(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118974822 A
(43)申请公布日 2024.11.15

(21)申请号 202280095063.2
(22)申请日 2022.10.18
(30)优先权数据
17/726,244 2022.04.21 US
(85)PCT国际申请进入国家阶段日
2024.10.18
(86)PCT国际申请的申请数据
PCT/US2022/047027 2022.10.18
(87)PCT国际申请的公布数据
WO2023/204841 EN 2023.10.26
(71)申请人 谷歌有限责任公司
地址 美国
(72)发明人 马丁·博伊姆尔
特胡尚·阿马拉西里瓦德纳
罗伯托·皮拉奇尼
蒋禄卡·马提尼
(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
专利代理师 朴金丹 周亚荣
(51)Int.Cl.
G10L 15/22(2006.01)
G06F 3/16(2006.01)
G06F 16/332(2006.01)
G06F 40/40(2006.01)
G06T 13/40(2006.01)
G10L 13/033(2006.01)
G10L 15/183(2006.01)

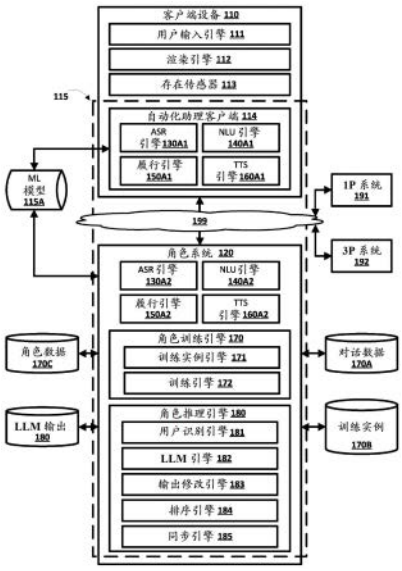
权利要求书8页 说明书35页 附图10页

(54)发明名称

基于指派给自动化助理的给定角色动态地适配给定助理输出

(57)摘要

各实现方式涉及基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配给定助理输出。在一些实现方式中,可以生成所述给定助理输出并且随后基于指派给所述自动化助理的所述给定角色对其进行适配。在其他实现方式中,可以生成特定于所述给定角色的所述给定助理输出,并且不必随后针对所述给定角色适配所述给定助理输出。值得注意的是,所述给定助理输出可以包括将合成以向用户可听呈现的文本内容流,以及用于控制客户端设备的显示和/或用于控制所述自动化助理的可视化表示的视觉提示流。各种实现方式利用大语言模型(LLM)或先前利用LLM生成的输出来反映所述给定助理输出中的所述给定角色。



CN 118974822 A

1.一种由一个或多个处理器实现的方法,所述方法包括:

接收捕获客户端设备的用户的口头话语的音频数据流,所述音频数据流由所述客户端设备的一个或多个麦克风生成,并且所述口头话语指向至少部分地在所述客户端设备处执行的自动化助理的实例;

基于处理所述音频数据流,生成响应于所述口头话语的给定助理输出,其中所述给定助理输出包括:(i)文本内容流;以及(ii)视觉提示流,所述视觉提示流用于响应于所述口头话语控制所述客户端设备的显示和/或用于控制被视觉地渲染以经由所述客户端设备的所述显示呈现给所述用户的所述自动化助理的所述实例的可视化表示;以及

基于由所述用户从多个不同的助理角色中指派给所述自动化助理的所述实例的给定助理角色,修改响应于所述口头话语的所述给定助理输出以生成修改后的给定助理输出,其中所述修改后的给定助理输出包括:(i)与所述文本内容流不同的修改后的文本内容流;以及(ii)与所述视觉提示流不同的修改后的视觉提示流,所述修改后的视觉提示流用于响应于所述口头话语控制所述客户端设备的所述显示和/或用于控制所述自动化助理的所述实例的所述可视化表示;以及

响应于接收到捕获所述客户端设备的所述用户的所述口头话语的所述音频数据流:

使捕获与所述修改后的文本内容流相对应的合成语音的合成语音音频数据被可听地渲染以经由所述客户端设备的一个或多个扬声器呈现给所述用户;以及

使所述修改后的视觉提示流被利用以控制所述客户端设备的所述显示和/或用以控制所述自动化助理的所述实例的所述可视化表示。

2.如权利要求1所述的方法,还包括:

使与所述修改后的文本内容流相对应的所述合成语音的所述可听渲染和所述修改后的视觉提示流在控制所述客户端设备的所述显示和/或在控制所述自动化助理的所述实例的所述可视化表示中的所述利用同步,以呈现给所述用户。

3.如权利要求2所述的方法,还包括:

在使捕获与所述修改后的文本内容流相对应的所述合成语音的所述合成语音音频数据被可听地渲染以呈现给所述用户之前:

用一个或多个视觉提示时间戳注释所述修改后的文本内容流,所述一个或多个视觉提示时间戳指示所述视觉提示流何时被利用以控制所述客户端设备的所述显示和/或用于控制所述自动化助理的所述实例的所述可视化表示,

其中使与所述修改后的文本内容流相对应的所述合成语音的所述可听渲染和所述视觉提示流在控制所述客户端设备的所述显示和/或在控制所述自动化助理的所述实例的所述可视化表示中的所述利用同步是基于所述一个或多个视觉提示时间戳。

4.如权利要求3所述的方法,其中所述一个或多个视觉提示时间戳至少包括开始视觉提示时间戳和停止视觉提示时间戳,所述开始视觉提示时间戳指示包括在所述视觉提示流中的给定视觉提示何时将开始被利用以控制所述客户端设备的所述显示和/或用于控制所述自动化助理的所述实例的所述可视化表示,所述停止视觉提示时间戳指示包括在所述视觉提示流中的所述给定视觉提示何时将停止被利用以控制所述客户端设备的所述显示和/或用于控制所述自动化助理的所述实例的所述可视化表示。

5.如任一前述权利要求所述的方法,其中基于处理所述音频数据流生成响应于所述口

头话语的所述给定助理输出包括：

使用自动语音辨识(ASR)模型来处理捕获所述口头话语的所述音频数据流以生成ASR输出流；

使用自然语言理解(NLU)模型来处理所述ASR输出流以生成NLU输出流；以及

至少基于所述NLU输出流来确定响应于所述口头话语的所述给定助理输出。

6.如权利要求5所述的方法，其中基于所述NLU输出流来确定响应于所述口头话语的所述给定助理输出包括：

使用大语言模型(LLM)处理所述ASR输出流和/或所述NLU输出流以确定包括在所述给定助理输出中的所述文本内容流和所述视觉提示流。

7.如权利要求5所述的方法，其中基于所述NLU输出流来确定响应于所述口头话语的所述给定助理输出包括：

使用先前使用基于所述口头话语的先前实例的大语言模型(LLM)生成的LLM输出，处理所述ASR输出流和/或所述NLU输出流以确定包括在所述给定助理输出中的所述文本内容流和所述视觉提示流。

8.如权利要求5所述的方法，其中基于所述NLU输出流来确定响应于所述口头话语的所述给定助理输出包括：

基于所述ASR输出流和/或所述NLU输出流，生成一个或多个结构化请求；

向一个或多个第一方智能体和/或一个或多个第三方智能体传输所述一个或多个结构化请求；以及

基于响应于所述一个或多个结构化请求而接收到的内容，确定包括在所述给定助理输出中的所述文本内容流和所述视觉提示流。

9.根据任一前述权利要求所述的方法，其中基于指派给所述自动化助理的所述实例的所述给定助理角色来修改响应于所述口头输出的所述给定助理输出以生成所述修改后的给定助理输出包括：

从一个或多个数据库获得特定于指派给所述自动化助理的所述实例的所述给定角色的给定角色数据；以及

处理所述文本内容流和所述视觉提示流以及特定于指派给所述自动化助理的所述实例的所述给定角色的所述角色数据以生成与所述文本内容流不同的所述修改后的文本内容流和与所述视觉提示流不同的所述修改后的视觉提示流。

10.如权利要求9所述的方法，其中特定于指派给所述自动化助理的所述实例的所述给定角色的所述角色数据包括特定于指派给所述自动化助理的所述实例的所述给定角色的给定角色词元和/或特定于指派给所述自动化助理的所述实例的所述给定角色的给定嵌入。

11.如权利要求9或权利要求10所述的方法，其中处理所述文本内容流和所述视觉提示流以及特定于指派给所述自动化助理的所述实例的所述给定角色的所述角色数据以生成与所述文本内容流不同的所述修改后的文本内容流和与所述视觉提示流不同的所述修改后的视觉提示流包括：

使用大语言模型(LLM)来处理所述文本内容流和所述视觉提示流以及特定于指派给所述自动化助理的所述实例的所述给定角色的所述角色数据以生成与所述文本内容流不同

的所述修改后的文本内容流和与所述视觉提示流不同的所述修改后的视觉提示流。

12.如权利要求9或权利要求10所述的方法,其中处理所述文本内容流和所述视觉提示流以及特定于指派给所述自动化助理的所述实例的所述给定角色的所述角色数据以生成与所述文本内容流不同的所述修改后的文本内容流和与所述视觉提示流不同的所述修改后的视觉提示流包括:

使用先前使用基于所述口头话语的先前实例的大语言模型(LLM)生成的LLM输出,处理所述文本内容流和所述视觉提示流以及特定于指派给所述自动化助理的所述实例的所述给定角色的所述角色数据以生成与所述文本内容流不同的所述修改后的文本内容流和与所述视觉提示流不同的所述修改后的视觉提示流。

13.根据任一前述权利要求所述的方法,其中指派给所述 自动化助理的所述实例的所述给定角色与以下相关联:

多个不同词汇中的第一词汇,被利用以修改所述文本内容流以生成所述修改后的文本内容流,

多个不同的韵律属性集合中的第一韵律属性集合,被利用以生成捕获与所述修改后的文本内容流相对应的所述合成语音的所述合成语音音频数据,以被可听地渲染以呈现给所述用户,以及/或者

多个不同的视觉提示集合中的第一视觉提示集合,被利用以修改所述视觉提示流以生成所述修改后的视觉提示流。

14.如权利要求13所述的方法,其中与所述文本内容流不同的所述修改后的文本内容流是使用所述第一词汇修改的,并且其中与所述视觉提示流不同的用于响应于所述口头话语控制所述客户端设备的所述显示和/或用于控制所述自动化助理的所述实例的所述可视化表示的所述修改后的视觉提示流是使用所述第一视觉提示集合修改的。

15.如权利要求14所述的方法,还包括:

使用文本转语音(TTS)模型并且基于所述第一韵律属性集合,处理所述修改后的文本内容流以生成所述合成语音音频数据。

16.如权利要求13或权利要求14所述的方法,还包括:

接收捕获附加客户端设备的附加用户的附加口头话语的附加音频数据流,所述附加音频数据流由所述附加客户端设备的一个或多个附加麦克风生成,所述附加口头话语指向至少部分地在所述附加客户端设备处执行的所述自动化助理的附加实例,并且所述附加口头话语与所述口头话语相同;

基于处理所述附加音频数据流,生成响应于所述附加口头话语的所述给定助理输出,其中所述给定助理输出包括:(i)所述文本内容流;以及(ii)所述视觉提示流,所述视觉提示流用于响应于所述附加口头话语控制所述附加客户端设备的附加显示和/或用于控制被视觉地渲染以经由所述附加客户端设备的所述附加显示呈现给所述附加用户的所述自动化助理的所述附加实例的附加可视化表示;以及

基于由所述附加用户从所述多个不同的助理角色中指派给所述自动化助理的所述附加实例的除所述给定助理角色之外的给定附加助理角色,修改响应于所述附加口头话语的所述给定助理输出以生成修改后的给定附加助理输出,其中所述修改后的给定附加助理输出包括:(i)与所述文本内容流不同并且与所述修改后的文本内容流不同的修改后的附加

文本内容流;以及(ii)与所述视觉提示流不同并且与所述修改后的视觉提示流不同的修改后的附加视觉提示流,所述修改后的附加视觉提示流用于响应于所述附加口头话语控制所述附加客户端设备的所述附加显示和/或用于控制所述自动化助理的所述附加实例的所述附加可视化表示;以及

响应于接收到捕获所述附加客户端设备的所述附加用户的所述附加口头话语的所述附加音频数据流:

使捕获与所述修改后的附加文本内容流相对应的附加合成语音的附加合成语音音频数据被可听地渲染以经由所述附加客户端设备的一个或多个附加扬声器呈现给所述附加用户;以及

使所述修改后的附加视觉提示流被利用以控制所述附加客户端设备的所述附加显示和/或用以控制所述自动化助理的所述附加实例的所述附加可视化表示。

17.根据权利要求16所述的方法,其中指派给所述自动化助理的所述附加实例的所述给定附加角色与以下相关联:

所述多个不同词汇中除所述第一词汇之外的第二词汇,被利用以修改所述附加文本内容流以生成所述修改后的附加文本内容流,

所述多个不同的韵律属性集合中的除所述第一韵律属性集合之外的第二韵律属性集合,被利用以生成捕获与所述修改后的附加文本内容流相对应的所述附加合成语音的所述附加合成语音音频数据,以被可听地渲染以呈现给所述附加用户,以及/或者

所述多个不同的视觉提示集合中的除所述第一视觉提示集合之外的第二视觉提示集合,被利用以修改所述附加视觉提示流以生成所述修改后的附加视觉提示流。

18.如任一项前述权利要求所述的方法,还包括:

接收捕获所述客户端设备的附加用户的附加口头话语的附加音频数据流,所述附加音频数据流由所述客户端设备的所述一个或多个附加麦克风生成,所述附加口头话语指向至少部分地在所述客户端设备处执行的所述自动化助理的所述实例,并且所述附加口头话语与所述口头话语相同;

基于处理所述附加音频数据流,生成响应于所述附加口头话语的所述给定助理输出,其中所述给定助理输出包括:(i)所述文本内容流;以及(ii)所述视觉提示流,所述视觉提示流用于响应于所述附加口头话语控制所述客户端设备的所述显示和/或用于控制被视觉地渲染以经由所述客户端设备的所述显示呈现给所述附加用户的所述自动化助理的所述实例的附加可视化表示;以及

基于由所述附加用户从所述多个不同的助理角色中指派给所述自动化助理的所述附加实例的除所述给定助理角色之外的给定附加助理角色,修改响应于所述附加口头话语的所述给定助理输出以生成修改后的给定附加助理输出,其中所述修改后的给定附加助理输出包括:(i)与所述文本内容流不同并且与所述修改后的文本内容流不同的修改后的附加文本内容流;以及(ii)与所述视觉提示流不同并且与所述修改后的视觉提示流不同的修改后的附加视觉提示流,所述修改后的附加视觉提示流用于响应于所述附加口头话语控制所述客户端设备的所述显示和/或用于控制所述自动化助理的所述实例的所述附加可视化表示;以及

响应于接收到捕获所述客户端设备的所述附加用户的所述附加口头话语的所述附加

音频数据流：

使捕获与所述修改后的附加文本内容流相对应的附加合成语音的附加合成语音音频数据被可听地渲染以经由所述客户端设备的所述一个或多个扬声器呈现给所述附加用户；以及

使所述修改后的附加视觉提示流被利用以控制所述客户端设备的所述显示和/或用以控制所述自动化助理的所述实例的所述附加可视化表示。

19.如任一前述权利要求所述的方法,其中所述客户端设备的所述用户在初始配置所述自动化助理的所述实例的自动化助理帐户的同时或在与所述自动化助理的所述实例的自动化助理应用的助理设置交互的同时将所述给定角色指派给所述自动化助理的所述实例。

20.如任一前述权利要求所述的方法,其中所述修改后的视觉提示流用于响应于所述口头话语控制所述客户端设备的所述显示。

21.如权利要求20所述的方法,其中所述修改后的视觉提示流还用于控制所述自动化助理的所述实例的所述可视化表示。

22.如权利要求20所述的方法,其中用于响应于所述口头话语 控制所述客户端设备的所述显示的所述修改后的视觉提示流包括一个或多个显示动画,所述一个或多个显示动画使所述客户端设备的所述显示在所述合成语音音频数据被可听地渲染以呈现给所述用户的同时被动态地适配。

