(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116249952 A
(43)申请公布日 2023.06.09

(21)申请号 202180067547.1
(22)申请日 2021.09.17
(30)优先权数据
63/089,507 2020.10.08 US
17/308,593 2021.05.05 US
(85)PCT国际申请进入国家阶段日
2023.03.31
(86)PCT国际申请的申请数据
PCT/US2021/071503 2021.09.17
(87)PCT国际申请的公布数据
WO2022/076963 EN 2022.04.14
(71)申请人 高通股份有限公司
地址 美国加利福尼亚
(72)发明人 T·沙赫巴齐米尔扎哈桑洛
R·G·阿尔维斯 E·维瑟
金莱轩
(74)专利代理机构 永新专利商标代理有限公司 72002
专利代理师 张殿慧
(51)Int.Cl.
G06F 1/3206(2006.01)

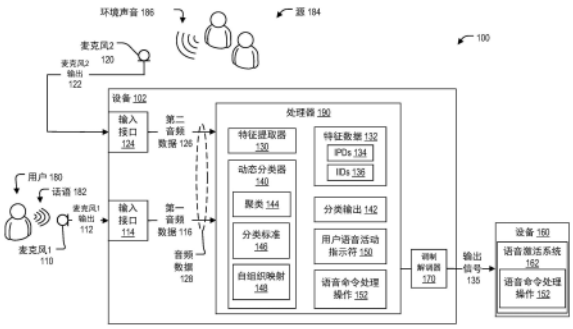
权利要求书3页 说明书24页 附图11页

(54)发明名称

使用动态分类器的用户语音活动检测

(57)摘要

一种设备包括被配置为存储指令的存储器以及被配置为执行指令的一个或多个处理器。所述一个或多个处理器被配置为执行所述指令以接收音频数据,所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据。所述一个或多个处理器还被配置为执行所述指令,以将音频数据提供给动态分类器。动态分类器被配置为生成与音频数据相对应的分类输出。所述一个或多个处理器还被配置为执行所述指令,以至少部分地基于所述分类输出来确定所述音频数据是否对应于用户语音活动。



CN 116249952 A

1.一种设备,包括:

存储器,其被配置为存储指令;以及

一个或多个处理器,其被配置为执行所述指令以进行以下操作:

接收音频数据,所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据;

将所述音频数据提供给动态分类器,所述动态分类器被配置为生成与所述音频数据相对应的分类输出;以及

至少部分地基于所述分类输出,来确定所述音频数据是否对应于用户语音活动。

2.根据权利要求1所述的设备,还包括所述第一麦克风和所述第二麦克风,其中,所述第一麦克风耦合到所述一个或多个处理器并且被配置为捕获用户的话语,并且其中,所述第二麦克风耦合到所述一个或多个处理器并且配置为捕获环境声音。

3.根据权利要求1所述的设备,其中,所述分类输出基于在所述第一音频数据和所述第二音频数据之间的增益差、在所述第一音频数据和所述第二音频数据之间的相位差、或其组合。

4.根据权利要求1所述的设备,其中,所述一个或多个处理器还被配置为:基于所述第一音频数据和所述第二音频数据来生成特征数据,其中,将所述音频数据作为所述特征数据提供给所述动态分类器,并且其中,所述分类输出基于所述特征数据。

5.根据权利要求4所述的设备,其中,所述特征数据包括:

在所述第一音频数据和所述第二音频数据之间的至少一个耳间相位差;以及

在所述第一音频数据和所述第二音频数据之间的至少一个耳间强度差。

6.根据权利要求4所述的设备,其中,所述一个或多个处理器被配置为:还基于所述特征数据的至少一个值的符号或幅度中的至少一者,来确定所述音频数据是否对应于所述用户语音活动。

7.根据权利要求4所述的设备,其中,所述一个或多个处理器还被配置为:在生成所述特征数据之前,将所述第一音频数据和所述第二音频数据变换到变换域,并且其中,所述特征数据包括针对多个频率的耳间相位差和针对多个频率的耳间强度差。

8.根据权利要求4所述的设备,其中,所述动态分类器被配置为:基于所述音频数据中表示的声音是否来源于相比于靠近所述第二麦克风,更靠近所述第一麦克风的源,自适应地对特征数据集进行聚类。

9.根据权利要求1所述的设备,其中,所述一个或多个处理器还被配置为:基于所述音频数据,更新所述动态分类器的聚类操作。

10.根据权利要求1所述的设备,其中,所述一个或多个处理器还被配置为:更新所述动态分类器的分类决策标准。

11.根据权利要求1所述的设备,其中,所述动态分类器包括自组织映射。

12.根据权利要求1所述的设备,其中,所述动态分类器还被配置为:接收音频数据集合的序列,并且至少部分地基于所述序列中的先前音频数据集合,自适应地对所述序列的集合进行聚类。

13.根据权利要求1所述的设备,其中,所述一个或多个处理器还被配置为:响应于确定所述音频数据对应于所述用户语音活动,来启动语音命令处理操作。

14.根据权利要求13所述的设备,其中,所述一个或多个处理器被配置为:生成唤醒信号或中断中的至少一者,以启动所述语音命令处理操作。

15.根据权利要求14所述的设备,其中,所述一个或多个处理器还包括:

包括所述动态分类器的常开功率域;以及

包括语音命令处理单元的第二功率域,并且其中,所述唤醒信号被配置为将所述第二功率域从低功率模式转换为激活所述语音命令处理单元。

16.根据权利要求1所述的设备,还包括耦合到所述一个或多个处理器的调制解调器,所述调制解调器被配置为:响应于基于所述动态分类器确定所述音频数据对应于所述用户语音活动,将所述音频数据发送给第二设备。

17.根据权利要求1所述的设备,其中,所述一个或多个处理器被集成在包括所述第一麦克风和所述第二麦克风的头戴式设备中,并且其中,所述头戴式设备被配置为:当用户佩戴时,将所述第一麦克风放置为比所述第二麦克风更靠近所述用户的嘴,以便与在所述第二麦克风处相比,在所述第一麦克风处以更大的强度和更小的延迟来捕捉所述用户的话语。

18.根据权利要求1所述的设备,其中,所述一个或多个处理器被集成在移动电话、平板计算机设备、可穿戴电子设备、相机设备、虚拟现实头戴式设备、或增强现实头戴式设备中的至少一者中。

19.根据权利要求1所述的设备,其中,所述一个或多个处理器被集成在运载工具中,所述运载工具还包括所述第一麦克风和所述第二麦克风,并且其中,所述第一麦克风被放置为捕获所述运载工具的操作者的话语。

20.一种语音活动检测的方法,包括:

在一个或多个处理器处接收音频数据,所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据;

在所述一个或多个处理器处,将所述音频数据提供给动态分类器,以生成与所述音频数据相对应的分类输出;以及

在所述一个或多个处理器处并且至少部分地基于所述分类输出,来确定所述音频数据是否对应于用户语音活动。

21.根据权利要求20所述的方法,其中,所述分类输出基于在所述第一音频数据和所述第二音频数据之间的增益差、在所述第一音频数据和所述第二音频数据之间的相位差、或其组合。

22.根据权利要求20所述的方法,其中,所述动态分类器包括自组织映射。

23.根据权利要求20所述的方法,其中,确定所述音频数据是否对应于所述用户语音活动还基于与所述音频数据相对应的特征数据的至少一个值的符号或幅度中的至少一者。

24.根据权利要求20所述的方法,还包括:响应于确定所述音频数据对应于所述用户语音活动,来启动语音命令处理操作。

25.根据权利要求24所述的方法,还包括:生成唤醒信号或中断中的至少一者,以启动所述语音命令处理操作。

26.根据权利要求20所述的方法,还包括:响应于基于所述动态分类器确定所述音频数据对应于所述用户语音活动,将所述音频数据发送给第二设备。

27.一种包括指令的非暂时性计算机可读介质，所述指令在由一个或多个处理器执行时，使得所述一个或多个处理器进行以下操作：

接收音频数据，所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据；

将所述音频数据提供给动态分类器，以生成与所述音频数据相对应的分类输出；以及

至少部分地基于所述分类输出，来确定所述音频数据是否对应于用户语音活动。

28.根据权利要求27所述的非暂时性计算机可读介质，其中，所述分类输出基于在所述第一音频数据和所述第二音频数据之间的增益差、在所述第一音频数据和所述第二音频数据之间的相位差、或其组合。

29.一种装置，包括：

用于接收音频数据的单元，所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据；

用于在动态分类器处，生成与所述音频数据相对应的分类输出的单元；以及

用于至少部分地基于所述分类输出，来确定所述音频数据是否对应于用户语音活动的单元。

30.根据权利要求29所述的装置，其中，所述分类输出基于在所述第一音频数据和所述第二音频数据之间的增益差、在所述第一音频数据和所述第二音频数据之间的相位差、或其组合。

使用动态分类器的用户语音活动检测

[0001] 相关申请的交叉引用

[0002] 本申请要求享受共同拥有的2020年10月8日提交的美国暂时专利申请No.63/089,507和2021年5月5日提交的美国非暂时专利申请No.17/308,593的优先权,故以引用方式将这些申请中的每一份的全部内容并入本文。

技术领域

[0003] 概括地说,本公开内容涉及自身语音活动检测。

背景技术

[0004] 技术的进步导致了更小和更强大的计算设备。例如,目前存在各种各样的便携式个人计算设备,其包括诸如移动和智能电话、平板设备和膝上型计算机之类的无线电话,它们体积小、重量轻且易于用户携带。这些设备可以通过无线网络传输语音和数据分组。此外,许多这样的设备并入了诸如数码相机、数码摄像机、数码录音机和音频文件播放器之类的其它功能。另外,此类设备可以处理包括软件应用程序的可执行指令,比如可以用于访问互联网的网络浏览器应用程序。因此,这些设备可以包括显著的计算能力。

[0005] 此类计算设备通常包含从一个或多个麦克风接收音频信号的功能。例如,音频信号可以表示麦克风捕获的用户语音、麦克风捕获的外部声音或其组合。为了说明起见,头戴式设备可以包括自身语音活动检测,以努力区分用户的语音(例如,佩戴头戴式设备的人所说的语音)和来源于其它源的语音。例如,当包括头戴式设备的系统支持关键词激活时,自身语音活动检测可以减少“虚警”,即基于来自附近人的语音(称为“非用户语音”)启动一个或多个组件或操作的激活。减少这样的虚警,提高了设备的功耗效率。然而,执行音频信号处理以区分用户语音和非用户语音也会消耗功率,并且提高设备区分用户语音与非用户语音的准确性的传统技术也往往会增加设备的功耗和处理资源需求。

发明内容

[0006] 根据本公开内容的一种实现,一种设备包括:被配置为存储指令的存储器、以及被配置为执行所述指令的一个或多个处理器。所述一个或多个处理器被配置为执行所述指令以接收音频数据,所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据。所述一个或多个处理器还被配置为执行所述指令以将所述音频数据提供给动态分类器。所述动态分类器被配置为生成与所述音频数据相对应的分类输出。所述一个或多个处理器进一步被配置为执行所述指令以至少部分地基于所述分类输出,来确定所述音频数据是否对应于用户语音活动。

[0007] 根据本公开内容的另一种实现,一种方法包括:在一个或多个处理器处接收音频数据,所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据。该方法还包括:在所述一个或多个处理器处,将所述音频数据提供给动态分类器,以生成与所述音频数据相对应的分类输出。该方法还包括:在所

述一个或多个处理器处并且至少部分地基于所述分类输出，来确定所述音频数据是否对应于用户语音活动。

[0008] 根据本公开内容的另一种实现，一种包括指令的非暂时性计算机可读介质，当所述指令由一个或多个处理器执行时，使得所述一个或多个处理器接收音频数据，所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据。当所述指令由一个或多个处理器执行时，进一步使得所述一个或多个处理器将所述音频数据提供给动态分类器，以生成与所述音频数据相对应的分类输出。当所述指令由一个或多个处理器执行时，还使得所述一个或多个处理器至少部分地基于所述分类输出，来确定所述音频数据是否对应于用户语音活动。

[0009] 根据本公开内容的另一种实现，一种装置包括：用于接收音频数据的单元，所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据。该装置还包括：用于在动态分类器处，生成与所述音频数据相对应的分类输出的单元。该装置还包括：用于至少部分地基于所述分类输出，来确定所述音频数据是否对应于用户语音活动的单元。

[0010] 在了解了包括下面部分的整个申请之后，本公开内容的其它方面、优点和特征将变得显而易见：附图说明、具体实施方式和权利要求书。

附图说明

[0011] 图1是根据本公开内容的一些例子，可操作用于执行自身语音活动检测的系统的特定说明性方面的框图。

[0012] 图2是根据本公开内容的一些例子，与自身语音活动检测相关联的操作的说明性方面的图。

[0013] 图3是根据本公开内容的一些例子，可操作用于执行自身语音活动检测的系统的说明性方面的框图。

[0014] 图4是根据本公开内容的一些例子，图1的系统的组件操作的说明性方面的图。

[0015] 图5根据本公开内容的一些例子，示出了包括用于检测用户语音活动的动态分类器的集成电路的例子。

[0016] 图6是根据本公开内容的一些例子，包括用于检测用户语音活动的动态分类器的移动设备的图。

[0017] 图7是根据本公开内容的一些例子，包括用于检测用户语音活动的动态分类器的头戴式设备的图。

[0018] 图8是根据本公开内容的一些例子，包括用于检测用户语音活动的动态分类器的可穿戴电子设备的图。

[0019] 图9是根据本公开内容的一些例子，包括用于检测用户语音活动的动态分类器的语音控制扬声器系统的图。

[0020] 图10是根据本公开内容的一些例子，包括用于检测用户语音活动的动态分类器的相机的图。

[0021] 图11是根据本公开内容的一些例子的包括用于检测用户语音活动的动态分类器的头戴式设备(例如，虚拟现实或增强现实头戴式设备)的图。

[0022] 图12是根据本公开内容的一些例子,包括用于检测用户语音活动的动态分类器的运载工具的第一例子的图。

[0023] 图13是根据本公开内容的一些例子,包括用于检测用户语音活动的动态分类器的运载工具的第二例子的图。

[0024] 图14A是根据本公开内容的一些例子,可以由图1的设备执行的自身语音活动检测的方法的特定实现的图。

[0025] 图14B是根据本公开内容的一些例子,可以由图1的设备执行的自身语音活动检测的方法的另一种特定实现的图。

[0026] 图15是根据本公开内容的一些例子,可操作用于执行自身语音活动检测的设备的特定说明性示例的框图。

具体实施方式

[0027] 自身语音活动检测(“SVAD”)可以减少“虚警”(即,非用户语音导致一个或多个组件或操作的激活),可以通过在检测到虚警时防止这些组件或操作激活来提高设备的功耗效率。然而,用于提高SVAD准确性的传统音频信号处理技术,在执行改进的准确度技术的同时也增加了设备的功耗和处理资源。由于SVAD处理通常是连续操作的,即使当设备处于低功率或休眠模式时,由于使用传统SVAD技术减少虚警而导致的功耗降低也可能被与SVAD处理本身相关联的功耗增加部分地或完全地抵消。

[0028] 公开了使用动态分类器进行自身语音活动检测的系统和方法。例如,在头戴式设备实现中,可以从定位成捕获用户语音的第一麦克风和定位成捕获外部声音的第二麦克风(例如,用于执行降噪和回声消除)接收音频信号。可以对音频信号进行处理,以提取包括耳间相位差(“IPD”)和耳间强度差(“IID”)的频域特征集。

[0029] 动态分类器对提取的频域特征集进行处理,并生成指示特征集分类的输出。动态分类器可以执行特征数据的自适应聚类,以及特征数据空间的两个最具鉴别性的类别之间的决策边界的调整,以在对应于用户语音活动的特征集和对应于其它音频活动的特征集之间进行区分。在一个说明性示例中,使用自组织映射来实现动态分类器。

[0030] 动态分类器能够使用提取的特征集进行区分,以主动响应和适应各种条件,例如:高度非平稳情况下的环境条件;麦克风不匹配;用户头戴式设备佩戴的变化;不同的用户头部相关传递函数(“HRTF”);非用户信号的到达方向(“DOA”)跟踪;麦克风在整个频谱上的本底噪声、偏置和灵敏度;或其组合。在一些实现中,动态分类器实现了自适应特征映射,该自适应特征映射能够响应于这样的变化,并且减少或最小化所使用的阈值参数的数量和客户的头戴式设备调谐的量。在一些实现中,与提供可比较准确度的传统SVAD系统相比,动态分类器能够在变化的条件下以高准确度并且以相对较低的功耗来有效地区分用户语音活动和其它音频活动。

