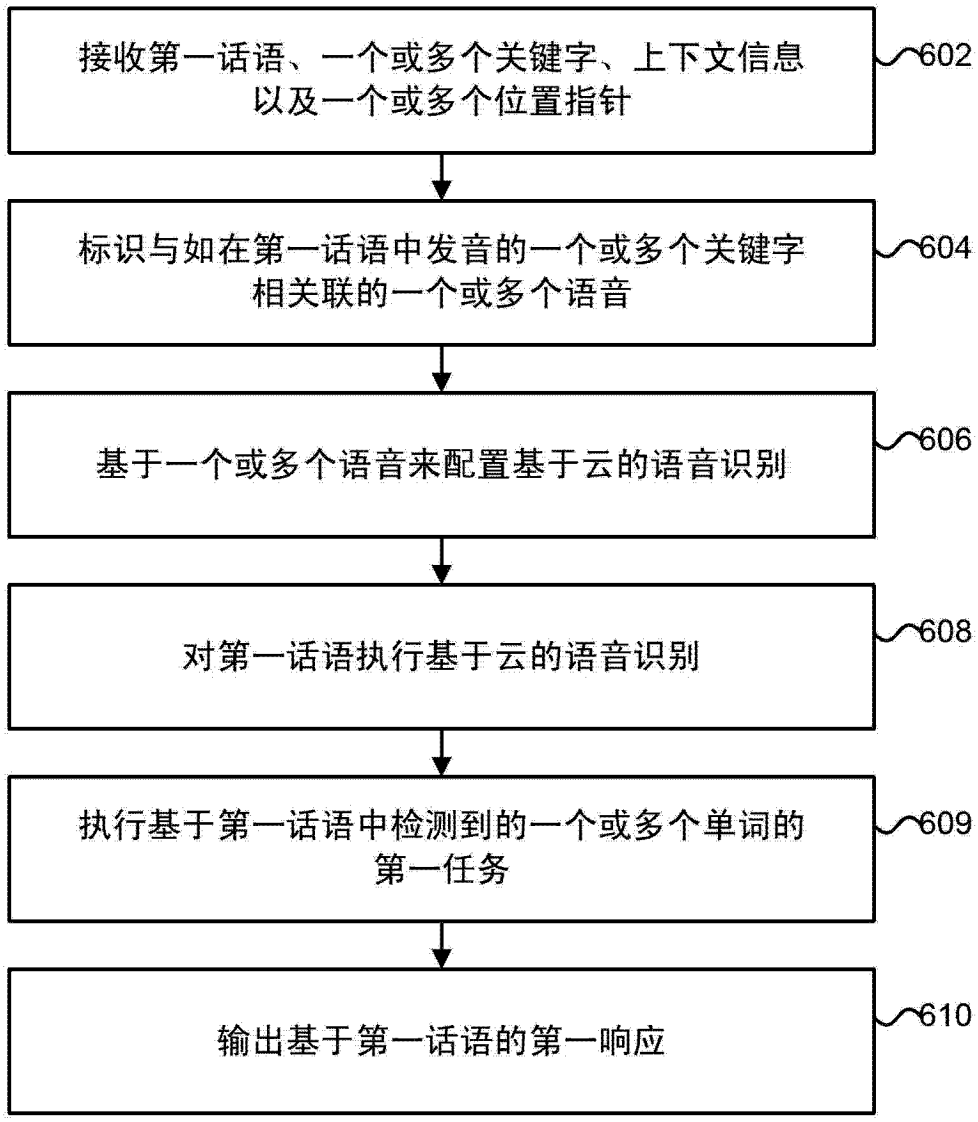

接收第一话语、一个或多个关键字、上下文信息以及一个或多个位置指针
602
标识与如在第一话语中发音的一个或多个关键字相关联的一个或多个语音
604
基于一个或多个语音来配置基于云的语音识别
606
对第一话语执行基于云的语音识别
608
执行基于第一话语中检测到的一个或多个单词的第一任务
609
输出基于第一话语的第一响应
610

