(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113874936 A
(43)申请公布日 2021.12.31

(21)申请号 202080032257.9
(22)申请日 2020.03.17
(30)优先权数据
16/398,836 2019.04.30 US
(85)PCT国际申请进入国家阶段日
2021.10.29
(86)PCT国际申请的申请数据
PCT/US2020/023054 2020.03.17
(87)PCT国际申请的公布数据
WO2020/222925 EN 2020.11.05
(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 吉冈拓也 A·施特尔克 陈卓
D·B·迪米特利亚迪斯 曾南山
秦莉娟 W·I·欣托恩 黄学东
(74)专利代理机构 北京市金杜律师事务所 11256
代理人 马明月
(51)Int.Cl.
G10L 15/26(2006.01)
G10L 19/018(2013.01)
G10L 21/0216(2013.01)
H04L 12/18(2006.01)
G06N 3/08(2006.01)
G06N 3/04(2006.01)
G06F 40/279(2020.01)

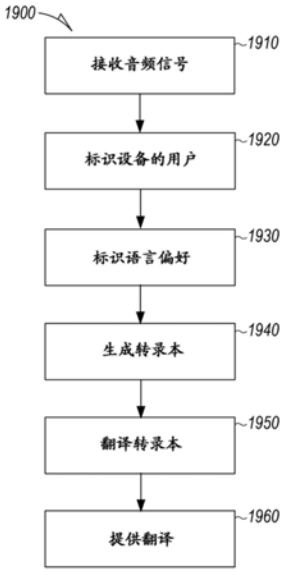
权利要求书1页 说明书22页 附图15页

(54)发明名称

用于优化分布式系统中的用户偏好的定制输出

(57)摘要

提供了用于基于分布式系统中的用户偏好提供定制输出的系统和方法。在示例实施例中，会议服务器或系统从智能会议所涉及的多个分布式设备接收音频流。会议系统标识与多个分布式设备中的分布式设备相对应的用户，并且确定用户的偏好语言。来自所接收的音频流的转录本被生成。会议系统将转录本翻译为用户的偏好语言，以形成所翻译的转录本。所翻译的转录本被提供给用户的分布式设备。



CN 113874936 A

1.一种计算机实施的方法,包括:

从智能会议中所涉及的多个分布式设备接收音频流;

标识与所述多个分布式设备中的分布式设备相对应的用户;

确定所述用户的偏好语言;

由硬件处理器,从所接收的所述音频流生成转录本;

将所述转录本翻译为所述用户的所述偏好语言,以形成经翻译的转录本;以及

将所述经翻译的转录本提供给所述分布式设备。

2.根据权利要求1所述的方法,其中提供所述经翻译的转录本包括:提供具有经翻译的文本的所述转录本。

3.根据权利要求1所述的方法,其中提供所述经翻译的转录本包括:将所述经翻译的转录本的文本转换为语音。

4.根据权利要求1所述的方法,其中提供所述经翻译的转录本包括:针对所述转录本的每个经翻译的话语提供发言者身份。

5.根据权利要求1所述的方法,其中确定所述用户的所述偏好语言包括:访问先前针对所述用户所建立的用户偏好,所述用户偏好指示所述偏好语言。

6.根据权利要求1所述的方法,其中所述智能会议是临时会议,所述方法还包括:

比较所述音频流,以确定所述音频流表示来自所述临时会议的声音;以及

响应于所述比较确定所述音频流表示来自所述临时会议的声音,生成会议实例以处理所述音频流。

7.根据权利要求1所述的方法,还包括:

对所接收的所述音频流执行连续语音分离,以将来自同时说话的不同发言者的语音分离到单独的音频通道中,所述生成所述转录本基于所分离的所述音频通道。

8.根据权利要求1所述的方法,其中标识所述用户包括:

接收捕获所述用户的视频信号;以及

将所述用户的所存储的图像与所述视频信号相匹配,以标识所述用户。

9.根据权利要求1所述的方法,其中标识所述用户包括:

将所述用户的所存储的语音签名与来自所述音频流的语音进行匹配。

10.根据权利要求1所述的方法,其中标识所述用户包括:

获取与所述分布式设备相关联的用户标识符。

11.一种存储指令的机器可读介质,在由机器的一个或多个硬件处理器执行时,所述指令使所述机器执行权利要求1至10中任一项的所述方法。

12.一种设备,包括:

一个或多个硬件处理器;以及

存储器设备,所述存储器设备被耦合到所述处理器并且在其上存储有程序,在由所述一个或多个硬件处理器执行时,所述程序使所述一个或多个硬件处理器执行权利要求1至10中任一项的所述方法。

用于优化分布式系统中的用户偏好的定制输出

背景技术

[0001] 提前计划的会议可以利用在会议之前或在会议开始时设置的一个或多个会议工具以记录对话，并且生成归属于发言者的转录本。这种现有的会议工具位于会议桌上的设备，多个固定位置的发言者位于设备的不同侧。该设备可以具有塔状或圆锥状形状，并且可以具有可以被用于标识和追踪会议中的人的相机或与其相关联。语音到文本算法可以被用于创建转录本。音频波束形成可以与固定发言者的已知地点以及与会者的视频一起使用，以对转录本中的语音执行归属。

附图说明

[0002] 图1是根据示例实施例的多个用户之间的会议的透视图。

[0003] 图2是根据示例实施例的用于在会议中使用的用户设备的框图。

[0004] 图3是图示了根据示例实施例的、在具有关联分布式设备的两个用户之间发起智能会议的计算机实施的方法的流程图。

[0005] 图4是图示了根据示例实施例的通过使用会议码将分布式设备添加到智能会议的计算机实施的方法的流程图。

[0006] 图5是图示了根据示例实施例的、将其他设备添加到智能会议的计算机实现的方法的流程图。

[0007] 图6是图示了根据示例实施例的、检测到正在发生临时会议的计算机实现的方法的流程图。

[0008] 图7是图示了根据示例实施例的响应于用户离开会议从用户设备和其他设备移除音频通道的计算机实现的方法的流程图。

[0009] 图8是图示了根据示例实施例的认证设备用于将来自设备的音频流添加到由会议服务器实例处理的音频通道的计算机实现的方法的流程图。

[0010] 图9是根据示例实施例的用于针对多个用户之间的会议生成转录本的系统的高级流程框图。

[0011] 图10是图示了根据示例实施例的、包括来自分布式设备的音频流的信息的分布式会议服务器处理的详细流程框图。

[0012] 图11是图示了根据示例实施例的、在智能会议期间同步从多个分布式设备接收的多个音频通道的计算机实施的方法的流程图。

[0013] 图12是图示了根据示例实施例的、在分布式设备智能会议中分离重叠语音的计算机实现的方法的流程图。

[0014] 图13是图示了根据示例实施例的、在处理期间在多个所选择的点处融合音频流的计算机实现的方法的流程图。

[0015] 图14A和14B图示了根据示例实施例的示例周围环境捕获设备。

[0016] 图15图示了根据示例实施例的麦克风阵列的示例放置。

[0017] 图16图示了根据示例实施例的具有周围环境捕获设备的人工智能(AI)系统。

[0018] 图17是图示了根据示例实施例的减少通过网络发送给会议服务器用于生成转录本的音频流的数目的计算机实现的方法的流程图。

[0019] 图18是图示了根据示例实施例的用于使用来自分布式设备的视频和音频通道两者、视听数据以提供更好的发言者标识的计算机实现的方法的流程图。

[0020] 图19是图示了根据示例实施例的用于基于用户偏好定制输出的计算机实现的方法的流程图。

[0021] 图20是实现一个或多个示例实施例的计算机系统的示意性框图。

具体实施方式

[0022] 在以下描述中，参照形成其部分并且通过图示的方式示出可以被实践的具体实施例的附图。这些实施例以足够细节描述以使本领域技术人员能够实践本发明，并且要理解的是，其他实施例可以被利用，并且结构、逻辑和电的改变可以在不脱离本发明的范围的情况下进行。因此，示例实施例的以下描述不是限制意义的，并且本发明的范围由所附权利要求限定。

[0023] 在一个实施例中，本文所描述的功能或算法可以在软件中实现。该软件可以包括存储在计算机可读介质或计算机可读存储设备上的计算机可执行指令，诸如一个或多个非暂态存储器或其他类型的基于硬件的存储设备，无论是本地的还是联网的。进一步地，这种功能对应于模块，该模块可以是软件、硬件、固件或其任何组合。多个功能可以在一个或多个模块中根据需要执行，并且所描述的实施例仅是示例。软件可以在数字信号处理器、ASIC、微处理器或者在计算机系统(诸如，个人计算机、服务器或其他计算机系统)上操作的其他类型的处理器，将这种计算机系统转变为专门编程的机器的上执行。

[0024] 功能性可以被配置为使用例如软件、硬件、固件等来执行操作。例如，短语“被配置为”可以指要实现关联功能性的硬件元件的逻辑电路结构。短语“被配置为”还可以指要实现固件或软件的关联功能性的编码设计的硬件元件的逻辑电路结构。术语“模块”是指可以使用任何合适的硬件(例如，处理器等)、软件(例如，应用等)、固件或者硬件、软件和固件的任何组合来实施的结构元件。术语“逻辑”涵盖用于执行任务的任何功能性。例如，流程图中所图示的每个操作都对应于用于执行该操作的逻辑。操作可以使用软件、硬件、固件等来执行。术语“组件”、“系统”等可以指执行中的计算机有关的实体、硬件和软件、固件或其组合。组件可以是在处理器上运行的过程、对象、可执行文件、程序、函数、子例程、计算机或者软件和硬件的组合。术语“处理器”可以指硬件组件，诸如计算机系统的处理单元。

[0025] 此外，所要求保护的主题可以被实现为方法、装置或制品，其使用标准编程和工程技术以生产软件、固件、硬件或其任何组合以控制计算设备实施所公开的主题。本文使用的术语“制品”旨在涵盖从任何计算机可读存储设备或介质可访问的计算机程序。

[0026] 被称为用户的个人可以在任何时间开始对话或会议。如果会议已经被安排，则可以进行布置以记录对话并且创建对话的转录本用于稍后参考。然而，临时会议通常不涉及这种准备。停止会议或以其他方式花时间设置方法来记录对话并且布置要被创建的转录本可能会分散注意力，或者在会议期间可能不会被考虑。另外，临时会议通常在会议室外发生。在这些情况下，专门针对会议设计的记录设备不可用。

[0027] 在对话期间，对话的音频可以由用户携带的设备来捕获，称为分布式设备。在示例

实施例中，所捕获的音频信号通过无线通道被传输给会议系统，以识别多个用户正在进行的被称为会议的对话，该对话可能已经或可能没有被计划。如果会议是计划外的，则其被称为临时会议。

[0028] 响应于已检测到或以其他方式布置的会议，会议实例在会议系统上生成，以识别来自正在说话的用户的语音(speech)并且生成会议的转录本。来自多个分布式设备的多个语音信号作为单独的音频通道接收，并且被用于生成转录本。分布式设备可以包括个人用户设备(例如，智能电话)以及其他设备，包括数字助理、相机和能够在对话范围内接收音频和/或视频的任何类型的设备。

[0029] 在一些实施例中，会议可以经由会议应用利用单次按下单个设备上的按钮来创建。其他设备和具有设备的用户可以通过经由会议应用按下在其用户设备上所呈现的按钮或通过在不使用时被招募(例如，房间中存在的现有会议设备)来加入会议。会议参与者可以通过语音指纹，通过是参与设备的所有者，通过面部识别或者通过在任何点经由其设备上的会议应用手动添加用户(例如，针对远程参与者)来推断(例如，被标识)。

[0030] 存在会议可以被建立的许多方式。在一些实施例中，诸如智能电话的分布式设备与相应的用户相关联，并且包括用于将从设备上的麦克风所接收的音频流式传输到会议系统或服务器的会议应用。从附近设备所接收的音频将具有基于周围环境噪声和/或设备附近所生成的任何声音的组合的音频签名。响应于两个用户设备经由它们相应的音频流(音频通道)提供类似的音频签名，会议系统识别可能正在发生会议，并且创建会议实例以处理所所接收的音频。用户可以经由其会议应用提示以加入会议。备选地，诸如地点信息、先前交互、日历信息或最近电子邮件交互的其他信息可以被用于确认两个用户或第三用户应该被添加到会议实例。

[0031] 在其他实施例中，音频水印由用户设备中的一个或多个用户设备生成。音频水印包括音频签名，或者音频签名可以被单独检测。音频水印可以是频率高于用户的正常听力范围的声音模式，诸如20Khz或更高，或者可以只是对于用户来说难以察觉的声音，以免干扰对话。在其他实施例中，水印可以是完全可听和可识别的。在一些实施例中，水印可以被选择，以由期望确保会议实例在对话期间被创建的用户发送。水印由范围内的分布式设备接收，并且自动或可选地被添加到会议实例。水印声音范围内的设备也可以将其音频流作为附加音频通道添加到会议实例。

[0032] 在一些实施例中，会议码被生成，并且被发送给其他用户以将其添加到计划的或临时的会议。会议码也可以在安排的会议之前选择，并且在会议邀请中使用。在从用户设备接收到会议码时，一旦被实例化，会议系统就将来自这种用户设备的音频流添加到会议。

[0033] 示例实施例提供了用于基于分布式系统中所提供的用户偏好来提供定制输出的系统和方法。在示例实施例中，会议服务器或系统从智能会议所涉及的多个分布式设备接收音频流。会议系统标识与多个分布式设备中的分布式设备相对应的用户，并且确定用户的偏好语言。当会议发生时，来自所接收的音频流的转录本被生成。会议系统将转录本翻译为用户的偏好语言，以形成所翻译的转录本。所翻译的转录本被提供给用户的分布式设备。在示例实施例中，所翻译的转录本在会议发生时实时(或近实时)提供。所翻译的转录本可以经由文本提供(例如，显示在用户的设备上)或作为音频输出(例如，经由发言者、助听器、耳机)。在一些实施例中，代替翻译或除了翻译之外，其他类型的变换可以被应用于原始转

录本、所翻译的转录本或所翻译的语音音频。

[0034] 图1是多个用户之间的会议100的透视图。第一用户110具有第一设备115，其包括麦克风以捕获包括语音的音频语音。第二用户120具有也能够捕获音频(包括语音)的第二设备125。在一个示例会议100中，用户可以坐在桌子130旁。

[0035] 第一设备115和第二设备125(也被称为“多个分布式设备”或“多个分布式设备”)将所捕获的音频传输给用于处理和生成转录本的会议服务器135。会议可以是临时的，因为它是计划外的。例如，用户可能在休息时遇到彼此，或者碰巧在走廊里相遇并且决定谈论他们正在进行的项目。会议应用(也称为“会议app”)可以在第一设备115和第二设备125两者上运行。会议app可以被用于向会议服务器135提供音频。

[0036] 会议服务器135检测到两个设备都在发送：具有类似音频签名的音频、类似的音频水印、由两个设备提供的类似会议码或者指示用户之间正在进行的讨论的其他信息。作为响应，会议服务器135生成会议实例，以处理所接收的音频并且生成转录本。

[0037] 在各种实施例中，水印可以是具有仅高于人类听觉范围的能量的任何类型的声音，该范围大约为20kHz，或者以其他方式是听不见的、难以察觉的或不分散注意力的，其标识对应于会议100的会议实例或会议码。在其他实施例中，水印可以是对会议码或会议实例的其他标识进行编码的声音。

[0038] 会议100可以涉及多于两个人，无论是计划的还是临时的。具有第三设备145的第三用户140也可以加入会议100。第三设备145还向分布式会议服务器135提供音频。音频通过所描述的相同机制中的一个或多个机制被识别为涉及会议100用于识别涉及会议100的前两个用户/设备。