[0031] 下面参考附图来描述本公开内容的特定方面。在本说明书中,通过共同的附图标记来表示共同的特征。如本文所使用的,各种术语仅用于描述特定的实现,并且并不旨在对实施方式进行限制。例如,单数形式的“一个(a)”、“某个(an)”和“该(the)”也旨在包括复数形式,除非上下文另外明确地说明。此外,本文所描述的一些特征在一些实施方式中是单数的,而在其它实施方式中是复数的。为了说明起见,图1描绘了包括一个或多个处理器(图1

中的“处理器”190)的设备102,这表明在一些实施方式中设备102包括单个处理器190,而在其它实施方式中设备102包括多个处理器190。为了便于本文引用起见,通常将这些特征引入为“一个或多个”特征,并且随后以单数形式来提及,除非正在描述与多个这些特征相关的方面。

[0032] 还应当理解的是,术语“包含”、“含有”和“包含的”可以与“包括”、“具有”或“包括的”互换地使用。此外,应当理解,术语“其中(wherein)”可以与“在其中(where)”互换地使用。如本文所使用的,“示例性”可以表示示例、实施方式和/或方面,并且不应被解释为限制性的或表示优选或优选的实施方式。如本文所使用的,用于修改元素(例如,结构、组件、操作等)的序数项(例如,“第一”、“第二”、“第三”等)本身并不表示元素相对于另一个元素的任何优先级或顺序,而是仅仅将元素与具有相同名称的另一个元素区分开来(只是为了使用序数项)。如本文所使用的,术语“集合”是指特定元素的一个或多个,而术语“多个”是指特定元素的多个(例如,两个或更多个)。

[0033] 如本文所使用的,“耦合”可以包括“通信地耦合”、“电耦合”或“物理耦合”,并且还可以(或替代地)包括其任何组合。两个设备(或组件)可以通过一个或多个其它设备、组件、电线、总线、网络(例如,有线网络、无线网络、或它们的组合)等来直接地或间接地耦合(例如,通信地耦合、电耦合、或物理耦合)。作为说明性而非限制性的举例,电耦合的两个设备(或组件)可以包括在同一设备或不同设备中,并且可以通过电子设备、一个或多个连接器或电感耦合进行连接。在一些实现中,例如在电通信中通信耦合的两个设备(或组件)可以经由一个或多个电线、总线、网络等等,直接或间接地发送和接收信号(例如,数字信号或模拟信号)。如本文所使用的,“直接耦合”可以包括在没有居间部件的情况下进行耦合(例如,通信地耦合、电耦合或物理耦合)的两个设备。

[0034] 在本公开内容中,诸如“确定”、“计算”、“估计”、“移位”、“调整”等等之类的术语可以用于描述如何执行一个或多个操作。应当注意,这些术语不应被解释为限制性的,并且可以利用其它技术来执行类似的操作。此外,如本文所指代的,可以互换地使用“生成”、“计算”、“估计”、“使用”、“选择”、“访问”和“确定”。例如,“生成”、“计算”、“估计”或“确定”参数(或信号),可以指代主动地生成、估计、计算或确定参数(或信号),也可以指代使用、选择或访问已生成的参数(或信号)(例如,由另一个组件或设备生成的参数(或信号))。

[0035] 参考图1,其公开了被配置为使用动态分类器来执行自身语音活动检测的系统的特定说明性方面,并且总体上用100表示。系统100包括耦合到第一麦克风110、第二麦克风120和第二设备160的设备102。设备102被配置为使用动态分类器140,对麦克风110、120捕获的声音执行自身语音活动检测。为了说明起见,在设备102对应于头戴式设备的实施方式中,第一麦克风110(例如,“主”麦克风)可以被配置为主要捕获设备102的用户的话语(例如,定位在设备102的佩戴者的嘴附近的麦克风),并且第二麦克风120(例如,“辅助”麦克风)可以被配置为主要捕获环境声音,例如定位在佩戴者的耳朵附近。在其它实现中,例如当设备102对应于可能在多个人附近的独立语音助手(例如,其包括带有麦克风的扬声器,如参考图11进一步描述的)时,设备102可以被配置为检测来自最靠近主麦克风的人的语音作为自身语音活动,即使与头戴式设备实现方式相比,人可能相对远离主麦克风。如本文所使用的,术语“自身语音活动检测”与“用户语音活动检测”可互换地使用,以指示区分设备102的用户的语音(例如,语音或话语)(例如,”用户语音活动“)与非源自设备的用户的声音

(例如,“其它音频活动”)。

[0036] 设备102包括第一输入接口114、第二输入接口124、一个或多个处理器190和调制解调器170。第一输入接口114耦合到处理器190,并且被配置为耦合到第一麦克风110。第一输入接口114被配置为从第一麦克风110接收第一麦克风输出112,并将第一麦克风输出112作为第一音频数据116提供给处理器190。

[0037] 第二输入接口124耦合到处理器190,并被配置为耦合到第二麦克风120。第二输入接口124被配置为接收来自第二麦克风120的第二麦克风输出122,并将第二麦克风输出122提供给处理器190作为第二音频数据126。

[0038] 处理器190耦合到调制解调器170,并包括特征提取器130和动态分类器140。处理器被配置为接收音频数据128,音频数据128包括与第一麦克风110的第一输出112相对应的第一音频数据116和与第二麦克风120的第二输出122相对应的第二音频数据126。处理器190被配置为在特征提取器130处,处理音频数据128以生成特征数据132。

[0039] 在一些实现中,处理器190被配置为在生成特征数据132之前,处理第一音频数据116和第二音频数据126。在一个例子中,处理器190被配置为对第一音频数据116和第二音频数据126执行回声消除、噪声抑制或两者。在一些实现中,处理器190被配置为在生成特征数据132之前,将第一音频数据116和第二音频数据126变换(例如,傅立叶变换)到变换域。

[0040] 处理器190被配置为基于第一音频数据116和第二音频数据126,生成特征数据132。根据一些方面,特征数据132包括在第一音频数据116和第二音频数据126之间的至少一个耳间相位差134,以及在第一音频数据116和第二音频数据126之间的至少一个耳间强度差136。在特定的例子中,特征数据132包括用于多个频率的耳间相位差(IPD)134和用于多个频率的耳间强度差(IID)136。

[0041] 处理器190被配置为:在动态分类器140处,处理特征数据132以生成特征数据132的分类输出142。在一些实现中,动态分类器140被配置为基于音频数据128中表示的声音是否来源于比靠近第二麦克风120更靠近第一麦克风110的源,自适应地对特征数据132的集合(例如,样本)进行聚类。例如,动态分类器140可以被配置为接收特征数据132的样本序列,并在包含IID和IPD频率值的特征空间中,自适应地对样本进行聚类。

[0042] 动态分类器140还可以被配置为调整特征空间的两个最具鉴别性的类别之间的决策边界,以区分与用户语音活动(例如,用户180的话语182)相对应的特征数据集合和与其它音频活动相对应的特征数据集合。为了说明起见,动态分类器140可以被配置为将输入的特征数据分类为两个类别中的一个(例如,类别0或类别1),其中这两个类别之一对应于用户语音活动,而这两个类别中的另一个对应于其它音频活动。分类输出142可以包括具有两个值之一的单个比特或标志:第一值(例如,“0”)用于指示特征数据132对应于两个类别中的一个;或者第二值(例如,“1”)用于指示特征数据132对应于两个类别中的另一个。

[0043] 在一些实现中,动态分类器140执行聚类和矢量量化。例如,聚类包括:减少(例如,最小化)聚类内的平方和,其定义为$min\sum_{i=1}^{n}\sum_{x_j\in C_i}p_i\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|^2$,其中$C_i$表示聚类i,$p_i$表示分配给聚类i的权重,$x_j$表示特征空间中的节点j,$\mu_i$表示聚类i的质心。聚类权重$p_i$可能是概率的,例如先验聚类分布;可能性的,例如分配给每个聚类的可能性的置信测度;或者根据任何其它因素决定,该因素将强制对不同的簇施加某种形式的非均匀偏置。矢量量化包括:通过下

式将输入矢量量化为量化权重矢量，来减少(例如，最小化)误差：$min\sum_{i=1}^{n}\sum_{x_j\in C_i}p_i\|x_j-w_i\|^2$，其中$w_i$表示量化权重向量i。

[0044] 在一些实现中，动态分类器140被配置为执行竞争学习，其中量化单元进行竞争以吸收特征数据132的新样本。然后在新样本的方向上调整获胜单元。例如，可以分离地或随机地初始化每个单元的权重向量。对于接收到的特征数据的每个新样本，作为非限制性示例，例如基于欧几里得距离或内积相似度，来确定哪个权重向量最接近新样本。然后，可以在新样本的方向上，移动最接近新样本的权重向量(“获胜者”或最佳匹配单元)。例如，在Hebbian学习中，获胜者加强了他们与输入的相关性(例如，通过与两个节点的输入的乘积成比例地调整两个节点之间的权重)。

[0045] 在一些实现中，动态分类器140包括突触前片中的局部簇，局部簇连接到突触后片中的局部簇，并且通过Hebbian学习来加强相邻神经元之间的互连，以加强相关刺激之间的连接。动态分类器140可以包括Kohonen自组织映射，其中，将输入连接到突触后片或映射中的每个神经元。学习导致对映射进行局部化，使得不同的吸收场响应于输入空间(例如，特征数据空间)的不同区域。

[0046] 在一种特定的实现中，动态分类器140包括自组织映射148。自组织映射140可以通过初始化权重向量来操作，然后对于每个输入t(例如，每个接收到的特征数据集132)，根据$v(t)=\arg\min_i\|x(t)-w_i(t)\|$来确定获胜单元(或者蜂窝或神经元)，以找到获胜者v(t)，作为与输入x(t)具有最小距离(例如，欧几里得距离)的单元。对获胜单元及其邻居的权重进行更新，例如根据$\Delta w_i(t)=\alpha(t)l(v,i,t)[x(t)-w_v(t)]$，其中$\Delta w_i(t)$表示单位i的变化，α(t)表示学习参数，l(v,i,t)表示获胜单元周围的邻域函数(例如，高斯径向基函数)。在一些实现中，可以使用内积或另一种度量作为相似性度量，而不是使用欧几里得距离。

[0047] 在一些实现中，动态分类器140包括Kohonen自组织映射的变型，以适应语音样本序列，例如参考图4进一步描述的。在一个例子中，动态分类器140可以例如根据时间Kohonen映射来实现时间序列处理，其中为每个单元定义具有时间常数建模衰减(“D”)的激活函数，并将其更新为$U_i(t,D)=D.U_i(t-1,D)-\frac{1}{2}\|x(t)-w_i(t)\|^2$，并且获胜单元是具有最大活动性的单元。再举一个例子，动态分类器140可以例如根据递归自组织映射来实现递归网络，其使用差分向量y而不是平方的范数：$y_i(t,\gamma)=(1-\gamma)y_i(t-1,\gamma)+\gamma(x(t)-w_i(t))$，其中γ表示值在0和1之间的遗忘因子，将获胜单元确定为具有最小差向量$v(t)=\arg\min_i\|y_i(t,\gamma)\|$的单元，并将权重更新为$\Delta w_i(t)=\alpha(t)l(v,i,t)[x(t)-y_v(t,\gamma)]$。

[0048] 在一些实现中，处理器190被配置为基于特征数据132来更新动态分类器140的聚类操作144，并且更新动态分类器的分类决策标准146。例如，如上所述，处理器190被配置为基于音频数据128的传入样本来调整用户语音活动和其它音频活动之间的聚类和决策边界，使得动态分类器140能够基于用户180的变化条件、环境、其它条件(例如，麦克风放置或调整)、或其任意组合，来调整操作。

[0049] 尽管将动态分类器140示出为包括自组织映射148，但是在其它实现中，动态分类器140可以结合不是自组织映射148或在自组织映射148之外的一种或多种其它技术来生成分类输出142。作为非限制性示例，动态分类器140可以包括具有无监督配置、无监督自动编

码器、Hopfield网络的在线变体、在线聚类或其组合的受限玻尔兹曼机。作为另一个非限制性示例，动态分类器140可以被配置为执行主分量分析(例如，将一组正交方向向量顺序地拟合到特征空间中的特征向量样本，其中将每个方向向量被选择为使投影到特征空间的方向向量上的特征向量样本的方差最大化)。作为另一个非限制性示例，动态分类器140可以被配置为执行独立分量分析(例如，确定特征空间中的特征向量样本的一组附加子分量，假设这些子分量是在统计上彼此独立的非高斯信号)。

[0050] 处理器190被配置为至少部分地基于分类输出142，来确定音频数据128是否对应于用户语音活动，并生成指示是否检测到用户语音活动的用户语音活动指示符150。例如，尽管分类输出142可以指示是否将特征数据132分类为两个类别中的一个(例如，类别“0”或类别“1”)，但是分类输出142可能不指示哪个类别对应于用户语音活动，哪个类别对应其它音频活动。例如，基于动态分类器140是如何初始化的以及已经用于更新动态分类器140的特征数据，在一些情况下，具有值“0”的分类输出142指示用户语音活动，而在其它情况下，具有值“0”的分类输出指示其它音频活动。处理器190可以进一步基于特征数据132的至少一个值的符号或幅度中的至少一个来确定两个类别中的哪一个指示用户语音活动，以及两个类别中的哪一个指示其它音频活动，如参考图2进一步描述的。

[0051] 为了说明起见，话语182从用户180的嘴到第一麦克风110和到第二麦克风120的声音传播导致相位差(由于话语182在第二麦克风110之前到达第一麦克风110)和信号强度差，可以在特征数据132中检测到该相位差和信号强度差，并且可以与来自其它音频源的声音的相位和信号强度差区分开。可以根据特征数据132中的IPD 134和IID 136来确定相位和信号强度差，并且使用相位和信号强度差将分类输出142映射到用户语音活动或其它音频活动。处理器190可以生成指示音频数据128是否对应于用户语音活动的用户语音活动指示符150。

[0052] 在一些实现中，处理器190被配置为响应于确定音频数据128对应于用户语音活动，来启动语音命令处理操作152。在说明性示例中，语音命令处理操作152包括语音激活操作，例如关键字或关键字短语检测、声纹认证、自然语言处理、一个或多个其它操作或其任意组合。再举一个例子，处理器190可以处理音频数据128以执行关键词检测的第一阶段，并且在经由语音命令处理操作152启动对音频数据128的进一步处理(例如，在包括更强大的语音活动识别和语音识别操作的第二检测阶段)之前，可以使用用户语音活动指示符150来确认所检测到的关键词是由设备102的用户180说出的，而不是由附近的人说出的。

[0053] 调制解调器170耦合到处理器190，并被配置为实现与第二设备160的通信(例如，经由无线传输)。在一些例子中，调制解调器170被配置为响应于基于动态分类器140确定音频数据128对应于用户语音活动，而将音频数据128发送到第二设备160。例如，在设备102对应于无线耦合到第二设备160的头戴式设备的实现方式中(例如，到移动电话或计算机的蓝牙连接)，设备102可以将音频数据128发送到第二设备160，以在第二设备160的语音激活系统162处执行语音命令处理操作152。在该例子中，设备102卸载使用第二设备160的更大的处理资源和功率资源来执行的更昂贵的处理(例如，语音命令处理操作152)。

[0054] 在一些实现中，设备102对应于一种或多种类型的设备，或者被包括在一种或多种类型的设备中。在说明性示例中，处理器190集成在包括第一麦克风110和第二麦克风120的头戴式设备中。头戴式设备被配置为：当用户180佩戴时，将第一麦克风110定位为比第二麦

克风120更靠近用户的嘴，以便与第二麦克风110相比，在第一麦克风110处以更大的强度和更小的延迟捕获用户180的话语182，如参考图7进一步描述的。在其它例子中，处理器190集成在以下中的至少一个中：如参考图6所述的移动电话或平板计算机设备、如参考图8所述的可穿戴电子设备、如参考图9所述的声控扬声器系统、如参考图10所述的相机设备、或者如参考图11所述的虚拟现实头戴式设备、混合现实头戴式设备或增强现实头戴式设备。在另一个说明性示例中，处理器190集成到还包括第一麦克风110和第二麦克风120的运载工具中，如参考图12和图13进一步描述的。