23.如任一前述权利要求所述的方法,其中所述修改后的视觉提示流用于控制所述自动化助理的所述实例的所述可视化表示。

24.如权利要求23所述的方法,其中所述修改后的视觉提示流还用于响应于所述口头话语控制所述客户端设备的所述显示。

25.如权利要求23所述的方法,其中用于控制所述自动化助理的所述实例的所述可视化表示的所述修改后的视觉提示流包括在所述合成语音音频数据被可听地渲染以呈现给所述用户的同时由所述自动化助理的所述实例的所述可视化表示执行的一个或多个动画物理手势动作。

26.一种由一个或多个处理器实现的方法,所述方法包括：

接收捕获客户端设备的用户的口头话语的音频数据流,所述音频数据流由所述客户端设备的一个或多个麦克风生成,并且所述口头话语指向至少部分地在所述客户端设备处执行的自动化助理的实例；

基于处理所述音频数据流并且使用给定大语言模型(LLM),生成响应于所述口头话语并且特定于从多个不同的角色中指派给所述自动化助理的所述实例的给定角色的给定助理输出,其中所述给定助理输出包括:(i)文本内容流,所述文本内容流特定于所述自动化助理的所述实例的所述给定角色;以及(ii)视觉提示流,所述视觉提示流用于响应于所述口头话语控制所述客户端设备的显示和/或用于控制被视觉地渲染以经由所述客户端设备的所述显示呈现给所述用户的所述自动化助理的所述实例的可视化表示并且特定于指派给所述自动化助理的所述实例的所述给定角色;以及

响应于接收到捕获所述客户端设备的所述用户的所述口头话语的所述音频数据流：

使捕获与文本内容流相对应的合成语音的合成语音音频数据被可听地渲染以经由所

述客户端设备的一个或多个扬声器呈现给所述用户;以及

使所述视觉提示流被利用以控制所述客户端设备的所述显示和/或用以控制所述自动化助理的所述实例的所述可视化表示。

27.如权利要求26所述的方法,还包括:

从多个不同的角色中识别由所述用户指派给所述自动化助理的所述实例的所述给定角色;以及

从多个不同的LLM中选择与由所述用户指派给所述自动化助理的所述给定角色相关联的给定LLM。

28.如权利要求26所述的方法,还包括:

从多个不同的角色中识别由所述用户指派给所述自动化助理的所述实例的所述给定角色;以及

选择特定于指派给所述自动化助理的所述实例的所述给定角色的给定角色数据;并且

其中所述给定角色数据是在生成所述给定助理输出时使用所述给定LLM处理的。

29.如权利要求28所述的方法,其中特定于指派给所述自动化助理的所述实例的所述给定角色的所述角色数据包括特定于指派给所述自动化助理的所述实例的所述给定角色的给定角色词元和/或特定于指派给所述自动化助理的所述实例的所述给定角色的给定嵌入。

30.一种由一个或多个处理器实现的方法,所述方法包括:

接收捕获客户端设备的用户的口头话语的音频数据流,所述音频数据流由所述客户端设备的一个或多个麦克风生成,并且所述口头话语指向至少部分地在所述客户端设备处执行的自动化助理的实例;

基于处理所述音频数据流,生成响应于所述口头话语的给定助理输出,其中所述给定助理输出包括:(i)文本内容流;以及(ii)视觉提示流,所述视觉提示流用于响应于所述口头话语控制所述客户端设备的显示和/或用于控制被视觉地渲染以经由所述客户端设备的所述显示呈现给所述用户的所述自动化助理的所述实例的可视化表示;以及

基于由所述用户从多个不同的助理角色中指派给所述自动化助理的所述实例的给定助理角色,修改响应于所述口头话语的所述给定助理输出以生成修改后的给定助理输出,其中所述修改后的给定助理输出包括:(i)所述文本内容流;以及(ii)与所述视觉提示流不同的修改后的视觉提示流,所述修改后的视觉提示流用于响应于所述口头话语控制所述客户端设备的所述显示和/或用于控制所述自动化助理的所述实例的所述可视化表示;以及

响应于接收到捕获所述客户端设备的所述用户的所述口头话语的所述音频数据流:

使捕获与文本内容流相对应的合成语音的合成语音音频数据被可听地渲染以经由所述客户端设备的一个或多个扬声器呈现给所述用户;以及

使所述修改后的视觉提示流被利用以控制所述客户端设备的所述显示和/或用以控制所述自动化助理的所述实例的所述可视化表示。

31.一种由一个或多个处理器实现的方法,所述方法包括:

从与自动化助理相关联的开发者并且针对能够指派给所述自动化助理的给定角色,接收与将被利用以相对于文本内容流控制客户端设备的显示和/或用以相对于所述文本内容流控制所述自动化助理的实例的可视化表示的一个或多个视觉提示相关联的开发者输入;

至少基于所述开发者输入,生成将用于进一步训练特定于多个不同的角色中的所述给定角色的给定大语言模型(LLM)的实例的给定角色训练实例,所述给定LLM先前被训练为生成所述文本内容流;

至少基于所述给定角色训练实例,训练所述给定LLM的所述实例;以及

使所述给定LLM的所述实例被利用以随后处理捕获指向被指派所述给定角色的所述自动化助理的所述实例的口头话语的音频数据。

32.如权利要求31所述的方法,还包括:

从与所述自动化助理相关联的所述开发者并且针对能够指派给所述自动化助理的给定附加角色,接收与将被利用以相对于附加文本内容流控制所述客户端设备的所述显示和/或用于相对于所述附加文本内容流控制所述自动化助理的附加实例的附加可视化表示的一个或多个附加视觉提示相关联的附加开发者输入;

至少基于所述附加开发者输入,生成将被利用以进一步训练特定于所述多个不同的角色中的所述给定附加角色的所述给定LLM的附加实例的给定附加角色训练实例;

至少基于所述给定附加角色训练实例,训练所述给定附加LLM的所述附加实例;以及

使所述给定附加LLM的所述附加实例被利用以随后处理捕获指向被指派所述给定附加角色的所述自动化助理的所述附加实例的附加口头话语的附加音频数据。

33.如权利要求31或权利要求32所述的方法,其中所述开发者输入用一个或多个视觉提示时间戳注释所述文本内容流,所述一个或多个视觉提示时间戳指示所述视觉提示流何时用于相对于所述文本内容流控制所述客户端设备的所述显示和/或用于相对于所述文本内容流控制所述自动化助理的所述实例的所述可视化表示。

34.如权利要求33所述的方法,其中所述一个或多个视觉提示时间戳至少包括开始视觉提示时间戳和停止视觉提示时间戳,所述开始视觉提示时间戳指示包括在所述视觉提示流中的给定视觉提示何时将开始被利用以相对于所述文本内容流控制所述客户端设备的所述显示和/或用以相对于所述文本内容流控制所述自动化助理的所述实例的所述可视化表示,所述停止视觉提示时间戳指示包括在所述视觉提示流中的所述给定视觉提示何时将停止被利用以相对于所述文本内容流控制所述客户端设备的所述显示和/或用以相对于所述文本内容流控制所述自动化助理的所述实例的所述可视化表示。

35.如权利要求31所述的方法,其中所述开发者输入相对于所述文本内容流修改所述客户端设备的所述显示处的屏幕动画,和/或使所述自动化助理的所述实例的所述可视化表示相对于所述文本内容流执行一个或多个动画物理手势动作。

36.一种由一个或多个处理器实现的方法,所述方法包括:

从在线多媒体储存库获得视频内容,所述视频内容包括所述视频的可听内容的音频数据流和所述视频的视觉内容的视觉数据流;

使用自动语音辨识模型,处理所述视频的所述可听内容的所述音频数据流,以生成与在所述视频的所述可听内容的所述音频数据流中捕获的一个或多个口头话语相对应的文本内容流;

使用一个或多个运动跟踪机器学习模型,处理所述视频的所述视觉内容的所述视觉数据流以生成视觉提示流;

基于处理所述音频数据流并且基于处理所述视频数据流,生成将用于进一步训练特定

于多个不同的角色中的给定角色的体现在所述视频内容中的给定大语言模型(LLM)的实例的给定角色训练数据实例;

至少基于所述给定角色训练实例,训练所述给定LLM的所述实例;以及

使所述给定LLM的所述实例被利用以随后处理捕获指向被指派所述给定角色的自动化助理的实例的附加口头话语的附加音频数据。

37.一种系统,包括:

至少一个处理器;以及

存储器,所述存储器存储指令,所述指令在被执行时使所述至少一个处理器执行与权利要求1至36中任一项相对应的操作。

38.一种非暂时性计算机可读存储介质,所述非暂时性计算机可读存储介质存储指令,所述指令在被执行时使至少一个处理器执行与权利要求1至36中任一项相对应的操作。

基于指派给自动化助理的给定角色动态地适配给定助理输出

背景技术

[0001] 人类可以与在本文中称为“自动化助理”(也称为“聊天机器人”、“交互式个人助理”、“智能个人助理”、“个人语音助理”、“对话智能体”等)的交互式软件应用进行人机对话。自动化助理通常依赖组件流水线来解释和响应口头话语。例如,自动语音辨识(ASR)引擎可以处理与用户的口头话语相对应的音频数据以生成ASR输出,诸如口头话语的ASR假设(即,词项和/或其他标记的序列)。此外,自然语言理解(NLU)引擎可以处理ASR输出(或触摸/键入输入)以生成NLU输出,诸如用户在提供口头话语(或触摸/键入输入)时表达的请求(例如,意图)以及可选地与意图相关联的参数的槽值。最终,NLU输出可以由各种履行组件处理以生成履行输出,诸如响应于口头话语的响应内容和/或可以响应于口头话语而执行的一个或多个动作。

[0002] 通常,这些自动化助理使用上述组件流水线来对口头话语进行响应。例如,这些自动化助理可以使用各种文本转语音(TTS)技术使可听内容被提供用于向用户进行可听呈现,诸如对查询的响应、代表用户执行一个或多个动作的确认等。此外,这些自动化助理可以另外或替代地使视觉内容被提供用于向用户进行视觉呈现,诸如显示包括用户所请求的信息的信息卡、用于能够在实现这些自动化助理的客户端设备处访问的各种应用的各种界面等。虽然已经对在这些对话会话期间由这些自动化助理提供用于呈现给用户的可听内容进行了改进(例如,以更准确地反映人类之间的自然对话),但是对在这些对话会话期间由这些自动化助理提供用于视觉呈现的视觉内容的改进仍然相对停滞。

[0003] 例如,假设给定用户将“Good morning (早上好)”的口头话语指向在给定用户的给定客户端设备处实现的给定自动化助理。在该示例中,还假设该口头话语使给定自动化助理问候给定用户,向给定用户提供天气预报的概述,并向给定用户提供新闻标题,诸如通过“Good morning John, the weather today is 65 and sunny, here is your news [NEWS] (早上好,约翰,今天气温65度,天气晴朗,以下是您的新闻[新闻])”的合成语音的可听呈现。值得注意的是,给定自动化助理可以使给定客户端设备的显示器显示与系统语音相对应的文本内容、与天气预报相关联的信息卡、与新闻标题相关联的信息卡等等。然而,在该示例中,由给定自动化助理提供用于呈现给给定用户的视觉内容只是确认在可听内容中提供的信息。因此,被提供用于呈现给给定用户的视觉内容缺乏人类在交流时可能表达的视觉丰富性,诸如面部表情、身体语言、动画等。

[0004] 这种缺乏人类表达的视觉丰富性的一种解决方案包括为这些自动化助理提供可视化表示(例如,化身(avatar))。然而,这些自动化助理的当前可视化表示在它们可以向用户传达的信息的视觉丰富性方面相对有限。继续以上示例,给定自动化助理的给定可视化表示可以在给定客户端设备的显示上持续存在,并且可采用“CONTEXT[GREETING] = ACTION[WAVE]”的硬编码规则来使给定自动化助理的给定可视化表示在说“早上好”时向给定用户挥手。然而,人类在交流时可能表达的视觉丰富性的范围实际上可能是无限的。因此,本领域需要旨在改进在这些对话会话期间由这些自动化助理提供用于呈现给用户的视觉内容的技术。

发明内容

[0005] 本文所述的各实现方式涉及使自动化助理能够基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配给定助理输出。给定助理输出可以包括例如对应文本内容流和对应视觉提示流。对应文本内容流可以包括响应于对应口头话语并且将被合成以向提供对应口头话语的用户进行可听呈现的文本内容。此外,对应视觉提示流可以包括用于响应于对应口头话语而控制客户端设备(例如,在其处实现自动化助理的实例)的显示和/或用于控制自动化助理的实例的可视化表示的指令。例如,对应视觉提示流可以包括使客户端设备的显示动态地适配的显示动画、由自动化助理的可视化表示执行的动画物理运动手势和/或以其他方式控制显示和/或自动化助理的实例的可视化表示的任何其他指令。给定角色可以由例如特定于给定角色并且用于生成对应文本内容流的给定词汇、特定于给定角色并且用于合成对应文本内容流以向用户进行可听呈现的给定韵律属性集合和/或包括特定于给定角色的一些视觉提示(例如,通常与自动化助理的实例的可视化表示相关联的动画物理运动手势)并且包括在多个不同的角色中的多个角色中共用的一些视觉提示(例如,挥手、某些面部表情等)的给定视觉提示集合体现。自动化助理的可视化表示可以是例如表示自动化助理的实例的动画化身或实体,并且可以基于例如真实人类、虚构角色、动画对象和/或动物和/或其他可视化表示。

[0006] 各实现方式可以接收捕获用户的口头话语的音频数据流,并且基于处理音频数据流生成响应于口头话语的给定助理输出。给定助理输出可以包括文本内容流和视觉提示流,该视觉提示流用于响应于口头话语控制客户端设备的显示和/或用于控制被视觉地渲染以经由客户端设备的显示呈现给用户的自动化助理的实例的可视化表示。各实现方式可以使捕获与文本内容流相对应的合成语音的合成语音音频数据被可听地渲染以经由客户端设备的扬声器呈现给用户,并且可以使视觉提示流用于控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示。值得注意的是,响应于口头话语的给定助理输出可以特定于用户指派给客户端设备的实例的给定角色。

[0007] 在那些实现方式的一些版本中,各实现方式可以使用自动语音辨识(ASR)模型处理捕获口头话语的音频数据流以生成ASR输出流。此外,各实现方式可以使用自然语言理解(NLU)模型处理ASR输出流以生成NLU输出流。此外,各实现方式可以至少基于NLU输出流确定响应于口头话语的给定助理输出。值得注意的是,这些实现方式中的给定助理输出是使用典型的自动化助理流水线生成的,并且可以不包括视觉提示流,或者可以仅包括非常基本的视觉提示流,诸如要提供用于向用户进行视觉呈现的静态图形或信息卡。因此,各实现方式可以修改给定助理输出以生成修改后的给定助理输出,该修改后的给定助理输出针对指派给自动化助理的给定角色定制给定助理输出。例如,各实现方式可以使用特定于给定角色和/或利用特定于给定角色的给定角色数据的大语言模型(LLM) (例如,一个或多个transformer模型,诸如Meena、RNN和/或任何其他LLM)来处理给定助理输出,以生成修改后的给定助理输出。而且,例如,各实现方式可以确定先前基于由用户提供的相同或相似的口头话语生成的先前生成的LLM输出,以生成修改后的给定助理输出。

[0008] 如本文所用,以离线方式利用的先前生成的(例如,在接收口头话语之前生成的)LLM输出和/或以在线方式生成的(例如,响应于接收到口头话语生成的) LLM输出可以包括一个或多个概率分布。例如,在确定如本文所述的文本内容流时,这些LLM输出可以包括一

个或多个词汇中的一个或多个单词和/或短语的序列上的对应概率分布。文本内容流可以基于包括在概率分布中的概率从一个或多个单词和/或短语中选择。在一些实现方式中,一个或多个词汇可以包括特定于指派给自动化助理的给定角色的词汇。在附加或替代实现方式中,一个或多个词汇可以包括通用词汇,但是用于包括在文本内容流中的文本内容的选择可以偏向特定于给定角色的一个或多个单词和/或短语。

