



(12) 发明专利申请

(10) 申请公布号 CN 105324811 A

(43) 申请公布日 2016. 02. 10

(21) 申请号 201480026549. 6

(74) 专利代理机构 上海专利商标事务所有限公司 31100

(22) 申请日 2014. 05. 09

代理人 顾嘉运

(30) 优先权数据

13/892,094 2013. 05. 10 US

(51) Int. Cl.

G10L 15/26(2006. 01)

(85) PCT国际申请进入国家阶段日

G06F 1/16(2006. 01)

2015. 11. 10

(86) PCT国际申请的申请数据

PCT/US2014/037410 2014. 05. 09

(87) PCT国际申请的公布数据

W02014/182976 EN 2014. 11. 13

(71) 申请人 微软技术许可有限责任公司

地址 美国华盛顿州

(72) 发明人 D·麦克洛克 A·L·李

A·B·史密斯—基普尼斯

J·W·普鲁姆 A·戴维

M·O·黑尔 J·科尔

H·M·朗格拉克

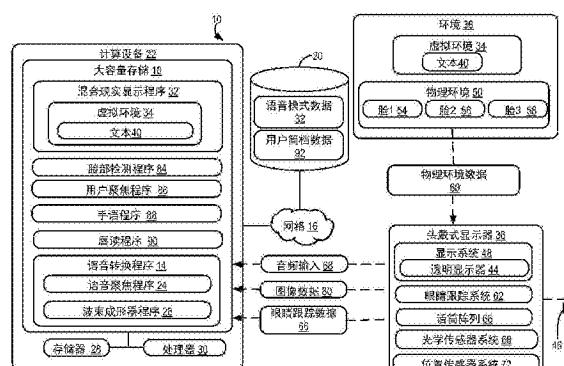
权利要求书2页 说明书10页 附图6页

(54) 发明名称

语音到文本转换

(57) 摘要

公开了涉及将来自环境的音频输入转换成文本的各实施例。例如，在一个公开的实施例中，语音转换程序接收来自头戴式显示设备的话筒阵列的音频输入。从环境中捕捉图像数据，并且从图像数据中检测一个或多个可能的脸。眼睛跟踪数据被用于确定用户聚焦于的目标脸。波束成形技术被应用于音频输入的至少一部分以标识与目标脸相关联的目标音频输入。这些目标音频输入被转换成通过头戴式显示设备的透明显示器来显示的文本。



1. 一种语音转换系统，包括：

操作上连接到计算设备的头戴式显示设备，所述头戴式显示设备包括：

包括透明显示器的显示系统；

用于跟踪用户的眼睛的注视的眼睛跟踪系统；

包括被刚性安置在所述头戴式显示设备上以接收音频输入的多个话筒的话筒阵列；以及

用于捕捉图像数据的一个或多个图像传感器；

由所述计算设备的处理器执行的脸部检测程序，所述脸部检测程序被配置为从所述图像数据检测一个或多个可能的脸；

由所述计算设备的处理器执行的用户聚焦程序，所述用户聚焦程序被配置为使用来自所述眼睛跟踪系统的眼睛跟踪数据来确定用户聚焦于的目标脸；以及

由所述计算设备的处理器执行的语音转换程序，所述语音转换程序被配置为使用应用于来自所述话筒阵列的音频输入的至少一部分的波束成形技术来标识与所述目标脸相关联的目标音频输入以供用于语音到文本转换。

2. 如权利要求 1 所述的语音转换系统，其特征在于，所述脸部检测程序还被配置为确定与目标脸相关联的身份，并且所述语音转换程序还被配置为：

将所述目标音频输入转换成文本；以及

通过头戴式显示设备的透明显示器显示所述文本。

3. 如权利要求 2 所述的语音转换系统，其特征在于，显示的文本被地理定位在所述环境内。

4. 如权利要求 2 所述的语音转换系统，其特征在于，所述语音转换程序被进一步配置成：

访问对应于与所述目标脸相关联的所述身份的语音模式数据；以及

使用所述语音模式数据来将所述目标音频输入转换成所述文本。

5. 如权利要求 2 所述的语音转换系统，其特征在于，显示的文本被标记给对应于所述身份的人。

6. 一种用于将来自环境的音频输入转换成文本的方法，所述音频输入是在头戴式显示设备的话筒阵列处被接收到的，包括：

从所述环境中捕捉图像数据；

从所述图像数据中检测一个或多个可能的脸；

使用来自所述头戴式显示设备的眼跟踪系统的眼跟踪数据来确定用户所聚焦于的目标脸；以及

使用应用于来自所述话筒阵列的音频输入的至少一部分的波束成形技术来标识与所述目标脸相关联的目标音频输入以供用于语音到文本转换。

7. 如权利要求 6 所述的方法，其特征在于，进一步包括：

从所述环境中接收来自一个或多个外部源的附加音频输入；以及

使用所述附加音频输入来标识所述目标音频输入。

8. 如权利要求 6 所述的方法，其特征在于，进一步包括：

从所述图像数据中标识一个或多个手语字母和单词；

将所述字母和单词转换成符号文本；以及  
通过所述头戴式显示设备的透明显示器显示所述符号文本。

9. 如权利要求 6 所述的方法，其特征在于，进一步包括：

从所述目标脸的所述图像数据标识所述目标脸的一个或多个嘴唇和舌头的移动；

将所述移动转换成唇读文本；以及

通过所述头戴式显示设备的透明显示器显示所述唇读文本。

10. 如权利要求 6 所述的方法，其特征在于，所述多个话筒包括全向话筒，并且还包括：  
标识在一个或多个所述全向话筒处接收到语音的位置；以及

使用应用于在所述一个或多个全向话筒处接收到的语音的波束成形技术标识与所述位置相关联的目标音频输入。

## 语音到文本转换

[0001] 背景

[0002] 具有听觉障碍的人可以使用一种或多个技术来理解源自另一个人或设备的可听到的语音和 / 或其它声音。例如，在讲话者正在讲话且有听觉障碍的人可以看到讲话者的嘴巴的情况下，这个人可以使用唇读技术来理解该语音的内容。然而，要使用这样的技术需要这个人学习唇读技术。而且，在这个人对讲话者的查看受到限制或阻挡的情况下，这样的技术不能够提供令人满意的协助。

[0003] 另一种可能性是由第三方将语音翻译成特定的手语，该手语可以被掌握该手语的人理解。第三方还可以将语音改写成可由这个人阅读的书面形式。然而，使第三方可用于执行这样的翻译和改写强加了相当大的限制。

[0004] 另一个方案可以使用语音识别技术来接收语音、解释语音和将语音可视地呈现给有听觉障碍的人。然而，当讲话者没有清楚直接地对着接收话筒讲话和 / 或当背景噪声过多时，这种技术的准确度通常变坏。因此，并且尤其在嘈杂和拥挤的环境中，这样的技术可能是不切实际的且少有帮助。而且，有听力能力的人也可能遇到涉及许多人和 / 或过多噪声的情形，例如社交聚会、贸易展等，在这些情形中难以或不可能听到另一个人的语音。

[0005] 概述

[0006] 本文公开了与语音转换系统有关的各实施例。例如，一个公开的实施例提供了一种用于将来自环境的音频输入转换成文本的方法。所述方法包括捕捉来自环境的图像数据并从该图像数据检测一个或多个可能的脸。来自头戴式显示设备的眼睛跟踪系统的眼睛跟踪数据被用于确定用户所聚焦于的目标脸。

[0007] 一种波束成形技术可以被应用到来自头戴式显示设备的话筒阵列的音频输入中以标识与目标脸相关联的目标音频输入。所述方法包括将目标音频输入转换成文本。所述方法还包括通过头戴式显示设备的透明显示器来显示文本。

[0008] 提供该概述以便以简化形式介绍概念的选集，所述概念在以下详细描述中被进一步描述。本概述并不旨在标识所要求保护主题的关键特征或必要特征，也不旨在用于限制所要求保护主题的范围。而且，所要求保护的主题不限于解决该公开的任一部分中所注的任何或全部缺点的实现方式。