[0055] 在操作期间，第一麦克风110被配置为捕获用户180的话语182，第二麦克风120被配置为捕获环境声音186。在一个例子中，第一麦克风110和第二麦克风120捕获来自设备102的用户180的话语182。因为第一麦克风110更靠近用户180的嘴，所以与第二麦克风120相比，第一麦克风110以更高的信号强度和更小的延迟捕获用户180的语音。在另一个例子中，第一麦克风110和第二麦克风120可以捕获来自一个或多个声源184的环境声音186(例如，两个附近的人之间的对话)。基于声源184相对于第一麦克风110和第二麦克风120的位置和距离，在第一麦克风110处捕获环境声音186和在第二麦克风110处捕获环境声音186之间的信号强度差和相对延迟，将与来自用户180的话语182的信号强度差和相对延迟不同。

[0056] 在处理器190处，例如通过执行回声消除、噪声抑制、频域变换等等来处理第一音频数据116和第二音频数据126。在特征提取器130处，对所得到的音频数据进行处理以生成包括IPD 134和IID 136的特征数据132。将特征数据132输入到动态分类器140以生成分类输出142，处理器190将该分类输出142解释为用户语音活动或其它声音活动。处理器190生成用户语音活动指示符150，例如指示音频数据128对应于用户语音活动的“0”值，或者指示音频数据128对应于其它音频活动的“1”值(反之亦然)。

[0057] 可以使用用户语音活动指示符150来确定是否在设备102处启动语音命令处理操作152。替代地或另外地，可以使用用户语音活动指示符150来确定是否启动向第二设备160生成输出信号135(例如，音频数据128)，以在语音激活系统162处进行进一步处理。

[0058] 此外，结合生成分类输出142，基于特征数据132更新动态分类器140，例如通过将获胜单元以及其邻居的权重调整为更类似于特征数据132、更新聚类操作144、分类标准146或其组合。用此方式，动态分类器140自动地适应用户语音的变化、环境的变化、设备102或麦克风110、120的特性的变化、或者其组合。

[0059] 因此，与传统的自身语音活动检测技术相比，系统100通过使用动态分类器140以相对低的复杂度、低功耗和高准确度来区分用户语音活动和其它音频活动，从而提高了自身语音活动检测的性能。通过减少或消除由用户执行的校准并增强用户的体验，自动地适应用户和环境的变化提供了改进的益处。

[0060] 尽管在一些实现中，处理器190以特征提取器130生成的特征数据132(例如，频域数据)的形式，向动态分类器140提供音频数据128，但在其它实现中，省略了特征提取器。在一个例子中，处理器190将音频数据128作为音频样本的时间序列，提供给动态分类器140，并且动态分类器140处理音频数据128以生成分类输出142。在说明性实现中，动态分类器140被配置为从音频数据128确定频域数据(例如，生成特征数据132)，并使用提取的频域数据来生成分类输出142。

[0061] 尽管将第一麦克风110和第二麦克风120示出为耦合到设备102，但在其它实现中，

第一麦克风110或第二麦克风120中的一个或两个可以集成在设备102中。尽管示出了两个麦克风110、120，但在其它实现中，可以包括被配置为捕获用户语音的一个或多个其它麦克风、被配置为捕获环境声音的一个或多个麦克风、或者二者。尽管将系统100示出为包括第二设备160，但在其它实现中，可以省略第二设备，并且设备102可以执行被描述为在第二设备处执行的操作。

[0062] 图2是可以由图1的设备102(例如，处理器190)执行的与自身语音活动检测相关联的操作200的说明性方面的图。对输入202执行特征提取204，以生成特征数据206。在一个例子中，输入202对应于音频数据128，由特征提取器130执行特征提取204，并且特征数据206对应于特征数据132。

[0063] 动态分类器208对特征数据206进行操作，以生成分类输出210。在一个例子中，动态分类器208对应于动态分类器140，并且被配置为基于特征数据206执行无监督实时聚类，其具有针对分类输出210中的语音激活类的“自身”与“其它”标记的高动态决策边界。例如，动态分类器208可以将特征空间划分为两个类别，一个类别与用户语音活动相关，另一个类别则与其它声音活动相关。分类输出210可以包括哪个类别与特征数据206相关联的二进制指示符。在一个例子中，分类输出210对应于分类输出142。

[0064] 自身/其它关联操作212基于分类输出210和验证输入216来生成自身/其它指示符218。验证输入216可以提供将分类输出210的每个类别与用户语音活动(例如，“自身”)或其它声音活动(例如“其它”)相关联的信息。例如，可以基于至少一个先前验证标准214来生成验证输入216，所述先前验证标准例如比较相位差的符号230(例如，一个或多个IPD 134在一个或多个特定频率范围上的值，其指示哪个麦克风更接近由输入202表示的音频的源)，比较强度差的幅度232(例如，一个或多个IID 136在一个或多个特定频率范围上的值，其指示音频源到分开的麦克风的相对距离)或其组合。例如，自身/其它关联可以确定“0”的分类输出210值对应于在一个或多个相关频率范围中表现出负号230的特征数据206，或者在一个或者多个相关的频率范围中呈现出小于阈值量的幅度232，或者两者都对应，因此可以填充一个表，使得“0”对应于“其它”，“1”对应着“自身”。

[0065] 自身/其它关联操作212导致生成自身/其它指示符218(例如，具有第一值(例如“0”)以指示用户语音活动或具有第二值(例如，“1”)以指示其它声音活动的二进制指示符，反之亦然)。控制操作220中的唤醒/中断响应于自身/其它指示符218，以生成到语音命令处理224的信号222。例如，信号222可以具有第一值(例如，“0”)，以指示将在输入202、特征数据206或两者上执行语音命令处理224，以在输入202对应于用户语音活动时执行进一步的语音命令处理(例如，执行关键词检测、语音认证或两者)，或者可以具有第二值(例如，“1”)以指示当输入202对应于其它声音活动时，语音命令处理224不执行语音命令处理。

[0066] 动态分类(例如参考图1的动态分类器140和图2的动态分类器208所描述的)有助于提高SVAD的准确性，目的是仅在用户说话时才做出响应，并在其它干扰(例如，外部语音)到达时始终抑制响应，最大化自身关键字接受率(“SKAR”)和其它关键字拒绝率(“OKRR”)。通过使用动态分类，规避或以其它方式减少与传统SVAD处理相关联的各种挑战。例如，通过实现动态分类来规避或减少的传统SVAD处理挑战包括：噪声和回声条件(在严重条件下，可能导致错误的唤醒和闯入)、麦克风失配和灵敏度、语音激活引擎依赖性、不同的用户头部相关传递函数(HRTF)、不同的耳机硬件效果、用户行为驱动的遮挡和隔离水平的变化、用户

对其它语音活动的特征相似性、对语音激活的最终负面影响、以及用户语音开始的响应延迟。为了说明起见，传统的SVAD高度依赖于内部/外部麦克风校准和灵敏度、干扰语音的到达方向、头戴式设备适配和隔离的变化、以及特征的非平稳统计，这些可以通过动态分类的操作来适应。

[0067] 动态分类的使用，使得能够使用提取的特征数据206进行区分，以主动响应和适应各种条件，例如：高度非平稳情况下的环境条件；麦克风不匹配；用户头戴式设备配件的变化；不同的用户头部相关传递函数；非用户信号的到达方向(“DOA”)跟踪；以及麦克风在整个频谱上的本底噪声、偏置和灵敏度。动态分类能够实现自适应特征映射，该自适应特征映射能够响应这样的变化，并减少或最小化所使用的阈值参数的数量和客户的耳机调谐量。

[0068] 图3是根据本公开内容的一些例子，可操作以执行自身语音活动检测的系统的说明性方面的框图，其中处理器190包括常开功率域303和第二功率域305(例如，按需功率域)。在一些实现中，自身语音活动检测器320的第一级340和缓冲器360被配置为以常开模式进行操作，并且自身语音活动探测器320的第二级350被配置为按需模式进行操作。

[0069] 常开功率域303包括缓冲器360、特征提取器130和动态分类器140。缓冲器360被配置为存储第一音频数据116和第二音频数据126以便可访问以供自身语音活动检测器320的组件处理。

[0070] 第二功率域305包括自身语音活动检测器320的第二级250中的语音命令处理单元370，并且还包括激活电路330。在一些实现中，语音命令处理单元370被配置为执行图1的语音命令处理操作152或图2的语音命令处理224。

[0071] 自身语音活动检测器320的第一级240被配置为生成唤醒信号322或中断324中的至少一个，以启动语音命令处理单元370处的语音命令处理操作152(或语音命令处理224)。在一个例子中，唤醒信号322被配置为将第二功率域305从低功率模式332转换到活动模式334，以激活语音命令处理单元370。在一些实现中，唤醒信号322、中断324或两者对应于图2的信号222。

[0072] 例如，激活电路330可以包括或耦合到电源管理电路、时钟电路、头部开关或脚部开关电路、缓冲器控制电路或其任何组合。激活电路330可以被配置为例如通过选择性地施加或升高第二级350的电源电压、第二功率域305的电源电压或二者，来启动第二级350的上电。再举一个例子，激活电路330可以被配置为选择性地选通或不选通到第二级350的时钟信号，例如在不移除电源的情况下防止或启用电路操作。

[0073] 将自身语音活动检测器320的第二级350生成的检测器输出352提供给应用354。应用354可以被配置为基于检测到的用户语音，来执行一个或多个操作。为了说明起见，应用354可以对应于语音接口应用、集成辅助应用、车辆导航和娱乐应用或家庭自动化系统，作为说明性而非限制性的示例。

[0074] 通过基于在自身语音活动检测器320的第一级340处理音频数据的结果，选择性地激活第二级350，可以减少与自身语音活动检测、语音命令处理或两者相关联的总功耗。

[0075] 图4是根据本公开内容的一些例子，图1的系统的组件的操作的说明性方面的图。特征提取器130被配置为接收音频数据样本的序列410(例如，音频数据128的连续捕获的帧的序列)，其示出为第一帧(F1)412、第二帧(F2)414和包括第N帧(FN)416(其中N是大于2的整数)的一个或多个其它帧。特征提取器130被配置为输出特征数据集的序列420，该特征数

据集的序列包括第一集合422、第二集合424、以及包括第N集合426的一个或多个其它集合。

[0076] 动态分类器140被配置为接收特征数据集的序列420，并至少部分地基于序列420中的特征数据的先前集合(例如，第一集合422)，自适应地对序列420的集合(例如，第二集合424)进行聚类。作为说明性的非限制性的示例，可以将动态分类器140实现为时间Kohonen映射或递归自组织映射。

[0077] 在操作期间，特征提取器130对第一帧412进行处理以生成特征数据的第一集合422，并且动态分类器140对特征数据的第一集合422进行处理以生成分类输出序列430的第一分类输出(C1)432。特征提取器130对第二帧414进行处理以生成特征数据的第二集合424，并且动态分类器140对特征数据的第二集合424进行处理，以基于特征数据的第二集合424并且至少部分地基于特征数据第一集合422来生成第二分类输出(C2)434。这样的处理继续，包括特征提取器130处理第N帧416以生成特征数据的第N集合426，并且动态分类器140对特征数据的第N集合426进行处理以生成第N分类输出(CN)436。第N分类输出436基于特征数据的第N集合426，并且至少部分地基于序列420的特征数据的先前集合中的一者或多者。

[0078] 通过基于特征数据的一个或多个先前集合进行动态分类，可以提高动态分类器140对可能跨越多个音频数据帧的语音信号的分类精度。

[0079] 图5将设备102的实现500描绘为包括一个或多个处理器190的集成电路502。集成电路502还包括音频输入504(例如，一个或多个总线接口)，以使能够接收音频数据128以进行处理。集成电路502还包括信号输出512(例如，总线接口)，以使得能够发送输出信号(例如，用户语音活动指示符150)。集成电路502能够实现作为系统中的组件的自身语音活动检测，该系统包括麦克风，例如，如图6所示的移动电话或平板设备、图7所示的头戴式设备、图8所示的可穿戴电子设备、图9所示的声控扬声器系统、图10所示的相机、图11所示的虚拟现实头戴式设备、混合现实头戴式设备、增强现实头戴式设备、或者如图12或图13所示的运载工具。

[0080] 图6描述了设备102是移动设备602(例如，电话或平板设备)的实现方式600，作为说明性的非限制性的示例。移动设备602包括被放置为主要捕获用户的语音的第一麦克风110、被放置为主要捕获环境声音的多个第二麦克风120和显示屏604。处理器190的组件(包括特征提取器130和动态分类器140)集成在移动设备602中，并且使用虚线来表示移动设备602的用户通常不可见的内部组件。尽管将处理器190示出为包括特征提取器130，但是在其它实现中(例如，当动态分类器140被配置为在处理第一音频数据116和第二音频数据126期间提取特征数据时)，省略了特征提取器130，如参考图1所描述的。在特定的例子中，动态分类器140操作以检测用户语音活动，然后对用户语音活动进行处理以在移动设备602处执行一个或多个操作，例如启动图形用户界面或以其它方式在显示屏604处显示与用户语音相关联的其它信息(例如，通过集成的“智能助理”应用程序)。

[0081] 图7描述了设备102是头戴式设备702的实现方式700。头戴式设备702包括被放置为主要捕获用户的语音的第一麦克风110、以及被放置为主要捕获环境声音的第二麦克风120。处理器190的组件(其包括特征提取器130和动态分类器140)集成在头戴式设备702中。在特定的例子中，动态分类器140操作以检测用户语音活动(这可以使得头戴式设备702在头戴式设备702处执行一个或多个操作)、将与用户语音活动相对应的音频数据发送到第二

设备(没有示出)(例如,图1的第二设备160)进行进一步处理、或其组合。尽管将处理器190示出为包括特征提取器130,但在其它实现中,例如当动态分类器140被配置为在处理第一音频数据116和第二音频数据126期间提取特征数据时,省略了特征提取器130,如参考图1所描述的。

[0082] 图8描述了设备102是可穿戴电子设备802的实现方式800,其中将可穿戴电子设备802示出为“智能手表”。特征提取器130、动态分类器140、第一麦克风110和第二麦克风120集成到可穿戴电子设备802中。在特定的例子中,动态分类器140进行操作以检测用户语音活动,然后对用户语音活动进行处理以在可穿戴电子设备802处执行一个或多个操作,例如启动图形用户界面或以其它方式在可穿戴电子设备802的显示屏804处显示与用户语音相关联的其它信息。为了说明起见,可穿戴电子设备802可以包括显示屏,该显示屏被配置为基于可穿戴电子装置802检测到的用户语音来显示通知。在特定的例子中,可穿戴电子设备802包括触觉设备,该触觉设备响应于用户语音活动的检测而提供触觉通知(例如,振动)。例如,触觉通知可以使得用户看着可穿戴电子设备802,以看到显示的指示检测到用户说出的关键词的通知。因此,可穿戴电子设备802可以提醒听力受损的用户或佩戴头戴式设备的用户检测到用户的语音活动。尽管将可穿戴电子设备802示出为包括特征提取器130,但是在其它实现中,例如当动态分类器140被配置为在第一音频数据116和第二音频数据126的处理期间提取特征数据时,省略了特征提取器130,如参考图1所描述的。

[0083] 图9是设备102是无线扬声器和语音激活设备902的实现方式900。无线扬声器和语音激活设备902可以具有无线网络连接,并且被配置为执行辅助操作。包括特征提取器130和动态分类器140的处理器190、第一麦克风110、第二麦克风120或其组合被包括在无线扬声器和语音激活设备902中。尽管将处理器190示出为包括特征提取器130,但是在其它实现中,例如,当动态分类器140被配置为在处理第一音频数据116和第二音频数据126期间提取特征数据时,省略了特征提取器130,如参考图1所描述的。无线扬声器和语音激活设备902还包括扬声器904。在操作期间,响应于经由动态分类器140的操作接收到被识别为用户语音的口头命令,无线扬声器和语音激活设备902可以执行辅助操作,例如经由语音激活系统162(例如,集成辅助应用)的执行。辅助操作可以包括调节温度、播放音乐、开灯等等。例如,响应于接收到关键字或关键短语(例如,“hello assistant”)之后的命令,执行辅助操作。