[0009] 而且,例如,在确定如本文所述的视觉提示流时,这些LLM输出可以包括表示由自动化助理的实例的可视化表示可执行的一个或多个动画物理运动手势和/或可以由客户端设备的显示实现的一个或多个显示动画的词元(token)序列上的对应概率分布。视觉提示流可以基于包括在概率分布中并且针对词元序列的概率从一个或多个动画物理运动手势和/或一个或多个显示动画中选择。在一些实现方式中,一个或多个动画物理运动手势和/或一个或多个显示动画可以包括特定于指派给自动化助理的给定角色的动画物理运动手势。在附加或替代实现方式中,一个或多个动画物理运动手势可以包括一般动画物理运动手势,但是用于包括在文本提示流中的动画物理运动手势的选择可以偏向特定于给定角色的一个或多个动画物理运动手势。

[0010] 在那些实现方式的附加或替代版本中,各实现方式可以使用LLM处理捕获口头话语的音频数据流、口头话语的ASR输出流、口头话语的NLU输出流和/或在其中接收到口头话语的对话会话的场境(context)(如果有的话),以生成给定助理输出。在这些实现方式中,各实现方式随后可以不修改给定助理输出,因为它可以通过利用LLM特定于给定角色生成。例如,LLM的实例可能已被预先训练为生成给定角色的给定助理输出,使得多个不同的角色中的每个角色可以与LLM的对应实例相关联。而且,例如,LLM对于能够指派给自动化助理的实例的多个角色可能是通用的,但LLM可以在生成给定助理输出时另外处理特定于给定角色的给定角色数据,诸如给定角色词元、给定角色嵌入、给定角色向量和/或可以用于定制使用对于多个角色通用的LLM生成的给定助理响应的其他数据。

[0011] 在各种实现方式中,各实现方式可以使与文本内容流相对应的合成语音的可听渲染和视觉提示流在控制客户端设备的显示和/或在控制自动化助理的实例的可视化表示中的利用同步,以呈现给用户。在那些实现方式的一些版本中,诸如当随后修改给定助理输出时,修改后的文本内容流可以用指示视觉提示流何时用于控制显示的一个或多个对应视觉提示时间戳来注释。例如,一个或多个对应视觉提示时间戳可以包括指示何时应该开始利用包括在视觉提示流中的一个或多个视觉提示的对应开始视觉提示时间戳、指示何时应该暂停利用包括在视觉提示流中的一个或多个视觉提示的对应暂停视觉提示时间戳、指示何时应该恢复利用包括在视觉提示流中的一个或多个视觉提示的对应恢复视觉提示时间戳、指示何时应该停止利用包括在视觉提示流中的一个或多个视觉提示的对应结束视觉提示时间戳和/或其他视觉提示。

[0012] 通过使用本文所述的技术,可以实现一个或多个技术优点。作为一个非限制性示例,本文所述的技术使得自动化助理不仅能够提供被合成以呈现给用户的更稳健且场境相关的文本内容流,而且能够提供用于控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示的稳健且场境相关的视觉提示流。因此,通过利用本文所述的LLM,用户与自动化助理之间的对话会话可以更好地与用户共鸣。因此,可以减少用户重复口头话语的实例的数量和/或对话会话失败的实例的数量,从而减少在用户重复口头话语和/或对话

会话失败时消耗的计算和/或网络资源的数量。

[0013] 如本文所用,“对话会话”可以包括用户与自动化助理(并且在一些情况下,为其他人类参与者)之间的逻辑上独立的交流。自动化助理可以基于各种信号(诸如会话之间的时间流逝、会话之间的用户场境(例如,位置、安排会议之前/期间/之后等)的改变、用户与客户端设备之间除用户与自动化助理之间的对话之外的一个或多个干预交互的检测(例如,用户切换应用一会儿、用户走开然后返回到独立的语音激活产品)、会话之间的客户端设备的锁定/睡眠、用于与自动化助理接口的客户端设备的改变等)来区分与用户的多个对话会话。值得注意的是,在给定对话会话期间,用户可以使用各种输入模态(包括但不限于口头输入、键入输入和/或触摸输入)与自动化助理交互。

[0014] 以上描述仅仅是为了举例说明而作为本文所公开的一些实现方式的概述来提供。在本文中更详细地描述了那些实现方式和其他实现方式。

[0015] 应当理解,本文所公开的技术可以在客户端设备上本地实现,由经由一个或多个网络连接到客户端设备的服务器远程实现,或者两者兼而有之。

附图说明

[0016] 图1描绘了展示本公开的各个方面并且在其中可以实现本文所公开的实现方式的示例环境的框图。

[0017] 图2描绘了示出根据各种实现方式的基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配给定助理输出的示例方法的流程图。

[0018] 图3描绘了示出根据各种实现方式的基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配给定助理输出的另一示例方法的流程图。

[0019] 图4描绘了示出根据各种实现方式的训练大语言模型以用于基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配指派给该自动化助理的给定助理输出的示例方法的流程图。

[0020] 图5描绘了示出根据各种实现方式的生成用于训练用于基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配指派给该自动化助理的给定助理输出的大语言模型的角色训练实例的示例方法的流程图。

[0021] 图6A和图6B示出了根据各种实现方式的基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配在其处实现该自动化助理的客户端设备的显示的非限制性示例。

[0022] 图7A和图7B示出了根据各种实现方式的基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配该自动化助理的可视化表示的非限制性示例。

[0023] 图8描绘了根据各种实现方式的计算设备的示例架构。

具体实施方式

[0024] 现在转到图1,描绘了展示本公开的各个方面并且在其中可以实现本文所公开的实现方式的示例环境100的框图。示例环境100包括客户端设备110和角色系统120。在一些实现方式中,角色系统120可以在客户端设备110处本地实现。在附加或替代实现方式中,角色系统120可以从如图1所描绘的客户端设备110远程地(例如,在远程服务器处)实现。在这

些实现方式中,客户端设备110和角色系统120可以经由一个或多个网络199 (诸如一个或多个有线或无线局域网(“LAN”,包括Wi-Fi LAN、网状网络、蓝牙、近场通信等)或广域网(“WAN”,包括互连网))彼此通信地耦合。

[0025] 客户端设备110可以是例如以下中的一个或多个:台式计算机、膝上型计算机、平板电脑、移动电话、车辆的计算设备(例如,车载通信系统、车载娱乐系统、车载导航系统)、独立交互式扬声器(可选地具有显示)、智能家电(诸如智能电视)和/或用户的包括计算设备的可穿戴装置(例如,用户的具有计算设备的手表、用户的具有计算设备的眼镜、虚拟或增强现实计算设备)。可以提供附加和/或替代客户端设备。

[0026] 客户端设备110可以执行自动化助理客户端114。自动化助理客户端114的实例可以是与客户端设备110的操作系统分离(例如,安装在操作系统“之上”)的应用,或者可以替代地直接由客户端设备110的操作系统实现。自动化助理客户端114可以与在客户端设备110处本地实现或远程实现的并且经由如图1所描绘的网络199中的一个或多个网络调用的角色系统120交互。自动化助理客户端114 (并且可选地通过其与其他远程系统(例如,服务器)的交互)可以形成从用户的角度看来似乎是自动化助理115的逻辑实例的内容,用户可以在对话会话期间与该自动化助理进行人机对话。自动化助理115的实例在图1中描绘,并且由包括客户端设备110的自动化助理客户端114和自然对话系统120的虚线包围。因此,应当理解,与在客户端设备110上执行的自动化助理客户端114互动的用户实际上可以与他或她自己的自动化助理115的逻辑实例(或在家庭或其他用户组之间共享的自动化助理115的逻辑实例)互动。为了简洁和简单起见,如本文所用的自动化助理115将指在客户端设备110上本地执行和/或在可以实现角色系统120的一个或多个远程服务器处远程执行的自动化助理客户端114。

[0027] 在各种实现方式中,客户端设备110可以包括用户输入引擎111,该用户输入引擎被配置为检测由客户端设备110的用户使用一个或多个用户界面输入设备提供的用户输入。例如,客户端设备110可以配备有生成音频数据流(诸如捕获用户的口头话语和/或客户端设备110的环境中的其他声音的音频数据流)的一个或多个麦克风。另外或替代地,客户端设备110可以配备有一个或多个视觉组件,这些视觉组件被配置为生成捕获在视觉组件中的一个或多个视觉组件的视场中检测到的图像、视频和/或某些移动(例如,手势)的视觉数据流。另外或替代地,客户端设备110可以配备有一个或多个触敏组件(例如,键盘和鼠标、触控笔、触摸屏、触摸面板、一个或多个硬件按钮等),这些触敏组件被配置为生成与指向客户端设备110的触摸输入相对应的信号(例如,在客户端设备110包括触摸屏显示的实现方式中)。

[0028] 在各种实现方式中,客户端设备110可以包括渲染引擎112,该渲染引擎被配置为使用一个或多个用户界面输出设备提供用于向客户端设备110的用户进行可听和/或视觉呈现的内容。例如,客户端设备110可以配备有一个或多个扬声器,这些扬声器使得能够提供可听内容以经由客户端设备110向用户进行可听呈现。另外或替代地,客户端设备110可以配备有显示器或投影仪,该显示器或投影仪使得能够提供视觉内容以经由客户端设备110向用户进行视觉呈现。

[0029] 在各种实现方式中,客户端设备110可以包括一个或多个存在传感器113,这些存在传感器被配置为在对应用户的批准下提供指示检测到的存在、特别是人类存在的信号。

在那些实现方式中的一些实现方式中,自动化助理115可以至少部分地基于用户在客户端设备110处(或在与客户端设备110的用户相关联的另一计算设备处)的存在来识别客户端设备110 (或与客户端设备110的用户相关联的另一计算设备)以满足口头话语。可以如下满足口头话语:通过在客户端设备110和/或与客户端设备110的用户相关联的其他计算设备处渲染给定助理输出(例如,经由渲染引擎112)、通过使客户端设备110和/或与客户端设备110的用户相关联的其他计算设备被控制和/或通过使客户端设备110和/或与客户端设备110的用户相关联的其他计算设备执行任何其他动作以满足口头话语。如本文所述,自动化助理115可以利用基于存在传感器113确定的数据而基于用户在哪里或最近在哪里来确定客户端设备110 (或其他计算设备),并且仅向客户端设备110 (或那些其他计算设备)提供对应命令。在一些附加或替代实现方式中,自动化助理115可以利用基于存在传感器113确定的数据来确定任何用户(任何用户或特定用户)当前是否接近客户端设备110 (或其他计算设备),并且可以可选地基于接近客户端设备110 (或其他计算设备)的用户来抑制向和/或从客户端设备110 (或其他计算设备)提供数据。

[0030] 存在传感器113可以有各种形式。例如,客户端设备110可以利用上文关于用户输入引擎111描述的用户界面输入组件中的一个或多个用户界面输入组件来检测用户的存在。另外或替代地,客户端设备110可以配备有其他类型的基于光的存在传感器113,诸如测量从其视场内的物体辐射的红外(“IR”)光的被动红外(“PIR”)传感器。

[0031] 另外或替代地,在一些实现方式中,存在传感器113可以被配置为检测与人类存在或设备存在相关联的其他现象。例如,在一些实施方式中,客户端设备110可以配备有存在传感器113,该存在传感器检测例如由用户携带/操作的其他计算设备(例如,移动设备、可穿戴计算设备等)和/或其他计算设备发射的各种类型的无线信号(例如,波,诸如无线电波、超声波、电磁波等)。例如,客户端设备110可以被配置为发射人类无法感知的波,诸如超声波或红外波,这些波可由其他计算设备检测(例如,经由超声/红外接收器,诸如具有超声能力的麦克风)。

[0032] 另外或替代地,客户端设备110可以发射其他类型的人类无法感知的波,诸如无线电波(例如,Wi-Fi、蓝牙、蜂窝等),这些波可由用户携带/操作的其他计算设备(例如,移动设备、可穿戴计算设备等)检测并且用于确定用户的特定位置。在一些实现方式中,可以使用GPS和/或Wi-Fi三角测量例如基于去往/来自客户端设备110的GPS和/或Wi-Fi信号来检测人的位置。在其他实现方式中,客户端设备110可以单独地或共同地使用其他无线信号特性(诸如飞行时间、信号强度等的)基于由用户携带/操作的其他计算设备发射的信号来确定特定用户的位置。另外或替代地,在一些实现方式中,客户端设备110可以执行说话人识别(SID)以根据其语音来辨识用户和/或可以执行面部识别(FID)以根据捕获用户的面部的视觉数据来辨识用户。

[0033] 在一些实现方式中,然后可以例如由客户端设备110的存在传感器113 (以及可选地,GPS传感器、Soli芯片和/或客户端设备110的加速度计)来确定说话人的移动。在一些实现方式中,基于这种检测到的移动,可以预测用户的位置,并且当至少部分地基于客户端设备110和/或其他计算设备与用户的位置的接近度使任何内容在客户端设备110和/或其他计算设备处被渲染时,可以假设该位置是用户的位置。在一些实现方式中,可以简单地假设用户处于他或她与自动化助理115互动的最后位置,特别是如果自上次互动以来尚未过去

太多时间。

[0034] 此外,客户端设备110和/或角色系统120可以包括用于存储数据和/或软件应用的一个或多个存储器、用于访问数据并执行软件应用的一个或多个处理器和/或促进通过网络199中的一个或多个网络进行通信的其他组件。在一些实现方式中,软件应用中的一个或多个软件应用可以本地安装在客户端设备110处,而在其他实现方式中,软件应用中的一个或多个软件应用可以(例如,由一个或多个服务器)远程托管并且可以由客户端设备110通过网络199中的一个或多个网络进行访问。

[0035] 在一些实现方式中,由自动化助理115执行的操作可以经由自动化助理客户端114在客户端设备110处本地实现。如图1所示,自动化助理客户端114可以包括自动语音辨识(ASR)引擎130A1、自然语言理解(NLU)引擎140A1、履行(LLM)引擎150A1和文本转语音(TTS)引擎160A1。在一些实现方式中,由自动化助理115执行的操作可以分布在多个计算机系统上,诸如当角色系统120从如图1所描绘的客户端设备110远程实现时。在这些实现方式中,自动化助理115可以另外或替代地利用角色系统120的ASR引擎130A2、NLU引擎140A2、履行引擎150A2和TTS引擎160A2。

[0036] 这些引擎中的每个引擎可以被配置为执行一个或多个功能。例如,ASR引擎130A1和/或130A2可以使用存储在机器学习(ML)模型数据库115A中的流式ASR模型(例如,循环神经网络(RNN)模型、transformer模型和/或能够执行ASR的任何其他类型的ML模型)处理捕获口头话语并且由客户端设备110的麦克风生成的音频数据流,以生成对应ASR输出流。值得注意的是,流式ASR模型可以用于在生成音频数据流时生成对应ASR输出流。此外,NLU引擎140A1和/或140A2可以使用存储在ML模型数据库115A中的NLU模型(例如,长短期记忆(LSTM)、门控循环单元(GRU)和/或任何其他类型的RNN或能够执行NLU的其他ML模型)和/或基于语法的规则来处理对应ASR输出流,以生成对应NLU输出流。此外,履行引擎150A1和/或150A2可以使对应NLU输出流被处理以生成对应履行数据流。例如,自动化助理115可以生成一个或多个对应结构化请求并通过网络199中的一个或多个网络(或一个或多个应用编程接口(API))向一个或多个第一方(1P)系统191传输和/或通过网络中的一个或多个网络向一个或多个第三方(3P)系统192传输,并且从1P系统191和/或3P系统192中的一者或多者接收对应履行数据,以生成对应履行数据流。如本文所用,一个或多个1P系统191是指由开发和/或维护自动化助理115的同一实体开发和/或维护的任何系统,而一个或多个3P系统是指由与开发和/或维护自动化助理115的实体不同的实体开发和/或维护的任何系统。

[0037] 一个或多个对应结构化请求可以包括例如包括在对应NLU输出流中的NLU数据。对应履行数据流可以对应于例如被预测为响应于在由ASR引擎130A1和/或130A2处理的对应音频数据流中捕获的口头话语的对应给定助理输出。最后,TTS引擎160A1和/或160A2可以使用存储在ML模型数据库115A中的TTS模型来处理对应文本内容流(例如,由自动化助理115制定的文本),以生成包括计算机生成的合成语音的合成语音音频数据。对应文本内容流可以对应于例如一个或多个给定助理输出、修改后的给定助理输出中的一个或多个给定助理输出和/或本文描述的任何其他文本内容。