图 6

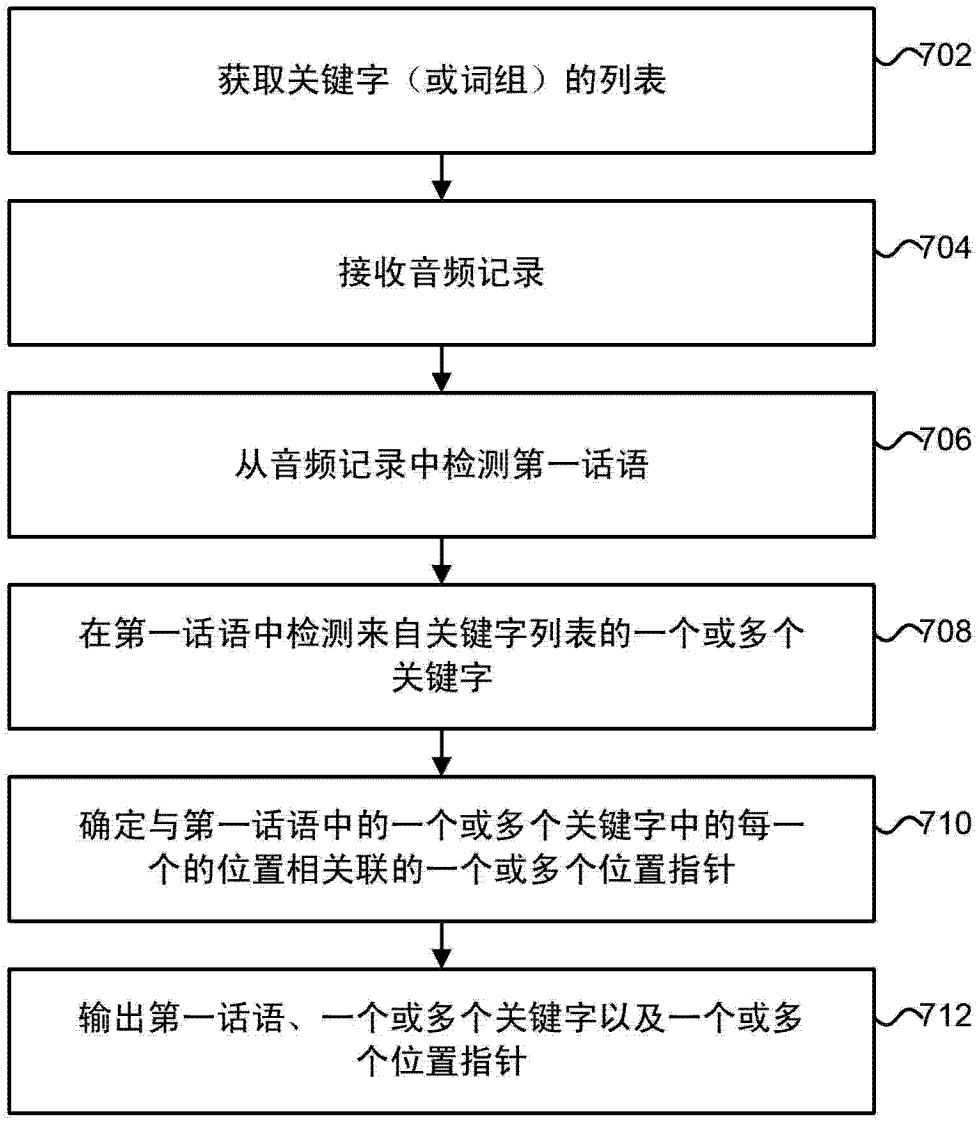

获取关键字（或词组）的列表
702
接收音频记录
704
从音频记录中检测第一话语
706
在第一话语中检测来自关键字列表的一个或多个关键字
708
确定与第一话语中的一个或多个关键字中的每一个的位置相关联的一个或多个位置指针
710
输出第一话语、一个或多个关键字以及一个或多个位置指针
712