[0084] 在一个说明性示例中,当无线扬声器和语音激活设备902靠近房间的墙壁(例如,靠近窗户)并且被布置为使得第一麦克风110被布置为与第二麦克风120相比更靠近房间的内部(例如,第二麦克风可以被布置为比第一麦克风110更靠近墙壁或窗户)时,可以将源自房间内部的语音识别为用户语音活动,而可以将源自房间外部的声音(例如,墙或窗户另一侧的人的语音)识别为其它音频活动。因为多个人可能在房间里,所以无线扬声器和语音激活设备902可以被配置为将来自多个人中的任何人的语音识别为用户语音活动(例如,无线扬声器和语音激活设备902可能有多个“用户”)。为了说明起见,动态分类器140可以被配置为将与源自房间内的语音相对应的特征数据识别为“自身语音”,即使说话的人可能离无线扬声器和语音激活设备902相对较远(例如,几米),并且比距离第二麦克风120更靠近第一麦克风110。在从房间中的多个人检测到语音的一些实现中,无线扬声器和语音激活设备902(例如,动态分类器140)可以被配置为将来自最靠近第一麦克风110的人的语音识别为用户语音活动(例如,与最近用户的自身语音)。

[0085] 图10描绘了设备102是与相机设备1002相对应的便携式电子设备的实现方式1000。特征提取器130和动态分类器140、第一麦克风110、第二麦克风120或其组合包括在相机设备1002中。在操作期间，响应于经由动态分类器140的操作而接收到被识别为用户语音的口头命令，相机设备1002可以执行响应于说出的用户命令的操作，例如，调整图像或视频捕获设置、图像或视频回放设置、或图像或视频捕捉指令，作为说明性示例。尽管将相机设备1002示出为包括特征提取器130，但在其它实现中，例如当动态分类器140被配置为在处理第一音频数据116和第二音频数据126期间提取特征数据时，省略特征提取器，如参考图1所描述的。

[0086] 图11描绘了设备102包括与扩展现实(“XR”)头戴式设备1102相对应的便携式电子设备(例如，虚拟现实(“VR”)、增强现实(“AR”)或混合现实(“MR”)头戴式设备)的实现方式1100。特征提取器130、动态分类器140、第一麦克风110、第二麦克风120或其组合集成到头戴式设备1102中。在特定的方面，头戴式设备1102包括被放置为主要捕获用户语音的第一麦克风110、以及被放置为主要捕获环境声音的第二麦克风120。可以基于从头戴式设备1102的第一麦克风110和第二麦克风120接收的音频信号，来执行用户语音活动检测。将视觉接口设备定位在用户的眼睛前方，以使得能够在佩戴头戴式设备1102时向用户显示增强现实或虚拟现实图像或场景。在特定的例子中，视觉接口设备被配置为显示指示在音频信号中检测到的用户语音的通知。尽管将头戴式设备1102示出为包括特征提取器130，但在其它实现中，例如当动态分类器140被配置为在第一音频数据116和第二音频数据126的处理期间提取特征数据时，省略了特征提取器130，如参考图1所描述的。

[0087] 图12描绘了设备102对应于或集成在运载工具1202内的实现方式1200，其中将运载工具1202示出为有人驾驶或无人驾驶的航空设备(例如，包裹递送无人机)。特征提取器130、动态分类器140、第一麦克风110、第二麦克风120或其组合集成到运载工具1202中。可以基于从运载工具1202的第一麦克风110和第二麦克风120接收的音频信号来执行用户语音活动检测，例如用于来自运载工具1202授权用户的递送指令。尽管将运载工具1202示出为包括特征提取器130，但是在其它实现中，例如当动态分类器140被配置为在处理第一音频数据116和第二音频数据126期间提取特征数据时，省略了特征提取器130，如参考图1所描述的。

[0088] 图13示出了设备102对应于运载工具1302、或集成在运载工具1302内的另一种实现方式1300，其中将运载工具1302示出为汽车。运载工具1302包括处理器190，处理器190包括特征提取器130和动态分类器140。尽管将运载工具1302示出为包括特征提取器130，但是在其它实现中，例如当动态分类器140被配置为在处理第一音频数据116和第二音频数据126期间提取特征数据时，省略了特征提取器130，如参考图1所描述的。运载工具1302还包括第一麦克风110和第二麦克风120。第一麦克风110定位成捕获运载工具1302的操作者的话语。可以基于从运载工具1302的第一麦克风110和第二麦克风120接收的音频信号来执行用户语音活动检测。在一些实现中，可以基于从内部麦克风(例如，第一麦克风110和第二麦克风120)接收的音频信号来执行用户语音活动检测，例如针对来自授权乘客的语音命令。例如，可以使用用户语音活动检测来检测来自运载工具1302的操作者的语音命令(例如，来自父母的将音量设置为5或设置自动驾驶运载工具的目的地的语音命令)，并且忽略另一乘客的语音(例如，来自孩子的将音量设置为10或其它乘客讨论另一位置)的语音命令。在一

些实现中，可以基于从外部麦克风(例如，第一麦克风110和第二麦克风120)(例如，运载工具的授权用户)接收的音频信号来执行用户语音活动检测。在特定实现中，响应于经由动态分类器140的操作接收到被识别为用户语音的口头命令，语音激活系统162基于在输出信号135中检测到的一个或多个关键词(例如，“解锁”、“启动发动机”、“播放音乐”、“显示天气预报”或另一语音命令)，例如通过经由显示器1320或一个或多个扬声器(例如，扬声器1310)提供反馈或信息，来启动运载工具1302的一个或多个操作。

[0089] 参考图14A，示出了用户语音活动检测的方法1400的具体实现。在特定的方面，由特征提取器130、动态分类器140、处理器190、设备102、图1的系统100或其组合中的至少一个，来执行方法1400的一个或多个操作。

[0090] 方法1400包括：在1402，在一个或多个处理器处接收音频数据，该音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据。例如，如参考图1所描述的，图1的特征提取器130接收音频数据128，该音频数据128包括与第一麦克风110的第一输出相对应的第一音频数据116和与第二麦克风126的第二输出相对应的第二音频数据126。

[0091] 方法1400包括：在1404，在一个或多个处理器处，基于第一音频数据和第二音频数据来生成特征数据。例如，如参考图1所描述的，图1的特征提取器130基于第一音频数据116和第二音频数据126来生成特征数据132。在另一个例子中，诸如图1的动态分类器140之类的动态分类器被配置为接收第一音频数据116和第二音频数据126，并且在处理第一音频数据和第二音频数据126期间提取特征数据132。

[0092] 方法1400包括：在1406，在一个或多个处理器的动态分类器处，生成特征数据的分类输出。例如，如参考图1所描述的，图1的动态分类器140生成特征数据132的分类输出142。

[0093] 方法1400包括：在1408，在一个或多个处理器处并且至少部分地基于分类输出，来确定音频数据是否对应于用户语音活动。例如，如参考图1所描述的，图1的处理器190至少部分地基于分类输出142，来确定音频数据128是否对应于用户语音活动。

[0094] 与传统的自身语音活动检测技术相比，方法1400通过使用动态分类器140以相对较低的复杂度、低功耗和高准确度，来区分用户语音活动和其它音频活动，从而提高了自身语音活动的检测性能。通过减少或消除由用户执行的校准并增强用户的体验，自动地适应用户和环境的变化提供了改进的益处。

[0095] 参考图14B，该图示出了用户语音活动检测的方法1450的特定实现。在特定的方面，由动态分类器140、处理器190、设备102、图1的系统100或其组合中的至少一个，执行方法1450的一个或多个操作。

[0096] 方法1450包括：在1452，在一个或多个处理器处，接收音频数据，该音频数据包括包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据。在一个例子中，图1的特征提取器130接收音频数据128，其中音频数据128包括与第二麦克风126的第二输出相对应的第一音频数据116和第二音频数据126，如参考图1所描述的。

[0097] 方法1450包括：在1454，在一个或多个处理器处，将音频数据提供给动态分类器，以生成与音频数据相对应的分类输出。在一个例子中，图1的特征提取器130基于第一音频数据116和第二音频数据126来生成特征数据132，并且由动态分类器140处理特征数据132

以生成分类输出142，如图1中所描述的并且根据图14A的方法1400。在另一个例子中，处理器190将第一音频数据116和第二音频数据126提供给动态分类器140，并且动态分类器140处理第一音频数据116和第二音频数据126以生成分类输出142。在一种说明性实现中，动态分类器140处理第一音频数据116和第二音频数据126以提取特征数据132，并基于特征数据132来确定分类输出142。

[0098] 方法1450包括：在1456，在一个或多个处理器处并且至少部分地基于分类输出，来确定音频数据是否对应于用户语音活动。例如，如参考图1所描述的，图1的处理器190至少部分地基于分类输出142来确定音频数据128是否对应于用户语音活动。

[0099] 与传统的自身语音活动检测技术相比，方法1450通过使用动态分类器140以相对较低的复杂度、低功耗和高准确度，来区分用户语音活动和其它音频活动，从而提高了自身语音活动的检测性能。通过减少或消除由用户执行的校准并增强用户的体验，自动地适应用户和环境的变化提供了改进的益处。

[0100] 可以通过现场可编程门阵列(FPGA)设备、专用集成电路(ASIC)、诸如中央处理单元(CPU)、DSP的处理单元、控制器、另一个硬件设备、固件设备或其任意组合，来实现图14A的方法1400、图14B的方法1450或其组合。举一个例子，可以由执行指令的处理器(例如，参照图15所描述的处理器)，来执行图14A的方法1400、图14B的方法1450或其组合。

[0101] 参考图15，描绘了设备的特定说明性实现的框图，并且总体上用1500表示。在各种实现方式中，设备1500可以具有比图15中所示的更多或更少的组件。在一种示例性实现中，设备1500可以对应于设备102。在一种示例性实现中，设备1500可以执行参照图1-14B描述的一个或多个操作。

[0102] 在特定的实现中，设备1500包括处理器1506(例如，中央处理单元(CPU))。设备1500可以包括一个或多个其它处理器1510(例如，一个或多个DSP)。在特定的方面，图1的处理器190对应于处理器1506、处理器1510或者其组合。处理器1510可以包括语音和音乐编码器-解码器(CODEC)1508，其包括语音译码器(“声码器”)编码器1536、声码器解码器1538、特征提取器130、动态分类器140或者其组合。

[0103] 设备1500可以包括存储器1586和CODEC 1534。存储器1586可以包括可以由一个或多个附加处理器1510(或处理器1506)执行的指令1556，以实现参考特征提取器130、动态分类器140或两者描述的功能。设备1500可以包括经由收发器1550耦合到天线1552的调制解调器170。

[0104] 设备1500可以包括连接到显示控制器1526的显示器1528。扬声器1592、第一麦克风110和第二麦克风120可以耦合到CODEC 1534。CODEC 1534可以包括数模转换器(DAC)1502、模数转换器(ADC)1504或两者。在特定的实现中，CODEC 1534可以从第一麦克风110和第二麦克风120接收模拟信号，使用模数转换器1504将模拟信号转换为数字信号，并将数字信号提供给语音和音乐编解码器1508。语音和音乐编解码器1508可以处理数字信号，并且特征提取器130和动态分类器140可以进一步处理数字信号。在特定的实现中，语音和音乐编解码器1508可以向CODEC 1534提供数字信号。CODEC 1534可以使用数模转换器1502，将数字信号转换为模拟信号，并且可以将模拟信号提供给扬声器1592。

[0105] 在一种特定的实现中，设备1500可以包括在封装系统或片上系统器件1522中。在一种特定的实现中，存储器1586、处理器1506、处理器1510、显示控制器1526、CODEC 1534和

调制解调器170包括在封装系统或片上系统器件1522中。在一种特定的实现中，输入设备1530和电源1544耦合到片上系统器件1522。此外，在一种特定的实现中，如图15中所示，显示器1528、输入设备1530、扬声器1592、第一麦克风110、第二麦克风120、天线1552和电源1544在片上系统器件1522外部。在一种特定实现中，显示器1528、输入设备1530、扬声器1592、第一麦克风110、第二麦克风120、天线1552和电源1544中的每一个可以耦合到片上系统器件1522的组件，例如接口(例如，第一输入接口114或第二输入接口124)或控制器。

[0106] 设备1500可以包括智能扬声器、扬声器条、移动通信设备、智能电话、蜂窝电话、膝上型计算机、计算机、平板设备、个人数字助理、显示设备、电视、游戏控制台、音乐播放器、收音机、数字视频播放器、数字视频光盘(DVD)播放器、调谐器、相机、导航设备、运载工具、头戴式设备，增强现实耳机、虚拟现实耳机、飞行器、家庭自动化系统、声控设备、无线扬声器和声控设备、便携式电子设备、汽车、运载工具、计算设备、通信设备、物联网(IoT)设备、虚拟现实(VR)设备、基站、移动设备、或者其任何组合。

[0107] 结合上面所描述的实现，一种装置包括用于接收音频数据的单元，其包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据。例如，用于接收的单元可以对应于第一输入接口114、第二输入接口124、特征提取器130、动态分类器140、处理器190、一个或多个处理器1510，一个或多个其它电路或组件，它们被配置为接收包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据、或其任意组合的音频数据。

[0108] 该装置还包括：用于基于第一音频数据和第二音频数据来生成特征数据的单元。例如，用于生成特征数据的单元可以对应于特征提取器130、动态分类器140、处理器190、一个或多个处理器1510、被配置为生成特征数据或其任何组合的一个或多个其它电路或组件。

[0109] 该装置还包括：用于在动态分类器处生成特征数据的分类输出的单元。例如，用于生成分类输出的单元可以对应于动态分类器140、处理器190、一个或多个处理器1510、被配置为在动态分类器处生成分类输出的一个或多个其它电路或组件、或者其任何组合。

[0110] 该装置还包括：用于至少部分基于分类输出，来确定音频数据是否对应于用户语音活动的单元。例如，用于确定的单元可以对应于动态分类器140、处理器190、一个或多个处理器1510、被配置为至少部分地基于分类输出来确定音频数据是否对应于用户语音活动的一个或多个其它电路或组件、或者其任何组合。

[0111] 结合所描述的实现，一种装置包括：用于接收音频数据的单元，其包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据。例如，用于接收的单元可以对应于第一输入接口114、第二输入接口124、特征提取器130、动态分类器140、处理器190、一个或多个处理器1510、被配置为接收包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据的音频数据的一个或多个其它电路或组件、或者其任意组合。

[0112] 该装置还包括：用于在动态分类器处，生成与音频数据相对应的分类输出的单元。例如，用于生成分类输出的单元可以对应于特征提取器130、动态分类器140、处理器190、一个或多个处理器1510、被配置为在动态分类器处生成分类输出的一个或多个其它电路或组件、或者它们的任意组合。

[0113] 该装置还包括:用于至少部分地基于分类输出,确定音频数据是否对应于用户语音活动的单元。例如,该用于确定的单元可以对应于动态分类器140、处理器190、一个或多个处理器1510、被配置为至少部分地基于分类输出来确定音频数据是否对应于用户语音活动的一个或多个其它电路或组件、或者它们的任意组合。

[0114] 在一些实现中,非暂时性计算机可读介质(例如,诸如存储器1586之类的计算机可读存储设备)包括指令(例如,指令1556),当该指令由一个或多个处理器(例如,一个或多个处理器1510或处理器1506)执行时,使所述一个或多个处理器接收音频数据(例如,音频数据128),该音频数据包括与第一麦克风(例如,第一麦克风110)的第一输出相对应的第一音频数据(如,第一音频数据116)和与第二麦克风(例如,第二麦克风120)相对应的第二输出的第二音频数据(例,第二音频数据126)。当这些指令由所述一个或多个处理器执行时,还使得所述一个或多个处理器将音频数据提供给动态分类器(例如,动态分类器140),以生成与音频信号相对应的分类输出(例如,分类输出142)。在一个例子中,当这些指令由一个或多个处理器执行时,使所述一个或多个处理器基于第一音频数据和第二音频数据来生成特征数据(例如,特征数据132),并在动态分类器处对特征数据进行处理。当这些指令由一个或多个处理器执行时,还使得所述一个或多个处理器至少部分地基于分类输出,来确定音频数据是否对应于用户语音活动。

[0115] 本公开内容包括以下示例。

[0116] 示例1、一种设备,包括一个或多个处理器,其被配置为接收音频数据,所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据;基于所述第一音频数据和所述第二音频数据来生成特征数据;在动态分类器处,对所述特征数据进行处理以生成所述特征数据的分类输出;并至少部分地基于所述分类输出,来确定所述音频数据是否对应于用户语音活动。

[0117] 示例2、根据示例1所述的设备,还包括所述第一麦克风和所述第二麦克风,其中所述第一麦克风耦合到所述一个或多个处理器并且被配置为捕获用户的话语,并且其中,所述第二麦克风耦合到所述一个或多个处理器并且配置为捕获环境声音。