[0038] 值得注意的是,存储在ML模型数据库115A中的ML模型可以是本地存储在客户端设备110处的设备上ML模型、从客户端设备远程地(例如,在远程服务器处)执行的远程ML模型或者客户端设备110和/或远程系统(例如,远程服务器)都能够访问的共享ML模型。在附加

或替代实现方式中,与一个或多个给定助理输出、修改后的给定助理输出中的一个或多个给定助理输出和/或本文描述的任何其他文本内容相对应的对应合成语音音频数据流可以预缓存在存储器或者能够由客户端设备110访问的一个或多个数据库中,使得自动化助理不需要使用TTS引擎160A1和/或160A2来生成对应合成语音音频数据。

[0039] 在各种实现方式中,对应ASR输出流可以包括例如被预测为与用户的在对应音频数据流中捕获的口头话语相对应的ASR假设流(例如,词项假设和/或转录假设)、包括在ASR假设流中的ASR假设中的每个ASR假设的一个或多个对应预测度量(例如,概率、对数似然和/或其他值)、被预测为与用户的在对应音频数据流中捕获的口头话语相对应的多个音素和/或其他ASR输出。在那些实现方式的一些版本中,ASR引擎130A1和/或130A2可以选择ASR假设中的一个或多个ASR假设作为与口头话语相对应的对应辨识文本(例如,基于对应预测度量选择的)。

[0040] 在各种实现方式中,对应NLU输出流可以包括例如带注释的辨识文本流,这些带注释的辨识文本流包括针对辨识文本的词项中的一个或多个(例如,全部)词项的辨识文本的一个或多个注释、包括在NLU输出流中的NLU输出的一个或多个对应预测度量(例如,概率、对数似然和/或其他值)和/或其他NLU输出。例如,NLU引擎140A1和/或140A2可以包括词性标注器(未描绘),该词性标注器被配置为利用其语法角色来注释词项。另外或替代地,NLU引擎140A1和/或140A2可以包括实体标注器(未描绘),该实体标注器被配置为注释辨识文本的一个或多个片段中的实体引用,诸如对人(包括例如文学角色、名人、公众人士等)、组织、位置(真实的和虚构的)等的引用。在一些实现方式中,关于实体的数据可以存储在一个或多个数据库中,诸如存储在知识图谱(未描绘)中。在一些实现方式中,知识图谱可以包括表示已知实体(并且在一些情况下,为实体属性)的节点,以及连接节点并且表示实体之间的关系的边。实体标注器可以按高粒度级别注释对实体的引用(例如,以使得能够识别对实体类别(诸如人)的所有引用)和/或按较低粒度级别注释对实体的引用(例如,以使得能够识别对特定实体(诸如特定人员)的所有引用)。实体标注器可以依赖于自然语言输入的内容来解析特定实体和/或可以可选地与知识图谱或其他实体数据库通信来解析特定实体。

[0041] 另外或替代地,NLU引擎140A1和/或140A2可以包括共指解析器(未描绘),该共指解析器被配置为基于一个或多个场境提示来对同一实体的引用进行分组或“聚类”。例如,可以使用共指解析器基于紧接在接收到自然语言输入“buy them (购买它们)”之前渲染的客户端设备通知中提及“theatre tickets (剧院门票)”来将输入“buy them (购买它们)”中的词项“them (它们)”解析为“buy theatre tickets (购买剧院门票)”。在一些实现方式中,NLU引擎140A1和/或140A2的一个或多个组件可以依赖于来自NLU引擎140A1和/或140A2的一个或多个其他组件的注释。例如,在一些实现方式中,实体标注器可以依赖于来自共指解析器的注释来注释对特定实体的所有提及。而且,例如,在一些实现方式中,共指解析器可依赖于来自实体标注器的注释来将对同一实体的引用进行聚类。

[0042] 尽管图1是相对于具有单个用户的单个客户端设备进行描述的,但应当理解,这是为了举例说明,并不意味着限制。例如,用户的一个或多个附加客户端设备也可实现本文所述的技术。例如,客户端设备110、一个或多个附加客户端设备和/或用户的任何其他计算设备可形成可采用本文所述的技术的设备生态系统。这些附加客户端设备和/或计算设备可与客户端设备110进行通信(例如,通过网络199)。作为另一示例,给定客户端设备可以由共

享设置中的多个用户(例如,用户组、家庭)使用。例如,共同位于家庭中的多个用户可以各自利用给定客户端设备,并且多个用户中的每个用户可以具有对多个用户中的每个用户私人的单独的自动化助理帐户。在这些示例中,自动化助理115可以利用本文所述的一种或多种用户识别技术来从多个用户中识别给定用户,并且相应地通过利用给定用户指派给自动化助理115的给定角色(例如,其可以不同于由多个用户中的另一用户指派给自动化助理115的另一角色)、利用特定于给定用户的信息(例如,给定用户的日历信息、来自给定用户的移动设备的移动设备信息(例如,指向给定用户的传入电子通信)等)和/或其他信息来为给定用户定制对话会话而定制任何对话会话。

[0043] 如本文所述,自动化助理115可以利用角色系统120来生成特定于多个不同的角色中指派给自动化助理115的给定角色的给定助理输出。给定助理输出中的每个给定助理输出可以包括例如基于特定于给定角色的词汇确定并且使用特定于给定角色的韵律属性集合(例如,使用TTS引擎160A1和/或160A2)合成的对应文本内容流,以及用于控制客户端设备110的显示(例如,如关于图6A和图6B所述)和/或用于控制与给定角色相关联的自动化助理115的实例的对应可视化表示的对应视觉提示流。能够指派给自动化助理的给定角色可以由特定于给定角色并且用于生成对应文本内容流的词汇、特定于给定角色的该韵律属性集合、对应视觉提示流和/或与给定角色相关联的自动化助理115的实例的对应可视化表示体现。因此,对于不同于给定角色的附加角色,这些中的一个或多个可以不同,以便以不同于给定角色的方式体现附加角色。

[0044] 在各种实现方式中,并且如图1所描绘,角色系统120可以另外或替代地包括角色训练引擎170和角色推断引擎180。角色训练引擎170可以包括例如训练实例引擎171和训练引擎172。此外,角色推断引擎180可以包括例如用户识别引擎181、LLM引擎182、输出修改引擎183、排序引擎184和同步引擎185。关于图2至图5更详细地描述了角色系统120的这些各种引擎。尽管在图1中描绘了特定引擎,但是应当理解,这是为了举例说明,并不意味着限制。例如,图1中描绘的各种引擎可以在各种实现方式中组合和/或省略。作为一个非限制性示例,LLM引擎182、输出修改引擎183、排序引擎184和/或同步引擎185中的一者或多者可以在其中特定于给定角色的给定LLM用于生成特定于给定角色的给定助理输出的实现方式中组合。作为另一非限制性示例,用户识别引擎181可以在其中客户端设备110对于用户(例如,用户的移动设备)私人的实现方式中省略。

[0045] 在一些实现方式中,角色训练实例引擎171可以基于可指派给自动化助理115的多个不同的角色的对话数据(例如,存储在对话数据数据库170A中)生成给定角色训练实例,并且将给定角色训练实例存储在角色训练引擎170能够访问的一个或多个数据库(例如,训练实例数据库170B)中。对话数据可以包括例如可以用于给定角色训练实例的任何数据,诸如捕获对应口头话语的对应音频数据流、基于处理对应音频数据流生成的对应ASR输出流、基于处理对应ASR输出流生成的对应NLU输出流、在其中接收到对应口头话语的对应对话会话的对应场境、捕获提供对应口头话语的人类或角色的对应视觉数据流、注释对话数据的对应开发者输入和/或可以用于给定角色训练实例的任何其他数据。本文更详细地(例如,关于图5)描述了基于该对话数据生成给定角色训练实例。

[0046] 在这些实现方式中,训练引擎172可以训练给定LLM (例如,存储在ML模型数据库115A中)的特定于多个不同的角色中的一个或多个角色的实例。例如,给定角色训练实例可

以由多个不同的角色中的每个角色索引(例如,在训练实例数据库170B中),使得多个不同的角色中的每个角色与对应给定角色训练实例集合相关联。因此,训练引擎172可以(例如,从训练实例数据库170B)获得用于给定角色的对应给定角色训练实例以训练给定LLM的特定于给定角色的实例,(例如,从训练实例数据库170B)获得用于附加角色的对应给定角色训练实例以训练给定LLM的特定于附加角色的附加实例,等等。本文更详细地(例如,关于图4)描述了训练给定LLM的特定于给定角色的实例。因此,在这些实现方式中,给定LLM的实例可以随后用于基于给定LLM的实例是专门针对用户已经指派给自动化助理115的实例的给定角色训练的而生成特定于该给定角色的给定助理响应。

[0047] 在附加或替代实现方式中,对于不同角色中的多个角色(例如,不同角色中的所有角色或其子集)通用的给定LLM可以用于生成给定助理响应。在这些实现方式中,给定LLM可以另外处理指派给自动化助理115的给定角色的给定角色数据(例如,存储在角色数据数据库170C中)。在这些实现方式中,给定角色数据可以对应于例如给定角色词元、给定角色嵌入、给定角色向量和/或可以用于体现特定于自动化助理115的实例的给定角色的其他数据。在这些实现方式中,给定角色数据可以由与自动化助理115和/或角色系统相关联的开发者定义和/或使用各种机器学习技术进行学习。因此,在这些实现方式中,给定LLM可以用于基于特定于用户已经指派给自动化助理115的实例的给定角色的给定角色数据而生成特定于该给定角色的给定助理响应。

[0048] 在一些实现方式中,用户识别引擎181可以用于确定提供在音频数据流中捕获的口头话语的用户的身份。用户识别引擎181可以基于处理音频数据流(例如,使用说话人识别(SID)模型)、处理捕获提供口头话语的用户的视觉数据流(例如,使用面部识别(FID)模型)、基于与在客户端设备110处活动的自动化助理相关联的自动化助理帐户和/或通过使用其他技术的其他方式来确定提供在音频数据流中捕获的口头话语的用户的身份。本文更详细地(例如,关于图2和图3)描述了识别提供在音频数据流中捕获的口头话语的用户。

[0049] 在一些实现方式中,LLM引擎182可以利用给定LLM的实例来生成要提供以直接呈现给用户的给定助理输出(例如,如关于图3所述)。在附加或替代实现方式中,LLM引擎182和/或输出修改引擎183可以随后利用给定LLM的实例来修改给定助理输出以生成要提供以直接呈现给用户的给定助理输出(例如,如关于图2所述)。值得注意的是,在这些实现方式中,LLM引擎182可以响应于接收到捕获口头话语的音频数据流而主动地处理数据。因此,这些实现方式可以以在线方式利用LLM。在附加或替代实现方式中,LLM引擎182可以利用先前基于在音频数据流中捕获的相同或相似的口头话语而生成(例如,如关于图2所述)的先前生成的LLM输出(例如,存储在LLM输出数据库180中)。

[0050] 在一些实现方式中,排序引擎184可以对由LLM引擎182和/或输出修改引擎183生成的多个助理输出进行排序,并且基于排序选择要提供以呈现给用户的给定助理输出。排序引擎184可以基于各种排序标准对多个助理输出进行排序。排序标准可以包括例如指示多个助理输出中的每个助理输出被预测为响应于口头话语的响应程度的一个或多个预测度量(例如,在生成ASR输出流时生成的ASR预测度量、在生成NLU输出流时生成的NLU预测度量等)、包括在对应NLU输出流中的一个或多个意图和/或其他排序标准。

[0051] 在一些实现方式中,同步引擎185可以用于同步给定助理输出,以呈现给客户端设备110的用户。例如,给定助理输出的将用于控制客户端设备110的显示和/或自动化助理

115的可视化表示的视觉提示流可以相对于给定助理输出的要被合成以向客户端设备110的用户进行可听呈现的文本内容流而定义。例如,同步可以用对应视觉提示时间戳注释文本内容流,这些视觉提示时间戳指示视觉提示何时应当用于控制客户端设备110的显示和/或自动化助理115的可视化表示。本文更详细地(例如,关于图2、图3、图6A、图6B、图7A和图7B)描述了同步给定助理输出以呈现给用户。

[0052] 现在转到图2,描绘了示出基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配给定助理输出的示例方法200的流程图。为了方便起见,方法200的操作参考执行这些操作的系统来描述。方法200的该系统包括一个或多个处理器、存储器和/或计算设备(例如,图1、图6A、图6B、图7A和图7B的客户端设备110、角色系统120和/或图8的计算设备810、一个或多个服务器和/或其他计算设备)的其他组件。此外,虽然方法200的操作以特定顺序示出,但这并不意味着限制。一个或多个操作可以被重新排序、省略和/或添加。

[0053] 在框252处,系统接收捕获客户端设备的用户的口头话语的音频数据流,该口头话语指向至少部分地在客户端设备处执行的自动化助理的实例。音频数据流可以例如经由客户端设备的一个或多个麦克风生成。在一些实现方式中,系统可以响应于自动化助理的实例被显式地调用(诸如基于在客户端设备处检测到特定单词或短语(例如,“Assistant (助理)”、“Hey Assistant (嘿,助理)”等)、基于在客户端设备处致动按钮(例如,客户端设备的硬件按钮、客户端设备的显示器的软件按钮)、基于在客户端设备处检测到特定手势和/或使用其他调用技术)来接收音频数据流。在其他实现方式中,系统可以在自动化助理的实例没有被显式地调用的情况下(诸如基于在用户注视客户端设备时接收到音频数据流和/或使用其他技术)来接收音频数据流。

[0054] 在各种实现方式中,系统可以实现一种或多种用户识别技术以识别提供口头话语的客户端设备的用户。例如,系统可以使用说话人识别(SID)模型(例如,文本相关(TD) SID模型和/或文本不相关(TI) SID模型)来处理音频数据流,以基于先前生成的说话人嵌入来识别客户端设备的用户。另外或替代地,系统可以处理由客户端设备的一个或多个视觉传感器生成的视觉数据流(例如,紧接在接收到音频数据流之前、在接收到音频数据流的同时和/或紧接在接收到音频数据流之后),使用面部识别(FID)模型而基于先前生成的面部嵌入来识别用户。另外或替代地,系统可以基于在客户端设备处活动的用户帐户(例如,与自动化助理的实例相关联的用户帐户)来识别用户。

[0055] 在框254处,系统基于处理音频数据流生成响应于口头话语的给定助理输出,该给定助理输出包括:(1)文本内容流;以及(2)视觉提示流。例如,系统可以使用ASR模型处理音频数据流,以生成ASR输出流,诸如与在音频数据流中捕获的口头话语相对应的一个或多个辨识词项。此外,系统可以使用NLU模型处理ASR输出流,以生成NLU输出流,诸如一个或多个意图以及与这些意图中的一个或多个意图相关联的一个或多个参数的对应槽值。

[0056] 在一些实现方式中,系统可以至少基于NLU输出流生成要向一个或多个1P系统和/或一个或多个3P系统传输的一个或多个结构化请求。响应于向一个或多个1P系统和/或一个或多个3P系统传输一个或多个结构化请求,系统可以从一个或多个1P系统和/或一个或多个3P系统接收响应内容。响应内容可以用于生成包括在给定助理输出中的文本内容流和视觉提示流。例如,如果口头话语对应于“Assistant,what’s the weather (助理,天气怎么样)”,则可以处理捕获口头话语的音频数据流以获得将被合成以向用户进行可听呈现的