[0039] 分布式设备的所有者/用户可以经由app来登记他/她自己以由会议服务器135识别。用户可以具有或创建称为语音拇指印或指纹的语音简档，以帮助会议服务器135将传入的语音声音与用户相关联。如果随机的人加入会议100，则会议服务器135识别出该人是未知的并且提示已经在会议中的用户中的一个或多个用户针对该人的姓名。备选地，会议服务器135在与会议中的已知用户相关联的组织中搜索数据库，以将该人与简档相匹配。如果该人未知或以其他方式标识，则该人在所生成的转录本中利用标签或标注来标识，诸如发言者1、发言者2等，如果该人稍后被命名，则更容易修改转录本。任何用户都可以在会议期间或之后的任何时间为发言者标签指派名称。已经在会议中的那些人的已知或频繁联系人可以被用于减少用于针对该人的初始检查的池/数据库，以优化标识该人的过程。

[0040] 可以有在会议100的音频或视觉范围内的附加设备，诸如数字助理148或专用会议设备150，两者都被示出在桌子130上，但是可以在会议100的音频范围内的任何地方。这种附加设备可以被连接至分布式会议服务器135，并且将它们的音频流添加到会议实例，以进一步增强在会议服务器135上运行的会议实例的音频和语音到文本的处理能力。这种附加设备可以由会议服务器135检测到，并且如上所述被添加到会议，或者可以作为添加到会议的选项被呈现给用户中的一个或多个用户。

[0041] 相机155或其他图像捕获设备可以具有涵盖会议100(或会议100的部分)的视野。会议服务器135知道相机155在会议100附近，并且可以向用户中的一个或多个用户提供指示，从而提供从相机155获得信息并且将信息提供给会议实例的选项，以进一步增强转录本的处理和提供。例如，相机155可以被用于检测哪个用户正在说话，或至少提供用户可能在

任何特定时间点说话的信息。

[0042] 图2是用于在会议中使用的用户设备200的框图。参与会议的其他设备可能具有类似的组件集。设备200包括至少一个麦克风210和用于执行被存储在存储器225上的会议app220的处理器215。收发器230被用于将音频和/或视频从相机235流式传输给分布式会议服务器135。用户设备200还可以具有显示屏，诸如触摸屏240，其部分被示出。

[0043] 可能参与会议的设备经由日历条目、当前地点、NFC(使电话靠得非常近)、蓝牙®广告以及经由会议码或可以被生成并且与会议100相关联的其他代码的直接邀请来标识。

[0044] 会议服务器135可以经由多个会议实例同时处理多个会议。每个会议实例可以包括会议标识符，诸如会议码、正在流式传输音频的设备的标识、正在参与会议的用户的标识(经由用户关联的设备)，或通过面部识别、语音识别或识别用户的其他方式以其他方式由会议服务器135识别。

[0045] 图3是图示了在具有关联分布式设备的两个用户之间发起智能会议的方法300的流程图。在操作310处，音频水印经由与第一分布式设备相关联的麦克风在第一分布式设备处接收。在一个实施例中，音频水印在会议期间由与第二分布式设备相关联的发言者所传输。

[0046] 在操作320处，对应于所接收的音频水印的数据经由第一分布式设备被传输给分布式设备会议服务器。在一些实施例中，所接收的音频水印首先被转换为数字形式，其可以简单地是将音频水印直接转换为声音的数字表示，或者可以包括对音频水印进行解码以获得标识会议或发出音频水印的第二分布式设备的数据。

[0047] 在操作330处，从分布式会议服务器接收指示，该指示指出第一分布式设备已经被接受到分布式设备会议服务器上的会议实例。

[0048] 在操作340处，响应于所接收的指示，第一分布式设备将会议的音频流式传输给分布式设备会议服务器上的会议实例。所接收的指示可以包括标识要使用的通信通道的信息，或者音频流可以简单地标识会议服务器用于将音频流引导到正确的会议实例的流式传输设备。

[0049] 图4是图示了通过使用会议码将分布式设备添加到智能会议的方法400的流程图。在一些实施例中，如方法300中讨论的，会议码被编码在水印中。在操作410处，会议码由第一分布式用户设备针对用户之间的会议来接收或生成。第一分布式用户设备可以从执行会议实例的会议服务器接收代码，或者第一分布式用户设备经由在第一分布式用户设备上运行的会议app生成会议码。

[0050] 在操作420处，代码被发送给第二分布式用户设备。代码可以经由电子邮件、文本或电子发送数据的其他方式发送，或者可以被编码为可听信号(音频水印)并且以声学方式传输给其余的参与设备，诸如经由用户设备中的一个用户设备(诸如第一分布式用户设备)的扬声器。

[0051] 第二分布式用户向会议服务器会议实例提供会议码，由此在操作430处，会议码被用于标识至少一个第二分布式用户设备。在操作440处，第二分布式用户设备将音频从第一分布式用户设备和第二分布式用户设备流式两者传输给会议服务器会议实例。

[0052] 会议可以是多个用户或多个用户设备之间的临时会议，并且会议码在临时会议开始之后生成。要注意的是，也可能存在没有关联用户设备的用户正在参与会议。其他用户设

备和不与用户相关联的设备可以基于检测到的设备地点来标识。来自这种设备的数据可以通过向(多个)用户提供其他附近设备的列表来将其数据流添加到会议实例,并且允许经由app的用户界面选择这种设备以添加到会议实例。可能参加会议的设备可以经由日历条目、当前地点、NFC(使电话靠得非常近)、蓝牙广告和直接邀请来标识。

[0053] 在其他实施例中,会议是多个用户或多个用户设备之间的所计划的会议,并且会议码在所计划的会议开始之前生成。会议码可以被发送给用户设备中的每个用户设备,并且由对应的app使用以向会议服务器会议实例标识设备,用于在会议期间添加来自这种设备的数据流。

[0054] 图5是将其他设备添加到智能会议的计算机实现的方法500。在操作510处,会议服务器从分布式设备组接收音频流,其中音频流包括在两个或多个用户的会议期间由这种分布式设备组所检测的语音。

[0055] 在操作520处,会议服务器从附加的或新的分布式设备接收与会议相对应的会议信息。新设备可以是刚加入会议的用户的用户设备,或者新设备可以是房间中或智能会议范围内的设备。

[0056] 在操作530处,附加分布式设备被添加到会议服务器会议实例。响应于添加附加分布式设备,来自附加分布式设备的信息流在操作540处被接收。

[0057] 图6是图示了检测到正在发生临时会议的计算机实现的方法600的流程图。在操作610处,音频流在会议服务器处从至少两个分布式设备接收,这些音频流是在两个用户之间的临时会议期间检测到的流式传输音频。

[0058] 在操作620处,音频流被比较以确定音频流代表来自临时会议的声音。例如,音频流可以通过计算两个信号之间的归一化交叉相关系数来比较。如果结果高于预定阈值,则音频流很可能来自同一临时会议。所选择的阈值可以是0和1之间的数字,并且可以基于在不同环境中的多个会议场景期间所进行的测试凭经验选择。选择可以被执行,以获得假阴性和假阳性的所期望的平衡。流来自同一会议的其他指示包括相同的设备的地点。其他指示包括过去有多次交互、在同一组织中的用户以及用户可能会面的其他指示。进一步的验证可以通过比较从音频流生成的文本来获得。

[0059] 一旦流被成功比较,会议ID/代码可以被生成,并且用于添加更多参与者。响应于与会议中已经存在的音频流成功比较的其他设备流式传输音频,其他参与者可以被添加。一旦设备被添加,该设备可以生成指示加入会议的信号,诸如ping。

[0060] 响应于确定音频流代表来自临时会议的声音,会议服务器在操作630处生成会议实例以处理音频流。在一些实施例中,用户在来自其相应设备的音频流被添加到会议实例之前被认证。认证可能基于来自会议app的用户确认、日历信息、组织图表、会议码的使用、与会议中已经存在的用户的联系/关系的程度以及认证的其他方式。

[0061] 在操作640处,音频流被处理以生成临时会议的转录本。在一个实施例中,会议服务器135检测设备和/或关联用户何时离开会议,并且从会议实例中移除来自该设备的音频流/通道。当与设备相关联的参与者离开会议时,会议服务器135检测到与会议中的设备相关联的音频信号不存在,并且将设备从会议中移除。备选方案包括用户经由会议app发信号通知离开、关闭会议app、检测设备的位置不再靠近会议位置、检测到设备地点不再在会议地点附近、检测到来自设备的视频流中没有对应的音频水印、检测到由设备所接收的音频

签名不再与其他设备音频流的音频签名相匹配或者对来自视频信号的图像执行图像识别，以检测用户正在离开或已经离开正在进行会议的会议室或区域。类似地，会议实例可以响应于剩余单个用户或剩余单个用户设备来结束。

[0062] 图7是图示了响应于对应用户离开会议而移除用户设备和其他设备的音频通道的计算机实现的方法700的流程图。在操作710处，从分布式设备组(从分布式设备会议接收音频)接收的对应多个音频通道上的多个音频信号由会议服务器实例处理。如上面所讨论的会议服务器实例在操作720处被用以检测与分布式设备组中的第一设备相关联的第一用户已经离开分布式设备会议。在操作730处，作为响应，第一分布式设备的音频通道从由会议服务器实例处理的多个音频通道中移除。

[0063] 图8是图示了认证设备并且将来自设备的音频流添加到由会议服务器实例处理的音频通道的计算机实现的方法800的流程图。方法800开始于在操作810处，在会议服务器处从多个分布式设备接收音频流，这些分布式设备在会议期间接收来自多个用户的语音。在操作820处，所接收的音频流经由在会议服务器上执行的会议实例处理，以基于音频流中所包括的语音生成转录本。

[0064] 在操作830处，信息在会议服务器处从与第一附加用户相关联的第一附加分布式设备接收，该信息对应于用户之间的会议。该信息可能对应于添加用户设备的请求，或者可能是通过注意来自这种设备的音频流包括水印或音频签名的暗示请求。

[0065] 在操作840处，第一附加分布式设备或关联用户被认证或以其他方式被授权加入会议。基于语音指纹、会议组织者接受、使用会议码和/或新代码、检测到的参与者设备的地点、设备ID和/或关联的用户ID与所授权的列表的比较、组织成员检查、使用非公开会议标志以要求组织者接受或以上一项或多项的组合，参与者可以被授权加入会议。要注意的是，方法800也可以被应用于加入会议的前两个设备，并且也可以被应用于不与用户直接关联的设备，诸如会议室中的会议助理类型的设备或具有会议视野的相机。

[0066] 响应于附加分布式设备或关联用户的认证，在操作850处，第一附加分布式设备将其音频流添加到会议实例。

[0067] 在一些实施例中，远程参与者可以经由诸如微软Skype或Teams等通信平台、电话拨入或任何其他电话会议应用被连接到会议中。如果像Skype等远程会议平台被使用，则会议可以通过跟随提前发送的链接加入。针对拨入，唯一的电话号码或访问代码(诸如，会议码)可以被共享。一旦远程音频通道被连接至针对会议的服务器，它的处理方式类似于来自会议区域的音频流。发言者ID是基于登录过程已知的。音频流可能是针对单个用户/发言者，这意味着除非免提电话由多个远程用户使用，否则不需要语音分离。会议中由免提电话播放并且由附近分布式设备检测到的音频可以从这种附近分布式设备的音频流中取消。

[0068] 图9是用于针对具有多个用户的会议生成转录本的系统900的高级流程框图。用户可以分别具有关联的(分布式)设备910、912、914，这些设备被配备有麦克风以捕获音频，包括会议中的各种用户的语音，并且将所捕获的音频作为音频信号提供给会议服务器，该会议服务器至少包括分别经由音频通道916、918和920的会议转录器925。不同的设备可能有稍微不同的时钟周期和不同的处理时延量。另外，每个设备到服务器的连接通道可能有不同的时延。因此，来自音频通道916、918和920的音频信号不一定是同步的。

[0069] 除了语音识别模块或功能之外，会议转录器925还包括同步模块或功能。根据一个

实施例，来自音频通道916、918和920的音频信号首先被同步然后被识别，从而产生与通道中的每个通道相关联的文本。然后识别输出被融合(通过融合930)或以其他方式处理以生成转录本940。然后转录本940可以随后被提供回用户。在其他实施例中，来自音频通道916、918和920的音频信号在语音识别之前被融合。融合后获得的音频信号被识别，从而产生单一版本的文本。在一些实施例中，转录本可以以非常小的延迟提供。

[0070] 在各种实施例中，结合发言者标识和被分类以标识发言者的转录本生成一起使用的将音频信号转换为文本由会议服务器135提供。由会议服务器135执行的功能包括同步、识别、融合和分类功能。虽然这种功能在图9中以特定顺序示出，但是在不同的实施例中，这些功能可以以不同的顺序执行。例如，融合可以在识别之前执行，并且也可以在下面描述的各种其他点处执行。

[0071] 图10是图示了包括来自分布式设备的音频流的、通常在方法1000中的信息的会议服务器处理的详细流程框图。多个音频数据流1005从多个分布式设备接收。流包括M个独立的数据分组序列。第m个序列的每个分组包含由第m个设备捕获的数字化音频信号的片段。所接收的分组被拆包，并且来自分组的数据被重组以创建多通道信号。多通道信号可以被表示为：$\{[x_0(t),\cdots,x_{M-1}(t)];t=0,1,\cdots\}$。

[0072] 多通道信号中的不同通道的数字化信号很可能不同步，因为许多分布式设备会受到数字信号处理差异、设备上软件时延差异以及信号传输速度差异的影响。所有这些差异可以合计，从而难以整合来自不同设备的信息以创建准确的转录本。流同步模块1015接收多通道信号，并且选择通道中的一个通道作为参考通道。不失一般性，第一通道可以被用作参考通道。针对参考通道，输出与输入相同(即，$y_0(t)=x_0(t)$)。针对第m个通道($0<m<M$)，$x_m(t)$和$x_0(t)$之间的未对准量被估计并且校正以生成$y_m(t)$。

[0073] 未对准程度可以通过使用针对非参考通道信号的滑动窗口来计算两个信号之间的归一化交叉相关系数并且拾取提供最大系数值的滞后来估计。这可以通过使用缓冲器以临时存储声学信号片段来实现，在这些声学信号片段上，交叉相关分析在参考通道和其他通道中的每个通道之间单独执行。代替归一化交叉相关，测量两个信号之间的对准程度的任何得分函数可以被使用。

[0074] 在一个实施例中，相邻同步周期之间的关系被考虑在内。未对准是由两个因素引起的：设备/通道依赖偏移和设备依赖时钟漂移。即使两个设备同时捕获声学事件，由于数字信号处理的差异、设备上软件时延差异、信号传输速度差异等，由单个设备捕获的信号可能会在不同的时间到达会议服务器。这是设备/通道依赖偏移。而且，由于制造可变性，不同的设备不可避免地具有略有不同的时钟。因此，即使两个设备声称支持例如16kHz采样率，由这些设备记录的信号也不是100％对准的，并且未匹配量会随着时间的推移而线性增长。这是设备依赖时钟漂移。设备/通道依赖偏移和设备依赖时钟漂移被表示为S和D。第k个同步周期的时间差被表示为S+kD。因此，S和D的估计提供了对未对准程度S+kD的稳健估计。

[0075] 未对准量可以通过使用上述交叉相关周期性地检测未对准并且校正针对这种所检测的未对准来校正。另外，为了减少所测量的未对准量，全局偏移(与设备/通道依赖)和设备依赖时钟漂移被计算以估计未对准程度。全局偏移可以被用以在通过交叉相关测量和校正未对准之前校正全局未对准。全局偏移可以被确定为随时间所测量的未对准的平均值，并且很可能是设备中的时钟漂移的结果。因此，根据一个实施例，未对准程度通过简单