[0118] 示例3、根据示例1所述的设备,其中,所述特征数据包括:在所述第一音频数据和所述第二音频数据之间的至少一个耳间相位差;以及在所述第一音频数据和所述第二音频数据之间的至少一个耳间强度差。

[0119] 示例4、根据示例3所述的设备,其中,所述一个或多个处理器进一步被配置为:在生成所述特征数据之前,将所述第一音频数据和所述第二音频数据变换到变换域,并且其中所述特征数据包括多个频率的耳间相位差和多个频率的耳间强度差。

[0120] 示例5、根据示例1所述的设备,其中,所述动态分类器被配置为基于所述音频数据中表示的声音是否来源于相比于靠近所述第二麦克风,更靠近所述第一麦克风的源,自适应地对特征数据集进行聚类。

[0121] 示例6、根据示例1所述的设备,其中,所述一个或多个处理器进一步被配置为基于所述音频数据,更新所述动态分类器的聚类操作。

[0122] 示例7、根据示例1所述的设备,其中,所述一个或多个处理器进一步被配置为更新所述动态分类器的分类决策标准。

[0123] 示例8、根据示例1所述的设备,其中,所述动态分类器包括自组织映射。

[0124] 示例9、根据示例1所述的设备，其中，所述动态分类器进一步被配置为接收特征数据集合的序列，并且至少部分地基于所述序列中的先前特征数据集合，自适应地对所述序列的集合进行聚类。

[0125] 示例10、根据示例1所述的设备，其中，所述一个或多个处理器进一步被配置为：进一步基于所述特征数据的至少一个值的符号或幅度中的至少一个，来确定所述音频数据是否对应于所述用户语音活动。

[0126] 示例11、根据示例1所述的设备，其中，所述一个或多个处理器被配置为响应于确定所述音频数据对应于所述用户语音活动，而启动语音命令处理操作。

[0127] 示例12、根据示例11所述的设备，其中，所述一个或多个处理器被配置为生成唤醒信号或中断中的至少一个，以启动所述语音命令处理操作。

[0128] 示例13、根据示例12所述的设备，其中，所述一个或多个处理器还包括：包括所述动态分类器的常开功率域；以及包括语音命令处理单元的第二功率域，并且其中，所述唤醒信号被配置为将所述第二功率域从低功率模式转换以激活所述语音命令处理单元。

[0129] 示例14、根据示例1所述的设备，还包括耦合到所述一个或多个处理器的调制解调器，所述调制解调器被配置为：响应于基于所述动态分类器而确定所述音频数据对应于所述用户语音活动，将所述音频数据发送到第二设备。

[0130] 示例15、根据示例1所述的设备，其中，所述一个或多个处理器集成在包括所述第一麦克风和所述第二麦克风的头戴式设备中，并且其中所述头戴式设备被配置为当用户佩戴时，将所述第一麦克风定位为比所述第二麦克风更靠近用户的嘴，以便与在所述第二麦克风处相比，在所述第一麦克风处以更大的强度和更小的延迟来捕捉所述用户的话语。

[0131] 示例16、根据示例1所述的设备，其中，所述一个或多个处理器集成在移动电话、平板计算机设备、可穿戴电子设备、相机设备、虚拟现实头戴式设备、或增强现实头戴式设备中的至少一个中。

[0132] 示例17、根据示例1所述的设备，其中，所述一个或多个处理器集成在运载工具中，所述运载工具还包括所述第一麦克风和所述第二麦克风，并且其中所述第一麦克风定位成捕获所述运载工具的操作者的话语。

[0133] 示例18、一种语音活动检测的方法，包括：在一个或多个处理器处接收音频数据，所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据；在所述一个或多个处理器处，基于所述第一音频数据和所述第二音频数据来生成特征数据；在所述一个或多个处理器的动态分类器处，生成所述特征数据的分类输出；在所述一个或多个处理器处并至少部分地基于所述分类输出，来确定所述音频数据是否对应于用户语音活动。

[0134] 示例19、根据示例18所述的方法，其中所述第一麦克风被配置为捕获用户的话语，并且其中，所述第二麦克风被配置为捕获环境声音。

[0135] 示例20、根据示例18所述的方法，其中，所述特征数据包括：在所述第一音频数据和所述第二音频数据之间的至少一个耳间相位差；以及在所述第一音频数据和所述第二音频数据之间的至少一个耳间强度差。

[0136] 示例21、根据示例20所述的方法，还包括：在生成所述特征数据之前，将所述第一音频数据和所述第二音频数据变换到变换域，其中所述特征数据包括多个频率的耳间相位

差和多个频率的耳间强度差。

[0137] 示例22、根据示例18所述的方法，还包括：由所述动态分类器基于所述音频数据中表示的声音是否来源于相比于靠近所述第二麦克风，更靠近所述第一麦克风的源，自适应地对特征数据集进行聚类。

[0138] 示例23、根据示例18所述的方法，还包括：基于所述特征数据，更新所述动态分类器的聚类操作。

[0139] 示例24、根据示例18所述的方法，还包括：更新所述动态分类器的分类决策标准。

[0140] 示例25、根据示例18所述的方法，其中，所述动态分类器包括自组织映射。

[0141] 示例26、根据示例18所述的方法，还包括：在所述动态分类器处，接收特征数据集合的序列，并且至少部分地基于所述序列中的先前特征数据集合，自适应地对所述序列的集合进行聚类。

[0142] 示例27、根据示例18所述的方法，其中，进一步基于所述特征数据的至少一个值的符号或幅度中的至少一个，来确定所述音频数据是否对应于所述用户语音活动。

[0143] 示例28、根据示例18所述的方法，还包括：响应于确定所述音频数据对应于所述用户语音活动，而启动语音命令处理操作。

[0144] 示例29、根据示例28所述的方法，还包括：生成唤醒信号或中断中的至少一个，以启动所述语音命令处理操作。

[0145] 示例30、根据示例29所述的方法，其中，所述唤醒信号被配置为将功率域从低功率模式转换以启动所述语音命令处理操作。

[0146] 示例31、根据示例18所述的方法，还包括：响应于基于所述动态分类器而确定所述音频数据对应于所述用户语音活动，将所述音频数据发送到第二设备。

[0147] 示例32、根据示例18所述的方法，其中，所述一个或多个处理器集成在包括所述第一麦克风和所述第二麦克风的头戴式设备中，并且其中所述头戴式设备被配置为当用户佩戴时，将所述第一麦克风定位为比所述第二麦克风更靠近用户的嘴，以便与在所述第二麦克风处相比，以更大的强度和更小的延迟来捕捉所述用户在所述第一麦克风处的话语。

[0148] 示例33、根据示例18所述的方法，其中，所述一个或多个处理器集成在移动电话、平板计算机设备、可穿戴电子设备、相机设备、虚拟现实头戴式设备、或增强现实头戴式设备中的至少一个中。

[0149] 示例34、一种包括指令的非暂时性计算机可读介质，当所述指令由一个或多个处理器执行时，使得所述一个或多个处理器用于：接收音频数据，所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据；基于所述第一音频数据和所述第二音频数据来生成特征数据；在动态分类器处，对所述特征数据进行处理以生成分类输出；并至少部分地基于所述分类输出，来确定所述音频数据是否对应于用户语音活动。

[0150] 示例35、根据示例34所述的非暂时性计算机可读介质，其中所述第一麦克风被配置为捕获用户的话语，并且其中，所述第二麦克风被配置为捕获环境声音。

[0151] 示例36、根据示例34所述的非暂时性计算机可读介质，其中，所述特征数据包括：在所述第一音频数据和所述第二音频数据之间的至少一个耳间相位差；以及在所述第一音频数据和所述第二音频数据之间的至少一个耳间强度差。

[0152] 示例37、根据示例34所述的非暂时性计算机可读介质，其中，当所述指令由所述一个或多个处理器执行时，进一步使所述一个或多个处理器在生成所述特征数据之前，将所述第一音频数据和所述第二音频数据变换到变换域，其中所述特征数据包括多个频率的耳间相位差和多个频率的耳间强度差。

[0153] 示例38、根据示例34所述的非暂时性计算机可读介质，其中，当所述指令由所述一个或多个处理器执行时，进一步使所述一个或多个处理器用于：由所述动态分类器基于所述音频数据中表示的声音是否来源于相比于靠近所述第二麦克风，更靠近所述第一麦克风的源，自适应地对特征数据集进行聚类。

[0154] 示例39、根据示例34所述的非暂时性计算机可读介质，其中，当所述指令由所述一个或多个处理器执行时，进一步使所述一个或多个处理器基于所述特征数据，更新所述动态分类器的聚类操作。

[0155] 示例40、根据示例34所述的非暂时性计算机可读介质，其中，当所述指令由所述一个或多个处理器执行时，进一步使所述一个或多个处理器更新所述动态分类器的分类决策标准。

[0156] 示例41、根据示例34所述的非暂时性计算机可读介质，其中，所述动态分类器包括自组织映射。

[0157] 示例42、根据示例34所述的非暂时性计算机可读介质，其中，当所述指令由所述一个或多个处理器执行时，进一步使所述一个或多个处理器用于：在所述动态分类器处，接收特征数据集合的序列，并且至少部分地基于所述序列中的先前特征数据集合，自适应地对所述序列的集合进行聚类。

[0158] 示例43、根据示例34所述的非暂时性计算机可读介质，其中，进一步基于所述特征数据的至少一个值的符号或幅度中的至少一个，来确定所述音频数据是否对应于所述用户语音活动。

[0159] 示例44、根据示例34所述的非暂时性计算机可读介质，其中，当所述指令由所述一个或多个处理器执行时，进一步使所述一个或多个处理器响应于确定所述音频数据对应于所述用户语音活动，而启动语音命令处理操作。

[0160] 示例45、根据示例34所述的非暂时性计算机可读介质，其中，当所述指令由所述一个或多个处理器执行时，进一步使所述一个或多个处理器生成唤醒信号或中断中的至少一个，以启动所述语音命令处理操作。

[0161] 示例46、根据示例45所述的非暂时性计算机可读介质，其中，所述唤醒信号被配置为将功率域从低功率模式转换以启动所述语音命令处理操作。

[0162] 示例47、根据示例34所述的非暂时性计算机可读介质，其中，当所述指令由所述一个或多个处理器执行时，进一步使所述一个或多个处理器用于：响应于基于所述动态分类器而确定所述音频数据对应于所述用户语音活动，将所述音频数据发送到第二设备。

[0163] 示例48、根据示例34所述的非暂时性计算机可读介质，其中，所述一个或多个处理器集成在包括所述第一麦克风和所述第二麦克风的头戴式设备中，并且其中所述头戴式设备被配置为当用户佩戴时，将所述第一麦克风定位为比所述第二麦克风更靠近用户的嘴，以便与在所述第二麦克风处相比，以更大的强度和更小的延迟来捕捉所述用户在所述第一麦克风处的话语。

[0164] 示例49、根据示例34所述的非暂时性计算机可读介质，其中，所述一个或多个处理器集成在移动电话、平板计算机设备、可穿戴电子设备、相机设备、虚拟现实头戴式设备、增强现实头戴式设备或运载工具中的至少一个中。

[0165] 示例50、一种装置包括：用于接收音频数据的单元，所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据；用于基于所述第一音频数据和所述第二音频数据来生成特征数据的单元；用于在动态分类器处，生成所述特征数据的分类输出的单元；用于至少部分地基于所述分类输出，来确定所述音频数据是否对应于用户语音活动的单元。

[0166] 示例51、根据示例50所述的装置，其中，所述特征数据包括：在所述第一音频数据和所述第二音频数据之间的至少一个耳间相位差；以及在所述第一音频数据和所述第二音频数据之间的至少一个耳间强度差。

[0167] 示例52、根据示例50所述的装置，还包括：用于在生成所述特征数据之前，将所述第一音频数据和所述第二音频数据变换到变换域的单元，其中所述特征数据包括多个频率的耳间相位差和多个频率的耳间强度差。

[0168] 示例53、根据示例50所述的装置，还包括：用于基于所述音频数据中表示的声音是否来源于相比于靠近所述第二麦克风，更靠近所述第一麦克风的源，自适应地对特征数据集进行聚类的单元。

[0169] 示例54、根据示例50所述的装置，还包括：用于基于所述特征数据，更新所述动态分类器的聚类操作的单元。

[0170] 示例55、根据示例50所述的装置，还包括：用于更新所述动态分类器的分类决策标准的单元。

[0171] 示例56、根据示例50所述的装置，其中，所述动态分类器包括自组织映射。

[0172] 示例57、根据示例50所述的装置，还包括：用于响应于确定所述音频数据对应于所述用户语音活动，而启动语音命令处理操作的单元。

[0173] 示例58、根据示例50所述的装置，还包括：用于生成唤醒信号或中断中的至少一个，以启动所述语音命令处理操作的单元。

[0174] 示例59、根据示例50所述的装置，还包括：用于响应于基于所述动态分类器而确定所述音频数据对应于所述用户语音活动，将所述音频数据发送到第二设备的单元。

[0175] 示例60、根据示例50所述的装置，其中，用于接收所述音频数据的单元、用于生成所述特征数据的单元、用于生成所述分类输出的单元、以及用于确定所述音频数据是否对应于所述用户语音活动的单元，集成在包括所述第一麦克风和所述第二麦克风的头戴式设备中，并且其中所述头戴式设备被配置为当用户佩戴时，将所述第一麦克风定位为比所述第二麦克风更靠近用户的嘴，以便与在所述第二麦克风处相比，以更大的强度和更小的延迟来捕捉所述用户在所述第一麦克风处的话语。

[0176] 示例61、根据示例50所述的装置，其中，用于接收所述音频数据的单元、用于生成所述特征数据的单元、用于生成所述分类输出的单元、以及用于确定所述音频数据是否对应于所述用户语音活动的单元，集成在移动电话、平板计算机设备、可穿戴电子设备、相机设备、虚拟现实头戴式设备、增强现实头戴式设备或运载工具中的至少一个中。

[0177] 示例62、一种设备包括：被配置为存储指令的存储器；以及一个或多个处理器，其

配置为执行所述指令以用于:接收音频数据,所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据;将所述音频数据提供给动态分类器,所述动态分类器被配置为生成与所述音频数据相对应的分类输出;并至少部分地基于所述分类输出,来确定所述音频数据是否对应于用户语音活动。

[0178] 示例63、根据示例62所述的设备,还包括所述第一麦克风和所述第二麦克风,其中所述第一麦克风耦合到所述一个或多个处理器并且被配置为捕获用户的话语,并且其中,所述第二麦克风耦合到所述一个或多个处理器并且配置为捕获环境声音。

[0179] 示例64、根据示例62或63所述的设备,其中,所述分类输出基于所述第一音频数据和所述第二音频数据之间的增益差、所述第一音频数据和所述第二音频数据之间的相位差、或其组合。

[0180] 示例65、根据示例62至64中的任何一项所述的设备,其中,所述一个或多个处理器进一步被配置为基于所述第一音频数据和所述第二音频数据来生成特征数据,其中,将所述音频数据作为所述特征数据提供给所述动态分类器,其中,所述分类输出基于所述特征数据。

[0181] 示例66、根据示例65所述的设备,其中,所述特征数据包括:在所述第一音频数据和所述第二音频数据之间的至少一个耳间相位差;以及在所述第一音频数据和所述第二音频数据之间的至少一个耳间强度差。

[0182] 示例67、根据示例65或66所述的设备,其中,所述一个或多个处理器被配置为进一步基于所述特征数据的至少一个值的符号或幅度中的至少一个,来确定所述音频数据是否对应于所述用户语音活动。

[0183] 示例68、根据示例65至67中的任何一项所述的设备,其中,所述一个或多个处理器进一步被配置为在生成所述特征数据之前,将所述第一音频数据和所述第二音频数据变换到变换域,并且其中所述特征数据包括针对多个频率的耳间相位差和针对多个频率的耳间强度差。

[0184] 示例69、根据示例65至68中的任何一项所述的设备,其中,所述动态分类器被配置为基于所述音频数据中表示的声音是否来源于相比于靠近所述第二麦克风,更靠近所述第一麦克风的源,自适应地对特征数据集进行聚类。