文本内容流(例如,“The weather today is rainy and 45 (今天有雨,气温45度)”)和视觉提示流(例如,与天气相关联的信息卡)。在那些实现方式的一些版本中,系统可以利用基于响应生成的文本内容流和视觉提示流作为给定助理输出。在那些实现方式的附加或替代版本中,系统还可以使用LLM (例如,以如本文所述的在线方式)或先前基于相同的口头话语或相似的口头话语生成的先前生成的LLM输出(例如,以如本文所述的离线方式)处理基于来自一个或多个1P系统和/或一个或多个3P系统的响应内容和/或来自一个或多个1P系统和/或一个或多个3P系统的响应内容生成的文本内容流和视觉提示流,以生成用作给定助理输出的文本内容流和视觉提示流(例如,“The weather today is rainy and 45, don't forget your umbrella or you might catch a cold (今天有雨,气温45度,别忘了带伞,不然你可能会感冒)”)。换句话说,在这些实现方式中,用作给定助理输出的文本内容流和视觉提示流可以与使用不包括任何LLM的典型自动化助理流水线生成的文本内容和视觉内容相对应,或者可以与使用典型自动化助理流水线生成但使用LLM或先前生成的LLM输出增强的文本内容和视觉内容相对应。

[0057] 在附加或替代实现方式中,系统可以使用LLM (例如,以如本文所述的在线方式)或先前生成的LLM输出(例如,以如本文所述的离线方式)处理音频数据流和/或NLU输出流,并且不生成任何结构化请求并向一个或多个1P系统和/或一个或多个3P系统传输。换句话说,在这些实现方式中,用作给定助理输出的文本内容流和视觉提示流可以基于音频数据流和/或NLU输出流直接生成。

[0058] 在框256处,系统基于由用户从多个不同的角色中指派给自动化助理的实例的给定助理角色修改给定助理输出以生成修改后的给定助理输出,该修改后的给定助理输出包括:(1)与文本内容流不同的修改后的文本内容流;以及/或者(2)与视觉提示流不同的修改后的视觉提示流。在生成修改后的给定助理输出时,系统还可以考虑用户在其中提供口头话语的对话会话的场境(如果有的话)。给定角色可以包括例如特定于给定角色并且用于修改文本内容流以生成修改后的文本内容流的给定词汇、特定于给定角色并且用于随后合成修改后的文本内容流以向用户进行可听呈现的给定韵律属性集合和/或特定于给定角色并且用于修改视觉提示流以生成修改后的视觉提示流的给定视觉提示集合。对话会话的场境可以基于一个或多个场境信号(包括例如一天中的时间、一周中的一天、客户端设备的位置、在客户端设备的环境中检测到的环境噪声、用户档案数据、软件应用数据、关于客户端设备的用户的已知环境的环境数据、用户与自动化助理之间的对话会话的对话历史和/或其他场境信号)来确定,并且可以以各种方式(例如,向量表示、语义词元表示和/或其他表示)来表示。值得注意的是,其他不同角色中的每个角色也可以与对应词汇、对应韵律属性集合和/或对应视觉提示集合相关联。

[0059] 在一些实现方式中,当用户最初配置客户端设备时,可以将给定角色指派给自动化助理的实例。例如,在设置客户端设备时,可能已经提示用户从多个不同的角色中选择要指派给自动化助理实例的给定角色。在附加或替代实现方式中,可以经由与自动化助理的实例相关联的自动化助理应用的设置将给定角色指派给自动化助理的实例。例如,用户可以导航到与自动化助理的自动化实例相关联的自动化助理应用的设置,并且能够从多个不同的角色中选择给定角色。在附加或替代实现方式中,可以基于包括在指向自动化助理的实例的口头话语中的语音命令将给定角色指派给自动化助理的实例。例如,用户可以提供

“talk to Walter (与沃尔特交谈)”(例如,其中“沃尔特”是对给定角色(例如,管家角色)的引用)或“pretend you are “Assistant, pretend you are Blackbeard (假装你是助理,假装你是黑胡子)” (例如,其中“黑胡子”是对给定角色(例如,海盗角色)的引用)的口头话语。在这些示例中,系统可以使用本文所述的各种组件(例如,ASR、NLU、履行等)来处理捕获口头话语的音频数据,以识别语音命令以将给定角色指派给自动化助理的实例。

[0060] 在一些实现方式中,可以将给定角色指派给客户端设备,使得与客户端设备相关联的多个不同用户与被指派给定角色的自动化助理的实例交互。在附加或替代实现方式中,可以将给定角色指派给客户端设备的用户,使得客户端设备的其他用户可以指派不同的角色以供自动化助理的实例在经由客户端设备与自动化助理交互时使用。在这些实现方式中,用户的身份(例如,如上文关于框252的操作所述确定的)可以用于确定将哪个角色指派给多个不同用户的自动化助理的实例。换句话说,可以将给定角色指派给客户端设备,使得当与客户端设备的每个用户交互时使用给定角色,或者可以将给定角色指派给客户端设备的特定用户(例如,提供在音频数据流中捕获的口头话语的用户),使得当与客户端设备的不同用户交互时使用不同的角色。值得注意的是,指派给自动化助理的实例的给定角色可以与如本文所述的可视化表示相关联,使得视觉提示流可以用于控制可视化表示和客户端设备的显示两者,而能够指派给自动化助理的其他角色可能缺乏可视化表示,使得视觉提示流仅可用于控制客户端设备的显示。

[0061] 在一些实现方式中,并且如框256A处所指示,修改后的给定助理输出可以使用LLM以在线方式生成(例如,响应于接收到口头话语并使LLM处理给定助理输出)。在这些实现方式中,系统可以使用一个或多个LLM以及可选地在其中提供口头话语的对话会话的场境来处理给定助理输出,以生成修改后的助理输出。例如,系统可以使图1的LLM引擎182使用一个或多个LLM来处理给定助理输出的文本内容流、给定助理输出的视觉提示流、在处理音频流(例如,ASR输出流、NLU输出流等)时生成的数据和/或用户在其中提供口头话语的对话会话的场境,以生成该组给定修改后的助理输出。换句话说,给定助理输出可以具有有限的词汇以及在用于合成语音的有限词汇和韵律属性方面的一般个性。然而,在生成修改后的给定助理输出时,系统将给定角色注入到自动化助理的实例中,从而向自动化助理的实例提供与给定助理输出相比作为在生成修改后的给定助理输出时使用一个或多个LLM的函数的特定于给定角色的大得多的词汇、在生成修改后的给定助理输出时与给定角色相关联的韵律属性方面的大得多的变化以及在生成修改后的给定助理输出时与给定角色相关联的稳健得多的交互式视觉提示。因此,修改后的给定助理输出可以更好地与参与与自动化助理的实例的对话会话的用户共鸣。

[0062] 在附加或替代实现方式中,并且如框256B处所指示,修改后的给定输出可以使用来自LLM的先前生成的LLM输出以离线方式生成(例如,响应于接收到口头话语,但不使LLM处理给定助理输出)。在这些实现方式中,系统可以确定在音频数据流中捕获的口头话语对应于先前已经针对其生成先前生成的LLM输出的先前口头话语,并且可选地确定用户在其中提供口头话语的对话会话的场境对应于先前口头话语的先前对话会话的先前场境。此外,系统可以获得先前生成的LLM输出,因为它可以基于先前口头话语和先前场境进行索引。类似地,这些技术使得自动化助理的实例具有与给定助理输出相比作为在生成修改后的给定助理输出时使用一个或多个LLM的函数的特定于给定角色的大得多的词汇、在生成

修改后的给定助理输出时与给定角色相关联的韵律属性方面的大得多的变化以及在生成修改后的给定助理输出时与给定角色相关联的稳健得多的交互式视觉提示。因此,修改后的给定助理输出可以更好地与参与与自动化助理的实例的对话会话的用户共鸣。

[0063] 在框258处,系统使捕获与文本内容流或修改后的文本内容流相对应的合成语音的合成语音音频数据被可听地渲染以呈现给用户。例如,系统可以使用TTS模型来处理文本内容流以生成捕获与文本内容流(例如,在文本内容流未被修改的情况下)或修改后的文本内容流(例如,在文本内容流被修改的情况下)相对应的合成语音的合成语音音频数据。值得注意的是,系统可以在生成合成语音音频数据时利用指派给给定角色的该组给定韵律属性来反映与给定角色相关联的给定音调、节奏、音高、语调和/或其他韵律属性。此外,系统可以使合成语音音频数据被可听地渲染以经由客户端设备的一个或多个扬声器呈现给用户。

[0064] 在框260处,系统使视觉提示流或修改后的视觉提示流用于控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示。在一些实现方式中,视觉提示流或修改后的视觉提示流用于控制客户端设备的显示。在这些实现方式中,视觉提示流或修改后的视觉提示流可以使客户端设备的显示在视觉上渲染一个或多个屏幕动画以呈现给用户(例如,如关于图6A和图6B所述)。在附加或替代实现方式中,视觉提示流或修改后的视觉提示流用于控制自动化助理的实例的可视化表示。自动化助理的实例的可视化表示可以是例如对应于动画人物(例如,真实的或虚构的)、角色(例如,管家、海盗、厨师)、对象(例如,动画助理点)、动物和/或任何其他可视化表示的化身。在这些实现方式中,视觉提示流或修改后的视觉提示流可以使自动化助理的可视化表示执行一个或多个动画物理手势动作。这些物理手势动作可以包括例如在所有可视化表示中通用的通用物理手势动作(例如,挥手、进入客户端设备的显示、退出客户端设备的显示等)、特定于指派给自动化助理的给定角色的角色特定的物理手势动作(例如,虚构角色的招牌动作)、可以可选地与上述通用或角色特定的物理手势动作耦合的由可视化表示描绘的情绪(例如,快乐、悲伤、愤怒等)、可以可选地与上述通用或角色特定的物理手势动作耦合的面部表情(例如,微笑、皱眉等)和/或其他物理手势动作。在这些实现方式中,视觉提示流或修改后的视觉提示流可以使自动化助理的实例的可视化表示在视觉上执行动画物理手势动作(例如,如关于图7A和图7B所述)。

[0065] 在各种实现方式中,系统可以使与视觉提示流或修改后的文本内容流相对应的合成语音的可听渲染和视觉提示流或修改后的视觉提示流在控制客户端设备的显示和/或在控制自动化助理的实例的可视化表示中的利用同步,以呈现给用户。值得注意的是,在生成给定助理输出和/或修改后的给定助理输出时利用LLM或LLM输出的实现方式中,该同步可以由以本文(例如,关于图4和图5)描述的方式训练的那些LLM (例如,经由图1的LLM引擎182)和/或由那些LLM生成的LLM输出自动执行。在其他实现方式中,该同步可以由专用同步引擎和/或系统的能够使视觉提示流或修改后的文本内容流和视觉提示流或修改后的视觉提示流的利用同步的其他组件(例如,经由图1的同步引擎185)执行。

[0066] 例如,假设用户提供“Hi Assistant (嗨,助理)”的口头话语,假设以上文关于框254和256的操作描述的方式处理捕获口头话语的音频,并且假设用户已经将与可视化表示相关联的给定角色指派给自动化助理的实例。在该示例中,进一步假设系统生成由<text_stream = “Hey there, how are you doing?”>的数据结构表示的文本内容流或修改后的

文本内容流,并且进一步假设系统生成由<visual_stream = hand lift_“Hey there”/body gesture>和<visual_stream = smile “how are you doing?”/face gesture>的数据结构表示的视觉提示流或修改后的视觉提示流。在该示例中,可以同步文本内容流或修改后的文本内容流或者视觉提示流或修改后的视觉提示流,从而产生<start = body gesture_hand lift / “Hey there” / end = body gesture_hand lift / start = face gesture_smile / “how are you doing?” / end = face gesture_smile>的同步数据结构。值得注意的是,在该示例中,视觉提示流或修改后的视觉提示流用视觉提示时间戳注释,这些视觉提示时间戳指示视觉提示流或修改后的视觉提示流何时将用于相对于文本内容流或修改后的文本内容流和/或相对于包括与文本内容流或修改后的文本内容流相对应的合成语音的合成语音音频数据控制自动化助理的实例的可视化表示。

[0067] 例如,同步数据结构包括对应的开始视觉提示时间戳,该开始视觉提示时间戳指示包括在视觉提示流中的给定视觉提示何时将开始用于控制自动化助理的实例的可视化表示,如关于文本内容的第一部分定义的“start = body gesture_hand lift”所指示的(例如,在“Hey there”被可听地呈现以呈现给用户的相同时间或阈值时间量内开始),以及如关于文本内容的第二部分定义的“start = face gesture_smile”所指示的(例如,在“how are you doing?”被可听地呈现以呈现给用户的相同时间或阈值时间量内开始)。此外,同步数据结构包括对应的停止视觉提示时间戳,该停止视觉提示时间戳指示包括在视觉提示流中的给定视觉提示何时将停止用于控制自动化助理的实例的可视化表示,如关于文本内容的第一部分定义的“end = body gesture_hand lift”所指示的(例如,在“Hey there”被可听地呈现以呈现给用户的相同时间或阈值时间量内停止),以及如关于文本内容的第二部分定义的“start = face gesture_smile”所指示的(例如,在“how are you doing?”被可听地呈现以呈现给用户的相同时间或阈值时间量内开始)。

[0068] 在这种情况下,这些时间戳可以由以本文描述的方式(例如,关于图4和图5)训练的那些LLM (例如,经由图1的LLM引擎182)为同步数据结构自动注释。同步可以由以本文描述的方式(例如,关于图4和图5)训练的那些LLM (例如,经由图1的LLM引擎182)自动执行。尽管上述情况是关于用于控制自动化助理的实例的可视化表示的视觉提示进行描述的,但是应当理解,这是为了举例说明,并不意味着限制,并且类似的视觉提示和一个或多个对应视觉提示时间戳可以另外或替代地用于控制客户端设备的显示(例如,如关于图6A和图6B所述)。在其他情况下,这些时间戳可以在使用TTS模型生成包括与文本内容流或修改后的文本内容流相对应的合成语音的合成语音音频数据时为同步数据结构自动注释,因为在TTS生成期间对应于文本内容流的词元或者对应于视觉提示流的词元是可用的。这确保视觉提示流与合成语音的呈现同步,而不仅仅是包括在合成语音中的基础文本内容流或修改后的文本内容流。

[0069] 在框262处,系统确定是否接收到附加音频数据流。如果在框262的迭代中,系统确定尚未接收到附加音频数据流,则系统继续在框262处监测附加音频数据流。如果在框262的迭代中,系统确定已经接收到附加音频数据流,则系统返回到框254,以基于处理附加音频数据流来生成给定附加助理输出。值得注意的是,附加音频数据流可以捕获来自用户或客户端设备的附加用户的附加口头话语(例如,使用上文关于框252的操作描述的各种技术确定)。因此,附加音频数据流的处理可以以与上述相同或相似的方式(例如,假设附加口头

话语由提供口头话语的用户提供)或其他方式(例如,假设附加口头话语由附加用户提供,并且附加用户已经将附加角色指派给自动化助理的实例)进行适配。

[0070] 在各种实现方式中,如果在框262的迭代中,系统确定在阈值持续时间(例如,10秒、30秒、3分钟、5分钟等)内尚未接收到附加音频数据流,则系统可以生成视觉提示流,以用于控制自动化助理的实例的可视化表示和/或用于控制客户端设备的显示。该视觉提示流可以被视觉地渲染以指示例如自动化助理的实例正在等待用户提供一个或多个附加口头话语以促进对话会话、客户端设备的一个或多个组件仍然处于活动状态以处理一个或多个附加口头话语和/或其他信息。值得注意的是,该视觉提示流可以可选地独立于任何合成语音音频数据并且独立于该视觉提示流可以被认为响应的任何口头话语而被视觉地渲染。

[0071] 值得注意的是,在图2的方法200的实现方式中,可以生成给定助理输出,随后使用各种后处理步骤将其适配于指派给自动化助理的实例的给定角色。然而,应当理解,这是为了举例说明,并不意味着限制。例如,并且如下文关于图3所述,可以生成给定助理输出并将其适配于指派给自动化助理的实例的给定角色,而无需这些后处理步骤。此外,尽管关于用户提供口头话语描述了图2的方法200的实现方式,但是应当理解,这也是为了举例说明,并不意味着限制。相反,应当理解,也可以响应于用户提供键入输入和/或触摸输入而利用图2的方法200的实现方式。