地考虑来自参考通道的差异来估计和校正。流同步可以以不同的间隔执行，诸如每30秒。小于或大于30秒的其他间隔可以在其他实施例中使用，因为网络时延可能会改变。

[0076] 流同步模块1015将多通道同步信号{[$y_0(t),\cdots,y_{M-1}(t)$];t＝0,1,…}提供给波束形成模块1020。波束形成模块1020用于分离重叠语音。当会议中的两个人同时说话时，会发生重叠语音。在识别语音并且将语音转换为文本之前，语音首先被分离到单独的通道中。因此，利用M个通道输入，输出是N个通道，并且被称为N个通道波束形成信号{[$z_0(t),\cdots,z_{N-1}(t)$];t＝0,1,…}。流同步模块1015充当第一融合点，其中多个输出被生成以保留输入信息的多样性。在没有语音重叠的情况下，这种融合是可选的。

[0077] 图11是图示了在智能会议期间同步从多个分布式设备所接收的多个音频通道的计算机实现的方法1100的流程图。在操作1110处，代表流式传输语音的音频信号从多个分布式设备接收，以生成多个音频通道。音频通道中的一个所选择的音频通道在操作1120处被指定为参考通道。

[0078] 一旦参考通道被指定，以下操作针对剩余音频通道中的每个音频通道执行。在操作1130处，与参考通道的时间差被确定。在操作1140处，每个剩余音频通道的时间通过将剩余音频通道与参考通道对准作为对应时间差的函数来校正。这可以通过简单地丢弃无关样本，附加零或使用重采样技术来完成。

[0079] 方法1100可以被周期性地执行以校正剩余音频通道的定时，诸如每30秒。在一个实施例中，方法1100还包括用于校正至少由分布式设备中的不同时钟所引起的全局偏移的其他操作。在操作1150处，全局偏移针对剩余音频通道中的每个音频通道来确定。然后，在针对所确定的时间差来校正每个剩余音频通道之前，在操作1160处，剩余音频通道通过每个对应的剩余音频通道全局偏移来校正。

[0080] 声学波束形成(或简称波束形成)是一种通过减少诸如来自多声道音频信号的背景噪声等不想要的声音来增强目标语音的技术。波束形成可以提高下游语音处理的准确性，诸如语音识别和发言者分类。

[0081] 针对具有从多个分布式设备流式传输的音频的智能会议，其相对于彼此的确切位置未知，传统的波束形成算法(诸如，延迟求和波束形成、超指向波束形成和差分波束形成)不起作用。这种算法依赖于关于麦克风设备布置的先验知识，而这种先验知识无法用于分布式设备。

[0082] 在一个实施例中，称为几何形状不可知的波束形成或盲波束形成的方法被用于执行针对分布式记录设备的波束形成。给定M个麦克风设备，对应于M个音频通道，语音和背景噪声的M维空间协方差矩阵被直接估计。矩阵分别捕获语音和噪声的空间统计数据。为了形成声束，M维空间协方差矩阵被反转。

[0083] 无论是传统的基于几何形状的波束形成还是盲波束形成，波束形成方法的缺点是它通常将信息流的数目从M减少到1，这意味着下游模块无法利用由空间分布式设备提供的声学多样性。为了生成M个波束形成信号并且保留声学多样性，留一法可以被采用。利用这种方法，第一输出信号是通过利用麦克风2-M执行波束形成来生成的。第二输出信号是利用麦克风1-M和3-M来生成的。这可以被重复M次，使得M个不同的输出信号被获得。针对每次波束形成，(M-1)维空间协方差矩阵被计算并且反转，这对计算的要求非常高。幸运的是，通过从原始M维逆矩阵导出所有(M-1)维逆矩阵，计算成本可以被显着降低。

[0084] 在一些实施例中，波束形成模块1020可以被配置为分离不同用户的重叠语音信号。这可以使语音识别和发言者归属更加准确。在一个实施例中，针对分布式麦克风记录系统的连续语音分离是经由使用置换不变训练或其变型(诸如，深度聚类或吸引子网络)来训练的神经网络来执行的。为了潜在地节省计算，重叠检测可以被使用以确定语音分离神经网络是否应该针对每个时间段执行。如果重叠语音未在所选择的时间段内检测到，则神经网络不会被执行，从而节省处理资源并且允许转录本更快速地实时产生。

[0085] 语音分离神经网络模型被执行，以针对分布式麦克风记录系统执行连续语音分离，其中输入麦克风的数目可以是任意的并且通过时间变化。该模型输出两个连续的语音流。当存在一个活动发言者时，输出流中的一个输出流将是无声的，而当两个发言者之间存在重叠语音时，每个发言者将占用不同的输出流。

[0086] 在示例实施例中，语音分离神经网络模型包含三个子模块：局部观察器、全局概括器和掩码重建器。多通道输入由这三个子模块连续处理。首先，相同的局部观察器被应用于每个输入麦克风。局部观察器包括堆叠的注意力层集，其将每个麦克风输入映射为高维表示，其中每个通道将交叉比较并且从所有其他通道中提取信息。两种不同类型的注意力被实现，即，自注意力和前馈注意力。

[0087] 接下来，全局概括器被应用于概括来自每个观察器的信息，以形成跨不同输入通道的全局表示。针对全局概括器的两个选项被考虑——平均池化和置换不变排序算法——其中每个通道的表示与置换不变损失进行比较以对准它们的局部置换和全局置换。当没有概括层时，网络会被缩减为通道式语音分离网络，其中每个通道都有自己的分离(即，通道之间没有全局分离协定)。

[0088] 最后，掩码重构器在针对任何任意时间的同时对两个掩码输出进行排序。掩码重建器包括长短期记忆网络的堆栈，并且从每个时间点的概括生成最终的两个通道输出。

[0089] 在从掩码重建器得到两通道输出后，置换不变训练目标函数在重建的掩码和干净参考之间应用，其中输出和干净参考的每个置换对的欧几里德距离被首先测量，然后最小距离和对应的置换被选择以更新神经网络。

[0090] 网络利用模拟的多通道数据来训练，其中输入通道的数目为针对每个样本随机挑选(例如，从2到10个通道)。利布里(Libri)语音数据集在模拟中被应用为源数据。在每个模拟句子中，来自两个随机用户/发言者的两个话语被首先选择。然后每个话语利用房间声学模拟来处理，其中房间脉冲响应来自具有随机房间和地点设置的图像方法。

[0091] 语音分离的一种变型是语音重叠检测，其中任务被简化为仅检测记录语音中的重叠区域。该算法以类似的方式操作，其中网络接收N个通道作为输入，并且连续输出两个通道作为输出。在重叠检测器中，网络不输出掩码。相反，网络输出两个一维指标函数，其中1意味着该通道中存在一个活动发言者，并且0意味着静音。因此，当存在两个活动发言者时，两个输出流将分别具有1作为输出。当存在一个活动发言者时，一个任意通道将有1作为输出，并且另一个将有0。网络也在网络输出(即，指标函数)和参考指标之间利用置换不变训练目标进行训练。

[0092] 图12是图示了在分布式设备智能会议中分离重叠语音的计算机实现的方法的流程图。在操作1210处，代表语音的音频信号经由与从对应的多个分布式设备所传输的流式传输音频相对应的多个音频通道接收。

[0093] 连续语音分离在操作1220处对所接收的音频信号执行，以将来自同时说话的不同发言者的语音分离到单独的音频通道中。在一个实施例中，操作1220处的语音分离由已训练的神经网络模型执行。神经网络模型使用置换不变训练或其变型进行训练。

[0094] 在操作1230处，所分离的音频通道被提供用于语音识别和转录本的生成。在一个实施例中，操作1230提供固定数目的单独输出通道。由于麦克风输入的数量可能会有所不同，而输出的数目是预先固定的，因此可能会存在其中有限数目的音频通道可以被容纳的实例，因为针对具有多个重叠发言者的每个音频通道，每个发言者会导致单独的音频通道。因此，如果输出音频通道的数目有限，则并非所有通道都可以分离发言者。

[0095] 图10中的波束形成模块1020的N个不同输出被提供给产生一系列senone后验概率的N个声学模型1025和1030。这种模型是众所周知的，并且通常是基于神经网络的。对来自分布式设备和/或波束形成器输出的多个音频通道中的每个音频通道使用声学模型针对每个senone提供N个得分。

[0096] 包括针对senones的得分在内的得分被提供给声学模型得分融合模块1035。单个输入通道的音频可以被常规地处理，以提供senones的序列及其后验概率。在将结果应用于多个语音识别(SR)解码器1040、1045之前，该结果使用模型得分融合模块1035组合。得分融合模块1035作为第二融合点操作，其组合多个信息源，并且同时生成多个输出以保留输入信息的多样性。两步骤过程涉及两个不同的神经网络(或分类器)：香草味声学模型和新的、更有针对性的声学模型。输出是senones数目的1倍的序列。要注意的是，得分融合模块1035使用声学模型(神经网络)的最后一层的输出作为输入。在其他实施例中，得分融合模块1035可以使用最后一层之前的任何层的输出。输入的大小可能与输出的大小不同。

[0097] 来自声学模型得分融合模块1035的多元音素(senones)序列被提供给SR解码器1040和1045，SR解码器1040和1045中的每个SR解码器利用标准语音识别处理来针对senones的每个片段提供n个最佳词语列表。开始时间和持续时间针对每个词语被提供。分段可以基于语音活动检测、发言者变化检测、固定间隔或一些其他合适的方法来执行。重新得分可以通过在解码器输出上使用神经网络语言模型(NNLM)来执行，以生成更好的n个最佳词语列表假设。

[0098] 多个发言者分类模块1050、1055接收SR解码器模块1040、1045的输出作为针对每个片段的N个最佳列表。在一种实现中，仅顶部词语序列假设被使用。第一操作以固定间隔提取发言者嵌入，诸如d向量(用于发言者验证的深度神经网络的隐藏层激活)。第二操作将词语序列因式分解为发言者同构的子片段。这可以利用凝聚聚类的变型、BIC(贝叶斯信息准则)或通过使用嵌入特征的其他方法来执行。第三操作通过比较子片段的发言者嵌入与候选发言者中的每个候选发言者的嵌入的接近度(例如，余弦相似度、负欧几里得距离)为上面所获得的子片段中的每个子片段指派发言者标识符。所得输出是将发言者标签指派给顶部SR假设的每个已识别词语。

[0099] 假设组合模块1060接收来自N个SR解码器模块1040、1045(例如，波束形成的音频通道)的n个最佳列表以及来自诸如波束形成/分离的音频通道的N个源的发言者识别输出作为输入。假设组合模块1060通过对来自每个通道的n个最佳得分进行缩放和归一化并且因此计算话语级后验概率来处理来自每个通道的n个最佳得分。n个最佳假设被对准到词语混淆网络中。通过添加与给定词语假设相关的话语级后验，词语级后验概率被获得。来自每

个通道的发言者识别输出被格式化为具有交替发言者和词语标签的混淆网络。词语标签来自1个最佳识别假设，其中发言者标签表示与语音片段相匹配的1个最佳或n个最佳发言者模型。针对发言者假设的后验概率表示归一化的发言者模型可能性。词语假设的后验被按比例缩小两个数量级，以免影响最终的词语识别，从而仅影响词语和发言者标签的正确对准。因此从每个通道获得的混淆网络在必要时被截断和/或串联，以覆盖相同的时间窗口，如在线处理约束所规定的那样。输出包括混淆网络(CN)，从而对词语和发言者假设两者及其后验概率进行编码。

[0100] 词语和发言者混淆网络根据最小编辑距离准则以及对所对准的节点之间的时间差异的惩罚来对准。这有效地将发言者和词语假设合并到单个网络中，从而对匹配标签的后验求和。如果需要，通过在每个位置处选取最高的后验标签，顶部发言者和词语假设从所组合的CN中读取。词语混淆网络可以从词语格而不是n个最佳列表构建，这取决于语音解码器的输出。

[0101] 来自组合模块1060的输出是第三次融合的结果，称为后期融合，以产生发言者用于生成会议的归属于发言者的转录本的文本和发言者识别。要注意的是，分别在波束形成模块1020和声学模型得分融合模块1035处的前两个融合步骤在各种实施例中是可选的。在一些实施例中，一个或多个音频通道可以被直接提供给声学模型得分模块1065，而无需波束形成或语音分离。语音识别然后经由SR解码器1070在一个或多个音频通道上执行，接着是发言者分类模块1075，输出被直接提供给组合模块1060。

[0102] 音频流可以在数字音频流同步之后通过几何形状不可知的波束成形或连续语音分离而早期融合。多个输出可以被生成，以保留输入信息多样性。后期融合可以在声学模型得分级别和/或文本级别/分类级别进行，以利用发言者信息和不同的模型假设。在一个实施例中，对词语或两个词语的后期融合通过使用固定时间窗口来执行。在一个实施例中，时间窗口对应于显着音频事件，并且可以被固定在例如两秒。这种时间窗口被选择为相当短，以能够提供具有低时延的实时(或近实时)转录本。

[0103] 在一个实施例中，实时转录本是基于短词序列生成的。数据的后期融合通过针对并行处理以产生短语的多个音频通道的语音识别来执行。从多个音频通道所导出的短语实时地组合。在一个实施例中，近似两秒的语音在假设组合模块1060处被组合。因此，音频流在它们被接收时被处理。两秒的非重叠滑动窗口被用于处理音频流，从而将会议系统135转录本生成的时延降低到接近于零。

[0104] 单个语音识别解码器连续输出一些结果，并且基于假设组合模块1060，结果被立即处理。特殊提供针对流同步模块1015处的单个系统的对准提供，否则最终结果可能包含相同事件的多个实例(由于未对准)。无论信号和/或语音识别输出对准如何，后处理步骤都会移除可能存在的任何重复项。对准可以在信号的词语级或样本级执行。还要注意的是，不同版本的音频由语音识别解码器接收。每个SR解码器可能会听到不同的声音。通过将SR结果(后期融合)与低时延组合，高度准确的转录本被产生。每个SR输出具有置信度的词语或两个词语。诸如两秒的时间足以获得一些显着输出——换言之，具有词语或两个词语的输出可以以某个置信度识别。固定的时间窗口(诸如，两秒)被发现效果更好。如果时间太短，则没有显着事件，并且如果时间太长，则时延变得太长并且转录本被延迟，使得转录本在会议期间的实用性降低。

[0105] 该方法的另一版本是等待音频流中的时间点，其中(1)所有流不包含高置信度的语音或(2)具有高置信度的单个词语假设。在这些地方，假设空间可以被压紧为单个假设，这使得在不因词语分段错误而损失准确性的情况下执行组合成为可能。

[0106] 转录本基于在1080中指示的输出被提供给会议参与者中的一个或多个会议参与者。单个会议转录本基于会议系统的输出提供。转录本由个人话语和关联介质组成(诸如，幻灯片或绘图照片)。每个话语都被指派有通用时间戳、归属发言者、关联文本和/或关联音频片段，其中音频是从来自所有参与客户端的同步输入流提取的。

[0107] 附加的介质或内容(诸如，图像、笔记和其他抽象对象)可以通过时间戳(例如，白板图片在时间t捕获并且上传)或没有具体时间戳的整个会议(例如，文件在会议之后上传并且与该会议实例相关联)内联与转录本相关联。所有与会者都可以访问会议和关联数据。临时会议可以由会议所有者、所有与会者或任何人查看和修改，取决于由创建会议的实体所设置的权限。附加服务(诸如，会议概况、动作项标识和主题建模)可以使用转录本和其他关联的会议数据来提供。