[0185] 示例70、根据示例62至69中的任何一项所述的设备,其中,所述一个或多个处理器进一步被配置为基于所述音频数据,更新所述动态分类器的聚类操作。

[0186] 示例71、根据示例62至70中的任何一项所述的设备,其中,所述一个或多个处理器进一步被配置为更新所述动态分类器的分类决策标准。

[0187] 示例72、根据示例62至71中的任何一项所述的设备,其中,所述动态分类器包括自组织映射。

[0188] 示例73、根据示例62至72中的任何一项所述的设备,其中,所述动态分类器进一步被配置为接收音频数据集合的序列,并且至少部分地基于所述序列中的先前音频数据集合,自适应地对所述序列的集合进行聚类。

[0189] 示例74、根据示例62至73中的任何一项所述的设备,其中,所述一个或多个处理器进一步被配置为响应于确定所述音频数据对应于所述用户语音活动,而启动语音命令处理操作。

[0190] 示例75、根据示例74所述的设备，其中，所述一个或多个处理器被配置为生成唤醒信号或中断中的至少一个，以启动所述语音命令处理操作。

[0191] 示例76、根据示例75所述的设备，其中，所述一个或多个处理器还包括：包括所述动态分类器的常开功率域；以及包括语音命令处理单元的第二功率域，并且其中，所述唤醒信号被配置为将所述第二功率域从低功率模式转换以激活所述语音命令处理单元。

[0192] 示例77、根据示例62至76中的任何一项所述的设备，还包括耦合到所述一个或多个处理器的调制解调器，所述调制解调器被配置为：响应于基于所述动态分类器而确定所述音频数据对应于所述用户语音活动，将所述音频数据发送到第二设备。

[0193] 示例78、根据示例62至77中的任何一项所述的设备，其中，所述一个或多个处理器集成在包括所述第一麦克风和所述第二麦克风的头戴式设备中，并且其中所述头戴式设备被配置为当用户佩戴时，将所述第一麦克风定位为比所述第二麦克风更靠近用户的嘴，以便与在所述第二麦克风处相比，以更大的强度和更小的延迟来捕捉所述用户在所述第一麦克风处的话语。

[0194] 示例79、根据示例62至77中的任何一项所述的设备，其中，所述一个或多个处理器集成在移动电话、平板计算机设备、可穿戴电子设备、相机设备、虚拟现实头戴式设备、或增强现实头戴式设备中的至少一个中。

[0195] 示例80、根据示例62至77中的任何一项所述的设备，其中，所述一个或多个处理器集成在运载工具中，所述运载工具还包括所述第一麦克风和所述第二麦克风，并且其中所述第一麦克风定位成捕获所述运载工具的操作者的话语。

[0196] 示例81、一种语音活动检测的方法包括：在一个或多个处理器处接收音频数据，所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据；在所述一个或多个处理器处，将所述音频数据提供给动态分类器，以生成与所述音频数据相对应的分类输出；并在所述一个或多个处理器处并且至少部分地基于所述分类输出，来确定所述音频数据是否对应于用户语音活动。

[0197] 示例82、根据示例81所述的方法，其中，所述分类输出基于所述第一音频数据和所述第二音频数据之间的增益差、所述第一音频数据和第二音频数据之间的相位差或其组合。

[0198] 示例83、根据示例81或82所述的方法，其中，所述动态分类器包括自组织映射。

[0199] 示例84、根据示例81至83中的任何一项所述的方法，其中，确定所述音频数据是否对应于所述用户语音活动进一步基于：与所述音频数据相对应的特征数据的至少一个值的符号或幅度中的至少一个。

[0200] 示例85、根据示例81至84中的任何一项所述的方法，还包括：响应于确定所述音频数据对应于所述用户语音活动，而启动语音命令处理操作。

[0201] 示例86、根据示例85所述的方法，还包括：生成唤醒信号或中断中的至少一个，以启动所述语音命令处理操作。

[0202] 示例87、根据示例81至86中的任何一项所述的方法，还包括：响应于基于所述动态分类器而确定所述音频数据对应于所述用户语音活动，将所述音频数据发送到第二设备。

[0203] 示例88、一种包括指令的非暂时性计算机可读介质，当所述指令由一个或多个处理器执行时，使得所述一个或多个处理器用于：接收音频数据，所述音频数据包括与第一麦

克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据;将所述音频数据提供给动态分类器,以生成与所述音频数据相对应的分类输出;并至少部分地基于所述分类输出,来确定所述音频数据是否对应于用户语音活动。

[0204] 示例89、根据示例88所述的非暂时性计算机可读介质,其中,所述分类输出基于所述第一音频数据和所述第二音频数据之间的增益差、所述第一音频数据和第二音频数据之间的相位差或其组合。

[0205] 示例90、一种装置包括:用于接收音频数据的单元,所述音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据;用于在动态分类器处,生成与所述音频数据相对应的分类输出的单元;用于至少部分地基于所述分类输出,来确定所述音频数据是否对应于用户语音活动的单元。

[0206] 示例91、根据示例90所述的装置,其中,所述分类输出基于所述第一音频数据和所述第二音频数据之间的增益差、所述第一音频数据和第二音频数据之间的相位差或其组合。

[0207] 本领域普通技术人员还应当理解,结合本文所公开实施方式描述的各种示例性的逻辑框、配置、模块、电路和算法步骤均可以实现成电子硬件、由处理器执行的计算机软件、或二者的组合。上面对各种示例性的部件、框、配置、模块、电路和步骤均围绕其功能进行了总体描述。至于这种功能是实现成硬件还是实现成处理器可执行指令,取决于特定的应用和对整个系统所施加的设计约束条件。熟练的技术人员可以针对每个特定应用,以变通的方式实现所描述的功能,但是,这种实现决策不应解释为背离本公开内容的保护范围。

[0208] 结合本文所公开实施方式描述的方法或者算法的步骤,可以直接体现为硬件、由处理器执行的软件模块或二者的组合。软件模块可以位于随机存取存储器(RAM)、闪存、只读存储器(ROM)、可编程只读存储器(PROM)、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)、寄存器、硬盘、移动硬盘、压缩光盘只读存储器(CD-ROM)、或者本领域已知的任何其它形式的非暂时性存储介质中。可以将一种示例性的存储介质连接至处理器,从而使该处理器能够从该存储介质读取信息,并可向该存储介质写入信息。在替代的方案中,存储介质也可以是处理器的组成部分。处理器和存储介质可以驻留在专用集成电路(ASIC)中。ASIC可以驻留在计算设备或用户终端中。替代地,处理器和存储介质可以作为分立组件驻留在计算设备或用户终端中。

[0209] 为使本领域普通技术人员能够实现或者使用所公开的方面,上面围绕所公开的方面进行了描述。对于本领域普通技术人员来说,对这些方面的各种修改是显而易见的,并且,本文所定义的原理也可以在不脱离本公开内容的保护范围的基础上适用于其它方面。因此,本公开内容并不限于本文所示出的方面,而是与如所附权利要求书所规定的原理和新颖性特征的最广范围相一致。

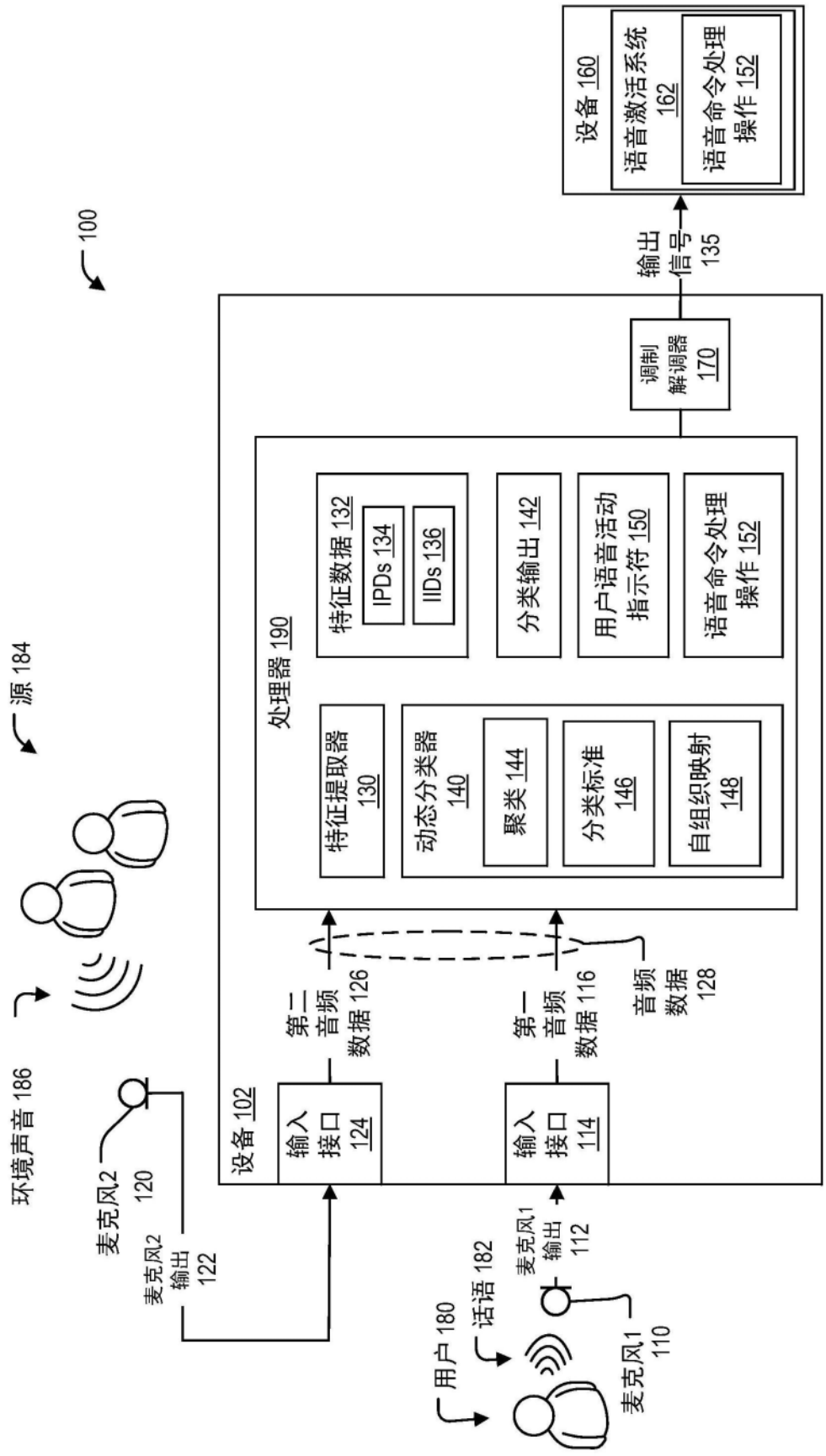

100
设备 160
语音激活系统 162
语音命令处理操作 152
输出信号 135
调制解调器 170
处理器 190
特征数据 132
IPDs 134
IIDs 136
分类输出 142
用户语音活动指示符 150
语音命令处理操作 152
特征提取器 130
动态分类器 140
聚类 144
分类标准 146
自组织映射 148
设备 102
输入接口 124
输入接口 114
第二音频数据 126
第一音频数据 116
音频数据 128
源 184
环境声音 186
麦克风2 120
麦克风2输出 122
麦克风1输出 112
麦克风1 110
用户 180
话语 182

图1

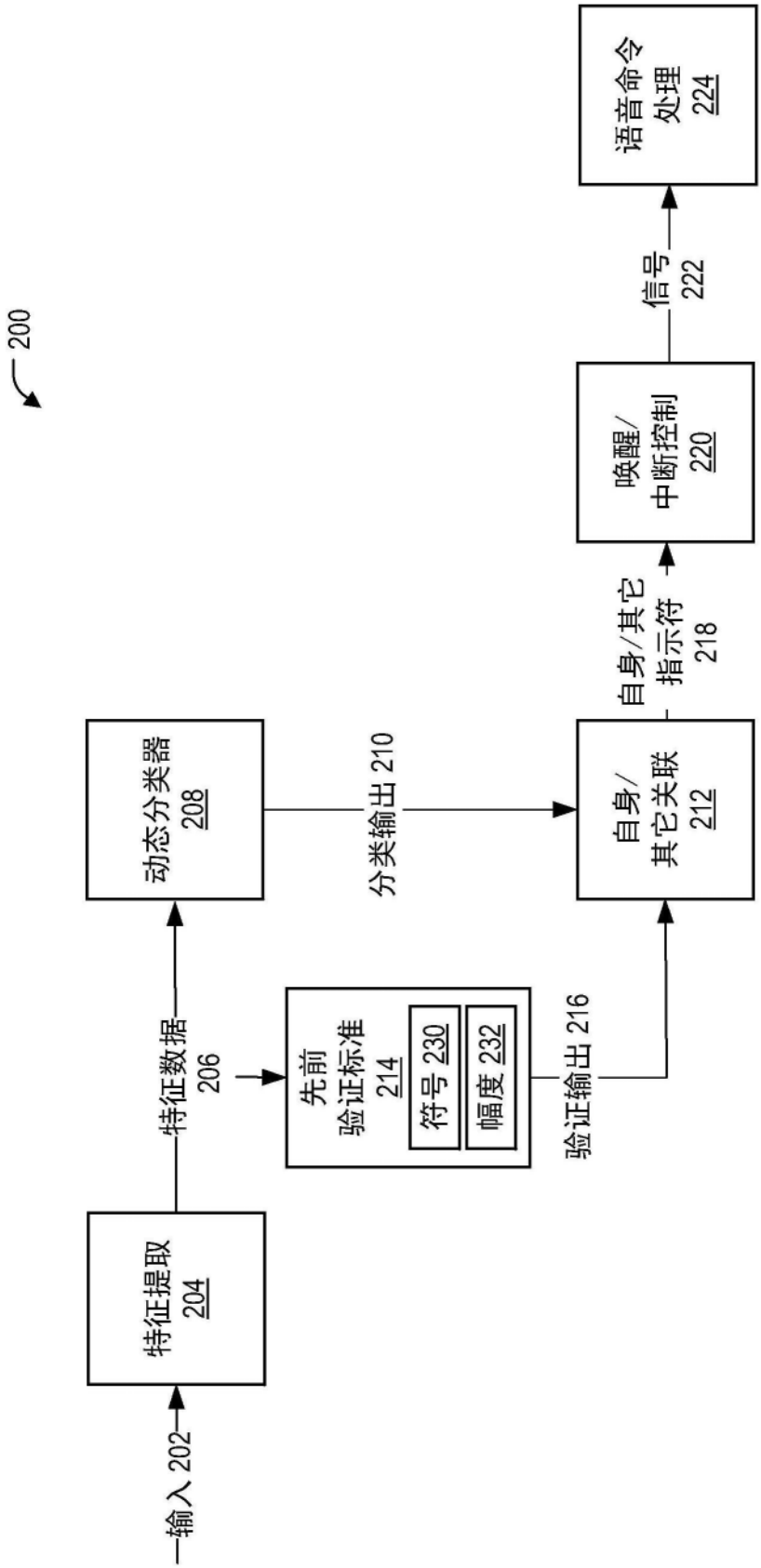

200
输入 202
特征提取 204
特征数据 206
动态分类器 208
分类输出 210
自身/其它关联 212
先前验证标准 214
符号 230
幅度 232
验证输出 216
自身/其它指示符 218
唤醒/中断控制 220
信号 222
语音命令处理 224