[0009] 附图简述

[0010] 图 1 是根据本公开的一实施例的语音转换系统的示意图。

[0011] 图 2 示出根据本公开的一实施例的示例头戴式显示设备。

[0012] 图 3 是房间中佩戴图 2 的头戴式显示设备的用户以及三个其他人的示意性透视图。

[0013] 图 4A 和 4B 示出了用于根据本公开的一实施例来将来自环境的音频输入转换成文本的方法的流程图。

[0014] 图 5 是计算设备的一实施例的简化示意图解。

[0015] 详细描述

[0016] 图 1 示出了语音转换系统 10 的一个实施例的示意图。语音转换系统 10 包括可被

存储在计算设备 22 的大容量存储 18 中的语音转换程序 14。如以下更详细描述的，语音转换程序 14 可以包括语音聚焦程序 24 和波束成形器程序 26。

[0017] 语音转换程序 14 可被加载到存储器 28 中并由计算设备 22 的处理器 30 执行以执行下文更为详细地描述的方法和过程中的一个或多个。又如下文更加详细描述的，大容量存储 18 还可以包括脸部检测程序 84、用户聚焦程序 86、手语程序 88 以及唇读程序 90。

[0018] 语音转换系统 10 包括混合现实显示程序 32，该混合现实显示程序 32 可生成用于经由显示设备（诸如头戴式显示器（HMD）设备 36）显示的虚拟环境 34 以创建混合现实环境。虚拟环境 34 可包括一个或多个虚拟对象。这样的虚拟对象可包括一个或多个虚拟图像（诸如三维全息图像）和其他虚拟对象（诸如二维虚拟对象）。如下文更加详细描述的，这样的虚拟对象可以包括已经从由 HMD 设备 36 所接收的目标音频输入中生成的文本 40。

[0019] 计算设备 22 可采用以下形式：台式计算设备，诸如智能电话、膝上型计算机、笔记本或平板计算机之类的移动计算设备，网络计算机，家庭娱乐计算机，交互式电视，游戏系统，或其他合适类型的计算设备。关于计算设备 22 的组件和计算方面的附加细节在下文中参考图 5 更详细地描述。

[0020] 计算设备 22 可使用有线连接来与 HMD 设备 36 在操作上连接，或可采用经由 WiFi、蓝牙或任何其他合适的无线通信协议的无线连接。例如，计算设备 22 可通信地耦合到网络 16。网络 16 可采取局域网（LAN）、广域网（WAN）、有线网络、无线网络、个域网、或其组合的形式，并且可包括因特网。

[0021] 如以下更详细描述的，计算设备 22 可经由网络 16 与一个或多个其它 HMD 设备和其它计算设备（诸如服务器 20）通信。另外，图 1 中示出的示例将计算设备 22 示为与 HMD 设备 36 分开的组件。将理解，在其他示例中，计算设备 22 可被集成到 HMD 设备 36 中。

[0022] 现在还参考图 2，提供了一副具有透明显示器 44 的可配戴眼镜形式的 HMD 设备 200 的示例。将明白，在其他示例中，HMD 设备 200 可以采取其他合适的形式，其中透明、半透明或不透明显示器被支撑在查看者的一只或两只眼睛前方。还将明白，图 1 中所示的 HMD 设备 36 可以采取 HMD 设备 200 的形式（如在下文更详细地描述的）或任何其他合适的 HMD 设备。另外，在本公开的范围内还可使用具有各种形状因子的许多其他类型和配置的显示设备。此类显示设备可包括但不限于手持式智能电话、平板计算机以及其他适当的显示设备。

[0023] 参考图 1 和 2，HMD 设备 36 包括显示系统 48 和使图像（诸如全息对象）能够被递送到用户 46 的眼睛的透明显示器 44。透明显示器 44 可被配置成向透过该透明显示器查看物理环境的用户 46 在视觉上增强该物理环境 50 的外观。例如，物理环境 50 的外观可以由经由透明显示器 44 呈现的图形内容（例如，一个或多个像素，每一像素具有相应色彩和亮度）来增强以创建混合现实环境。

[0024] 透明显示器 44 还可被配置成使用户能够透过显示虚拟对象表示的一个或多个部分透明的像素来查看物理环境 50 中的物理现实世界对象（诸如，脸 1 54、脸 2 56 和脸 3 58）。如图 2 所示，在一个示例中，透明显示器 44 可包括位于透镜 204 内的图像生成元件（诸如例如透视有机发光二极管（OLED）显示器）。作为另一示例，透明显示器 44 可包括在透镜 204 边缘上的光调制器。在这一示例中，透镜 204 可以担当光导以供将光从光调制器递送到用户的眼睛。这样的光导可使得用户能够感知位于物理环境 50 内的用户正在查看

的 3D 全息图像，同时还允许用户查看物理环境中的物理对象，由此创建混合现实环境。

[0025] HMD 设备 36 还可包括各种传感器和相关系统。例如，HMD 设备 36 可包括利用至少一个面向内的传感器 216 的眼睛跟踪系统 62。该面向内的传感器 216 可以是被配置成从用户的眼睛获取眼睛跟踪数据 66 形式的图像数据的图像传感器。假定用户已同意获取和使用这一信息，眼睛跟踪系统 62 可以使用这一信息来跟踪用户的眼睛的位置和 / 或运动。

[0026] 在一个示例中，眼睛跟踪系统 62 包括被配置成检测用户的每一个眼睛的注视方向的注视检测子系统。该注视检测子系统可被配置成以任何合适方式确定每一只用户眼睛的注视方向。例如，注视检测子系统可包括诸如红外光源等被配置成使得从用户的每一只眼睛反射闪光的一个或多个光源。一个或多个图像传感器然后可被配置成捕捉用户眼睛的图像。

[0027] 如从收集自图像传感器的图像数据确定的闪烁和瞳孔的图像可用于确定每一眼睛的光轴。使用该信息，眼睛跟踪系统 62 随后可以确定用户正在注视的方向和 / 或用户正注视着什么物理对象或虚拟对象。这样的眼睛跟踪数据 66 可随后被提供给计算设备 22。将理解，注视检测子系统可以具有任意适当数量和布置的光源和图像传感器。

[0028] HMD 设备 36 还可包括从物理环境 50 接收物理环境数据 60 的传感器系统。例如，HMD 设备 36 可包括利用至少一个面向外的传感器 212（如光学传感器）的光学传感器系统 68。面向外的传感器 212 可以检测其视野内的运动，如视野内的用户 46 或人或物理对象所执行的基于姿势的输入或其他运动。面向外的传感器 212 还可从物理环境 50 和该环境内的物理对象捕捉二维图像信息和深度信息。例如，面向外的传感器 212 可包括深度相机、可见光相机、红外光相机，和 / 或位置跟踪相机。

[0029] HMD 设备 36 可包括经由一个或多个深度相机的深度感测。在一个示例中，每一深度相机可包括立体视觉系统的左和右相机。来自这些深度相机中的一个或多个的时间分辨的图像可被彼此配准和 / 或与来自另一光学传感器（如可见光谱相机）的图像配准，且可被组合以产生深度分辨的视频。

[0030] 在其他示例中，结构化光深度相机可被配置成投影结构化红外照明并对从照明被投影到其之上的场景中反射的该照明进行成像。基于所成像的场景的各个区域内邻近特征之间的间隔，可构造该场景的深度图。在其他示例中，深度相机可以采取飞行时间深度相机的形式，其被配置成将脉冲的红外照明投影到该场景上以及检测从该场景反射的照明。可以理解，在本发明的范围内可使用任意其他合适的深度相机。

[0031] 面向外的传感器 212 可以捕捉用户 46 位于其中的物理环境 50 的图像。在一个示例中，混合现实显示程序 32 可包括使用这样的输入来生成对围绕该用户 46 的物理环境 50 进行建模的虚拟环境 34 的 3D 建模系统。