[0108] 图13是图示了在处理期间在多个所选择的点处融合音频流的计算机实现的方法1300的流程图。音频流在会议期间由多个分布式设备记录。方法1300由执行操作的一个或多个处理器执行。操作1310由在一个或多个处理器上执行的对应的语音识别系统对每个音频流执行语音识别，以生成话语级后验概率作为针对每个音频流的假设。在操作1320处，假设被对准和格式化为具有关联的词语级后验概率的词语混淆网络。操作1330通过执行生成归属于发言者的词语假设流的发言者标识算法来对每个音频流执行发言者识别。在操作1340处，发言者假设与针对每个音频流的关联的后验发言者标签后验概率和归属于发言者的假设一起被格式化为混淆网络。操作1350将来自所有音频流的词语和发言者混淆网络彼此对准，以合并后验概率并且对准词语和发言者标签。通过读取具有最高后验概率的词语和发言者标签的序列，最佳的归属于发言者的词语转录本在操作1360处被创建。

[0109] 在一个实施例中，当仅来自每个流的单个词语假设被生成时，甚至可能没有后验概率，并且其中简单投票在所有流之间使用，特殊近似版本被获得。

[0110] 方法1300的操作可以在应用于音频流的连续时间窗口上执行，使得处理被递增地执行以便能够实时地产生归属于发言者的词语识别假设。基于与针对每个音频流所生成的词语假设相关联的时间标记，输入假设被及时截断到应用于所有音频流的公共时间窗口。

[0111] 输入的发言者和/或词语假设流可以源自经由N个音频流中的K个的融合的输入音频流的多个部分组合，其中K<N。备选地，输入的发言者和/或词语假设流不是源自不同的音频流，而是源自应用于N个音频流中的K个的声学模型的多个部分组合，其转而可能由原始音频信号或音频信号的融合而导致。

[0112] 在又一实施例中，基于与针对每个音频流生成的词语假设相关联的时间标记，输入假设被及时截断到应用于所有音频流的公共时间窗口。N个原始音频信号中的K个的组合或音频信号的融合可以基于音频质量准则和/或基于发言者相对于分布式设备的相对位置。

[0113] 在一个实施例中，输入发言者和/或词语假设流源自经由融合N个音频流中的K个的输入音频流的多个部分组合，其中K<N。N个声学模型输出中的K个的组合可以基于输入信号的音频质量准则和/或基于发言者相对于分布式设备的相对位置。备选地，输入发言者

和/或词语假设流可以源自应用于N个音频流中的K个的声学模型的多个部分组合，其中K<N，这又由原始音频流或音频流的融合而导致。在又一实施例中，多个声学模型的输出可以被应用于N个音频流中的K个，其中K<N，其转而由原始音频流或音频流的融合而导致，这些音频流被组合作为M个语音识别解码器的输入。

[0114] 图14A和14B图示了示例周围环境捕获设备1410。在一个实施例中，周围环境捕获设备1410是圆柱形的，鱼眼相机1411位于设备1410的顶部并且相对于设备1410朝上。麦克风阵列1413被耦合至设备1410，其位于相机1411下方并且放置在圆柱体周围来以360°捕获音频。应该注意的是，图14A中的设备可能未按比例绘制。为了捕获最优的360°视觉(例如视频或静止图像)，可能期望鱼眼相机靠近地板或桌子表面1450。在实施例中，设备可以是矮并且宽的，以避免盲点在相机1411下方。在实施例中，鱼眼相机可以被放置在紧邻麦克风阵列1413的位置。

[0115] 捕获设备1410可以与分布式设备一起使用以捕获来自分布式设备会议的音频和视频。设备1410本身可以是分布式设备中的一个分布式设备。在一个实施例中，与语音相关联的用户的标识可以由捕获设备1410单独执行，或者从捕获设备1410所收集的信息流可以与从其他分布式设备收集的信息流一起使用以在各种实施例中生成归属于发言者的转录本。

[0116] 在图14B所图示的示例中，七个麦克风1423A至1423G被包括在麦克风阵列1413中。如所示出，六个麦克风1423A至1423F被放置在平面中的设备周围，并且与设备的中心或多或少等距，并且第七麦克风1423G被放置在中心。要理解的是，该设备可以由音频可穿透材料制成，诸如轻质织物、格栅或网格，并且麦克风1423不会被鱼眼相机1421或设备1420的其他结构部分阻挡，使得声音没有被阻碍。

[0117] 在一个实施例中，鱼眼相机可以离设备1420的底座近似30厘米，并且麦克风阵列1413可以被粘附至底座1430上方近似15厘米处。在操作时，设备1420可以位于或被粘附至环境中的地板或桌子1450上。由于设备1420被放置得更靠近地板，因此360°水平视野(HFOV)可能包括更多的环境。鱼眼相机1421通常朝上被粘附至设备1420，因此天花板可能在视野内。要理解的是，设备1420的其他形状、大小或配置以及鱼眼相机1421和麦克风阵列1423的放置可以被实现，并且进行一些适应以提供类似和不同两者的结果。

[0118] 在一个实施例中，用于音频捕获的声学参数取决于麦克风的规范而变化。针对实施例的声学规范的示例在下面在表1中示出。在实施例中，声学参数应用于整个音频子系统(例如捕获的脉冲编码调制(PCM)数据)而不仅仅是麦克风。所捕获的音频可以产生足够的语音识别准确性以用于AI应用。受益于本公开的本领域普通技术人员将了解，各种声学参数可以被用于实现语音识别准确性，并且表1中的示例参数出于说明性目的。

[0119]

灵敏度（1kHz 94dB SPL）	-26 +/- ≤0.1 dB FS
信噪比（SNR），包括电源和数字滤波器噪声	≥64 dB A
频率响应	50 - > 16kHz（+/-≤3 dB）
总谐波失真	≤1%（105 dB SPL） ≤5%（115 dB SPL）
方向性	全向（针对 50->16kHz 的灵敏度差异≤1 dB）
麦克风之间的变化	针对 50->16kHz 的灵敏度差异≤1 dB
持续时间	在以下情况不会永久损失性能: 最大 SPL ≥160 dB

[0120]

	最大冲击 ≥10,000g 温度范围−40°C to +80°C

[0121] 表1.示例声学参数

[0122] 图15图示了根据一个实施例的麦克风阵列1523的示例放置。在实施例中，该设备包括放置在同一平面中的七个麦克风。六个麦克风1523A至1523F以圆形或六边形图案被放置在平面中，距中心点近似4.25厘米。第七麦克风1523G被放置在中心点。在实施例中，七个麦克风的配置包括类似规范的麦克风。要理解的是，当麦克风不相似时，可能需要从麦克风阵列所接收的音频数据的附加处理以归一化或调整音频。在示例实现中，麦克风阵列1523可以包括端口朝上的七个数字微机电系统(MEMS)麦克风。要理解的是，当麦克风没有被吸音或阻挡组件(诸如，电路板或设备外壳)阻碍时，更好的性能可能会产生。

[0123] 在一个实施例中，类似的麦克风使用设备(未示出)中的相同时钟源来计时。音频的计时或时间戳可以辅助视听数据的同步和融合。

[0124] 周围环境捕获设备可以将所有麦克风信号抽取为16位16kHz PCM数据。在该上下文中，抽取是降低信号采样率的过程。针对自动语音识别，可能不需要高于8kHz的频带。因此，16kHz的采样率可能就足够了。抽取降低了比特率，而不会损害所需的准确性。在实施例中，捕获设备可以支持附加的位深度和采样频率。在实施例中，捕获设备可能不允许改变数据宽度和采样频率，以降低驱动器复杂性并且提高稳定性。麦克风可以使用任何适当的机械阻尼机制(例如，橡胶垫圈)来安装，以减少振动和噪声。

[0125] 要理解的是，麦克风阵列中可以存在更多或更少的麦克风。然而，较少的麦克风可

能会引入发言者地点的一些不确定性。附加的麦克风可以提供更高的音频确定性或分辨率,但代价是更多的硬件和附加的计算复杂性。

[0126] 在一个实施例中,音频扬声器位于设备的底部或底座以用于用户的音频反馈。音频扬声器可以被用于反馈公告或作为AI应用的集成部分。例如,在用于会议管理的AI应用中,用户可能会请求会议备忘录被读回到与会者。设备中的集成扬声器提供反馈或请求针对操作的指令或命令。如果口头命令无法被理解,则重复命令的请求可以通过音频扬声器播放。为了减少声学反馈,音频扬声器可能面向与麦克风阵列相反的方向。经由音频扬声器播放的音频可以作为附加的同步麦克风通道回送。

[0127] 参照回图14B,在实施例中,鱼眼相机1421接收360°HFOV和水平轴上方的至少95°垂直视野(VFOV)和水平轴下方的95°VFOV(导致190°VFOV)或近似200°对角线视野(DFOV)。在实践中,捕获设备1410可以被放置在桌子或地板上,因此不需要表面下方的垂直视图。因此,在本文的讨论中,VFOV被标识为近似95°,以指示设备水平基面上方的视野。

[0128] 在一个实施例中,鱼眼相机1421包括12兆像素(MP)的一个鱼眼传感器(例如,提供4K分辨率)。相机镜头可以相对于其图像传感器安装,使得光学中心与图像传感器的中心对准,并且光轴垂直于图像传感器。相机镜头与麦克风阵列的相对位置可以是固定的和已知的。具体地,光学中心可以与麦克风阵列的中心对准,并且光轴垂直于麦克风阵列。

[0129] 图16图示了具有上述周围环境捕获设备1610和会议服务器(称为云服务器1620)的AI系统1600。在示例中,用户1630与AI应用1623交互。要理解的是,AI应用1623可以驻留在云服务器1620或本地设备(未示出)上。视听数据可以由AI捕获设备1610以360°捕获。如上面讨论的,捕获设备1610可以包括提供360°HFOV和大约95°VFOV的鱼眼相机1611。捕获设备1610还可以包括麦克风阵列1613来以360°捕获音频。

[0130] 由相机1611接收的图像和视频流的视频压缩可以由设备上的处理器1615执行。视频模式以及压缩协议和准则可以由用户可选择的软件控制来控制。除了压缩之外,视听数据还可以通过加密来保护,以防止未经授权的人获得数据。在实施例中,压缩1618可以由设备上的电路系统执行,并且由软件开关控制。

[0131] 预处理1617(例如,基于图像内容裁剪图像或降噪)可以由处理器所执行的逻辑在压缩1618之前执行。在实施例中,预处理可以包括声学回声消除(AEC),以减少由耦合至设备的发言者1612引起的反馈、噪声和回声。

[0132] 在实施例中,用于关键字发现(KWS)的本地过程可以被包括在内,以便监听用于周围环境捕获设备的设备命令,诸如唤醒或关闭设备。本地KWS可能有利于召回率与精度,并且它可能基于所减小的麦克风阵列(例如,两个麦克风而不是完整阵列)。

[0133] 当AEC在设备1610上执行时,包括扬声器音频的声道可能不需要被发送给模型,以执行传感器融合1621。所压缩的视听数据可以由传输单元1619发送给云服务器1620。传输单元1619可以包括以下一项或多项:用于有线通信的网络接口卡,诸如以太网连接;使用无线协议的无线收发器,诸如WiFi®、蓝牙®、NFC;或其他通信部件。在实施例中,音频反馈可以经由无线通道中的一个无线通道被发送给设备。云服务器1620可以针对AI应用1623执行传感器融合1621。因此,压缩可以被执行,以减少经由传输单元1619传输给云的带宽。

[0134] 图17是图示了减少通过网络发送给会议服务器以用于生成转录本的音频流的数量的计算机实施的方法1700的流程图。方法1700开始于在操作1710处从检测来自多个用户

的会议的语音的多个(例如,三个或多个)麦克风接收多个音频通道。在操作1720处,活动发言者的方向被估计。语音分离模型被用于在操作1730处选择对应于主要麦克风和次级麦克风的两个通道,或者可以对应于所融合的音频通道。两个所选择的通道在操作1740处被发送给会议服务器,以生成智能会议转录本。通过减少发送给会议服务器的数据量,带宽可以被节省。由于所选择的数据可以说是最佳数据,因此几乎没有任何准确性丢失。

[0135] 在一个实施例中,麦克风由处于固定配置的设备支撑。固定配置可以包括具有被配置为包括多个用户的视野的相机。定位声源可以通过执行在来自相机的音频和视频通道上训练的模型来执行。例如,如果一个用户使用具有相机的膝上型计算机,则膝上型计算机可以提供音频和视频通道两者。音频通道可以相对于参考音频通道同步,并且相同的时间差可以被用于同步视频通道。图像识别可以在视频通道上使用以将用户标识为发言者,用于在产生转录本时进行分类。在又一实施例中,膝上型计算机执行图像处理以确定用户正在说话,并且在音频通道上提供将用户标识为发言者的标注。该标注然后可以被用于分类,而无需从膝上型计算机传输视频通道。

[0136] 在又一实施例中,麦克风与多个分布式设备相关联。分布式设备可以包括分别与多个用户相关联的无线设备。分布式设备中的至少一个设备可以包括提供用户中的至少一个用户的视频的相机。

[0137] 在又一实施例中,麦克风包括以固定配置支持的麦克风以及与关联于用户的分布式设备相关联的麦克风。该方法可以由在固定位置支持麦克风的设备或接收多个音频通道的边缘设备中的一个或多个设备来执行。语音分离模型可以在边缘设备上执行。

[0138] 在其他实施例中,客户端侧处理(在分布式设备、环境捕获设备和/或边缘服务器中的一个或多个设备上的处理)被用于减少会议服务器所需的计算资源以及减少用于处理来自分布式设备的分布式会议信息流的网络带宽量。除了如上所述减少经由网络发送给会议服务器的流数目之外,波束形成还可以在客户端侧执行以及生成音频水印和会议码。在其他实施例中,模型大小可以被减小和量化,以在客户端侧更好地运行。目标函数也可以被修改,以在客户端大小上更好地运行。代替输出语音掩码,声源定位可以用相应较少的计算使用。

[0139] 音频和视频通道两者都可以被用于使语音归属于用户,用于创建所分类的转录本。视听分类方法允许组合来自分布式传感器的语音标识、声源定位、面部追踪/标识和视觉活动发言者检测,以实现稳健的分类。

[0140] 图18是图示了用于使用来自分布式设备的视频和音频通道两者、视听数据以提供更好的发言者标识的计算机实施的方法1800的流程图。方法1800开始于在操作1810处,在会议服务器上从智能会议中所包括的多个分布式设备集接收信息流。在操作1820处,代表至少两个信息流中的至少两个用户的语音的音频信号被接收。在操作1830处,信息流中的至少一个用户的至少一个视频信号被接收。在操作1840中,所接收的音频和视频信号被用于根据所接收的音频和视频信号将所接收的音频信号中的语音与具体用户相关联。在操作1850处,智能会议的转录本被生成,其具有与语音相关联的用户的指示。

[0141] 在一个实施例中,多个分布式设备是与智能会议中的用户相关联的移动无线设备。移动无线设备可以包括麦克风和提供至少一个视频信号的相机。在其他实施例中,多个分布式设备包括具有以固定配置支持的多个麦克风的设备,每个麦克风提供所接收的音频

信号中的一个音频信号。该设备可以包括具有被配置为在智能会议中包括多个用户并且提供至少一个视频信号的视野的相机。

[0142] 在一个实施例中，融合模型在所接收的音频和视频信号上使用，以将具体用户与语音相关联。在实施例中，视听数据可以由会议服务器分析。视听数据首先可以在经由网络发送给会议服务器之前压缩。在另一实施例中，融合模型作为集成系统被耦合至捕获设备。本文的讨论出于说明目的而不是作为限制来描述会议服务器。