图2

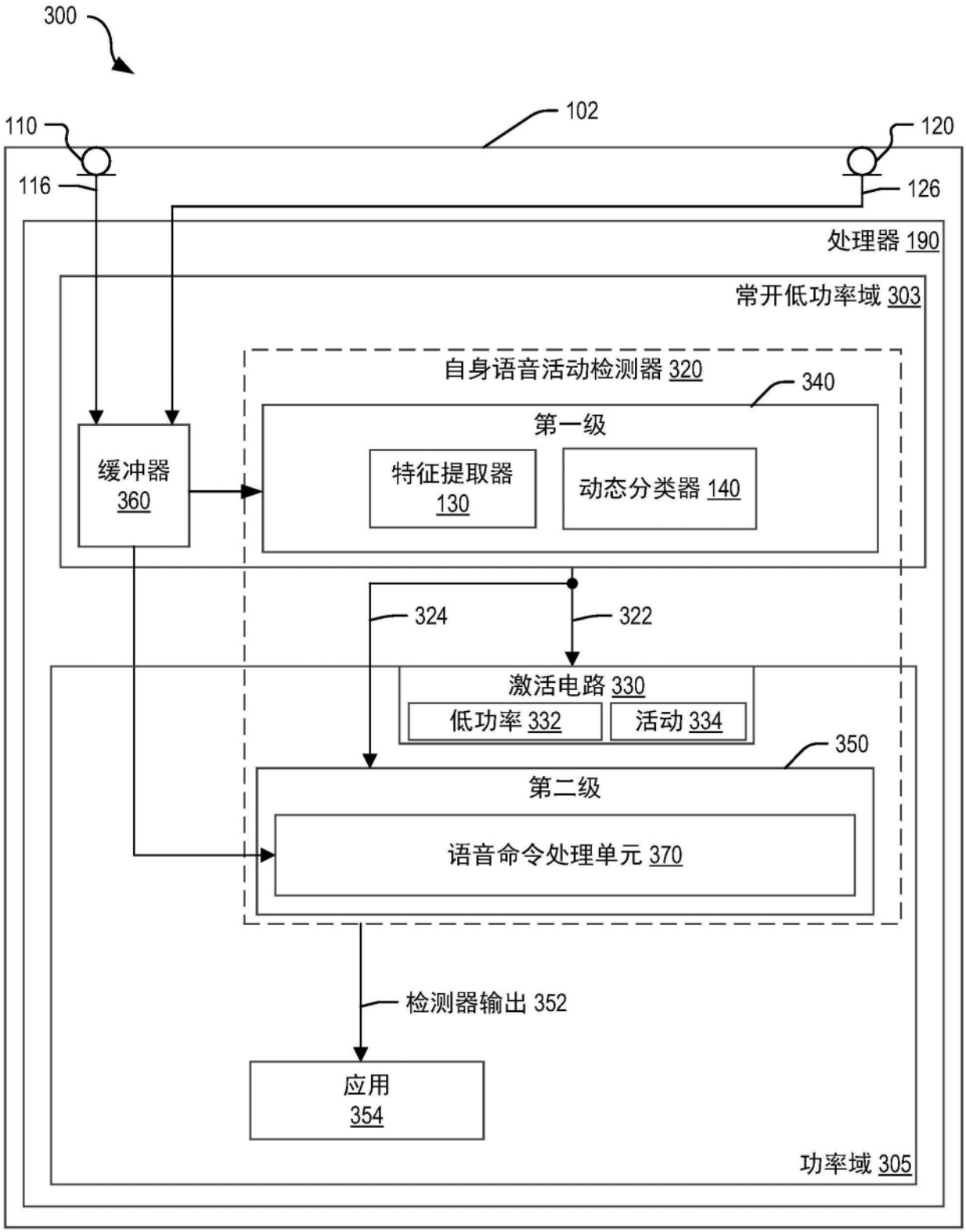

300
110
102
120
116
126
处理器 190
常开低功率域 303
自身语音活动检测器 320
340
第一级
缓冲器 360
特征提取器 130
动态分类器 140
324
322
激活电路 330
低功率 332
活动 334
350
第二级
语音命令处理单元 370
检测器输出 352
应用 354
功率域 305

图3

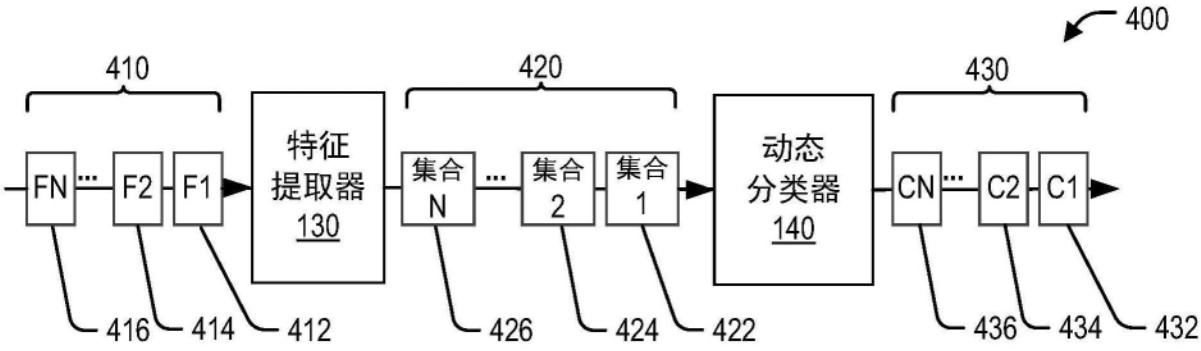

400
410
420
430
FN
F2
F1
特征提取器 130
集合 N
集合 2
集合 1
动态分类器 140
CN
C2
C1
416
414
412
426
424
422
436
434
432