[0032] HMD 设备 36 还可包括位置传感器系统 72，该位置传感器系统 72 利用一个或多个运动传感器 220 来实现对 HMD 设备的运动检测、位置跟踪和 / 或取向感测。例如，位置传感器系统 64 可被用来确定用户的头部的方向、速度和加速度。位置传感器系统 64 还可被用来确定用户的头部的姿态取向。在一个示例中，位置传感器系统 64 可包括配置成六轴或六自由度的位置传感器系统的惯性测量单元。这一示例位置传感器系统可以例如包括用于指示或测量 HMD 设备 36 在三维空间内沿三个正交轴（例如，x、y、z）的位置变化以及该 HMD 设备绕三个正交轴（例如，翻滚、俯仰、偏航）的取向变化的三个加速度计和三个陀螺

仪。

[0033] 位置传感器系统 64 还可以支持其他合适的定位技术,如 GPS 或其他全球导航系统。而且,尽管描述了位置传感器系统的具体示例,但将明白,可以使用其他合适的位置传感器系统。在一些示例中,运动传感器 220 还可以被用作用户输入设备,使得用户可以经由颈部和头部或者甚至身体的姿势来与 HMD 设备 36 交互。

[0034] HMD 设备 36 还可包括包含一个或多个刚性安装在 HMD 设备上的话筒的话筒阵列 66。在图 2 示出的示例中,提供了有 6 个话筒 224、228、232、236、240 以及 244 的阵列,当用户佩戴 HMD 设备 200 时,这些话筒被定位在用户头部周围的各个位置处。在一个示例中,所有的 6 个话筒 224、228、232、236、240 以及 244 可以是被配置为接收来自物理环境 50 的语音和其它音频输入的全向话筒。

[0035] 在另一个示例中,话筒 224、228、232 以及 236 可以是全向话筒,而话筒 240 和 244 可以是被配置为接收来自佩戴了 HMD 设备 200 的用户 46 的语音的单向话筒。还将理解的是,在其它示例中,HMD 设备 200 周围的话筒的数目、类型和 / 或位置可以是不同的,并且可以使用任何合适的数目、类型和布置的话筒。在又一个其他示例中,音频可经由 HMD 设备 36 上的一个或多个扬声器 248 被呈现给用户。

[0036] HMD 设备 36 还可包括具有与 HMD 设备的各传感器和系统通信的逻辑子系统和存储子系统的处理器 250,如在下文参考图 5 更详细地讨论的。在一个示例中,存储子系统可包括可由逻辑子系统执行的指令,用以接收来自传感器的信号输入并将此类输入转发到计算设备 22(以未经处理或经处理的形式)并且经由透明显示器 44 向用户呈现图像。

[0037] 要领会,HMD 设备 36 和相关的传感器以及上面描述的并在图 1 和 2 中解说的其他组件是作为示例来提供的。这些示例不旨在以任何方式进行限制,因为任何其他合适的传感器、组件,和 / 或传感器和组件的组合可被使用。因此,将理解,HMD 设备 36 可以包括未偏离本公开文本范畴的附加和 / 或替代的传感器、相机、话筒、输入设备、输出设备等。此外,HMD 设备 36 的物理配置及其各种传感器和子组件可以采取不偏离本公开文本范畴的各种不同形式。

[0038] 现参考图 3,现在将提供对语音转换系统 10 的示例用例和实施例的描述。图 3 提供了用户 304 的各个示意性图示,该用户 304 位于包括房间 308 的物理环境中并经由 HMD 设备 200 形式的 HMD 设备 36 体验混合现实环境。通过 HMD 设备 200 的透明显示器 44 观看房间 308,用户 304 可以具有包括具有脸 1 54 的第一人 316、具有脸 2 56 的第二人 320 以及具有脸 3 58 的第三人 324 的视野 312。壁挂式显示器 328 也可以在用户 304 的视野 312 内。

[0039] HMD 设备 200 的光学传感器系统 68 可以从房间 308 捕捉图像数据 80,包括表示一个或多个可能的脸(例如脸 1 54、脸 2 56 和脸 3 58)的图像数据。计算设备 22 的脸部检测程序 84 可以从图像数据 80 中检测脸 1 54、脸 2 56 和脸 3 58 的一个或多个。为了检测图像数据中的脸部图像,脸部检测程序 84 可使用任何合适的脸部检测技术和 / 或算法,包括但不限于局部二元图(LBP)、主分量分析(PCA)、独立分量分析(ICA)、进化追击(EP)、弹性束图匹配(EBGM)或其他合适的算法或算法组合。

[0040] 在一些示例中,用户 304 可具有听觉障碍,该听觉障碍使得理解语音变得困难,特别是在具有多个讲话者和 / 或显著的背景噪声的环境中。在如图 3 所示的示例中,第一人

316、第二人 320，以及第三人 324 中的每个可以同时在讲话。壁挂式显示器 328 还可以发出音频。所有这些音频输入 68 可以由 HMD 设备 200 的话筒 224、228、232、236、240 以及 244 接收。

[0041] 在一个示例中，用户 304 可以期望听到第一人 316 和 / 或与第一人 316 交谈。用户 304 可以注视第一人 316，如注视线 332 所指示的。对于用户注视的眼睛跟踪数据 66 可以由眼睛跟踪系统 62 捕捉并被提供给计算设备 22。使用眼睛跟踪数据 66 和图像数据 80，用户聚焦程序 86 可以确定用户 304 聚焦于第一人 316 的脸 1 54，该脸被指定为目标脸。

[0042] 可以确定第一人 316 的目标脸 1 54 相对于 HMD 设备 200 的位置。在一些示例中，眼睛跟踪数据 66、图像数据 80（例如由光学传感器系统 68 接收的深度信息和 / 或由位置传感器系统 72 生成的位置信息）可以被用于确定脸 1 54 的位置。

[0043] 使用目标脸 1 54 的位置，语音转换程序 14 可以使用波束成形器程序 26 将一种或多种波束成形技术应用于来自话筒阵列 66 的音频输入 68 的至少一个部分。或者可表示为，一种或多种波束成形技术可以被应用于音频输入 68 的源自目标脸 1 54 的位置的各部分。通过这种方式，波束成形器程序 26 可以标识与第一人 316 的脸 1 54 相关联的目标音频输入，其通常在图 3 中的 336 处指示。在图 3 中示出的示例中，目标音频输入可以对应于说着“*I'm speaking in the ballroom at 3:00 this afternoon*（我今天下午 3:00 在宴会厅演讲）”的第一人 316。

[0044] 在其它示例中，语音聚焦程序 24 可以利用在话筒阵列 66 中的每个话筒处接收到的音频输入 68 的时间差来确定接收各声音的方向。例如，语音聚焦程序 24 可以相对于 HMD 设备 200 的一个或多个全向话筒标识从其接收到讲话的位置。使用波束成形器程序 26，语音转换程序 14 可以随后将波束成形技术应用于语音并标识与接收语音的位置相关联的目标音频输入。

[0045] 在一些示例中，波束成形器程序 26 可以被配置为形成可用任何合适的方式确定的单个、定向自适应声音信号。例如，可以基于不随时间改变的波束成形技术、自适应波束成形技术或不随时间改变和自适应波束成形技术的组合来确定定向自适应声音信号。所得到的组合信号可以具有窄方向性模式，该模式可以以语音源的方向前进，例如第一人 316 的脸 1 54 的位置。将可以理解的是，可以使用任何合适的波束成形技术来标识与目标脸相关联的目标音频输入。

[0046] 继续参考图 1 和 3，语音转换程序 14 可以被配置成将目标音频输入转换成文本 40，并通过 HMD 200 的透明显示器 44 显示文本 40。在图 3 所示的示例中，可以将目标音频输入 336 转换成文本 40'，该文本由 HMD 设备 200 以文本气泡形式显示在第一人 316 的头部之上，因此，允许用户 304 很容易地将该文本与第一人相关联。

[0047] 在一个示例中，语音转换程序 14 可以向第一人 316 标记文本 40'，这样，文本气泡 340 在空间上被锚定到第一人，并在第一人移动时跟随第一人。在另一个示例中且如下文更加详细所述的，与目标脸（例如第一人 316 的脸 1 54）相关联的身份可以被确定，并且文本 40' 可以被标记给对应于该身份的人。