[0143] 会议服务器根据需要对数据进行解压缩、解码或解密。视听数据可以由AI应用利用LSTM模型进行融合和分析，例如标识或推断视听数据中的特征，诸如但不限于：音频方向；图像中的发言者地点；发言者移动；语音签名；面部签名；手势；和/或对象。在示例中，AI应用需要语音识别或面部识别。LSTM模型可以使用传感器数据利用特定于AI应用的数据进行训练。在实施例中，多于一个模型或分析引擎可以被使用，如上面讨论的。

[0144] 在实施例中，语音可以被标识，并且使用视频数据的手势识别可以被执行。LSTM模型可以使用所标识的语音和所识别的手势以提供数据的可能融合，并且将可能的结果发送给AI应用。在示例中，与语音命令组合的手势向AI应用提供具体的控制命令。在示例中，视频数据的分析指示眼睛注视或追踪眼睛移动，以推断用户正在看哪里。眼睛注视分析可能会产生用于AI应用的控制命令，并且可能会基于与音频数据的融合而有所不同。

[0145] 在实施例中，LSTM模型针对具体的AI应用进行训练，并且基于所融合的数据针对该应用提供控制或命令。在另一实施例中，LSTM模型可能更通用，并且向AI应用提供可能的相关数据，诸如具有发言者ID的每个发言者的音频流以及环境中的地点，以进行输入的进一步处理和解释。在该示例中，AI应用使用音频和视频流输入来导出适当的命令或执行动作。

[0146] 一个实施例利用具有12MP传感器的鱼眼相机。另一实施例包括红外(IR)或其他深度传感器以提供三维(3D)或深度信息。如果没有足够的深度传感器来覆盖整个HFOV，则深度信息可能无法以360°可用。捕获设备的变化可以被提供以适应广泛的用户可接受的各种价格点，或者用于不同的应用。例如，包括深度传感器或高分辨率传感器可能会增加设备的成本或复杂性，超出所选AI应用所需的程度。

[0147] 图19是图示了根据示例实施例的用于基于用户偏好定制输出的计算机实现的方法1900的流程图。方法1900中的操作由会议服务器或系统(例如，会议服务器135)使用上述组件来执行。因此，方法1900是参照会议服务器通过示例描述的。然而，应该了解的是，方法1900的至少一些操作可以被部署在各种其他硬件配置上，或者由驻留在网络环境中的其他地方的类似组件来执行。因此，方法1900不旨在被限于会议服务器。

[0148] 在操作1910中，会议服务器从多个分布式设备接收音频流。在示例实施例中，音频流包括在两个或多个用户的会议期间由多个分布式设备中的一个或多个分布式设备检测到的语音。在一些实施例中，会议是临时会议。在这些实施例中，服务器可以对所接收的音频流执行盲波束形成或连续语音分离，以将语音与背景噪声或同时说话的不同发言者分离到单独的音频通道中。在一些情况下，音频流被比较，以确定音频流表示来自(相同)临时会议的声音。然后会议实例被生成，以处理被标识为来自临时会议的音频流。

[0149] 在操作1920中，分布式设备中的一个分布式设备的用户的身份由会议服务器标识。在一个实施例中，用户基于由与会议相关联的相机(例如，相机155、相机1821)捕获的视

频信号来标识。视频信号被传输给会议服务器。会议服务器将来自视频信号的用户图像与已知(例如,已注册)用户的所存储的图像进行比较以确定匹配。如果所存储的图像与视频信号中的用户的捕获图像相匹配,然后用户被标识。在一个实施例中,用户的图像被存储或与用户的用户简档相关联。

[0150] 在备选实施例中,用户基于语音签名来标识。在该实施例中,来自音频流的语音被解析或分类,并且与已知用户的存储的语音签名进行比较。如果所存储的语音签名与来自音频流的所解析/分类的语音相匹配,那么用户被标识。在一个实施例中,用户的语音签名被存储或与用户的用户简档相关联。

[0151] 在操作1930中,所标识的用户的语言偏好被确定。在一些实施例中,所标识的用户的用户简档被访问。用户简档至少包括对用户语言的预定偏好。在一些情况下,预定偏好由用户建立(例如,明确指示)。在其他情况下,预定偏好基于与用户相关联的设备(例如,分布式设备,诸如蜂窝电话或膝上型计算机)的设备配置来确定。例如,设备可以被配置为以英文或中文运作。

[0152] 在操作1940中,会议服务器生成上述转录本。在示例实施例中,来自音频流的语音被转换为文本,以生成基于文本的转录本或数字转录本。在一个实施例中,如上面讨论的,实时转录本是基于短词序列生成的。在一些实施例中,数据的后期融合可以通过针对并行处理以产生短语的多个音频通道的语音识别来执行。从多个音频通道所导出的短语实时或近实时地组合。在一个实施例中,近似两秒的语音被组合。因此,音频流本质上是在它们被接收时处理的。几秒(例如,两秒)的非重叠滑动窗口被用于处理音频流,从而减少用于转录本生成的时延。

[0153] 在操作1950中,会议服务器根据用户的语言偏好翻译转录本。在一些实施例中,会议服务器从操作1940获取所生成的转录本,并且将所生成的转录本中的文本翻译为偏好语言的文本。在其他实施例中,会议服务器从操作1940获取所生成的转录本,并且将所生成的转录本转换为偏好语言的语音。更进一步地,一些实施例可以执行文本翻译和语音翻译两者。在示例实施例中,针对来自转录本的每个所翻译的话语的用户(例如,发言者)身份被提供有所翻译的转录本。在一些情况下,用户身份是从与分布式设备相关联的用户标识符获得的。

[0154] 在操作1960中,所翻译的转录本被提供给用户的设备(例如,分布式设备)。在一些实施例中,该设备包括被用于从用户捕获音频的相同设备。所翻译的转录本可以例如作为显示在设备的显示设备(例如,屏幕)上的文本或者作为通过使用文本到语音经由发言者设备(例如,听筒、助听器或扩音器)的语音音频来提供。在一些实施例中,分类结果也可以被提供。

[0155] 虽然图19的方法1900被描述为具有以特定顺序的操作,但是备选实施例可以以不同顺序的操作来执行方法1900。例如,标识用户(操作1920)和确定语言偏好(操作1930)可以在转录本被生成之后或被生成时(操作1940)并且在翻译转录本之前(操作1950)发生。

[0156] 图20是计算机系统2000的示意性框图,以实施和管理智能会议经由多个分布式设备、边缘设备和基于云的设备的处置,并且执行根据示例实施例的方法和算法。所有组件不需要在各种实施例中使用。

[0157] 计算机2000形式的一个示例计算设备包括处理单元2002、存储器2003、可移除存

储装置2010和不可移除存储装置2012。尽管示例计算设备被图示和描述为计算机2000，但是计算设备在不同实施例中可以是不同形式。例如，计算设备可以反而是智能电话、平板计算机、智能手表或者包括与关于图20图示和描述的元件相同或类似的元件的其他计算设备。诸如智能电话、平板计算机和智能手表等设备通常被统称为移动设备、分布式设备或用户设备。

[0158] 尽管各种数据存储元件被图示为计算机2000的一部分，但是存储装置也可以或者备选地包括经由网络(诸如，互联网)可访问的基于云的存储装置、基于服务器的存储装置或者智能存储设备(SSD)。还要注意的是，SSD可以包括处理器，在该处理器上，解析器可以被运行，从而允许通过I/O通道在SSD和主存储器之间传送所解析的滤波数据。

[0159] 存储器2003可以包括易失性存储器2014和非易失性存储器2008。计算机2000可以包括或访问计算环境，该计算环境包括各种计算机可读介质，诸如易失性存储器2014和非易失性存储器2008、可移除存储装置2010和不可移除存储装置2012。计算机存储装置包括随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM)和电可擦除可编程只读存储器(EEPROM)、闪存或其他存储器技术、光盘只读存储器(CD ROM)、数字通用盘(DVD)或者其他光盘存储装置、磁带盒、磁带、磁盘存储装置或其他磁性存储设备或者能够存储计算机可读指令的任何其他介质。

[0160] 计算机2000可以包括或访问计算环境，该计算环境包括输入接口2006、输出接口2004和通信接口2016。输出接口2004可以包括也可以用作输入设备的显示设备，诸如触摸屏。输入接口2006可以包括以下一项或多项：触摸屏、触摸板、鼠标、键盘、相机、一个或多个设备特定按钮、集成在计算机2000内或经由有线或无线数据连接耦合至计算机2000的一个或多个传感器和其他输入设备。计算机可以使用通信连接在联网环境中操作，以连接至一个或多个远程计算机，诸如数据库服务器。远程计算机可以包括个人计算机(PC)、服务器、路由器、网络PC、对等设备或者其他公共数据流网络开关等。通信连接可以包括局域网(LAN)、广域网(WAN)、蜂窝、Wi-Fi、蓝牙或其他网络。根据一个实施例，计算机2000的各种组件与系统总线2020连接。

[0161] 存储在计算机可读介质上的计算机可读指令由计算机2000的处理单元2002可执行(诸如，程序2018)。在一些实施例中，程序2018包括软件以实施一种或多种方法来实施会议app和会议服务器以及本文描述的模块、方法和算法。硬盘驱动器、CD-ROM和RAM是物品的一些示例，包括非暂时性计算机可读设备，诸如存储设备。术语计算机可读存储设备不包括载波到载波被认为过于短暂的程度。存储装置还可以包括联网存储装置，诸如存储区域网络(SAN)。计算机程序2018以及工作空间管理器2022可以被用于使处理单元2002执行本文所描述的一种或多种方法或算法。

[0162] 可执行指令和机器存储介质

[0163] 如本文使用的，术语“机器存储介质”、“设备存储介质”、“计算机存储介质”、“计算机可读存储介质”、“计算机可读存储设备”(统称为“机器存储介质”)是指相同的事物，并且可以在本公开中互换使用。该术语指的是存储可执行指令和/或数据的单个或多个存储设备和/或介质(例如集中式或分布式数据库和/或关联的缓存和服务器)以及包括多个存储装置或设备的基于云的存储系统或存储网络。因此，术语应该被理解为包括但不限于固态存储器以及光学和磁性介质，包括处理器内部或外部的存储器。机器存储介质、计算机存储

介质和/或设备存储介质的具体示例包括非易失性存储器，通过示例包括：半导体存储器设备，例如可擦除可编程只读存储器(EPROM)、电可擦除可编程只读存储器(EEPROM)、FPGA和闪存设备；磁盘，诸如内部硬盘和可移除盘；磁光盘；以及CD-ROM和DVD-ROM盘。术语机器存储介质、计算机存储介质和设备存储介质具体地排除了载波、调制数据信号和其他这种介质到这种介质被认为过于短暂的程度。其他这种介质也在下面讨论的术语“信号介质”下覆盖。在该上下文中，机器存储介质是非暂时性的。

[0164] 信号介质

[0165] 术语“信号介质”或“传输介质”应该被认为包括任何形式的调制数据信号、载波等。术语“调制数据信号”是指其特性中的一个或多个特性以这种方式被设置或改变为在该信号中对信息进行编码的信号。

[0166] 计算机可读介质

[0167] 术语“机器可读介质”、“计算机可读介质”和“设备可读介质”表示相同的事物，并且在本公开中可以互换使用。该术语被限定以包括机器存储介质和信号介质两者。因此，该术语包括存储设备/介质和载波/调制数据信号。

[0168] 示例

[0169] 示例1是一种用于基于分布式系统中的用户偏好提供定制输出的计算机实施的方法。该方法包括：从智能会议中所涉及的多个分布式设备来接收音频流；标识与多个分布式设备中的分布式设备相对应的用户；确定用户的偏好语言；由硬件处理器，从所接收的音频流生成转录本；将转录本翻译为用户的偏好语言以形成所翻译的转录本；以及将所翻译的转录本提供给分布式设备。

[0170] 在示例2中，示例1的主题可以可选地包括，其中提供所翻译的转录本包括提供具有所翻译的文本的转录本。

[0171] 在示例3中，示例1至示例2的主题可以可选地包括，其中提供所翻译的转录本包括将所翻译的转录本的文本转换为语音。

[0172] 在示例4中，示例1至示例3的主题可以可选地包括，其中提供所翻译的转录本包括：针对转录本的每个所翻译的话语提供发言者身份。

[0173] 在示例5中，示例1至示例4的主题可以可选地包括，其中确定用户的偏好语言包括访问先前针对用户所建立的、指示偏好语言的用户偏好。

[0174] 在示例6中，示例1至示例5的主题可以可选地包括，其中智能会议是临时会议，该方法还包括比较音频流以确定音频流代表来自临时会议的声音；以及响应于比较确定音频流代表来自临时会议的声音，生成会议实例以处理音频流。

[0175] 在示例7中，示例1至示例6的主题可以可选地包括对所接收的音频流执行连续语音分离，以将来自同时说话的不同发言者的语音分离到单独的音频通道中，生成转录本基于所分离的音频通道。

[0176] 在示例8中，示例1至示例7的主题可以可选地包括，其中标识用户包括接收捕获用户的视频信号；以及将用户的所存储的图像与视频信号进行比较以标识用户。

[0177] 在示例9中，示例1至示例8的主题可以可选地包括，其中标识用户包括将用户的所存储的语音签名与来自音频流的语音相匹配。

[0178] 在示例10中，示例1至示例9的主题可以可选地包括，其中标识用户包括获得与分

布式设备相关联的用户标识符。

[0179] 示例11是一种用于基于分布式系统中的用户偏好提供定制输出的机器存储介质。机器可读存储设备将一个或多个处理器配置为执行操作，该操作包括：从智能会议涉及的多个分布式设备接收音频流；标识与多个分布式设备中的分布式设备相对应的用户；确定用户的偏好语言；从所接收的音频流生成转录本；将转录本翻译为用户的偏好语言以形成所翻译的转录本；以及将所翻译的转录本提供给分布式设备。

[0180] 在示例12中，示例11的主题可以可选地包括，其中提供所翻译的转录本包括提供具有所翻译的文本的转录本。

[0181] 在示例13中，示例11至示例12的主题可以可选地包括，其中提供所翻译的转录本包括将所翻译的转录本的文本转换为语音。

[0182] 在示例14中，示例11至示例13的主题可以可选地包括，其中提供所翻译的转录本包括针对转录本的每个所翻译的话语提供发言者身份。

[0183] 在示例15中，示例11至示例14的主题可以可选地包括，其中确定用户的偏好语言包括访问先前针对用户所建立的、指示偏好语言的用户偏好。

[0184] 在示例16中，示例11至示例15的主题可以可选地包括，其中智能会议是临时会议，该方法还包括比较音频流以确定音频流代表来自临时会议的声音；以及响应于比较确定音频流代表来自临时会议的声音，生成会议实例以处理音频流。

[0185] 在示例17中，示例11至示例16的主题可以可选地包括，其中该操作还包括对所接收的音频流执行连续语音分离，以将来自同时说话的不同发言者的语音分离到单独的音频通道中，生成转录本基于所分离的音频通道。

[0186] 在示例18中，示例11至示例17的主题可以可选地包括，其中标识用户包括接收捕获用户的视频信号；以及将用户的所存储的图像与视频信号进行比较以标识用户。