图4

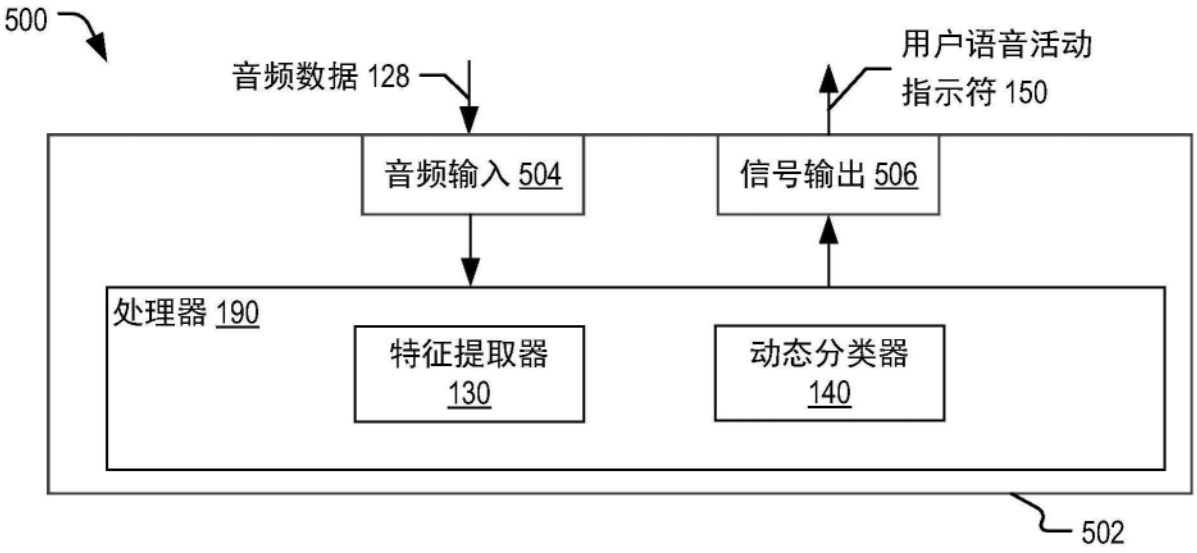

500
音频数据 128
用户语音活动指示符 150
音频输入 504
信号输出 506
处理器 190
特征提取器 130
动态分类器 140
502

图5

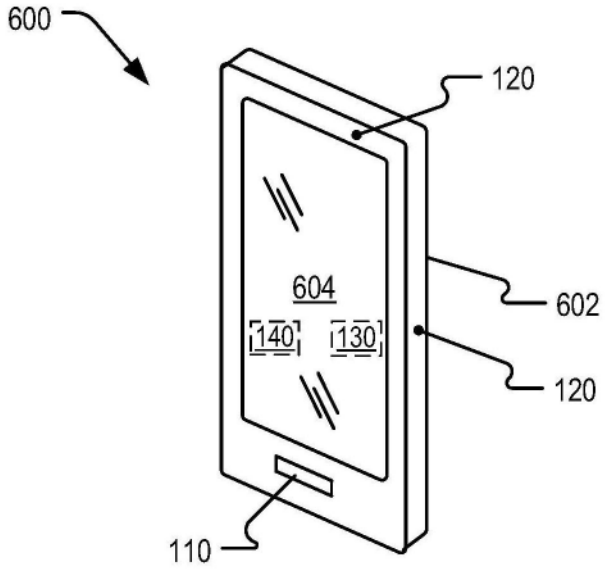


图6

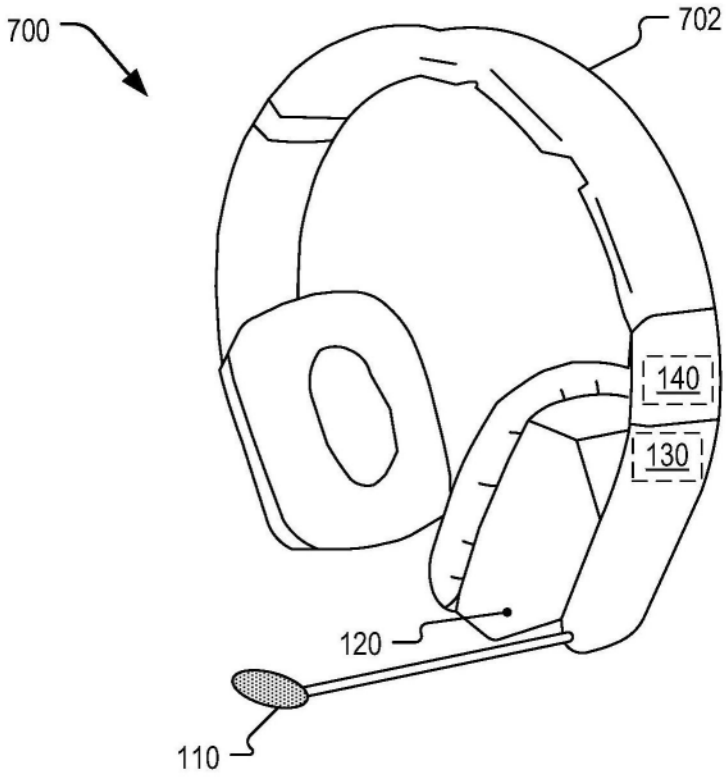


图7

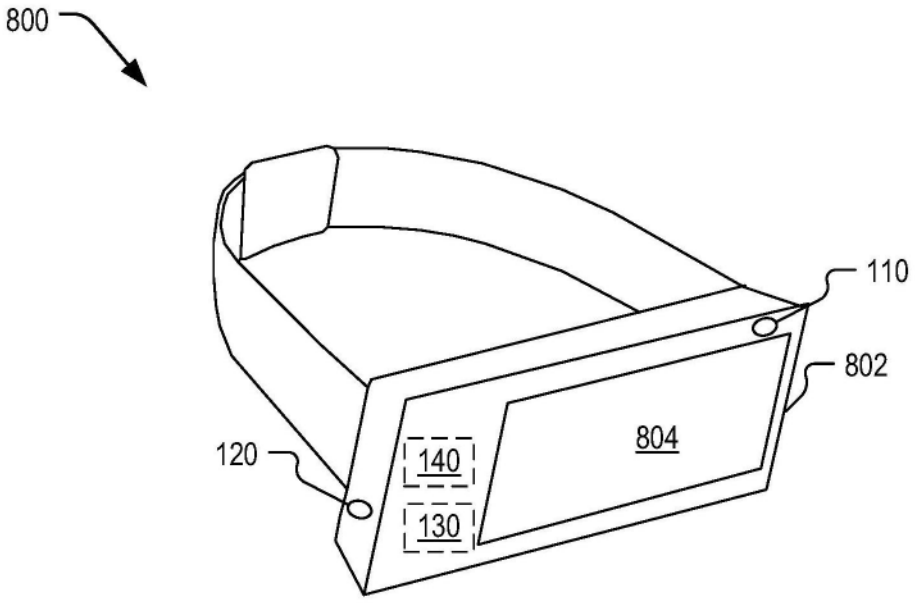

800
110
802
804
120
140
130

图8

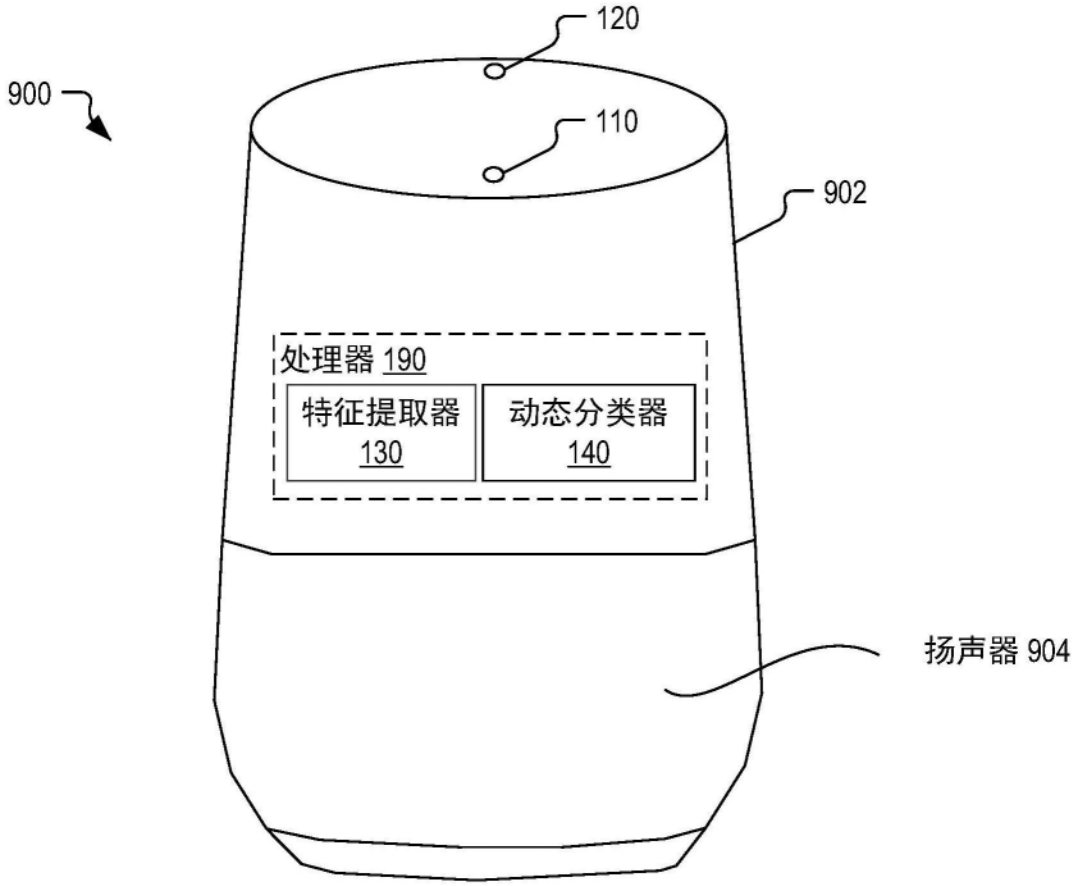

900
120
110
902
处理器 190
特征提取器 130
动态分类器 140
扬声器 904

图9

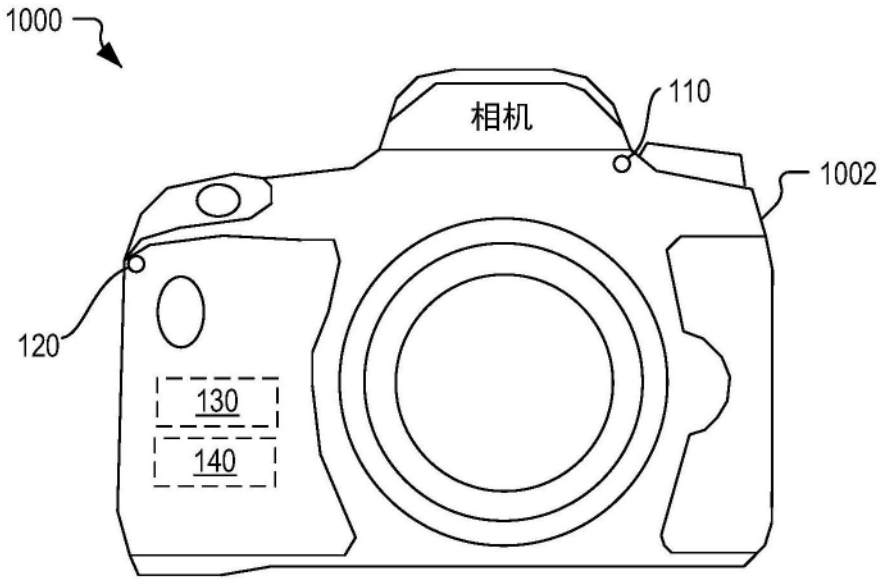


图10

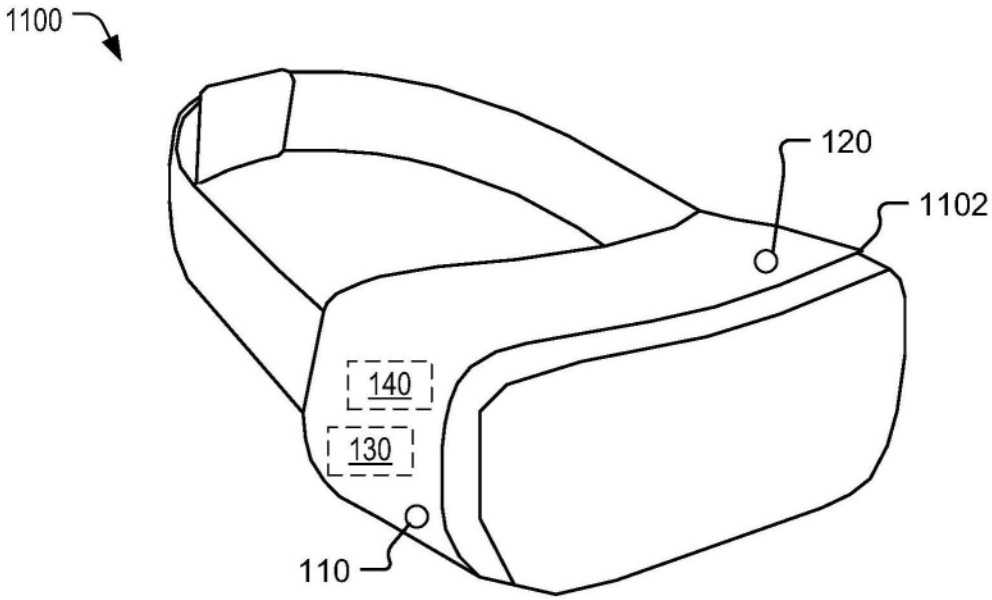


图11

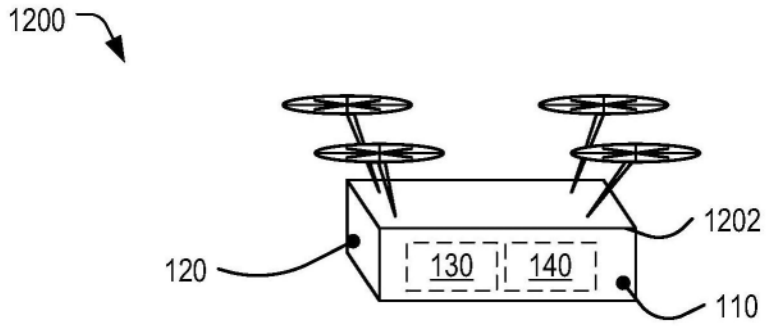


图12

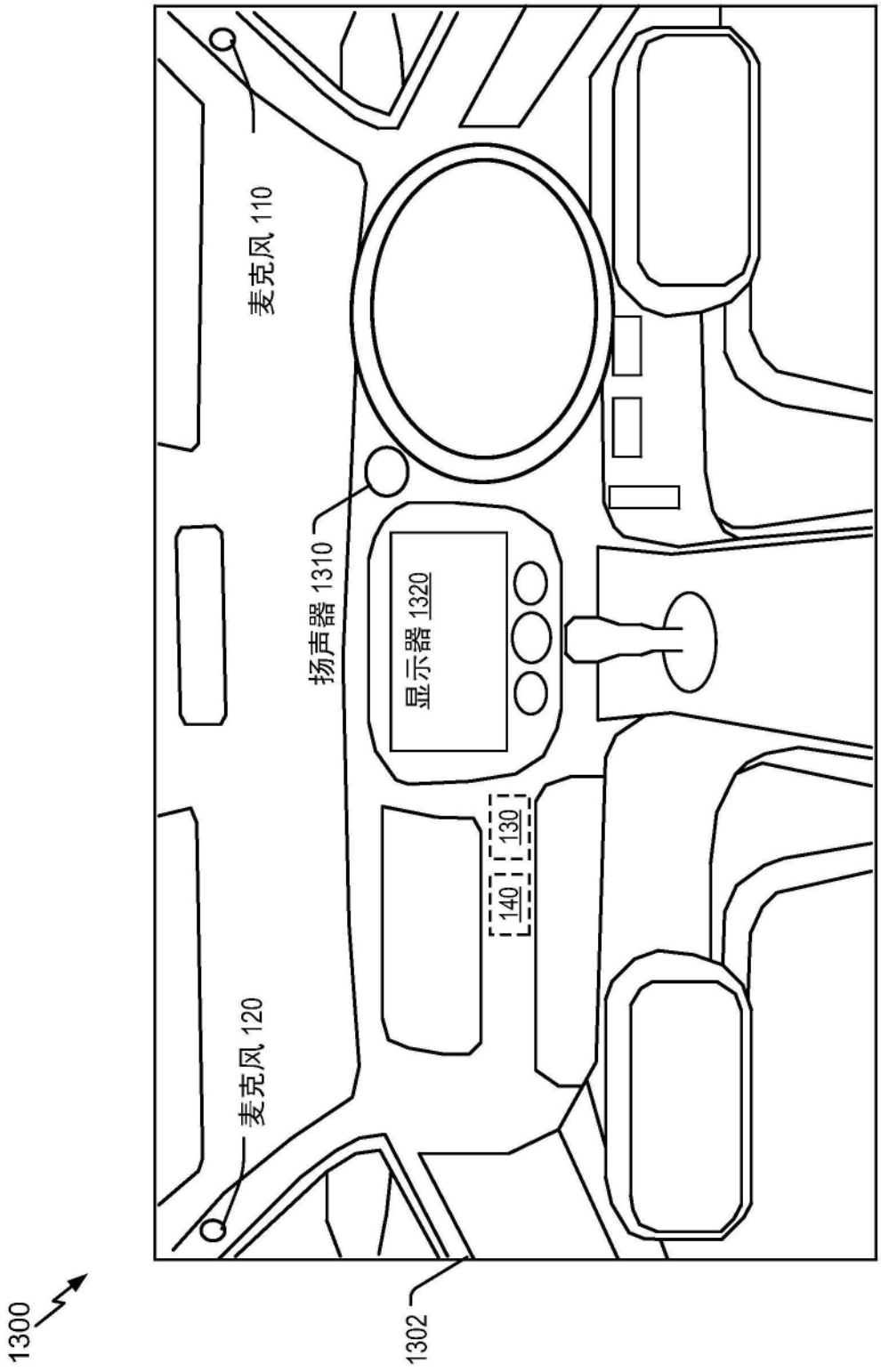


图13

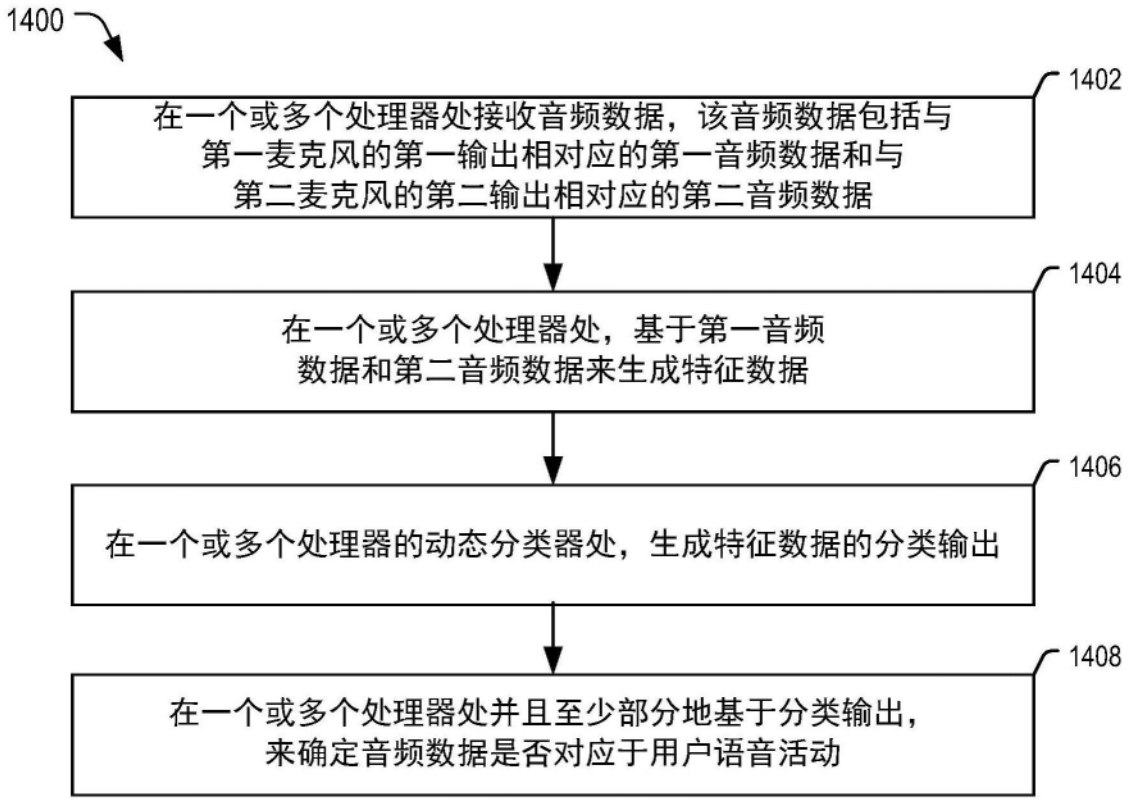

1400
1402
在一个或多个处理器处接收音频数据，该音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据
1404
在一个或多个处理器处，基于第一音频数据和第二音频数据来生成特征数据
1406
在一个或多个处理器的动态分类器处，生成特征数据的分类输出
1408
在一个或多个处理器处并且至少部分地基于分类输出，来确定音频数据是否对应于用户语音活动

图14A

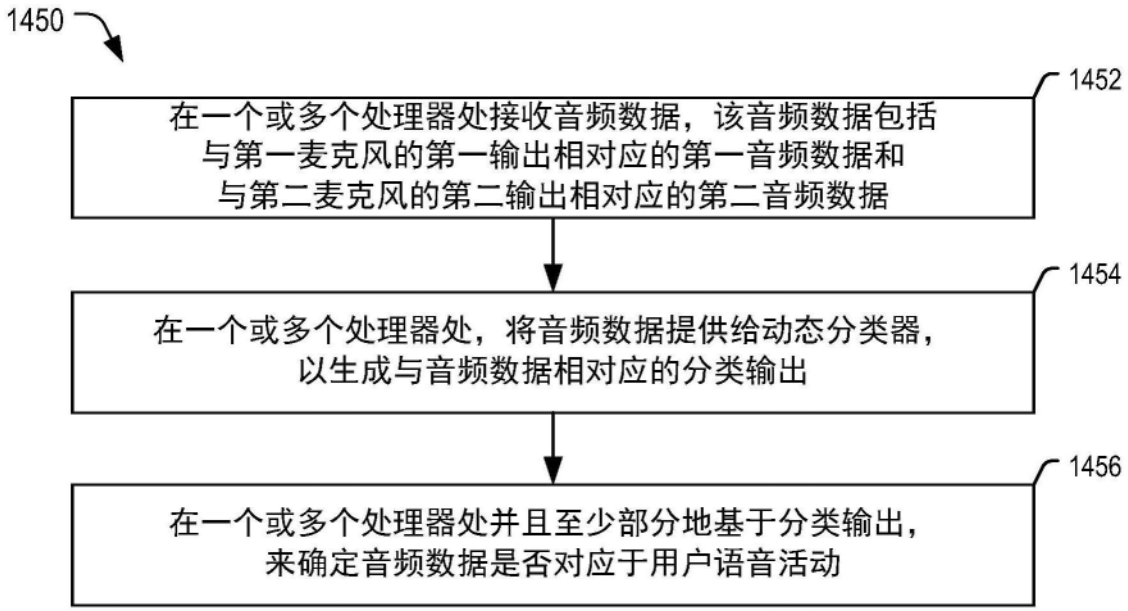

1450
1452
在一个或多个处理器处接收音频数据，该音频数据包括与第一麦克风的第一输出相对应的第一音频数据和与第二麦克风的第二输出相对应的第二音频数据
1454
在一个或多个处理器处，将音频数据提供给动态分类器，以生成与音频数据相对应的分类输出
1456
在一个或多个处理器处并且至少部分地基于分类输出，来确定音频数据是否对应于用户语音活动

图14B

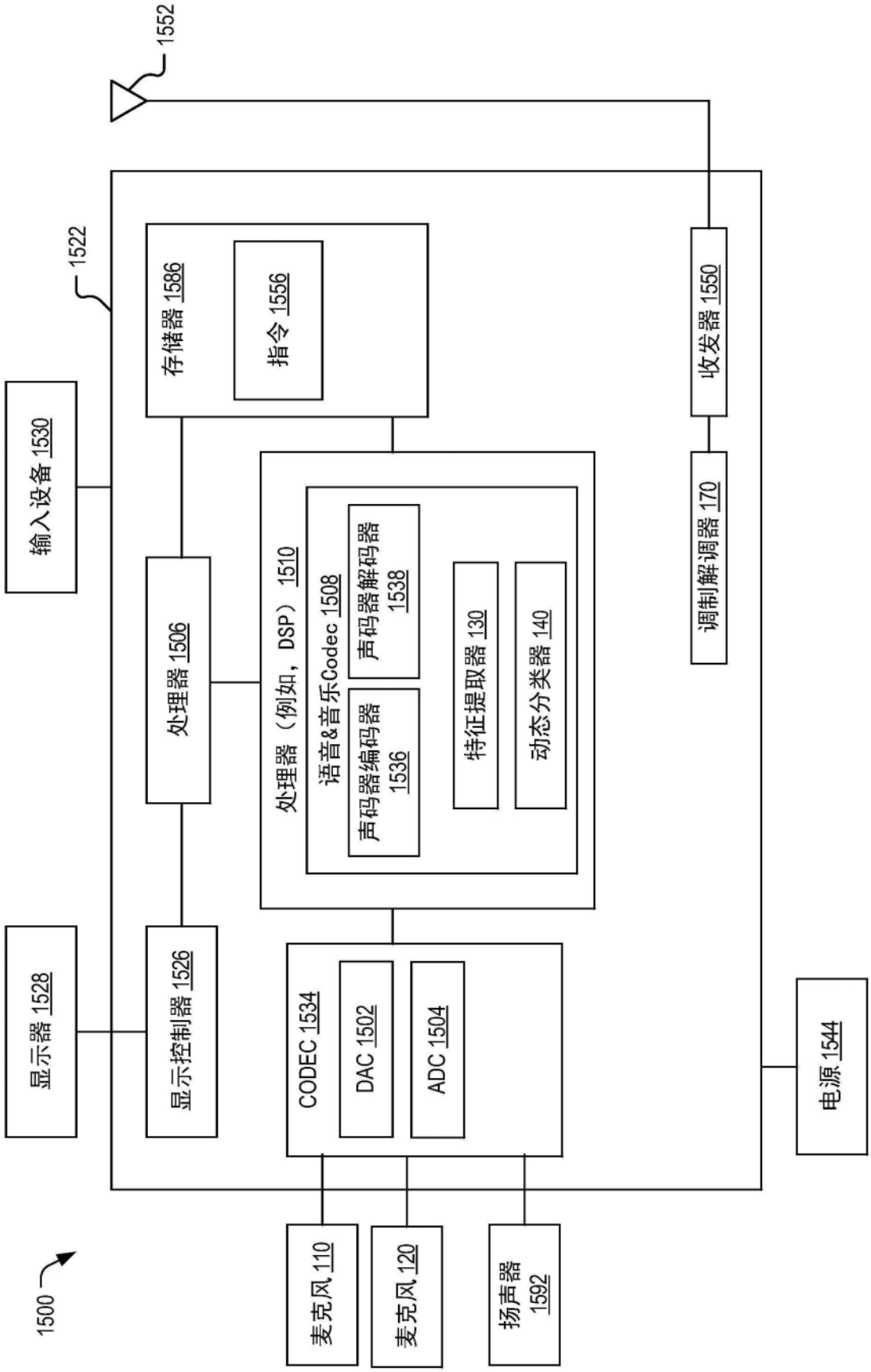

1552
1522
存储器 1586
指令 1556
输入设备 1530
处理器 1506
处理器（例如，DSP） 1510
语音&音乐Codec 1508
声码器解码器 1538
声码器编码器 1536
特征提取器 130
动态分类器 140
收发器 1550
调制解调器 170
显示器 1528
显示控制器 1526
CODEC 1534
DAC 1502
ADC 1504
电源 1544
麦克风 110
麦克风 120
扬声器 1592
1500

图15