图 7

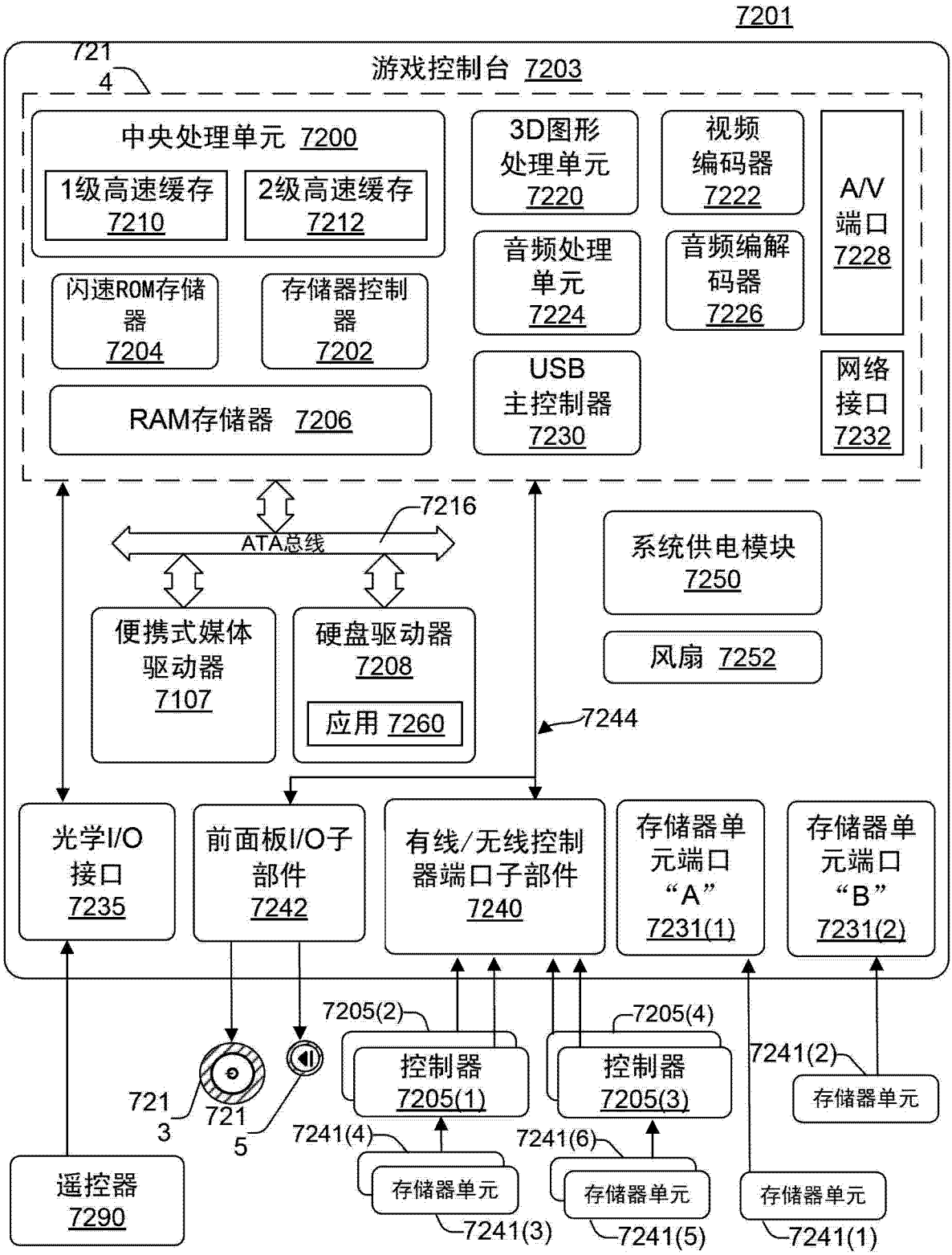

7201
7214
游戏控制台 7203
中央处理单元 7200
1级高速缓存 7210
2级高速缓存 7212
3D图形处理单元 7220
视频编码器 7222
A/V端口 7228
音频处理单元 7224
音频编解码器 7226
闪速ROM存储器 7204
存储器控制器 7202
USB主控制器 7230
网络接口 7232
RAM存储器 7206
7216
ATA总线
系统供电模块 7250
风扇 7252
便携式媒体驱动器 7107
硬盘驱动器 7208
应用 7260
7244
光学I/O接口 7235
前面板I/O子部件 7242
有线/无线控制器端口子部件 7240
存储器单元端口“A” 7231(1)
存储器单元端口“B” 7231(2)
7205(2)
7205(4)
控制器 7205(1)
控制器 7205(3)
7241(2)
存储器单元
7213
7215
7241(4)
7241(6)
存储器单元
存储器单元
存储器单元
遥控器 7290
7241(3)
7241(5)
7241(1)