[0048] 在另一个示例中，显示的文本 40' 可以被地理定位在房间 308 中。在一个示例中，当站立在房间 308 中时，用户 304 可以叙述“The WiFi signal in this room is very weak（这个房间中的 WiFi 信号非常弱）”，如在所显示的文本 40' 中所示”。该语音可以被

捕捉，并由语音转换系统 10 转换成文本。由于该叙述特别涉及房间 308，显示的文本 40”可以被地理定位到房间 308。因此，显示的文本 40”可以对于用户 304 仍然可见，且空间上锚定到房间 308。

[0049] 在一个示例中，显示的文本 40”可以保持被地理定位到房间 308 达预定时间帧。通过这种方式，无论用户 304 在该时间帧内的何时进入房间 308，文本 40”都将被显示给房间内的用户 304。在其他示例中，也可以通过位于房间 308 内的一个或多个其它用户的 HMD 设备 200 将文本 40”显示给该一个或多个其他用户。

[0050] 在其它示例中，可以由 HMD 设备 200 接收房间 308 中来自一个或多个外部源的附加音频输入。例如，第三人 324 也可以佩戴 HMD 设备 200’ 形式的 HMD 设备 36。HMD 设备 200’ 可以通过例如网络 16 通信地耦合到 HMD 设备 200。HMD 设备 200’ 可以接收来自房间 308 的附加音频输入，包括来自第一人 316 的第一人音频输入 348。HMD 设备 200’ 可将这些第一人音频输入 348 与和该输入有关的位置数据一起提供给 HMD 设备 200。HMD 设备 200’ 可以使用这些附加的音频输入来标识从第一人 316 接收的目标音频输入 336，和 / 或改进目标音频输入的语音到文本转换的质量和 / 或效率。

[0051] 如上所述，在一些示例中，脸部检测程序 84 可以被配置为确定与目标脸（例如第一人 316 的脸 1\_54）相关联的身份。在一个示例中，脸部检测程序 84 可访问服务器 20 上的用户简档数据 92 以将包括脸 1\_54 的图像数据与对应于第一人 316 的一个或多个图像以及用户简档信息进行匹配。将理解，脸部检测程序 84 可使用任何合适的脸部识别技术来将脸 1\_54 的图像数据与所存储的第一人 316 的图像进行匹配。

[0052] 在一些示例中，语音转换程序 14 可以利用与脸 1\_54 相关联的身份来访问对应于身份的语音模式数据 94。语音转换程序 14 可以随后使用该语音模式数据 94 来将目标音频输入 336 转换成文本 40。例如，语音模式数据 94 可以允许语音转换程序 14 更准确地和 / 或更有效地将目标音频输入 336 转换成文本 40。

[0053] 作为替换且如上所述，在一些示例中，显示的文本 40’ 可以被标记给对应于该身份的第一人 316。如此，语音转换程序 14 可以向第一人 316 标记文本 40’，这样，文本气泡 340 在空间上被锚定到第一人，并在第一人移动时跟随第一人。在一些示例中，也可以通过其它 HMD 设备（例如 HMD 设备 200’）来显示锚定到第一人 316 的文本 40’ 和文本气泡 340，其中该其它 HMD 设备也确定第一人的目标脸的身份。而且，在一些示例中，在第一人离开房间 308 之后，文本 40’ 和文本气泡 340 保持被锚定到第一人 316 且可通过一个或多个 HMD 设备看见。如此，在房间 308 之外碰到第一人 316 的其它人可以从观看文本 40’ 中受益。

[0054] 在其它示例中，手语程序 88 可以被配置为从图像数据 80 中标识手语字母和 / 或单词。再次参考图 3，在一个示例中，第二人 320 可以通过手语（例如美式手语）与第三人 324 进行交流。用户 304 可以注视第二人 320，如注视线 356 所示。如上所述，对应于用户注视的眼睛跟踪数据 66 可以被眼睛跟踪系统 62 捕捉并被用于确定用户 304 正聚焦于第二人 320 的脸 2\_56，或聚焦于在正在做出对应于字母或单词的手语手形的第二人的右手 360。

[0055] 使用图像数据 80，手语程序 88 可以标识对应于由用户右手 360 形成的手形的手语字母或单词。手语程序 88 可以将该字母或单词转换成符号文本。随后，可以通过 HMD 设备 200 的透明显示器 44 来显示该符号文本。在当前示例中，第二人的右手 360 正做着单词“disappointed(失望的)”手势。手语程序 88 可以解释该手形以及其它形成句子“I’m

disappointed with the lecture(我对演讲失望)”的手形。该句子可在位于第二人 320 的头部之上的文本气泡 362 中被显示为文本 40”。

[0056] 在其它示例中,唇读程序 90 可以被配置为从图像数据 80 中标识目标脸的一个或多个嘴唇和舌头的移动。再次参考图 3,在一个示例中,第一用户 316 可能正对用户 304 讲话。用户 304 可能正注视第一人 316,如注视线 332 所示。如上所述,对应于用户注视的眼睛跟踪数据 66 可以由眼睛跟踪系统 62 捕捉并被用于确定用户 304 聚焦于第一人 316 的脸 1 54。

[0057] 使用图像数据 80,唇读程序 90 可以标识脸 1 54 的一个或多个嘴唇和舌头的移动。唇读程序 90 可以将该移动转换成唇读文本。随后,可以通过 HMD 设备 200 的透明显示器 44 来显示该唇读文本。

[0058] 在其它示例中,语音转换程序 14 可以按跨多个全向话筒 224、228、232、236、240 以及 244 重复扫描的方式对在这些话筒处接收的音频输入 68 进行采样。例如且参考图 2,可以以从话筒 224 开始继续到话筒 240、228、232 和 234 并在话筒 236 处结束的方式从右向左地顺序采样音频输入 68。在每次这样的扫描期间,语音转换程序 14 可以分析在每个话筒处接收到的音频输入 68 以标识人类语音音频。使用这样的分析,语音转换程序 14 可以确定从其可以发出人类语音的一个或多个位置。

[0059] 在其它示例中,用户 304 的头 364 的头部位置数据,包括头部姿势和 / 或头朝向数据,可以被用户聚焦程序 86 用于确定用户聚焦于的目标脸。这种头部位置数据可以被单独或与如上所述的其它位置信息组合用来确定目标脸。

[0060] 在其它示例中,转换自目标音频输入的文本可以被保存在计算设备 22 的大容量存储 18 中和 / 或保存在一个或多个其它计算设备的存储子系统中。这样的被保存文本随后可以由 HMD 设备 36 和 / 或由其它计算设备来访问和显示。

[0061] 如上所述,在各种示例中,计算设备 22 可以与 HMD 设备 36 分开或被集成到 HMD 设备 36 中。将可以理解的是,在一些示例中,混合现实显示程序 32、脸部检测程序 84、用户聚焦程序 86、手语程序 88、唇读程序 90 和 / 或语音转换程序 14 以及上述的相关的办法和过程中的一个或多个可以位于在除计算设备 22 之外的计算设备上和 / 或在除计算设备 22 之外的计算设备上执行,例如诸如与计算设备 22 通过网络 16 通信耦合的服务器 20。

[0062] 图 4A 和 4B 示出了用于根据本发明的一实施例来将来自环境的音频输入转换成文本的方法 400 的流程图。在这个实施例中,在头戴式显示设备的话筒阵列处接收音频输入。参考以上描述并在图 1-3 中示出的语音转换系统 10 的软件和硬件组件来提供方法 400 的以下描述。可以理解,方法 400 还可在使用其他合适的硬件和软件组件的其他上下文中来执行。

[0063] 参考图 4A,在 402,方法 400 包括捕捉来自环境的图像数据。在 406,方法 400 包括从图像数据中检测一个或多个可能的脸。在 410,所述方法 400 包括使用来自头戴式显示设备的眼睛跟踪系统的眼睛跟踪数据来确定用户所聚焦的目标脸。在 414,所述方法 400 包括使用应用于来自话筒阵列的音频输入的至少一部分的波束成形技术来标识与目标脸相关联的目标音频输入。