[0072] 现在转到图3,描绘了示出基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配给定助理输出的另一示例方法300的流程图。为了方便起见,方法300的操作参考执行这些操作的系统来描述。方法300的该系统包括一个或多个处理器、存储器和/或计算设备(例如,图1、图6A、图6B、图7A和图7B的客户端设备110、角色系统120和/或图8的计算设备810、一个或多个服务器和/或其他计算设备)的其他组件。此外,虽然方法300的操作以特定顺序示出,但这并不意味着限制。一个或多个操作可以被重新排序、省略和/或添加。

[0073] 在框352处,系统接收捕获客户端设备的用户的口头话语的音频数据流,该口头话语指向至少部分地在客户端设备处执行的自动化助理的实例。系统可以以与上文关于图2的方法200的框252的操作相同或相似的方式接收口头音频数据流。此外,系统可以可选地以与上文关于图2的方法200的框252的操作描述的相同或相似的方式处理口头音频数据流,以识别提供口头话语的用户。

[0074] 在框354处,系统基于处理音频数据流并且使用给定LLM生成响应于口头话语并且特定于由用户从多个不同的角色中指派给自动化助理的实例的给定角色的给定助理输出,该给定助理输出包括:(1)文本内容流;以及(2)视觉提示流。在生成修改后的给定助理输出时,系统还可以考虑用户在其中提供口头话语的对话会话的场境(如果有的话)。如上文关于图2的方法200所指出,给定角色可以包括例如特定于给定角色并且用于修改文本内容流以生成修改后的文本内容流的给定词汇、特定于给定角色并且用于随后合成修改后的文本内容流以向用户进行可听呈现的给定韵律属性集合和/或特定于给定角色并且用于修改视觉提示流以生成修改后的视觉提示流的给定视觉提示集合。值得注意的是,其他不同角色中的每个角色也可以与对应词汇、对应韵律属性集合和/或对应视觉提示集合相关联。

[0075] 在一些实现方式中,并且如框354A处所指示,LLM可以特定于指派给自动化助理的实例的给定角色。在这些实现方式中,特定于给定角色的LLM可以以关于图4描述的方式进

行训练。例如,系统可以识别指派给自动化助理的实例的给定角色,并且可以从一个或多个数据库(例如,客户端设备的设备上存储装置和/或远离客户端设备但客户端设备能够通过一个或多个网络访问的存储装置)获得特定于指派给自动化助理的实例的给定角色的LLM。此外,系统可以使用ASR模型处理音频数据流,以生成ASR输出流,诸如与在音频数据流中捕获的口头话语相对应的一个或多个辨识词项。此外,系统可以使用NLU模型处理ASR输出流,以生成NLU输出流,诸如一个或多个意图以及与这些意图中的一个或多个意图相关联的一个或多个参数的对应槽值。在该示例中,系统可以使用LLM来处理ASR输出流、NLU输出流和/或用户在其中提供口头话语的对话会话的场境(如果有的话),以生成给定助理输出。因此,在框354处生成的给定助理输出可以特定于指派给自动化助理的实例的给定角色,并且不需要对给定助理输出进行任何附加后处理以针对指派给自动化助理的实例的给定角色对其进行定制,因为用于生成给定助理输出的LLM是专门针对给定角色进行训练的。

[0076] 在一些实现方式中,并且如框354B处所指示,LLM对于多个角色中的一个或多个角色可能是通用的。在这些实现方式中,系统可以使用LLM来处理ASR输出流、NLU输出流和/或用户在其中提供口头话语的对话会话的场境(如果有的话),以生成如上所述的给定助理输出。然而,在这些实现方式中,系统还可以使用LLM以及ASR输出流、NLU输出流和/或对话会话的场境来处理特定于指派给自动化助理的实例的给定角色的给定角色数据。给定角色数据可以对应于例如给定角色词元、给定角色嵌入和/或包括指派给自动化助理的实例的给定角色的表示的其他给定角色数据。如本文所述,给定角色数据可以由与自动化助理相关联的开发者策划,使用一个或多个LLM生成,和/或以其他方式使用各种技术学习。因此,即使LLM对于多个角色中的一个或多个角色是通用的,在框354处生成的给定助理输出可以特定于指派给自动化助理的实例的给定角色,并且不需要对给定助理输出进行任何附加后处理以针对指派给自动化助理的实例的给定角色对其进行定制,因为使用生成给定助理输出的LLM进行处理是通过利用给定角色数据来适配于给定角色的。

[0077] 在框356处,系统使捕获与文本内容流或修改后的文本内容流相对应的合成语音的合成语音音频数据被可听地渲染以呈现给用户。在框358处,系统使视觉提示流或修改后的视觉提示流用于控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示。框356和358的操作可以以分别关于图2的方法200的框258和260的操作描述的相同或相似的方式执行。在各种实现方式中,系统可以使与视觉提示流或修改后的文本内容流相对应的合成语音的可听渲染和视觉提示流或修改后的视觉提示流在控制客户端设备的显示和/或在控制自动化助理的实例的可视化表示中的利用同步,以呈现给用户(例如,如关于图2的方法200所述)。

[0078] 在框360处,系统确定是否接收到附加音频数据流。如果在框360的迭代中,系统确定尚未接收到附加音频数据流,则系统继续在框360处监测附加音频数据流。如果在框360的迭代中,系统确定已经接收到附加音频数据流,则系统返回到框354,以基于处理附加音频数据流来生成给定附加助理输出。系统可以接收附加音频数据流并且以与关于图2的方法200的框262的操作描述的相同或相似的方式对其执行后续处理和/或监测附加音频数据流。值得注意的是,在这些实现方式中并且与图2的方法200的实现方式相反,可以针对指派给自动化助理的实例的给定角色生成给定助理输出,而无需上文关于图2的方法200描述的各种后处理步骤。

[0079] 现在转到图4,描绘了示出训练大语言模型以用于基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配指派给该自动化助理的给定助理输出的示例方法400的流程图。为了方便起见,方法400的操作参考执行这些操作的系统来描述。方法400的该系统包括一个或多个处理器、存储器和/或计算设备(例如,图1、图6A、图6B、图7A和图7B的客户端设备110、角色系统120和/或图8的计算设备810、一个或多个服务器和/或其他计算设备)的其他组件。此外,虽然方法400的操作以特定顺序示出,但这并不意味着限制。一个或多个操作可以被重新排序、省略和/或添加。

[0080] 在框452处,系统生成将用于训练特定于多个不同的角色中的给定角色的给定LLM的实例的给定角色训练实例。给定LLM的实例可以对应于例如通用LLM,该通用LLM已经被训练为基于处理各种口头话语的对应ASR输出流、各种口头话语的对应NLU数据流和/或在其中接收到各种口头话语的对应对话会话的对应场境来至少生成文本内容流。然而,给定LLM的实例可能还没有被训练为生成特定于给定角色的文本内容流和/或可能还没有被训练为生成各种口头话语的视觉提示流。换句话说,给定LLM的实例可能已经用于使用通用大词汇来至少生成文本内容流,但是可能不被训练为生成使用特定于给定角色的给定角色大词汇的文本内容流和/或可能不被训练为生成用于控制被指派给定角色的自动化助理的实例在其处执行的客户端设备的显示和/或用于控制自动化助理的实例的可视化表示的视觉提示流。因此,以本文的方式描述的给定角色训练实例是用于使给定LLM的实例能够生成特定于给定角色的文本内容流和特定于给定角色的视觉提示流的一种技术。

[0081] 在一些实现方式中,并且如框452A处所指示,给定角色训练实例可以基于开发者输入而生成(例如,如关于图5的方法500A和500B所述)。在附加或替代实现方式中,并且如框452B处所指示,给定角色训练实例可以基于分析来自在线多媒体储存库的视频内容而生成(例如,如关于图5的方法500C所述)。在其他实现方式中,给定角色训练实例可以从一个或多个数据库获得,诸如其中给定角色训练实例由图1的3P系统192中的一个或多个3P系统生成的实现方式。

[0082] 在框454处,系统确定是否满足用于训练给定LLM的实例的一个或多个条件。用于训练给定LLM的实例的一个或多个条件包括可用于训练给定LLM的实例的训练实例的数量、一天中的时间、一周中的一天、自从给定LLM的实例先前经由图4的方法400的实例进行训练以来已经过去的持续时间、是否已经接收到用于训练给定LLM的实例的开发者输入和/或其他条件。如果在框454的迭代中,系统确定不满足用于训练给定LLM的一个或多个条件,则系统可以继续在框454处监测是否满足一个或多个条件。值得注意的是,在系统继续在框454处监测是否满足一个或多个条件的同时,系统可以继续生成附加给定角色训练实例以用于训练特定于给定角色的给定LLM的实例。如果在框454的迭代中,系统确定满足用于训练给定LLM的一个或多个条件,则系统可以进行到框456。

[0083] 在框456处,系统将给定LLM的实例训练为用于随后处理捕获指向被指派给定角色的自动化助理的实例的口头话语的音频数据。系统可以基于如何生成给定角色训练实例来适配如何训练给定LLM的实例,如关于图5的方法500A、500B和500C所述。在框458处,系统使给定LLM的实例用于随后处理捕获指向被指派给定角色的自动化助理的实例的口头话语的音频数据(例如,如分别关于图3的方法300和图4的方法400所述)。

[0084] 值得注意的是,系统可以返回到框452的操作,以生成特定于给定角色的附加给定

角色训练实例,以用于进一步训练特定于给定角色的给定LLM的给定实例。系统可以基于附加给定角色训练实例继续细化给定LLM的实例。此外,系统可以以并行方式或以串行方式执行图4的方法400的附加迭代,以训练特定于能够指派给自动化助理的实例的其他对应角色的给定LLM的附加实例。换句话说,可以利用图4的方法400的多次迭代来生成特定于能够指派给自动化助理的实例的多个不同的角色的给定LLM的对应实例。

[0085] 现在转到图5,描绘了示出生成用于训练用于基于多个不同的角色当中指派给自动化助理的给定角色动态地适配指派给该自动化助理的给定助理输出的大语言模型的角色训练实例的示例方法500A、500B和500C的流程图。为了方便起见,方法500A、500B和500C的操作参考执行这些操作的系统来描述。方法500A、500B和500C的该系统包括一个或多个处理器、存储器和/或计算设备(例如,图1、图6A、图6B、图7A和图7B的客户端设备110、角色系统120和/或图8的计算设备810、一个或多个服务器和/或其他计算设备)的其他组件。此外,虽然方法500A、500B和500C的操作以特定顺序示出,但这并不意味着限制。一个或多个操作可以被重新排序、省略和/或添加。此外,应当理解,方法500A、500B和500C中的每一者提供可以在图4的方法400的框452的操作的一次或多次迭代中实现的附加或替代技术。

[0086] 在一些实现方式中,在框552处,系统从与自动化助理相关联的开发者并且针对给定角色接收开发者输入,该开发者输入用一个或多个视觉提示时间戳注释文本内容流,这些视觉提示时间戳指示何时将相对于文本内容流使用视觉提示流。例如,假设文本内容流由<text_stream = “Hey there, how are you doing?”的数据结构表示。在该示例中,开发者输入可以将相关联的一个或多个视觉提示与文本内容流的一个或多个部分相关联,诸如提供包括由<visual_stream = hand lift_”Hey there”/body gesture> (以使自动化助理的实例的可视化表示在文本内容流的“Hey there”部分被可听地呈现时挥手)和<visual_stream = smile ”how are you doing?”/face gesture> (以使自动化助理的实例的可视化表示在文本内容流的“how are you doing”部分被可听地呈现时微笑)的数据结构表示的视觉提示的开发者输入。此外,在该示例或者其中一个或多个视觉提示已经与文本内容流相关联的其他示例中,开发者输入可以包括一个或多个视觉提示时间戳,这些视觉提示时间戳诸如通过定义何时应该开始利用、应该暂停、应该结束利用包括在视觉提示流中的视觉提示和/或将如何利用它们的其他指示来指示何时相对于文本内容流利用视觉提示流。换句话说,可以提供开发者输入以相对于诸如由如上文关于图2的方法200所述的<start = body gesture_hand lift / “Hey there” / end = body gesture_hand lift / start = face gesture_smile / “how are you doing?” / end = face gesture_smile>的同步数据结构表示的文本内容流同步视觉提示流的利用。

[0087] 在这些实现方式中,在框554处,系统至少基于开发者输入来生成给定角色训练实例。在这些实现方式的一些版本中,给定角色训练实例可以包括例如训练实例输入和训练实例输出。继续以上示例,训练实例输入可以包括例如可以产生文本片段流的一个或多个对应口头话语的ASR输出流、基于ASR输出流生成的NLU输出流和/或在其中接收到一个或多个对应口头话语的对应对话的一个或多个对应场境。训练实例输出可以包括例如文本内容流和相对于文本内容流定义的视觉提示流(例如,经由一个或多个视觉提示时间戳)。在这些实现方式中,在基于以关于框554的操作描述的方式生成的给定角色训练实例而训练给定LLM的实例时,系统可以使给定LLM的实例处理训练实例输入以生成预测输出。预测输出

可以包括例如预测文本内容流和相对于预测文本内容流定义的预测视觉提示流。在该示例中,系统可以将相对于预测文本内容流定义的预测视觉提示流与相对于基于开发者输入确定的训练实例输出的文本内容流定义的视觉提示流进行比较以生成一个或多个损失,并且给定LLM的实例可以基于这些损失中的一个或多个损失进行更新(例如,经由反向传播)。在该示例中,系统可以另外或替代地将预测文本内容流与由开发者注释的训练实例输出的文本内容流进行比较以生成一个或多个附加或替代损失,并且给定LLM的实例可以基于这些附加或替代损失中的一个或多个损失进行更新(例如,经由反向传播)。换句话说,系统可以使给定LLM的实例被训练为生成文本内容流以及特定于给定角色并且相对于视觉提示定义的视觉提示流。

[0088] 在这些实现方式的附加或替代版本中,给定角色训练实例可以包括例如文本内容流与视觉提示流(和/或一个或多个视觉提示时间戳)之间的映射。在这些实现中,在基于以关于框554的操作描述的方式生成的给定角色训练实例而训练给定LLM的实例时,系统可以将映射指派给文本内容流,使得当给定LLM的实例随后基于随后口头话语生成文本内容流时,可以获得视觉提示流和一个或多个视觉提示流时间戳,并将其用于控制客户端设备的显示和/或指派给给定角色的自动化助理的实例的可视化表示。换句话说,系统可以使给定LLM的实例被训练为生成文本内容流并且获得视觉提示流以及特定于给定角色并且相对于视觉提示定义的一个或多个视觉提示时间戳。

[0089] 在附加或替代实现方式中,在框556处,系统从与自动化助理相关联的开发者并且针对给定角色接收开发者输入,该开发者输入将屏幕动画修改为相对于文本内容流的视觉提示流。例如,再次假设文本内容流由<text_stream = “Hey there, how are you doing?”的数据结构表示。在该示例中,并且与上述框552的操作相反,开发者输入可以定义将用于相对于文本内容流控制自动化助理的实例的可视化表示的更高级别动画物理运动手势(例如,至少相对于上文关于框552的操作描述的数据结构的更高级别)和/或用于控制客户端设备的显示的更高级别屏幕动画。例如,开发者输入可以经由专用训练平台提供,该专用训练平台使得开发者能够经由一个或多个输入设备(例如,鼠标和键盘)相对于文本内容流修改或移动自动化助理的实例的可视化表示的动画主体部分,和/或修改或添加屏幕动画以用于控制客户端设备的显示。作为某个非限制性示例,开发者输入可以包括拖动可视化表示的手臂以挥手运动移动,提供“挥手”的输入以使可视化表示的手臂以挥手运动移动,提供“跳跃”的输入以使可视化表示以跳跃运动移动,拖动可视化表示的面部的一个或多个部分以使其微笑、皱眉等,提供“微笑”的输入以使可视化表示的手臂微笑,和/或经由专用训练平台的任何其他类型的开发者输入。值得注意的是,在这些实现方式中,开发者输入可能不需要用一个或多个视觉提示时间戳来显式地注释文本内容流。