[0187] 在示例19中，示例11至示例18的主题可以可选地包括，其中标识用户包括将用户的所存储的语音签名与来自音频流的语音相匹配。

[0188] 示例20是一种用于基于分布式系统中的用户偏好提供定制输出的设备。该系统包括一个或多个处理器和存储指令的存储设备，在由一个或多个硬件处理器执行时，该指令使一个或多个硬件处理器执行操作，该操作包括：从智能会议所涉及的多个分布式设备接收音频流；标识与多个分布式设备中的分布式设备相对应的用户；确定用户的偏好语言；从所接收的音频流生成转录本；将转录本翻译为用户的偏好语言以形成所翻译的转录本；以及将所翻译的转录本提供给分布式设备。

[0189] 尽管一些实施例已经在上面详细描述，但是其他修改也是可能的。例如，在附图中描绘的逻辑流程不需要所示的特定顺序或者相继顺序，以实现期望的结果。其他步骤可以被提供或者步骤可以从所描述的流程中消除，并且其他组件可以被添加到所描述的系统或者从所描述的系统中移除。其他实施例可以在以下权利要求的范围内。

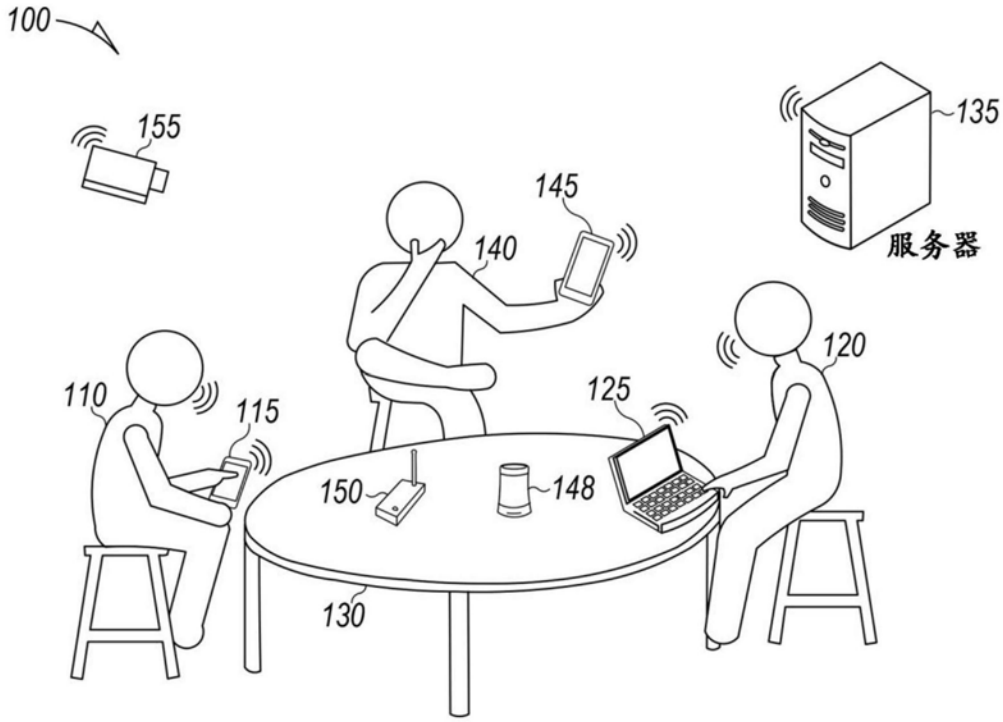

100
155
135
服务器
145
140
120
110
115
125
150
148
130

图1

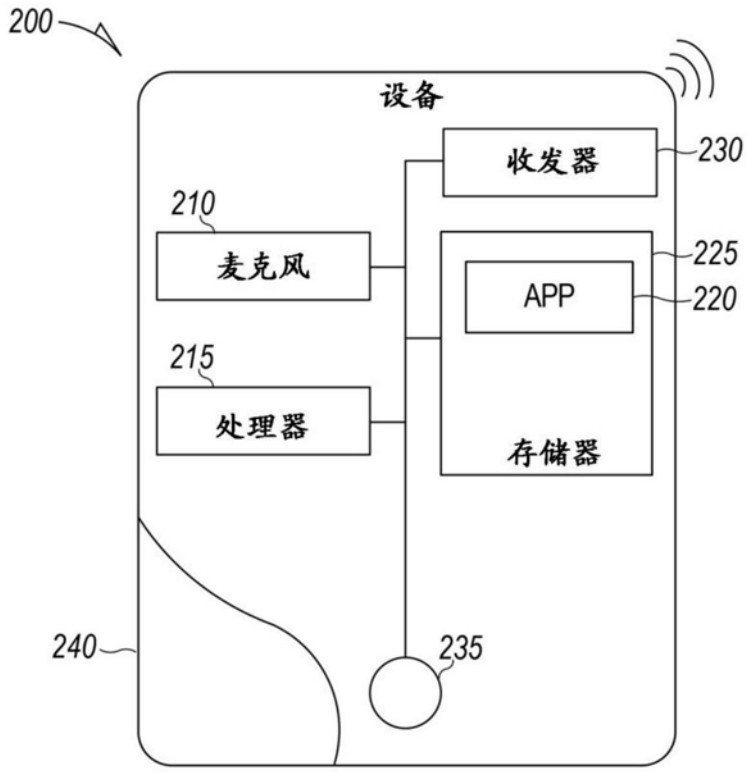

200
设备
收发器
230
210
麦克风
225
APP
220
215
处理器
存储器
240
235

图2

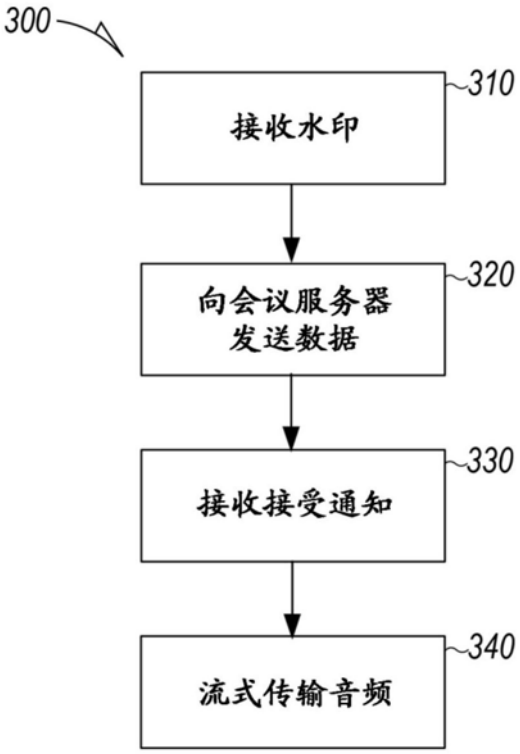

300
接收水印
310
向会议服务器
发送数据
320
接收接受通知
330
流式传输音频
340

图3

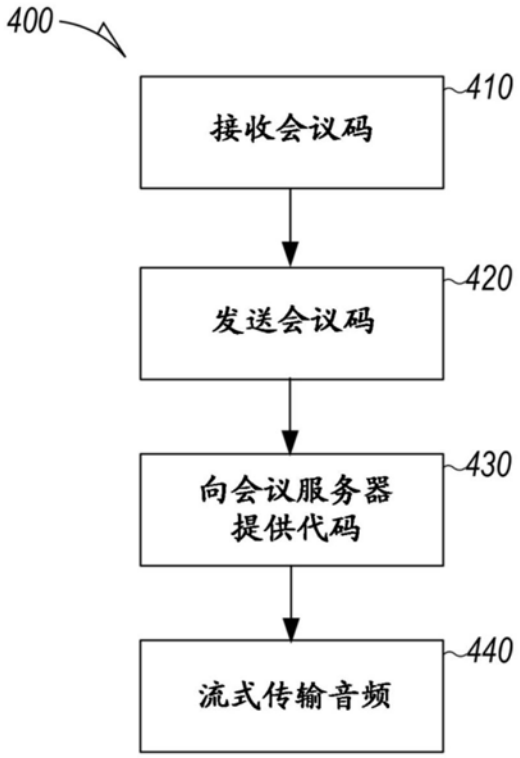

400
接收会议码
410
发送会议码
420
向会议服务器
提供代码
430
流式传输音频
440

图4

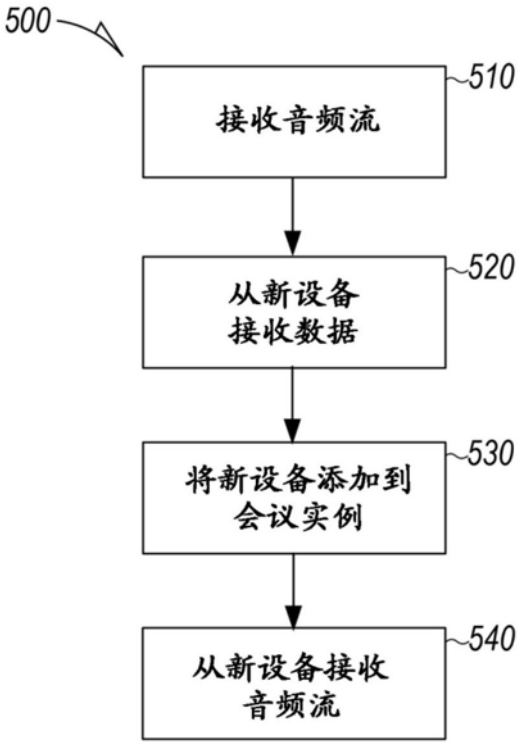

500
接收音频流
510
从新设备
接收数据
520
将新设备添加到
会议实例
530
从新设备接收
音频流
540

图5

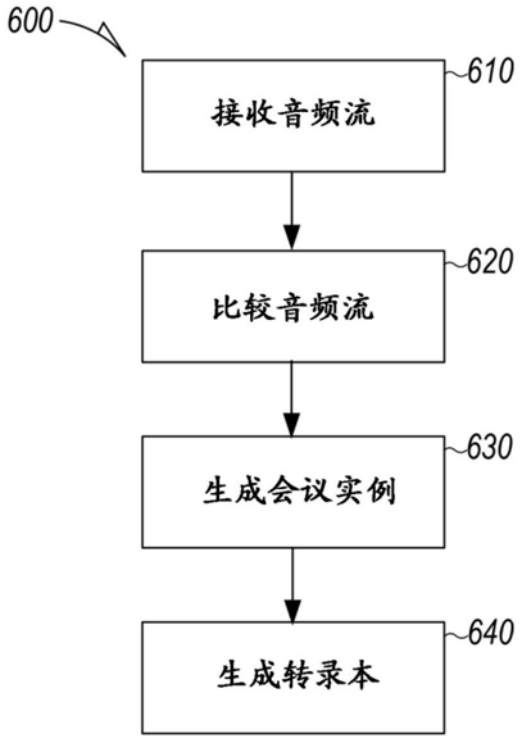

600
接收音频流
610
比较音频流
620
生成会议实例
630
生成转录本
640

图6

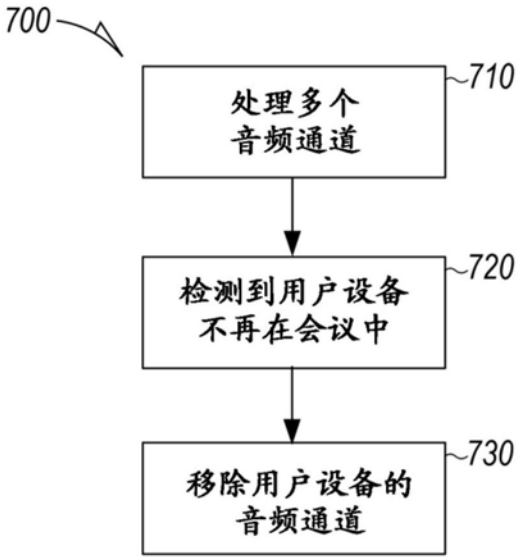

700
处理多个音频通道
710
检测到用户设备不再在会议中
720
移除用户设备的音频通道
730

图7

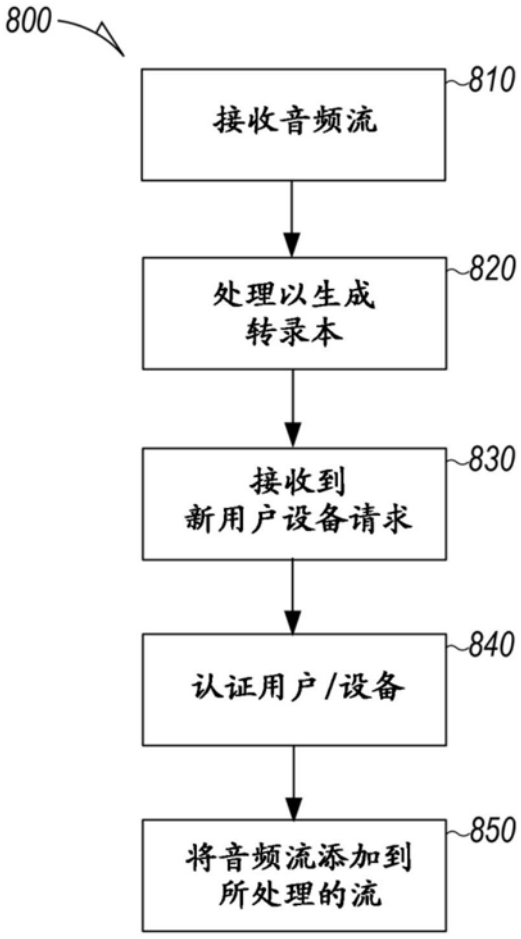

800
接收音频流
810
处理以生成转录本
820
接收到新用户设备请求
830
认证用户/设备
840
将音频流添加到所处理的流