[0064] 在 418,所述方法 400 还可包括从环境中接收来自一个或多个外部源的附加音频输入。在 422,所述方法 400 可以随后包括使用附加音频输入来标识目标音频输入。在 426,

所述方法 400 包括将目标音频输入转换成文本。在 430, 方法还可包括确定目标脸的身份。在 434, 方法 400 可包括访问对应于目标脸的身份的语音模式数据。在 438, 所述方法 400 可以包括使用该语音模式数据来将目标音频输入转换成文本。

[0065] 在 442, 所述方法 400 包括通过头戴式显示设备的透明显示器来显示文本。在 446, 方法 400 可以包括将显示的文本标记到对应于该身份的人。现在参考图 4B, 在 450, 方法 400 可包括将显示的文本地理定位在该环境内。在 454, 方法 400 可以进一步包括从图像数据中标识一个或多个手语字母和单词。在 458, 方法 400 可包括将字母和单词转换成符号文本。在 462, 方法 400 还可包括通过头戴式显示设备的透明显示器来显示符号文本。

[0066] 在 466, 方法 400 可以包括从目标脸的图像数据标识目标脸的一个或多个嘴唇和舌头的移动。在 470, 方法 400 可包括将这些移动转换成唇读文本。在 474, 方法 400 还可包括通过头戴式显示设备的透明显示器来显示唇读文本。在 478, 方法 400 可以包括, 当话筒包括全向话筒时, 标识在一个或多个全向话筒处接收到语音的位置。在 482, 方法 400 随后可以包括使用应用于在一个或多个全向话筒处接收到的语音的波束成形技术来标识与该位置相关联的目标音频输入。

[0067] 能够理解, 方法 400 是以举例方式提供的, 并且不旨在为限制性的。因此, 可以理解, 方法 400 可包括相比于图 4A 和 4B 中示出的那些步骤而言附加的和 / 或替换的步骤。并且, 可以理解, 方法 400 可用任何适当的次序执行。而且, 可以理解, 一个或多个步骤可从方法 400 中省略, 而不背离本发明的范围。

[0068] 图 5 示意性示出了可以执行上述方法和过程之中的一个或更多个的计算系统 500 的非限制性实施例。计算设备 22 可以采取计算系统 500 的形式。计算系统 500 以简化形式示出。应当理解, 可使用基本上任何计算机架构而不背离本公开的范围。在不同的实施例中, 计算系统 500 可以采取大型计算机、服务器计算机、台式计算机、膝上型计算机、平板计算机、家庭娱乐计算机、网络计算设备、移动计算设备、移动通信设备、游戏设备等等的形式。如上所述, 在一些示例中, 计算系统 500 可被集成到 HMD 设备中。

[0069] 如图 5 所示, 计算系统 500 包括逻辑子系统 504 和存储子系统 508。计算系统 500 可以任选地包括显示子系统 512、通信子系统 516、传感器子系统 520、输入子系统 522 和 / 或图 5 中未示出的其他子系统和组件。计算系统 500 还可包括计算机可读介质, 其中该计算机可读介质包括计算机可读存储介质和计算机可读通信介质。计算系统 500 还可以任选地包括其他用户输入设备, 诸如例如键盘、鼠标、游戏控制器, 和 / 或触摸屏等等。此外, 在某些实施例中, 此处所述的方法和过程可被实现为计算机应用、计算机服务、计算机 API、计算机库, 和 / 或包括一个或多个计算机的计算系统中的其他计算机程序产品。

[0070] 逻辑子系统 504 可包括被配置为执行一个或多个指令的一个或多个物理设备。例如, 逻辑子系统 504 可被配置为执行一个或多个指令, 该一个或多个指令是一个或多个应用、服务、程序、例程、库、对象、组件、数据结构、或其他逻辑构造的一部分。可实现这样的指令以执行任务、实现数据类型、变换一个或多个设备的状态、或以其他方式得到所希望的结果。

[0071] 逻辑子系统 504 可包括被配置成执行软件指令的一个或多个处理器。附加地或可替代地, 逻辑子系统可以包括被配置为执行硬件或固件指令的一个或多个硬件或固件逻辑机器。逻辑子系统的处理器可以是单核或多核, 且在其上执行的程序可被配置为并行或分

布式处理。逻辑子系统可以任选地包括遍布两个或更多设备分布的独立组件，所述设备可远程放置和 / 或被配置为进行协同处理。该逻辑子系统的一个或多个方面可被虚拟化并由以云计算配置进行配置的可远程访问的联网计算设备执行。

[0072] 存储子系统 508 可包括被配置为保持可由逻辑子系统 504 执行以实现此处所述的方法和过程的数据和 / 或指令的一个或多个物理持久设备。在实现此类方法和过程时，存储子系统 508 的状态可以被变换（例如，以保持不同的数据）。

[0073] 存储子系统 508 可以包括可移动介质和 / 或内置设备。存储子系统 508 可包括光学存储设备（例如，CD、DVD、HD-DVD、蓝光盘等）、半导体存储器设备（例如，RAM、EPROM、EEPROM 等）和 / 或磁性存储设备（例如，硬盘驱动器、软盘驱动器、磁带驱动器、MRAM 等）等等。存储子系统 508 可包括具有以下特性中的一个或多个特性的设备：易失性、非易失性、动态、静态、读 / 写、只读、随机存取、顺序存取、位置可寻址、文件可寻址，以及内容可寻址。

[0074] 在一些实施例中，可以将逻辑子系统 504 和存储子系统 508 的各方面集成在一个或多个共同设备中，通过该一个或多个共同设备，可以至少部分地实施本文所述的功能。这样的硬件逻辑组件可包括：例如，现场可编程门阵列（FPGA）、程序和应用专用集成电路（PASIC/ASIC）、程序和应用专用标准产品（PSSP/ASSP）、片上系统（SOC）系统以及复杂可编程逻辑设备（CPLD）。

[0075] 图 5 还示出以可移动计算机可读存储介质 524 形式的存储子系统 508 的一方面，该介质可以用于存储可执行以实现此处所述的方法和过程的数据和 / 或指令。可移动计算机可读存储介质 524 尤其是可以采取 CD、DVD、HD-DVD、蓝光盘、EEPROM 和 / 或软盘的形式。

[0076] 将明白，存储子系统 508 包括一个或多个物理持久设备。相反，在一些实施例中，本文描述的指令的各方面可以按暂态方式通过不由物理设备在至少有限持续时间期间保持的纯信号（例如电磁信号、光信号等）传播。此外，与本公开有关的数据和 / 或其他形式的信息可以经由计算机可读通信介质通过纯信号来传播。

[0077] 在被包括时，显示子系统 512 可用于呈现由存储子系统 508 保存的数据的视觉表示。由于以上所描述的方法和过程改变了由存储子系统 508 保持的数据，并由此变换了存储子系统的状态，因此同样可以转变显示子系统 512 的状态以在视觉上表示底层数据的改变。显示子系统 512 可包括利用几乎任何类型的技术的一个或多个显示设备。可以将此类显示设备与逻辑子系统 504 和 / 或存储子系统 508 一起组合在共享封装中，或者此类显示设备可以是外围触摸显示设备。显示子系统 512 可包括例如 HMD 设备 36 的显示系统 48 和透明显示器 44。

[0078] 在被包括时，通信子系统 516 可以被配置成将计算系统 500 与一个或多个网络和 / 或一个或多个其他计算设备可通信地耦合。通信子系统 516 可以包括与一个或多个不同通信协议兼容的有线和 / 或无线通信设备。作为非限制性示例，通信子系统 516 可被配置为经由无线电话网、无线局域网、有线局域网、无线广域网、有线广域网等进行通信。在一些实施例中，通信子系统可允许计算系统 500 经由诸如因特网之类的网络发送消息至其他设备和 / 或从其他设备接收消息。

[0079] 传感器子系统 520 可包括被配置成感测不同的物理现象（例如，可见光、红外光、声音、加速度、取向、位置等）的一个或多个传感器，如上所述。传感器子系统 520 例如可以被配置为向逻辑子系统 504 提供传感器数据。如上所述，此类数据可包括眼睛跟踪信息、图