[0090] 在这些实现方式中,在框558处,系统至少基于开发者输入来生成给定角色训练实例。在这些实现方式的一些版本中,给定角色训练实例可以包括例如训练实例输入和训练实例输出,如上(例如,关于框554的操作)所述。然而,在这些实现方式中,视觉提示流和/或一个或多个视觉提示时间戳可以通过将更高级别动画物理运动手势和/或屏幕动画转换成上述不同结构来从开发者输入中导出,以使得给定LLM的实例能够以与上文关于在框554的操作中对给定角色训练实例的处理相同或相似的方式进行训练。在这些实现方式的附加或替代版本中,给定角色训练实例可以包括例如文本内容流与更高级别动画物理运动手势

和/或屏幕动画之间的映射。在这些实现方式中,在基于以关于框558的操作描述的方式生成的给定角色训练实例而训练给定LLM的实例时,系统可以将映射指派给文本内容流,使得当给定LLM的实例随后基于随后口头话语生成文本内容流时,可以获得更高级别动画物理运动手势和/或屏幕动画,并将其用于控制客户端设备的显示和/或指派给给定角色的自动化助理的实例的可视化表示。

[0091] 在附加或替代实现方式中,在框560处,系统从在线多媒体储存库获得视频内容,该视频内容包括视频的可听内容的音频数据流和视频的视觉内容的视觉数据流。在线多媒体储存库可以由本文所述的一方(诸如图1的1P系统191和/或3P系统192中的一个或多个系统)托管。例如,在线多媒体储存库可以是使多个用户能够上载视频内容并与其他用户共享视频内容的公共在线多媒体储存库。而且,例如,在线多媒体储存库可以是仅某些用户能够访问的私有在线多媒体储存库。视频内容可以捕获例如将由与给定角色相关联的自动化助理的虚拟化表示模拟的一个或多个人物或角色。

[0092] 在这些实现方式中,在框562处,系统处理音频数据流以生成文本内容流,并处理视觉数据流以生成视觉提示流。例如,系统可以使用ASR模型来处理包括在视频内容中的音频数据流以生成ASR输出流,并且可以使用NLU模型来处理ASR输出流以生成NLU输出流。文本内容流可以至少如本文(例如,关于图2的方法200、图3的方法300等)所述基于ASR输出流和/或NLU输出流来确定。在该示例中,系统还可以使用一个或多个运动跟踪机器学习模型(例如,被训练为跟踪眼睛注视、嘴部运动、嘴唇运动、身体运动、身体姿势等的机器学习模型)来处理视觉数据流,以生成指示在视频内容中捕获的人物或角色在说话时如何视觉地表达他们自己的输出。在这些实现方式的一些版本中,视觉提示流可以对应于使用一个或多个运动跟踪机器学习模型生成的输出。在附加或替代实现方式中,可以进一步处理使用一个或多个运动跟踪机器学习模型生成的输出(例如,使用一个或多个分类机器学习模型)以确定更高级别动画物理运动手势和/或屏幕动画。值得注意的是,视频内容的可听内容和视觉内容由于都是视频内容的一部分而在初始处理时已经被同步。因此,与视频内容的可听内容和视觉内容两者相关联的一个或多个视频时间戳可以用于指派一个或多个视觉提示时间戳,这些视觉提示时间戳指示何时相对于文本内容流利用视觉提示流。

[0093] 在这些实现方式中,在框564处,系统至少基于文本内容流和视觉提示流来生成给定角色训练实例。在这些实现方式的一些版本中,给定角色训练实例可以包括例如训练实例输入和训练实例输出并且用于针对给定角色训练给定LLM的实例,如上(例如,关于框554的操作)所述。

[0094] 值得注意的是,这些不同方法500A、500B和500C提供了用于生成角色训练实例的不同技术。例如,图5的方法500A描述了经由开发者输入手动地定义给定角色训练实例的过程,其中开发者输入相对于文本内容流定义视觉提示流和/或一个或多个视觉提示时间戳。而且,例如,图5的方法500B描述了经由开发者输入半手动地定义给定角色训练实例的过程,其中开发者输入相对于使得能够生成给定角色训练实例的文本内容流以更高级别定义视觉提示流和/或一个或多个视觉提示时间戳。而且,例如,图5的方法500C描述了经由视频内容的处理自动地定义给定角色训练实例的过程,其中在生成给定角色训练实例时没有使用开发者输入。这些不同方法中的每一者可以单独或以任何组合使用,以生成给定角色训练实例和/或附加给定角色训练实例,以用于训练特定于给定角色的给定LLM的实例。例如,

方法500A和500B可以用于生成用于初始引导给定LLM的实例的给定角色训练实例,但是方法500C可以用于进一步训练和细化给定LLM的实例以减少与生成给定角色训练实例相关联的时间和成本。替代地,方法500C可以用于初始引导给定LLM的实例以减少与生成给定角色训练实例相关联的时间和成本,但是方法500A和500B可以用于进一步训练和细化给定LLM的实例以确保在后续使用给定LLM的实例时具有更高的准确度和精确度。

[0095] 尽管图5的方法500A、500B和500C描述了用于生成给定角色训练实例的特定技术,但是应当理解,这些特定技术是为了举例说明而提供的,并不意味着限制。此外,尽管方法500A、500B和500C总体上是关于生成用于训练给定LLM的实例的给定角色训练实例进行描述的,但是应当理解,这也是为了举例说明,并不意味着限制。例如,方法500A、500B和/或500C还可以用于生成附加给定角色训练实例,以用于训练特定于能够指派给自动化助理的实例的给定角色的给定LLM的实例。另外或替代地,方法500A、500B和/或500C还可以用于生成给定附加角色训练实例,以用于训练特定于也能够指派给自动化助理的实例的给定附加角色的给定LLM的附加实例。

[0096] 现在转到图6A和图6B,描绘了基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配在其处实现该自动化助理的客户端设备的显示的各种非限制性示例。为了举例说明,假设图6A和图6B中描绘的客户端设备是图1的客户端设备110的实例,并且客户端设备110包括显示190。显示190可以是例如包括各种部分的触摸屏显示。例如,显示190可以包括第一部分190A,该第一部分包括在客户端设备110处活动的用户帐户的指示(例如,如由显示190的第一部分190A的右手侧中的用户帐户符号所指示),以及各种组件何时在客户端设备110处活动的指示,诸如客户端设备110的一个或多个麦克风或语音处理组件(例如,如显示190的第一部分190A中的椭圆190A1所指示)。此外,显示190可以包括第二部分190B,该第二部分包括客户端设备110的用户与至少部分地在客户端设备110处实现的自动化助理的实例之间的对话的转录。此外,显示可以包括第三部分190C,该第三部分包括用于将被提供用于视觉呈现给用户的视觉内容的空间(例如,主屏幕)。

[0097] 尽管图6A和图6B中所示的显示190包括各种不同的部分,但是应当理解,这是为了举例说明,并不意味着限制。例如,显示的不同部分可以彼此重叠(例如,显示190的第二部分190B与显示190的第三部分190C重叠)和/或在某些情况下被省略(例如,当用户没有参与与自动化助理的实例的对话会话时,显示190的第二部分190B可以被省略)。此外,尽管在图6A和图6B中描绘了客户端设备110,但是还应当理解,这是为了举例说明,并不意味着限制,并且具有显示的附加或替代客户端设备可以实现本文所述的技术。此外,尽管关于图6A和图6B描述的示例示出了特定示例,但是应当理解,这也是为了举例说明,并不意味着限制,并且本文所述的技术可以用于在与各种不同主题相关的各种其他对话会话中控制客户端设备110的显示。

[0098] 为了在描述图6A和图6B时举例说明,假设用户提供“Hey Assistant, what time is it? (嘿,助理,现在几点了?)”的口头话语652,假设自动化助理的实例处理捕获口头话语652的音频数据流(例如,由客户端设备110的一个或多个麦克风生成)以使“Good morning [User]! It's 8:30 AM. Any fun plans today? (早上好[用户]!现在是早上8:30,今天有什么好玩的计划吗?)”的合成语音654被可听地渲染以经由客户端设备110的一个或多个扬声器并且响应于口头话语652呈现给用户,假设用户响应于合成语音654提供

“Yes, I’m thinking about going to the beach (有,我想去海滩)”的附加口头话语656,并且假设自动化助理的实例处理捕获口头话语656的音频数据流以使“Sounds fun! But if you’re going to Half Moon Bay again, expect rain and chilly temperatures (听起来很有趣!但如果你要再去半月湾,预计会下雨并且气温会很低)”的合成语音658被可听地渲染以经由客户端设备110的一个或多个扬声器并且响应于口头话语656呈现给用户。在该示例中,自动化助理的实例可以以本文(例如,关于图2的方法200和/或图3的方法300)描述的任何方式处理在对应音频数据流中捕获的对应口头话语以生成对应文本片段流,并且可选地利用以本文(例如,关于图5的方法400和/或图5的方法500A、500B和/或500C)描述的任何方式训练的一个或多个LLM。在一些实现方式中,自动化助理的实例可以使捕获该对话会话的转录被提供以用于在客户端设备110的显示190处视觉呈现给用户(例如,如图6A和图6B中的显示190的第二部分190B所示)。

[0099] 在该示例中,并且在处理捕获口头话语656的音频数据流(例如,由客户端设备110的一个或多个麦克风生成)时,自动化助理的实例可以利用视觉提示流来控制客户端设备110的显示190。例如,并且如图6A所示,在处理捕获口头话语656的音频数据流时生成的用于控制客户端设备110的显示190的视觉提示流可以使客户端设备的显示190视觉地呈现海滩场景,以在客户端设备110的显示190处呈现给用户(例如,如图6A中的显示190的第三部分190C所示),因为口头话语656指示用户计划去海滩。图6A中所示的海滩场景可以包括例如带有沙堡的沙滩、海浪汹涌且鱼在水中游动的海洋、鸟在空中飞翔的明亮晴朗天空和/或用户通常可以与海滩相关联的其他内容。值得注意的是,自动化助理的实例可以利用视觉提示流来控制客户端设备110的显示190,以相对于在合成语音658中捕获的文本内容流(诸如被合成以生成合成语音658的“Sounds fun! But if you’re going to Half Moon Bay again, expect rain and chilly temperatures (听起来很有趣!但如果你要再去半月湾,预计会下雨并且气温会很低)”的文本内容流)视觉地呈现海滩场景。在该示例中,用于生成合成语音的文本内容流658和用于控制客户端设备110的显示190的视觉提示流形成响应于口头话语656的给定助理输出。

[0100] 在图6A的实例中的视觉提示流可以包括例如要呈现的视觉内容的指示、应当如何呈现和/或布置视觉内容的指示、应当相对于在合成语音658中捕获的文本内容流呈现视觉内容的一个或多个视觉提示时间戳和/或用于指示自动化助理的实例显示什么、如何显示以及何时显示的其他信息。在一些实现方式中,根据视觉提示流被提供用于向用户进行视觉呈现的视觉内容可以是静态视觉内容。例如,波浪、鱼和鸟一旦被显示就不会移动,除非视觉提示流另外指示。在其他实现方式中,根据视觉提示流被提供用于向用户进行视觉呈现的视觉内容可以是动态视觉内容。例如,海浪可能看起来好像在移动并冲击着沙滩,鱼可能看起来好像在海洋中游动,并且鸟可能看起来好像在空中飞翔,如视觉提示流所指示。

[0101] 在图6A所示的实例中,自动化助理的实例可以使用例如一个或多个视觉提示时间戳来使海滩场景被视觉地提供以经由显示190呈现给用户,诸如紧接在合成语音658被可听地呈现之前的时间的实例和/或合成语音658最初被可听地呈现时的时间的实例。然而,当合成语音658最初被可听地呈现时,自动化助理的实例可以利用视觉提示流来使客户端设备190的显示190适配海滩场景,使其与用户与自动化助理的实例之间的对话会话更具场境相关性。

[0102] 值得注意的是,合成语音658指示半月湾(用户去海滩时通常会去的虚构海滩)会下雨并且气温会很低。该天气信息可以基于自动化助理的实例基于指示用户要去海滩的口头话语656提交天气信息查询而获得。因此,合成语音658也可能与用户与自动化助理的实例之间的对话会话场境相关。然而,图6A中描绘的海滩场景可能不准确地反映天气信息。因此,并且如图6B所示,在处理捕获口头话语656的音频数据流时生成的用于控制客户端设备110的显示190的视觉提示流可以使客户端设备的显示190视觉地适配海滩场景,以在客户端设备110的显示190处呈现给用户(例如,如图6B中的显示190的第三部分190C所示),因为天气信息指示会有雨并且气温会很低。图6B中所示的海滩场景可以包括例如带有沙堡的沙滩可以转变为移除沙堡,因为人们通常不会在雨中建造沙堡;海浪汹涌可能变得比图6B中所示的海浪更汹涌;并且鱼在水中游动可能消失,以反映由于下雨引起的更大涌浪和湍流的水;鸟在空中飞翔的明亮晴朗天空可以转变为倾盆大雨的阴沉天空;以及/或者用户通常可能与下雨海滩相关联的其他内容。值得注意的是,自动化助理的实例可以利用视觉提示流来控制客户端设备110的显示190,以相对于在合成语音658中捕获的文本内容流(诸如被合成以生成合成语音658的“Sounds fun! But if you're going to Half Moon Bay again, expect rain and chilly temperatures (听起来很有趣!但如果你要再去半月湾,预计会下雨并且气温会很低)”的文本内容流)视觉地呈现海滩场景。

[0103] 类似于图6A的示例,在图6B的示例中,自动化助理的实例可以使用例如一个或多个视觉提示时间戳来适配海滩场景以经由显示190呈现给用户,诸如当合成语音658的“But(但)”部分被可听地呈现时的时间实例和/或当合成语音658的“expect rain and chilly temperatures (预计会下雨并且气温会很低)”部分被可听地呈现时的时间实例。因此,当合成语音658被可听地呈现时,自动化助理的实例可以利用视觉提示流来使客户端设备190的显示190被控制以最初呈现图6A中所示的晴朗海滩场景,但是将晴朗海滩场景动态地适配为图6B中所示的下雨海滩场景,以与用户与自动化助理的实例之间的整个对话会话更具场境相关性。

[0104] 尽管图6A和图6B是上文关于用于控制客户端设备110的显示190的特定视觉提示流进行描述的,但是应当理解,这是为了举例说明,并不意味着限制。相反,应当理解,提供图6A和图6B的示例是为了说明除了特定文本内容流(例如,在合成语音658中捕获)之外,自动化助理的实例可以如何在响应于口头话语656时利用视觉提示流。而且,例如,视觉提示流可以另外或替代地用于控制自动化助理的实例的可视化表示(例如,如下文关于图7A和图7B所述)。此外,尽管用于控制客户端设备110的显示190的特定视觉提示流仅响应于口头话语656,但是应当理解,这是为了简洁,并不意味着限制。例如,可以相对于附加文本内容流(例如,在合成语音654中捕获)生成和利用附加特定视觉提示流,以动态地呈现与反映当前时间(例如,如由合成语音654指示的早上8:30)的时钟相关联的视觉内容。此外,尽管图6A和图6B未关于没有任何特定角色的自动化助理的实例进行描述,但是应当理解,这是为了举例说明,并不意味着限制。相反,应当理解,用于生成文本内容流的词汇可以偏向指派给自动化助理的实例的给定角色的给定角色词汇,并且用于生成合成语音的韵律属性集合可以特定于指派给自动化助理的实例的给定角色(例如,如关于图7A和图7B更详细地描述)。此外,尽管图6A和图6B是关于用户提供口头话语作为输入进行描述的,但是应当理解,这是为了举例说明,并不意味着限制。相反,应当理解,图6A和图6B的技术也可以用于用户

提供键入输入和/或键入输入和口头话语的混合的情况。

[0105] 现在转到图7A和图7B,描绘了基于多个不同的角色当中指派给自动化助理的给定角色来动态地适配自动化助理的可视化表示的各种非限制性示例。为了简洁起见,图7A和图7B中描绘的客户端设备是上文关于图6A和图6B描述的相同的客户端设备。为了在描述图7A时举例说明,假设客户端设备110的用户已经将管家角色指派给自动化助理的实例。为了在描述图7B时举例说明,假设客户端设备110的用户已经将海盗角色指派给自动化助理的实例。如本文所述,指派给自动化助理的这些不同角色可以影响由自动化助理的实例用于在用户与自动化助理的实例之间的对话会话期间生成文本片段流的词汇、由自动化助理的实例用于在用户与自动化助理的实例之间的对话会话期间生成捕获文本片段流的对应合成语音的韵律属性集合和/或由自动化助理的实例用于在用户与自动化助理的实例之间的对话会话期间生成视觉提示流的视觉提示集合。