850

图8

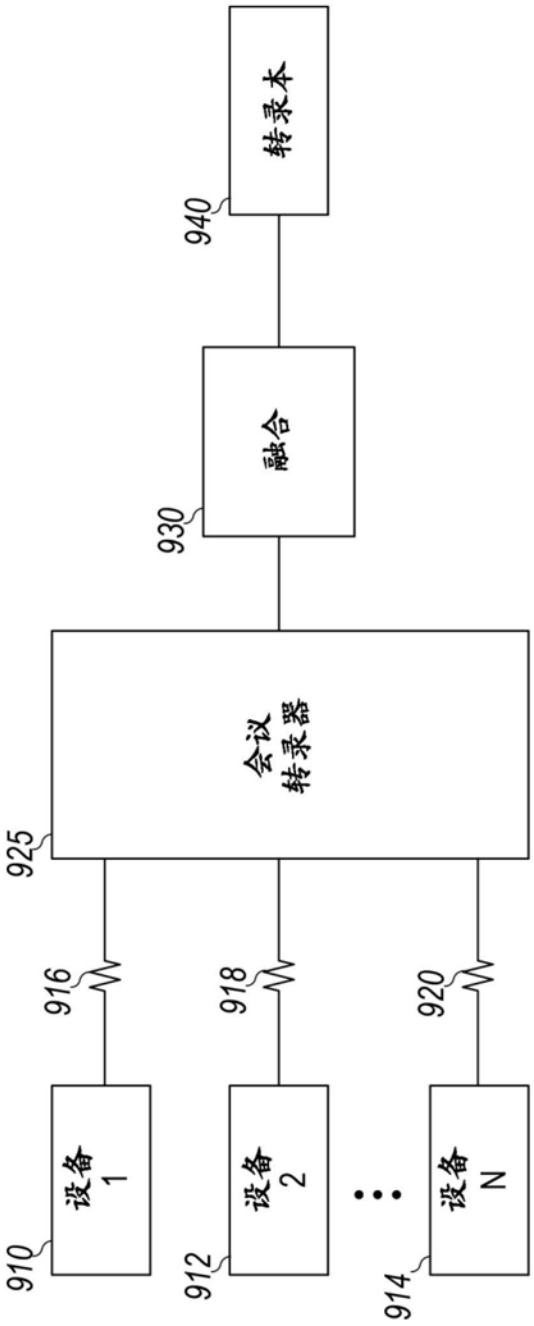


图9

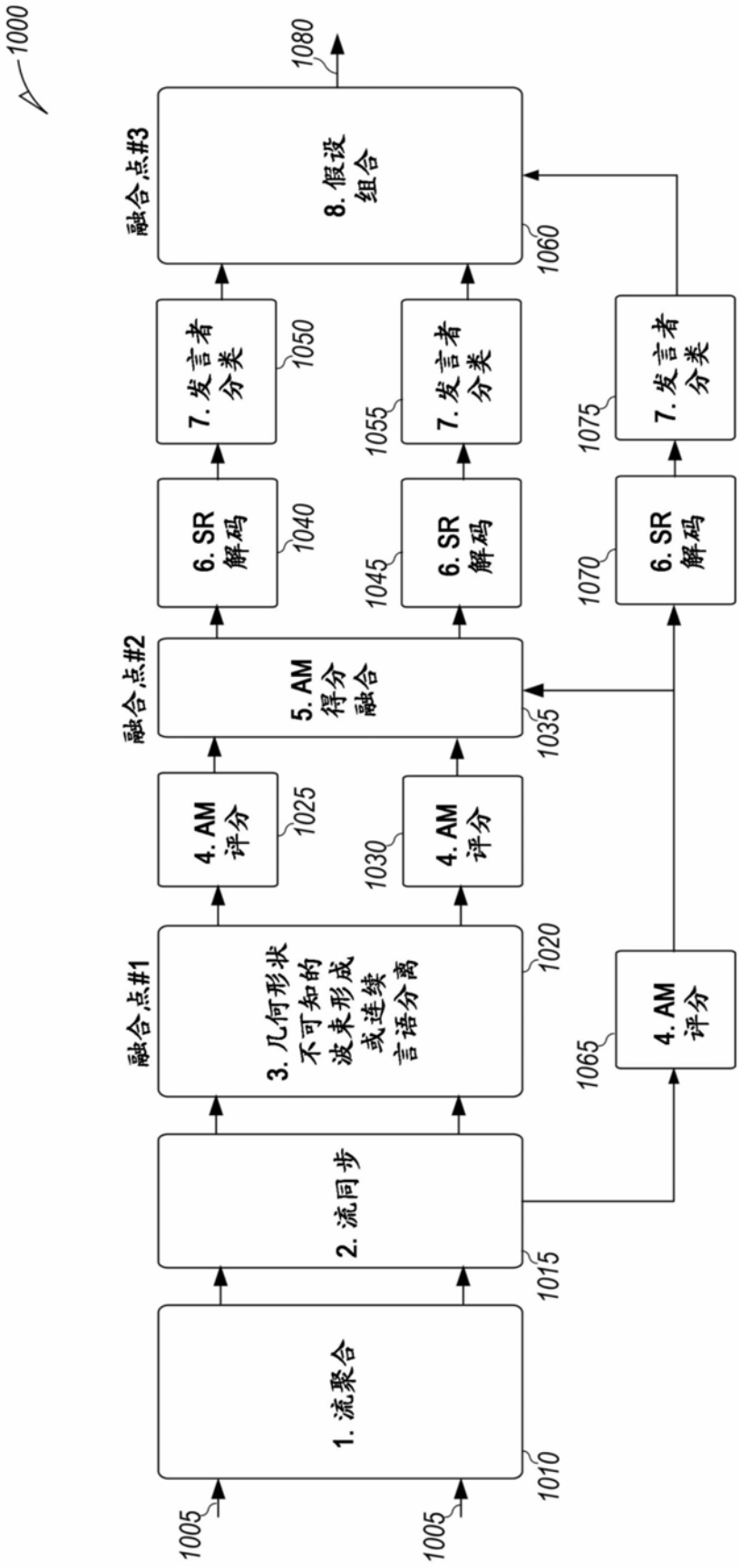

1000
1005
1005
1. 流聚合
1010
2. 流同步
1015
融合点#1
3. 几何形状不可知的波束形成或连续言语分离
1020
4. AM评分
1025
4. AM评分
1030
融合点#2
5. AM得分融合
1035
6. SR解码
1040
6. SR解码
1045
7. 发言者分类
1050
7. 发言者分类
1055
融合点#3
8. 假设组合
1060
1065
4. AM评分
1070
6. SR解码
1075
7. 发言者分类
1080

图10

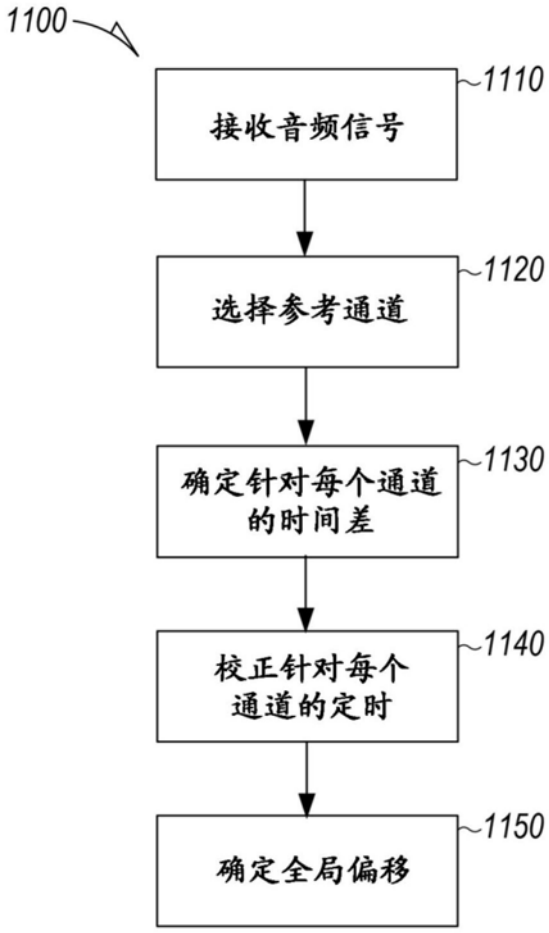

1100
接收音频信号
1110
选择参考通道
1120
确定针对每个通道的时间差
1130
校正针对每个通道的定时
1140
确定全局偏移
1150

图11

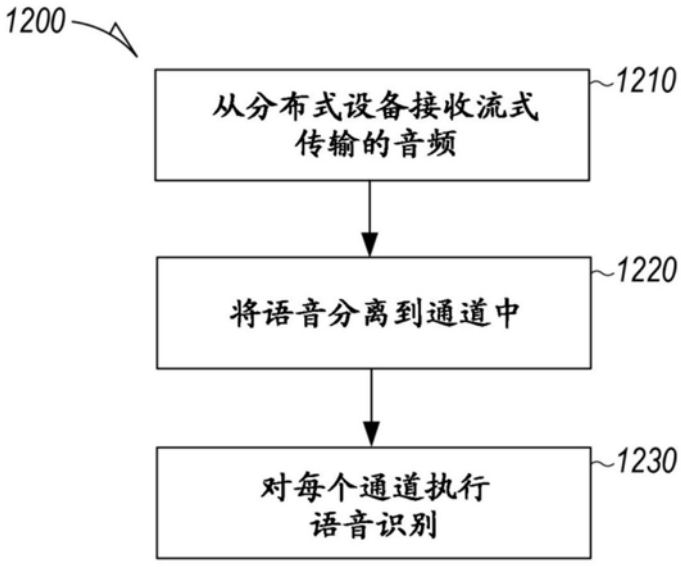

1200
从分布式设备接收流式传输的音频
1210
将语音分离到通道中
1220
对每个通道执行语音识别
1230

图12

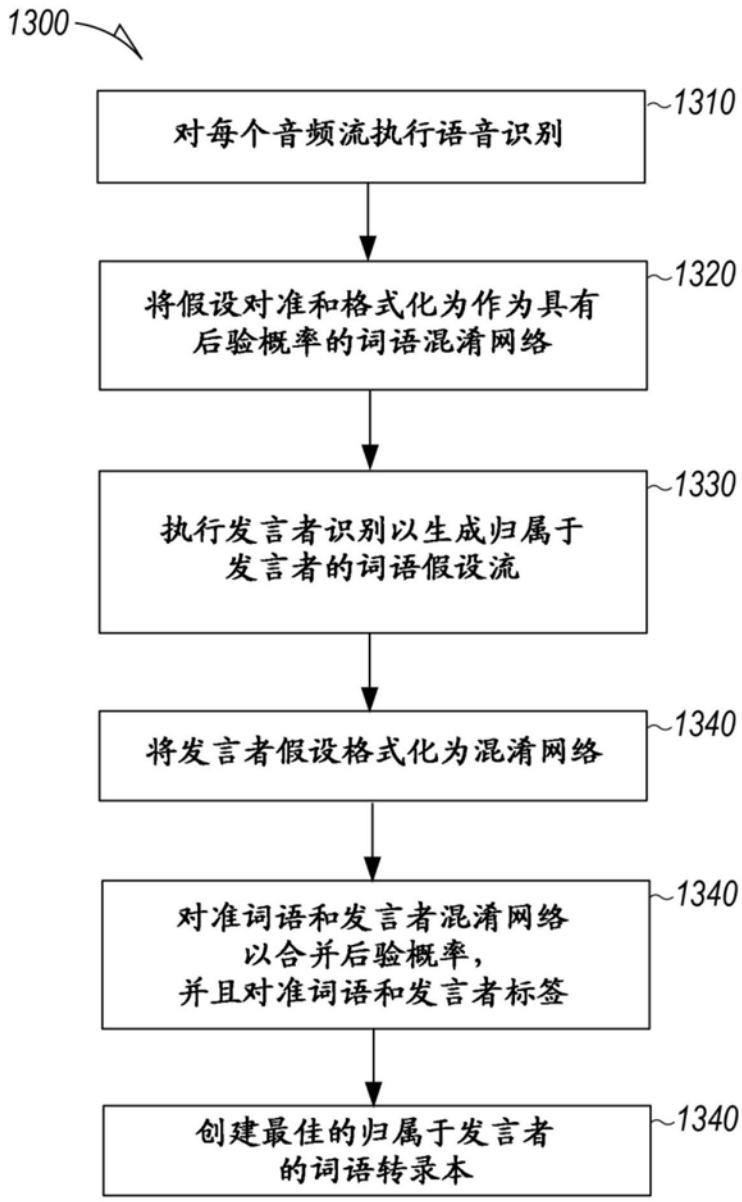

1300
对每个音频流执行语音识别
1310
将假设对准和格式化为作为具有后验概率的词语混淆网络
1320
执行发言者识别以生成归属于发言者的词语假设流
1330
将发言者假设格式化为混淆网络
1340
对准词语和发言者混淆网络以合并后验概率，并且对准词语和发言者标签
1340
创建最佳的归属于发言者的词语转录本
1340

图13

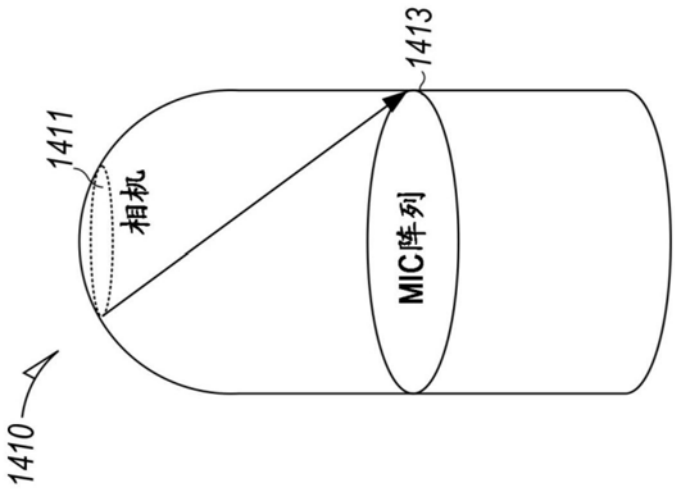

1413
1411
相机
MIC阵列
1410

图14A

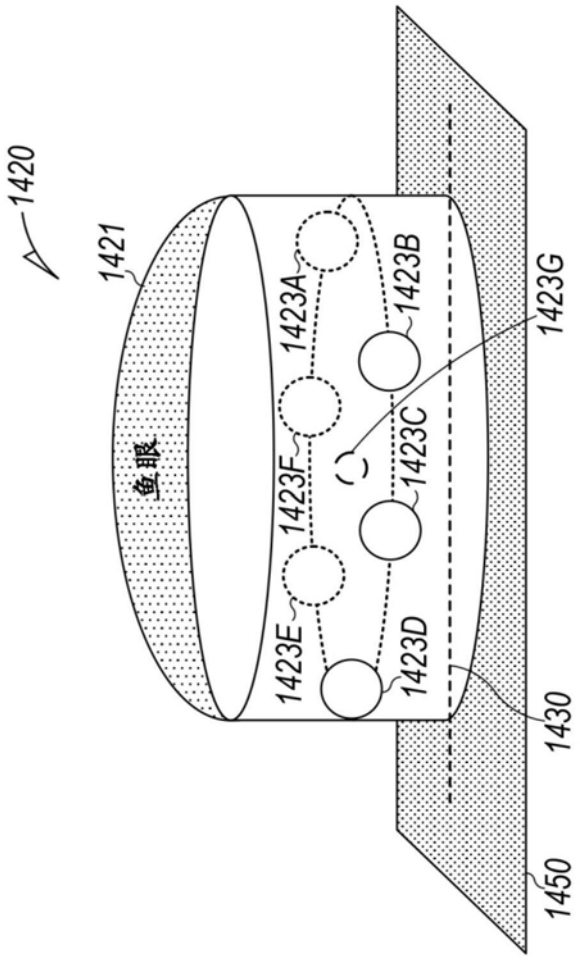

1420
1421
鱼眼
1423A
1423B
1423C
1423D
1423E
1423F
1423G
1430
1450

图14B

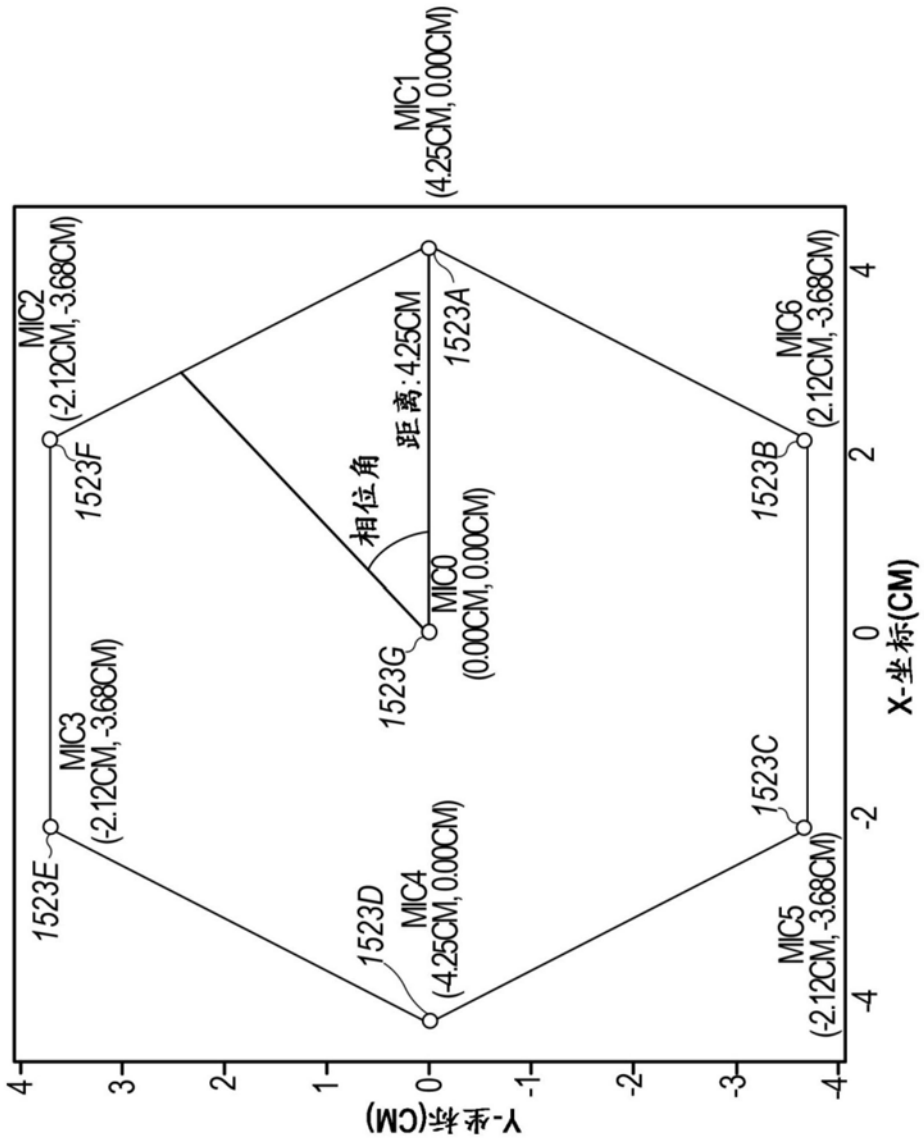

MIC1
(4.25CM,0.00CM)
MIC2
(-2.12CM,-3.68CM)
MIC3
(-2.12CM,-3.68CM)
MIC4
(-4.25CM,0.00CM)
MIC5
(-2.12CM,-3.68CM)
MIC6
(2.12CM,-3.68CM)
MIC0
(0.00CM,0.00CM)
1523A
1523B
1523C
1523D
1523E
1523F
1523G
距离:4.25CM
相位角
X-坐标(CM)
Y-坐标(CM)