像信息、音频信息、环境光信息、深度信息、位置信息、运动信息，用户位置信息和 / 或可被用来执行上述方法和过程的任何其他合适的传感器数据。

[0080] 在被包括时，输入子系统 522 可包括一个或多个传感器或用户输入设备（诸如游戏控制器、姿势输入检测设备、语音识别器、惯性测量单元、键盘、鼠标、或触摸屏）或与它们对接。在某些实施例中，输入子系统 522 可以包括所选的自然用户输入（NUI）部件或与其结合。这种部件可以是集成的或外围的，输入动作的转导和 / 或处理可以在板上或板外被处理。NUI 部件的示例可包括用于语言和 / 或语音识别的话筒；用于机器视觉和 / 或姿势识别的红外、色彩、立体显示和 / 或深度相机；用于运动检测和 / 或意图识别的头部跟踪器、眼睛跟踪器、加速计和 / 或陀螺仪；以及用于评估脑部活动的电场感测部件。

[0081] 术语“程序”可用于描述被实现来执行一个或多个特定功能的语音转换系统 10 的一个方面。在某些情况下，可以经由执行存储子系统 508 所保持的指令的逻辑子系统 504 来实例化这样的程序。将理解，可以从同一应用、服务、代码块、对象、库、例程、API、函数等实例化不同的程序。类似地，相同的模块可由不同的应用、服务、代码块、对象、例程、API、功能等来实例化。术语“程序”意在涵盖单个或成组的可执行文件、数据文件、库、驱动程序、脚本、数据库记录等。

[0082] 应该理解，此处所述的配置和 / 或方法在本质上是示例性的，并且这些具体实施例或示例不应被认为是局限性的，因为多个变体是可能的。此处描述的具体例程或方法可以表示任何数量的处理策略中的一个或多个。由此，所示出的各个动作可以按所示次序执行、按其他次序执行、并行地执行，或者在某些情况下被省略。同样，上述过程的次序可以改变。

[0083] 本公开的主题包括各种过程、系统和配置以及此处公开的其他特征、功能、动作和 / 或属性、以及它们的任一和全部等价物的所有新颖且非显而易见的组合和子组合。