图 8

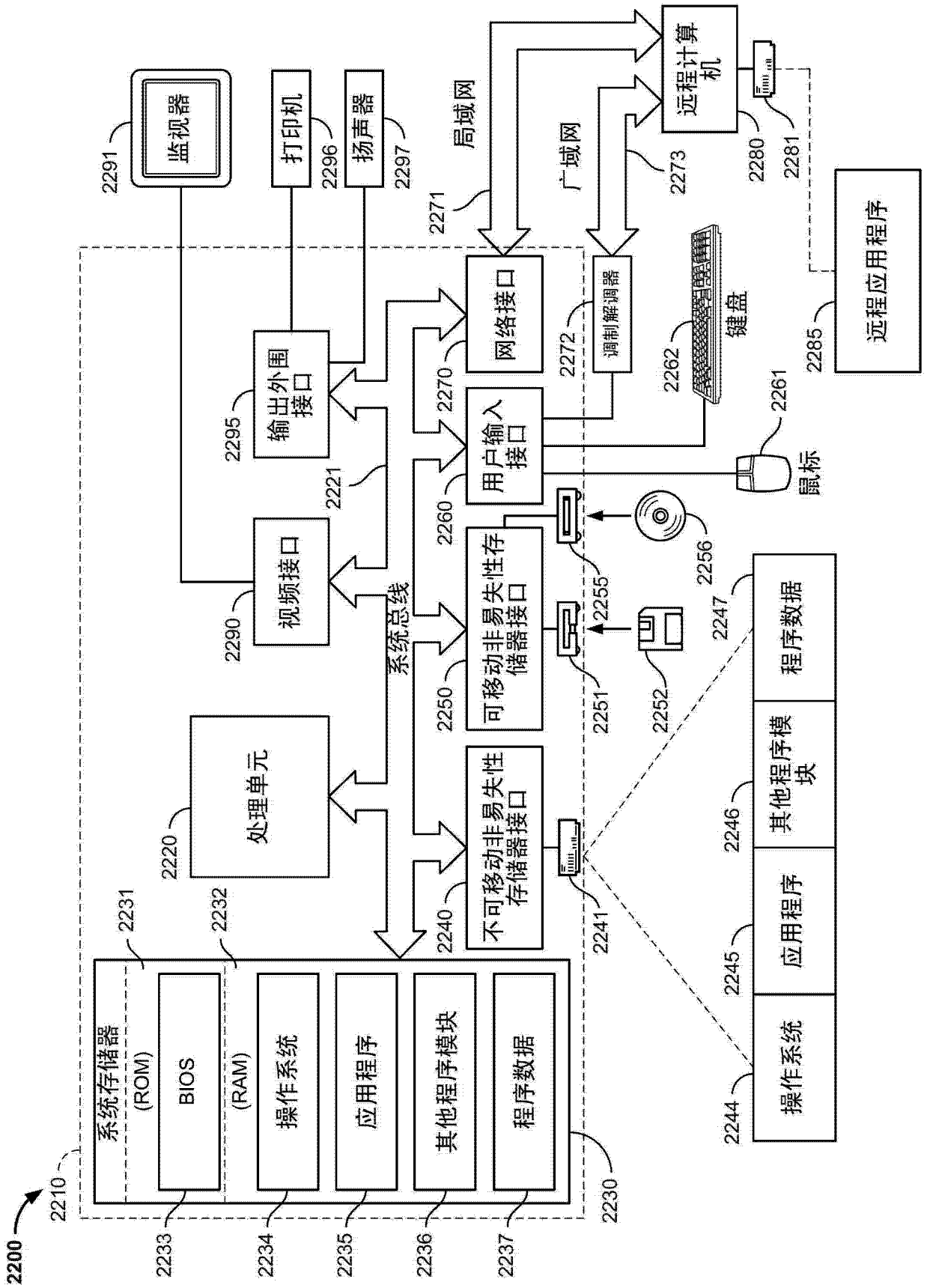

2200
2210
系统存储器
(ROM)
BIOS
(RAM)
操作系统
应用程序
其他程序模块
程序数据
2231
2232
2233
2234
2235
2236
2237
2230
处理单元
2220
系统总线
2221
视频接口
2290
输出外围接口
2295
打印机
2296
扬声器
2297
监视器
2291
网络接口
2270
用户输入接口
2260
2261
鼠标
2262
键盘
调制解调器
2272
局域网
2271
广域网
2273
远程计算机
2280
2281
远程应用程序
2285
可移动非易失性存储器接口
2250
2251
2252
2255
2256
不可移动非易失性存储器接口
2240
2241
2247
程序数据
其他程序模块
2246
应用程序
2245
操作系统
2244

图 9