图15

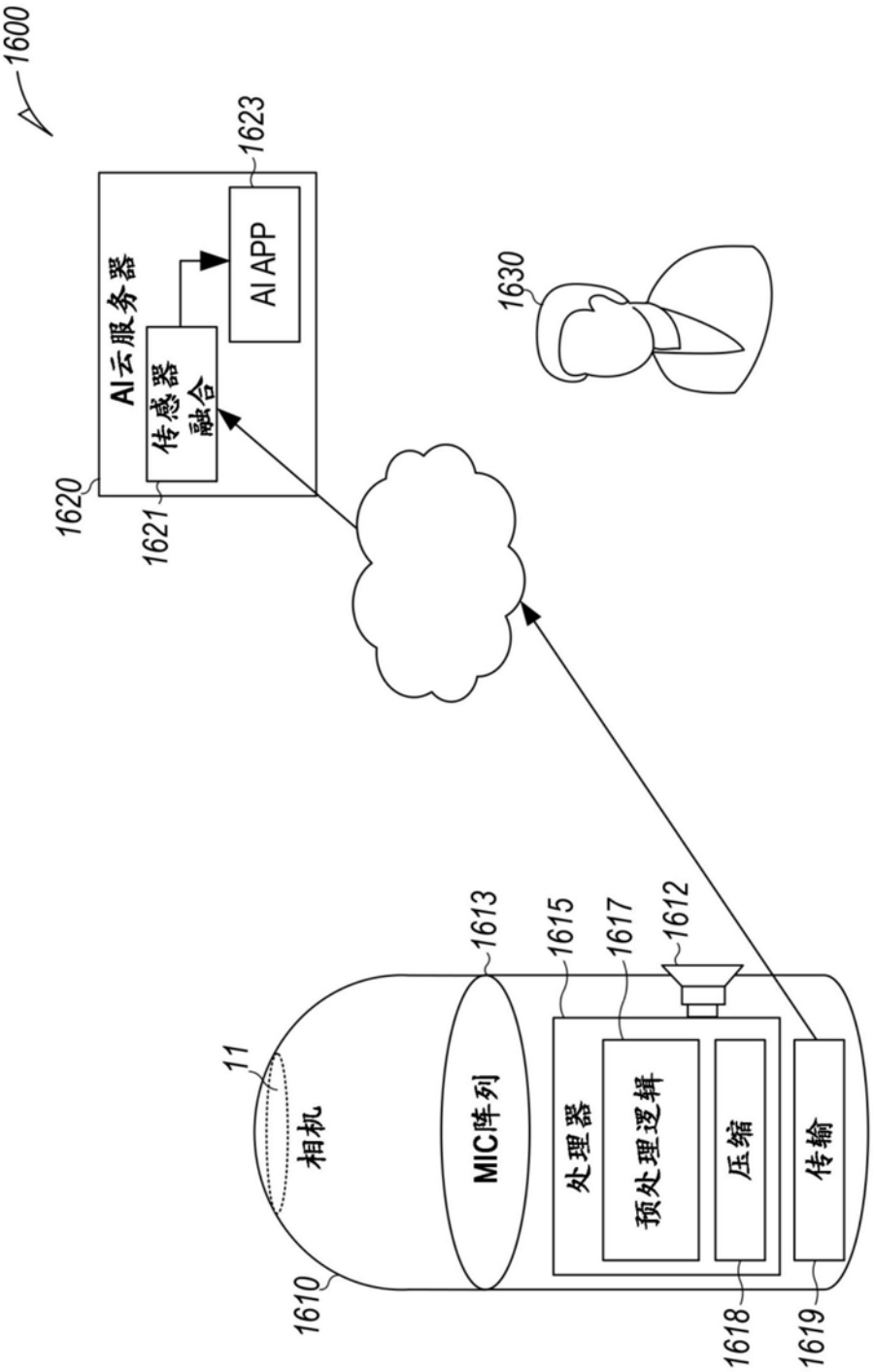

1600
1623
AI APP
AI云服务器
传感器融合
1620
1621
1630
1613
1615
1617
1612
11
相机
MIC阵列
处理器
预处理逻辑
压缩
传输
1610
1618
1619

图16

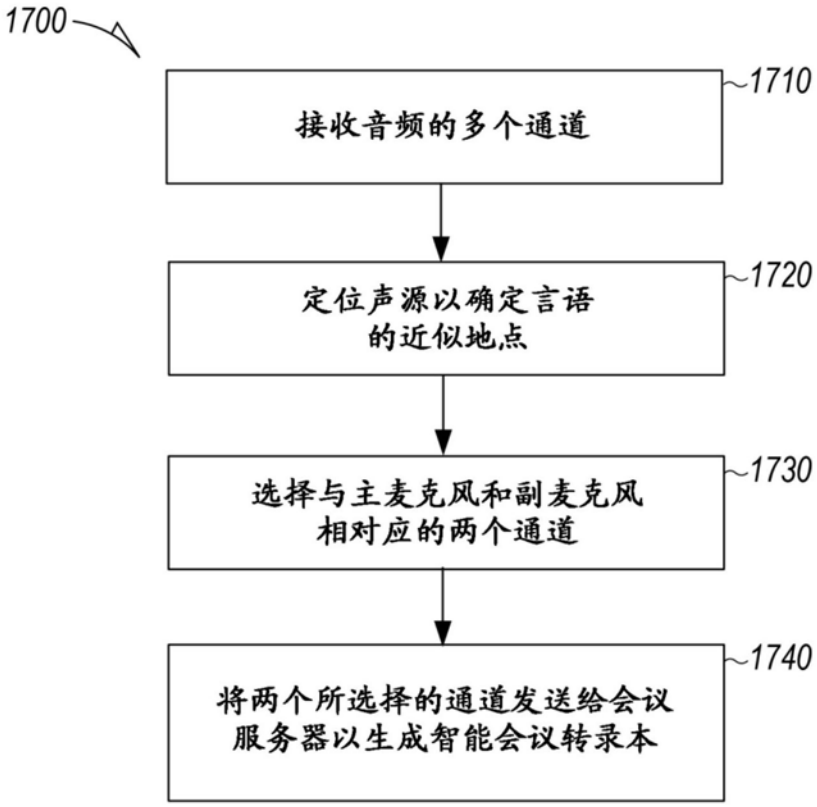

1700
接收音频的多个通道
1710
定位声源以确定言语的近似地点
1720
选择与主麦克风和副麦克风相对应的两个通道
1730
将两个所选择的通道发送给会议服务器以生成智能会议转录本
1740

图17

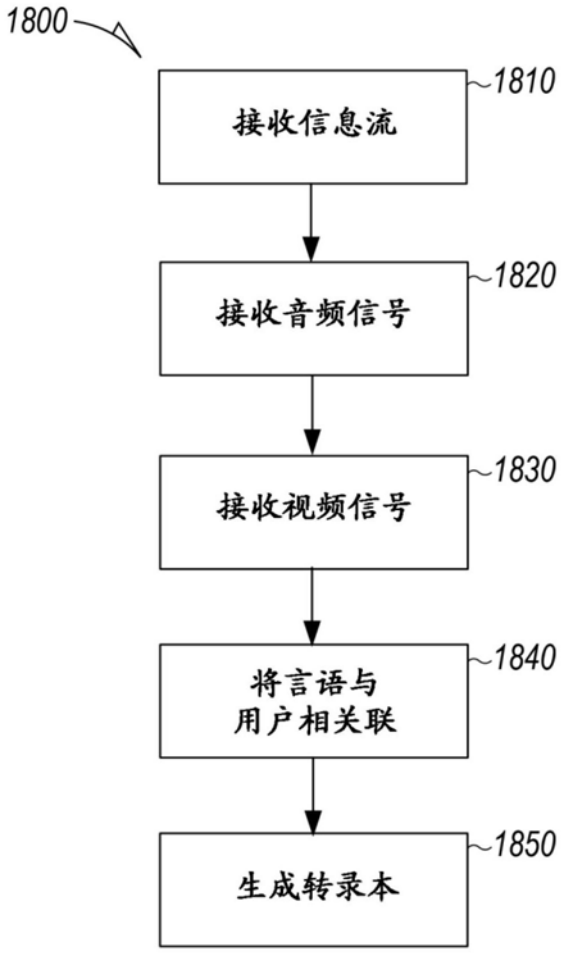

1800
接收信息流
1810
接收音频信号
1820
接收视频信号
1830
将言语与
用户相关联
1840
生成转录本
1850

图18

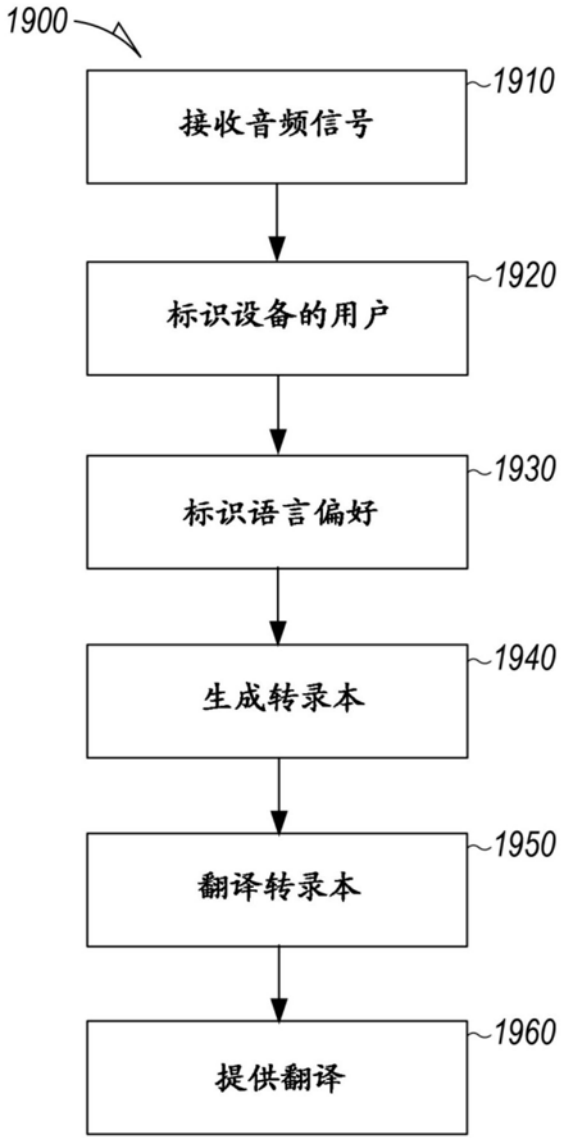

1900
接收音频信号
1910
标识设备的用户
1920
标识语言偏好
1930
生成转录本
1940
翻译转录本
1950
提供翻译
1960

图19

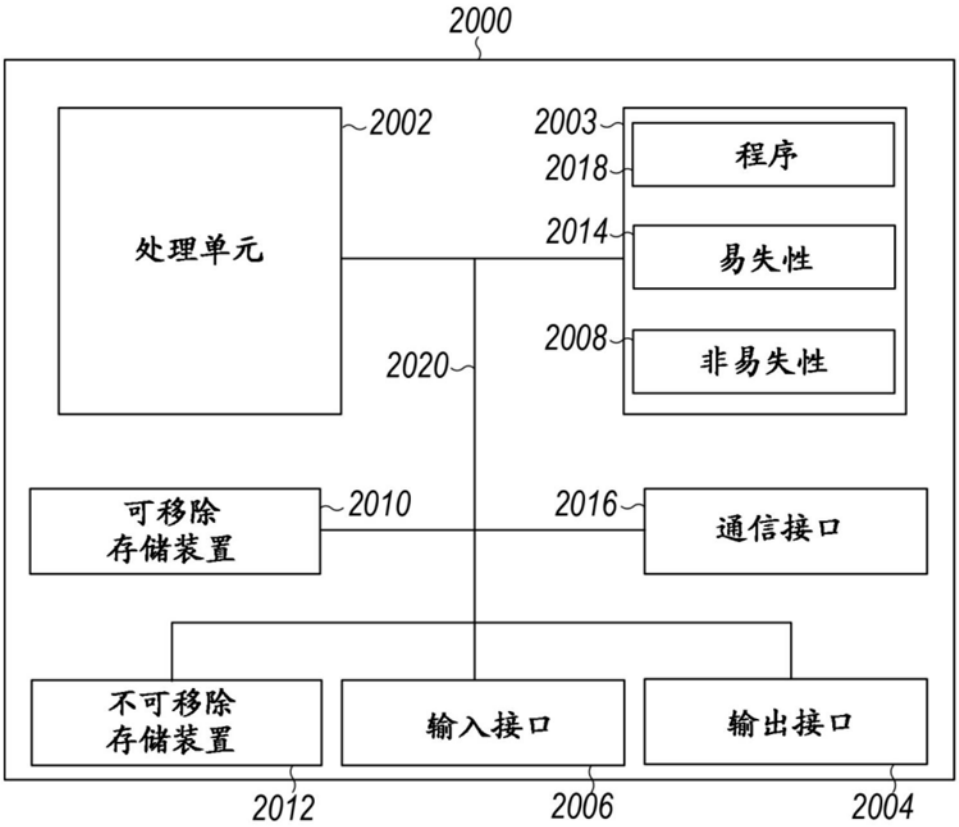

2000
处理单元
2002
2003
程序
2018
2014
易失性
2008
非易失性
2020
可移除
存储装置
2010
2016
通信接口
不可移除
存储装置
输入接口
输出接口
2012
2006
2004

图20