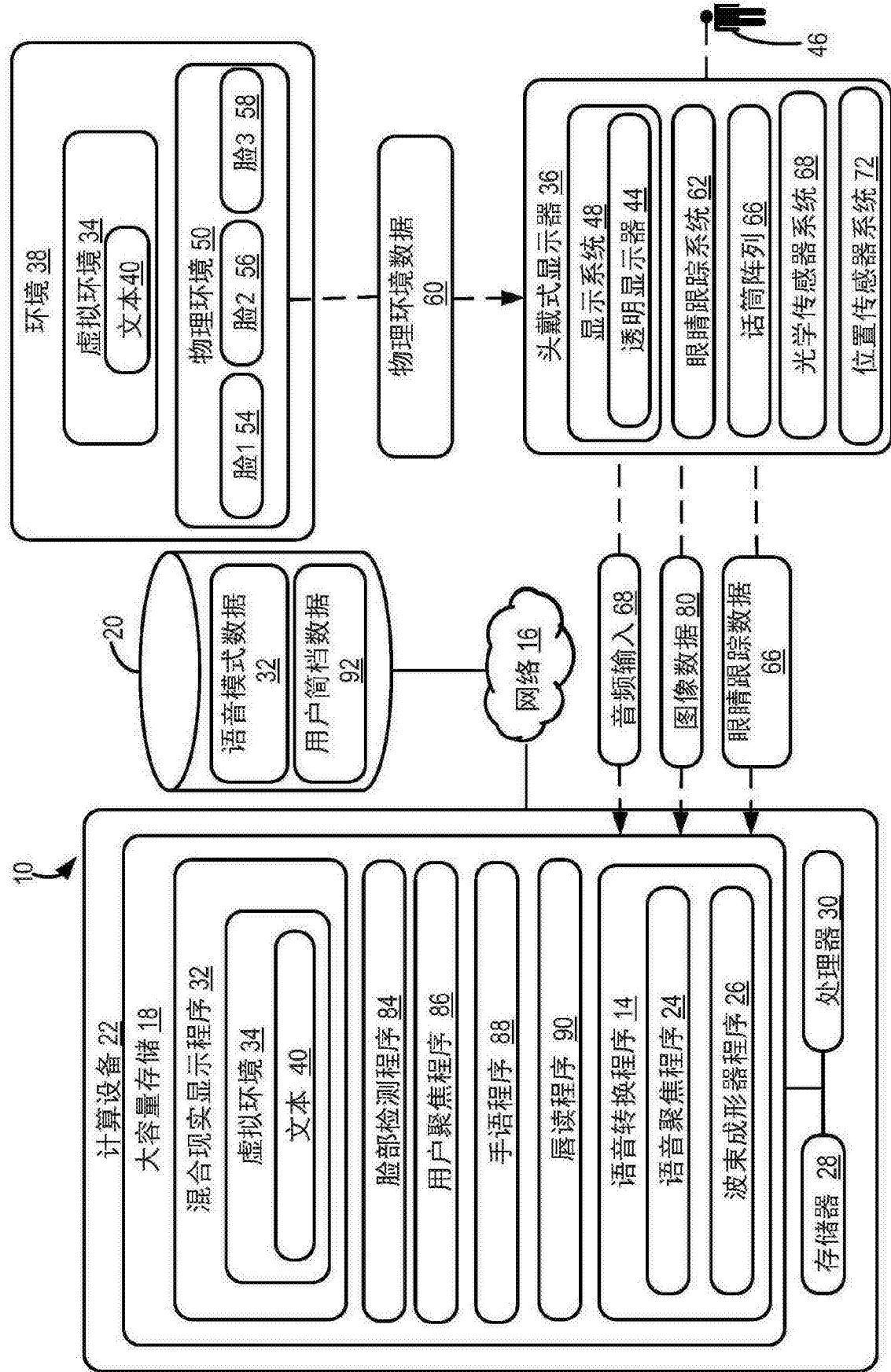


图 1

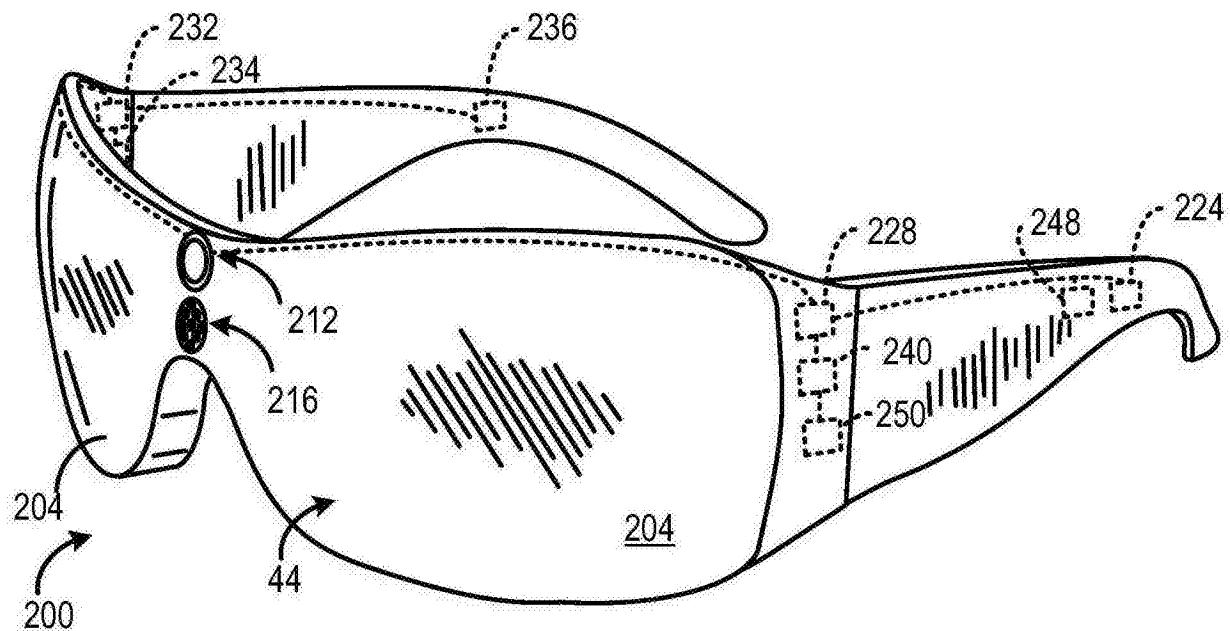


图 2

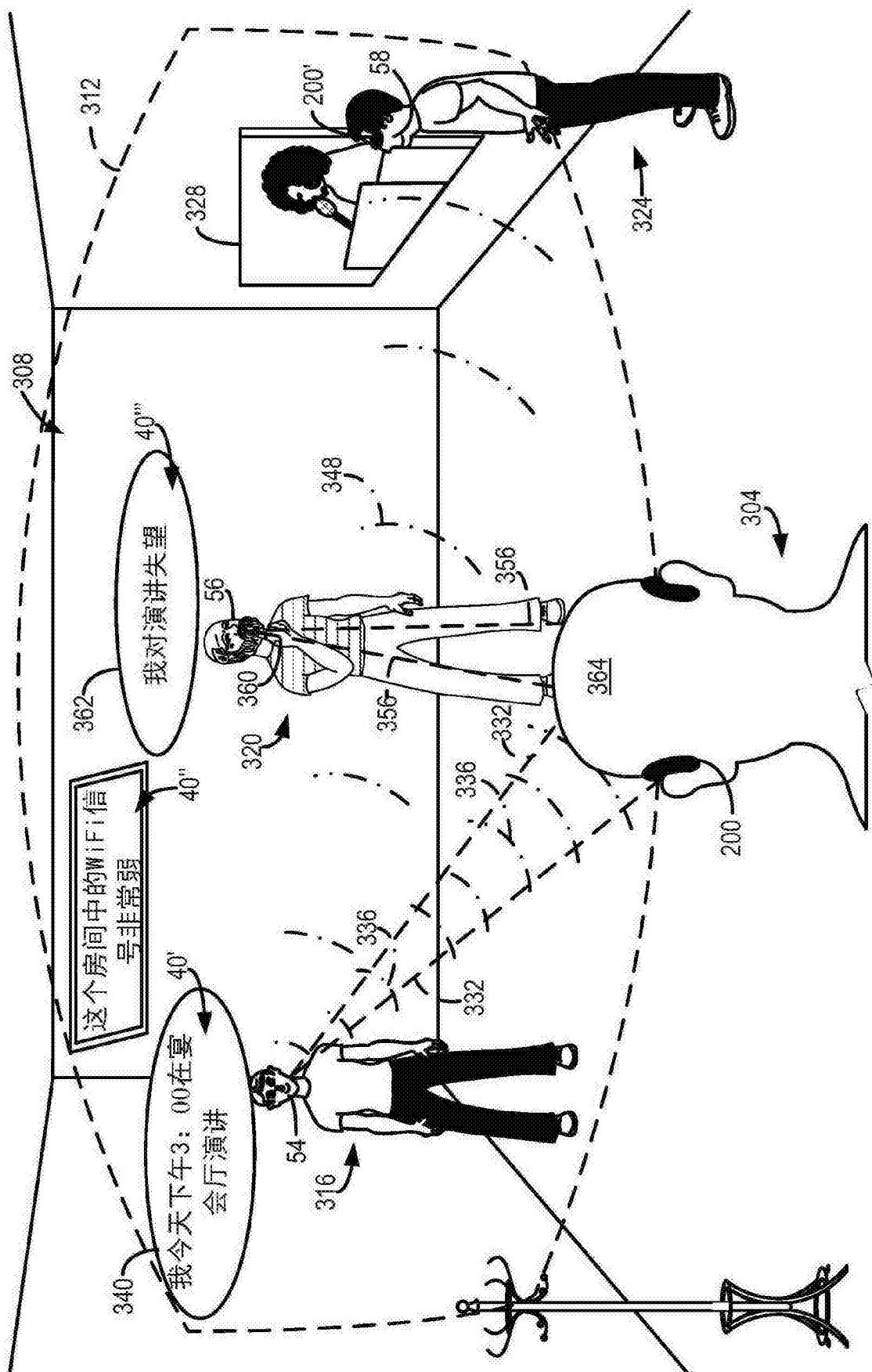


图 3

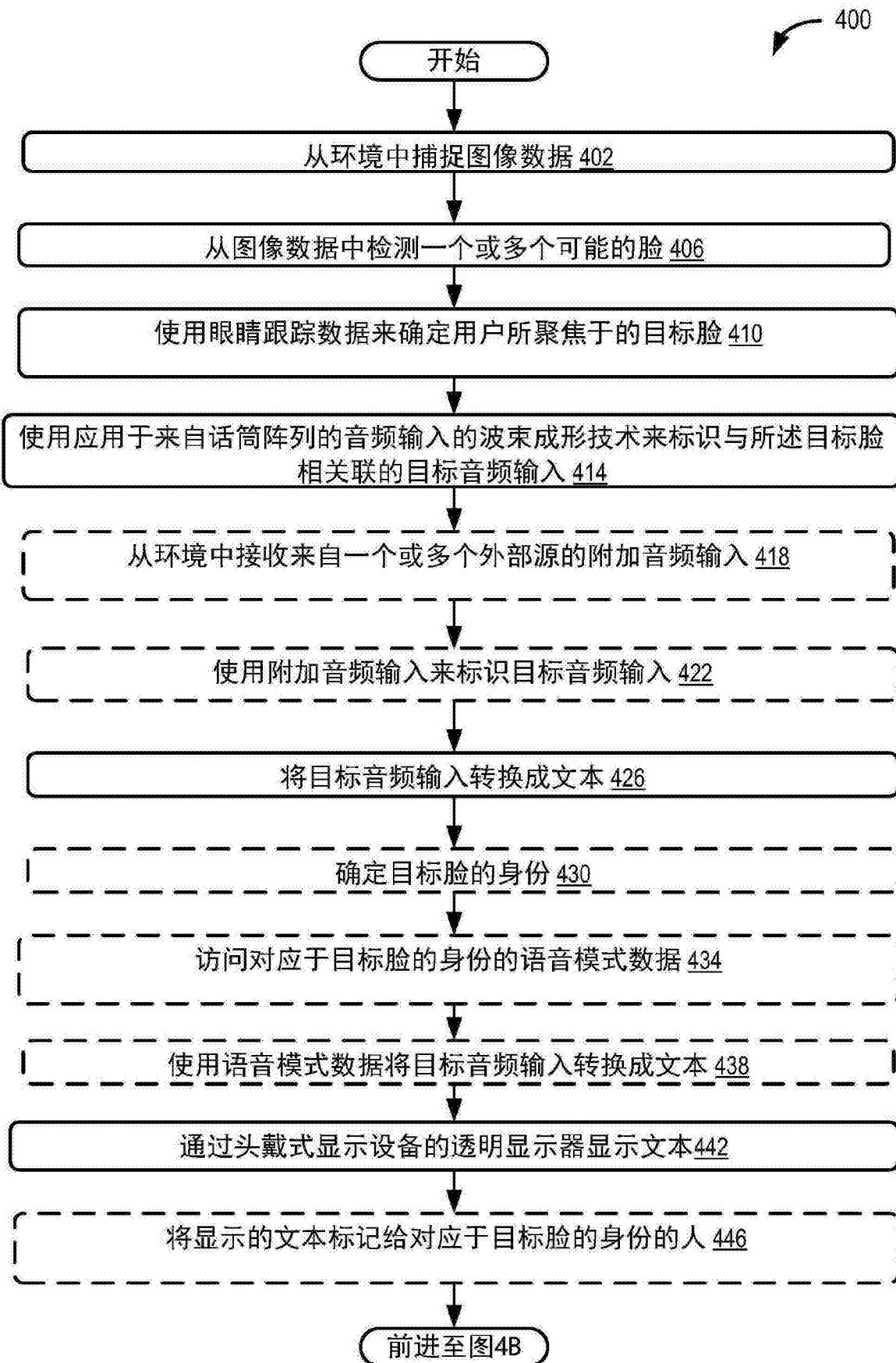


图 4A

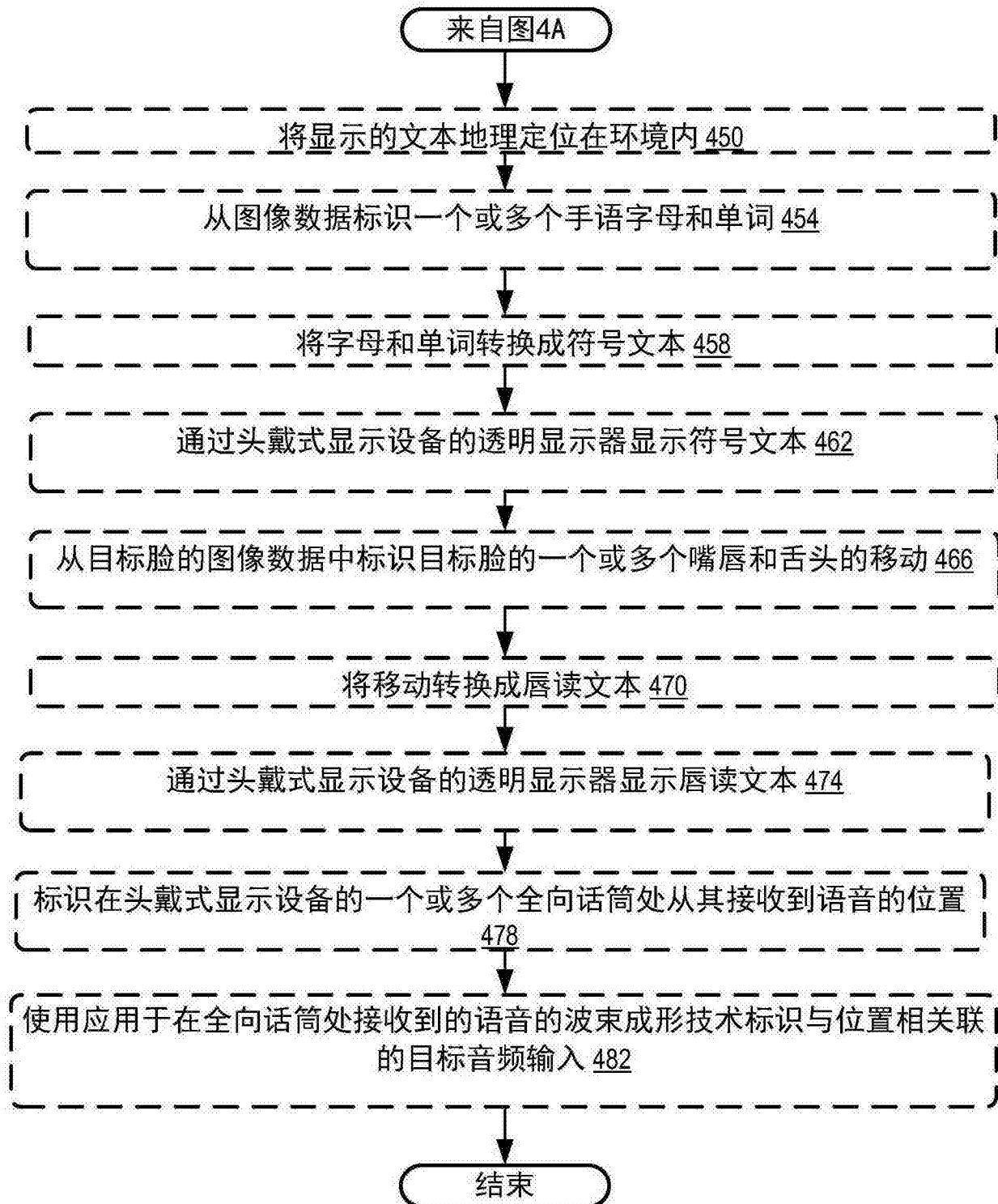


图 4B

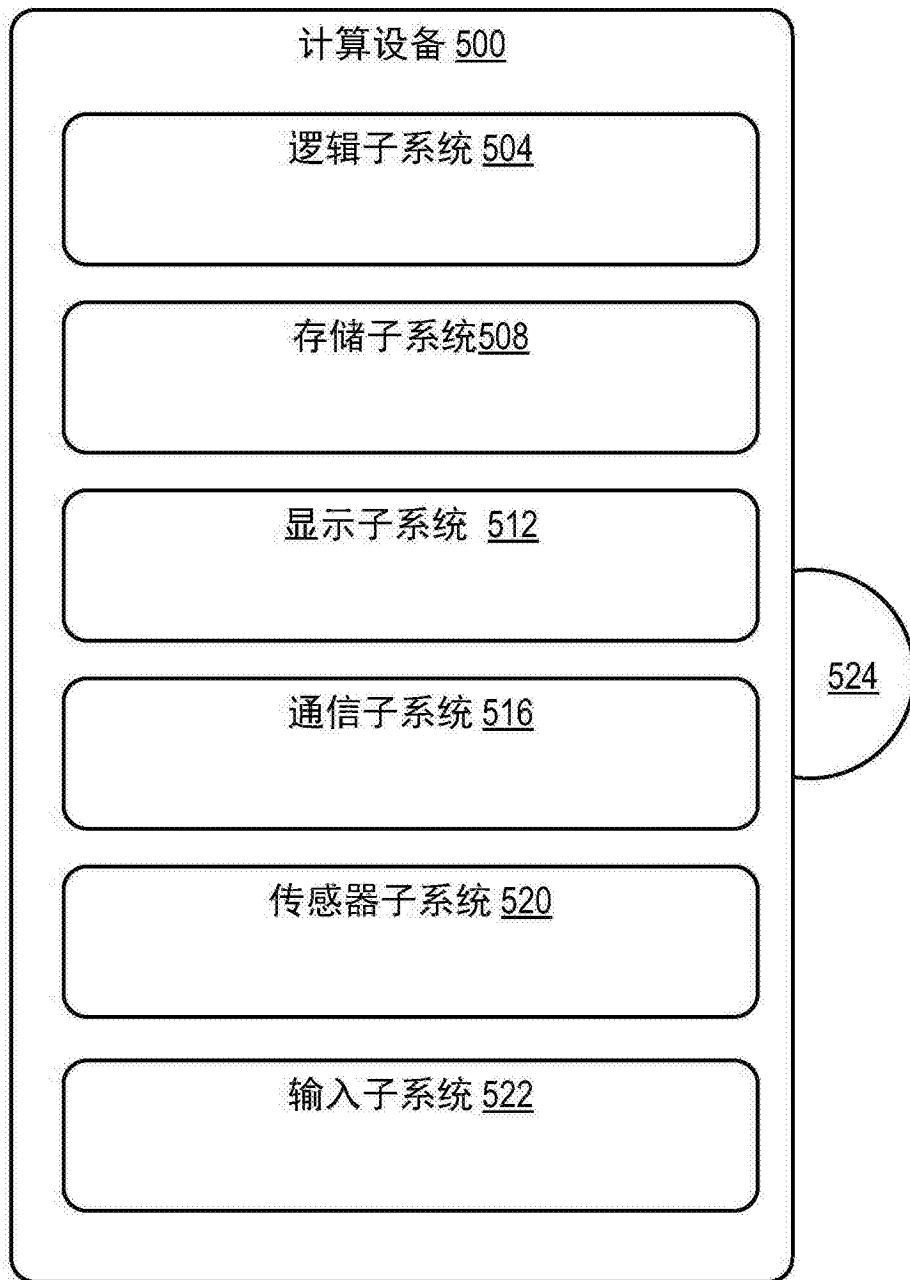


图 5