[0106] 具体参考图7A (例如,其中将管家角色指派给自动化助理),假设客户端设备110的用户提供“Hey Assistant, what time is it? (嘿,助理,现在几点了?)”的口头话语752,假设自动化助理的实例处理捕获口头话语752的音频数据流(例如,由客户端设备110的一个或多个麦克风生成)以使“Salutations [User]! It's 8:30 AM. How are you keeping on this fine morning? (问候[用户]!现在是早上8:30,这个美好的早晨你过得怎么样?)”的合成语音754A被可听地渲染以经由客户端设备110的一个或多个扬声器并且响应于口头话语752呈现给用户,假设用户响应于合成语音754A提供“I'm doing well and thinking about going to the beach (我感觉很好,今天想去海滩)”的附加口头话语756A,并且假设自动化助理的实例处理捕获口头话语756A的音频数据流以使“Excellent to hear sire, but I do caution you to take your umbrella and expect rain at Half Moon Bay (听起来很棒,先生,但我要提醒你带上雨伞,预计半月湾会下雨)”的合成语音758A被可听地渲染以经由客户端设备110的一个或多个扬声器并且响应于口头话语756A呈现给用户。

[0107] 具体参考图7B (例如,其中将海盗角色指派给自动化助理),假设客户端设备110的用户提供“Hey Assistant, what time is it? (嘿,助理,现在几点了?)”的口头话语752,假设自动化助理的实例处理捕获口头话语752的音频数据流(例如,由客户端设备110的一个或多个麦克风生成)以使“Good morning matey! It's 8:30 AM. Where are you setting sail to today? (早上好,伙计!现在是早上8:30,你今天要启航去哪里?)”的合成语音754B被可听地渲染以经由客户端设备110的一个或多个扬声器并且响应于口头话语752呈现给用户,假设用户响应于合成语音754B提供“I'm actually thinking about going to the beach today (我其实想今天去海滩)”的附加口头话语756B,并且假设自动化助理的实例处理捕获口头话语756B的音频数据流以使“It's going to be raining, but perfect weather to find some treasure at Half Moon Bay! (要下雨了,但这是去半月湾寻找宝藏的绝佳天气!)”的合成语音758B被可听地渲染以经由客户端设备110的一个或多个扬声器并且响应于口头话语756B呈现给用户。

[0108] 因此,在这些示例中,并且即使由客户端设备110的用户提供的口头话语在图7A和图7B中大体上相同,但用于生成图7A和图7B中的合成语音的对应文本片段流不同,并且用于控制不同角色的对应可视化表示的对应视觉提示流不同(例如,如由图7A中的自动化助

理的实例的管家可视化表示执行的第一手势和/或动画190A2A所指示,以及如由图7B中的自动化助理的实例的海盗可视化表示执行的第二手势和/或动画190A2B所指示)。在这些示例中,管家角色可以对应于更正式和保守的角色,而海盗角色可以对应于更冒险和前卫的角色。这些不同角色由与不同角色相关联的不同词汇、与不同角色相关联的不同的韵律属性集合和/或与不同角色相关联的不同的视觉提示集合来举例说明。

[0109] 例如,在其中将管家角色指派给自动化助理的实例的图7A中,自动化助理的实例可以生成对应文本内容流和对应视觉提示流作为特定于管家角色的对应给定助理输出。在一些实现方式中,自动化助理的实例可以通过利用各种后处理操作以及可选地通过利用特定于管家角色和/或利用特定于管家角色的管家角色数据的一个或多个LLM来生成特定于管家角色的对应给定助理输出(例如,如关于图2的方法200所述)。在附加或替代实现方式中,自动化助理的实例可以通过利用特定于管家角色和/或利用特定于管家角色的管家角色数据的一个或多个LLM来生成特定于管家角色的对应给定助理输出(例如,如关于图3的方法300所述)。

[0110] 例如,自动化助理的实例可以使用ASR模型来处理捕获口头话语752、756A的对应音频数据流以生成对应ASR输出流,并且可以使用NLU模型来处理对应ASR输出流以生成对应NLU输出流。可以处理对应音频数据流、对应ASR输出流、对应NLU输出流和/或对话会话的场境以生成对应给定助理输出(或对应修改后的给定助理输出)。在确定对应文本内容流时,使用LLM生成的对应输出或使用LLM生成的先前生成的输出(例如,先前基于与用户提供的那些相同或相似的口头话语生成)可以包括一个或多个词汇中的一个或多个单词和/或短语的序列上的对应概率分布,诸如特定于管家角色的管家角色词汇或可能偏向与管家角色相关联的单词和/或短语的通用词汇(例如,合成语音754A中的“Salutations (问候)…”和“How are you keeping on this fine morning? (这个美好的早晨你过得怎么样?)”以及合成语音758A中的“sire (先生)”和“I do caution (我要提醒你)”)。此外,在基于对应文本内容流生成对应合成语音754A和758A时,可以利用与管家角色相关联的韵律属性集合来确保音调、节奏、音高、节律和/或其他韵律属性反映管家角色的韵律属性。因此,自动化助理的实例不仅可以口头反映真实管家在对应文本内容流方面可以利用的术语,而且还可以口头反映真实管家在合成对应文本内容流方面可以利用的说话风格,从而反映管家角色的正式和保守个性。

[0111] 此外,在确定视觉提示流时,使用LLM生成的对应输出或使用LLM生成的先前生成的输出(例如,先前基于与用户提供的那些相同或相似的口头话语生成)可以包括词元序列上的对应概率分布,这些词元表示可以由可视化管家相对于对应文本内容流执行的一个或多个动画物理运动手势(例如,如图7A中显示190的第三部分190C所示)和/或相对于对应文本内容流执行的其他动画(例如,如上文参考图6A和图6B所述的用于控制显示190的指令),诸如可视化管家在“Salutations (问候)”时点头或抬起他的单片眼镜和/或挥手,或者在说“I do caution to take your umbrella (我要提醒你带上雨伞)”时拿着伞并打开伞。这些动画物理手势动作和/或其他动画可以以本文描述的任何方式和/或其他方式与对应文本内容流同步。因此,自动化助理的实例的可视化管家不仅可以在显示190上在外观方面视觉地反映真实管家,而且还可以执行动画物理手势运动,这些动画物理手势运动与真实管家的真实物理手势运动在可视化管家基于视觉提示流进行控制方面相匹配,从而进一步

反映管家角色的正式和保守个性。

[0112] 类似地,在其中将海盗角色指派给自动化助理的实例的图7B中,自动化助理的实例可以生成对应文本内容流和对应视觉提示流作为特定于海盗角色的对应给定助理输出。在一些实现方式中,自动化助理的实例可以通过利用各种后处理操作以及可选地通过利用特定于海盗角色和/或利用特定于海盗角色的海盗角色数据的一个或多个LLM来生成特定于海盗角色的对应给定助理输出(例如,如关于图2的方法200所述)。在附加或替代实现方式中,自动化助理的实例可以通过利用特定于海盗角色和/或利用特定于海盗角色的海盗角色数据的一个或多个LLM来生成特定于海盗角色的对应给定助理输出(例如,如关于图3的方法300所述)。

[0113] 例如,自动化助理的实例可以使用ASR模型来处理捕获口头话语752、756B的对应音频数据流以生成对应ASR输出流,并且可以使用NLU模型来处理对应ASR输出流以生成对应NLU输出流。可以处理对应音频数据流、对应ASR输出流、对应NLU输出流和/或对话会话的场境以生成对应给定助理输出(或对应修改后的给定助理输出)。在确定对应文本内容流时,使用LLM生成的对应输出或使用LLM生成的先前生成的输出(例如,先前基于与用户提供的那些相同或相似的口头话语生成)可以包括一个或多个词汇中的一个或多个单词和/或短语的序列上的对应概率分布,诸如特定于海盗角色的海盗角色词汇或可能偏向与海盗角色相关联的单词和/或短语的通用词汇(例如,合成语音754B中的“Ahoy matey (啊嗨,伙计)”和“Where are you setting sail today? (你今天要启航去哪里?)”以及合成语音758B中的“treasure (宝藏)”)。此外,在基于对应文本内容流生成对应合成语音754和758B时,可以利用与海盗角色相关联的韵律属性集合来确保音调、节奏、音高、节律和/或其他韵律属性反映海盗角色的韵律属性。因此,自动化助理的实例不仅可以口头反映真实海盗在对应文本内容流方面可以利用的术语,而且还可以口头反映真实海盗在合成对应文本内容流方面可以利用的说话风格,从而反映海盗角色的冒险和前卫个性。

[0114] 此外,在确定视觉提示流时,使用LLM生成的对应输出或使用LLM生成的先前生成的输出(例如,先前基于与用户提供的那些相同或相似的口头话语生成)可以包括可以由可视化海盗相对于对应文本内容流执行的动画物理运动手势(例如,如图7B中显示190的第三部分190C所示)和/或相对于对应文本内容流执行的其他动画(例如,如上文参考图6A和图6B所述的用于控制显示190的指令)的序列上的对应概率分布,诸如可视化海盗在说“Ahoy matey (啊嗨,伙计)”时上下跳跃和/或挥手,或者在说“perfect weather to find some treasure (寻找宝藏的绝佳天气)”时挖掘并打开宝箱。这些动画物理手势动作和/或其他动画可以以本文描述的任何方式和/或其他方式与对应文本内容流同步。因此,自动化助理的实例的可视化海盗不仅可以在显示190上在外观方面视觉地反映真实海盗,而且还可以执行动画物理手势运动,这些动画物理手势运动与真实海盗的真实物理手势运动在可视化海盗基于视觉提示流进行控制方面相匹配,从而进一步反映海盗角色的冒险和前卫个性。

[0115] 尽管图7A和图7B是关于特定角色(例如,相对于图7A的管家角色和在相对于图7B的海盗角色)以及关于特定角色的特定可视化表示进行描述的,但是应当理解,这是为了举例说明,并不意味着限制。相反,应当理解,可以被指派给自动化助理的实例的角色可以是几乎无限的。此外,尽管图7A和图7B的特定角色是关于特定文本片段流和特定视觉提示流进行描述的,但是应当理解,这也是为了举例说明,并不意味着限制。相反,应当理解,文本

片段流可以通过使用如本文所述的LLM基于几乎无限的词汇生成,并且视觉提示流可以包括角色特定的动画物理运动手势和/或动画(例如,可视化管家抬起他的单片眼镜、可视化海盗挖掘宝藏等)、与角色无关的动画物理运动手势和/或动画(例如,可视化管家微笑、可视化海盗微笑等)、针对不同场景的角色特定的情绪(例如,可视化管家对雨和温度持谨慎态度、可视化海盗对雨和温度兴奋)、用于控制如何相对于文本内容流利用视觉提示流的对应视觉提示时间戳等。此外,尽管图7A和图7B是关于基于视觉提示流控制自动化助理的实例的可视化表示而不控制显示190进行描述的(例如,如关于图6A和图6B所述),但是应当理解,这是为了简洁起见,并不意味着限制。相反,应当理解,这些技术可以组合。例如,图7A的可视化管家和图7B的可视化海盗可以看起来好像他们在针对图6B所示的下雨海滩场景适配的图6A所示的晴朗海滩场景中。此外,尽管图7A和图7B是关于用户提供口头话语作为输入进行描述的,但是应当理解,这是为了举例说明,并不意味着限制。相反,应当理解,图7A和图7B的技术也可以用于用户提供键入输入和/或键入输入和口头话语的混合的情况。

[0116] 现在转到图8,描绘了可以可选地用于执行本文所述的技术的一个或多个方面的示例计算设备810的框图。在一些实现方式中,客户端设备、基于云的自动化助理组件和/或其他组件中的一者或多者可以包括示例计算设备810的一个或多个组件。

[0117] 计算设备810通常包括经由总线子系统812与多个外围设备通信的至少一个处理器814。这些外围设备可以包括存储子系统824 (包括例如存储器子系统825和文件存储子系统826)、用户界面输出设备820、用户界面输入设备822和网络接口子系统816。输入和输出设备允许与计算设备810的用户交互。网络接口子系统816提供到外部网络的接口并且耦合到其他计算设备中的对应接口设备。

[0118] 用户界面输入设备822可以包括键盘、指向设备(诸如鼠标、轨迹球、触摸板或图形输入板、扫描仪、合并到显示器中的触摸屏)、音频输入设备(诸如话音辨识系统、麦克风)和/或其他类型的输入设备。一般来讲,术语“输入设备”的使用旨在包括将信息输入到计算设备810中或输入到通信网络上的所有可能类型的设备和方式。

[0119] 用户界面输出设备820可以包括显示子系统、打印机、传真机或非视觉显示器诸如音频输出设备。显示子系统可以包括阴极射线管(CRT)、平板设备诸如液晶显示器(LCD)、投影设备或用于创建可见图像的一些其他机构。显示子系统还可以诸如经由音频输出设备提供非视觉显示。一般来讲,术语“输出设备”的使用旨在包括将信息从计算设备810输出到用户或输出到另一机器或计算设备的所有可能类型的设备和方式。

[0120] 存储子系统824存储提供本文所述的一些或所有模块的功能性的编程和数据构造。例如,存储子系统824可以包括执行本文所公开的方法的选定方面以及实现图1中描绘的各种组件的逻辑。

[0121] 这些软件模块通常由处理器814单独或与其他处理器组合执行。存储子系统824中使用的存储器825可以包括多个存储器,包括用于在程序执行期间存储指令和数据的主随机存取存储器(RAM) 830和其中存储固定指令的只读存储器(ROM) 832。文件存储子系统826可以为程序和数据文件提供持久存储,并且可以包括硬盘驱动器、软盘驱动器以及相关联的可移除介质、CD-ROM驱动器、光盘驱动器或可移除介质盒。实现某些实现方式的功能性的模块可以由文件存储子系统826存储在存储子系统824中,或者存储在能够由处理器814访问的其他机器中。

[0122] 总线子系统812提供用于使计算设备810的各种组件和子系统按预期彼此通信的机构。尽管总线子系统812被示意性地示为单个总线,但是总线子系统812的替代实现方式可以使用多个总线。

[0123] 计算设备810可以是各种类型,包括工作站、服务器、计算集群、刀片服务器、服务器场或任何其他数据处理系统或计算设备。由于计算机和网络的性质不断变化,对图8中描绘的计算设备810的描述仅旨在作为用于说明一些实现方式的特定示例。计算设备810的许多其他配置可能具有比图8中描绘的计算设备更多或更少的组件。

[0124] 在其中本文所述的系统收集或以其他方式监测关于用户的个人信息或者可以利用个人信息和/或所监测的信息的情况下,可以为用户提供控制程序或功能是否收集用户信息(例如,关于用户的社交网络、社交行为或活动、职业、用户的偏好或用户的当前地理位置的信息)或者控制是否和/或如何从内容服务器接收可能与用户更相关的内容的机会。另外,某些数据在其存储或使用之前可能会以一种或多种方式进行处理,使得个人可识别信息被删除。例如,可以处理用户的身份,使得无法确定用户的个人可识别信息,或者在获得地理位置信息的情况下,可以将用户的地理位置泛化(诸如泛化到城市、邮政编码或州级别),使得无法确定用户的具体地理位置。因此,用户可以控制如何收集和/或使用关于用户的信息。

[0125] 在一些实现方式中,提供了一种由一个或多个处理器实现的方法,该方法包括:接收捕获客户端设备的用户的口头话语的音频数据流,该音频数据流由客户端设备的一个或多个麦克风生成,并且口头话语指向至少部分在客户端设备处执行的自动化助理的实例;以及基于处理音频数据流生成响应于口头话语的给定助理输出。给定助理输出包括:(i)文本内容流;以及(ii)视觉提示流,该视觉提示流用于响应于口头话语控制客户端设备的显示和/或用于控制被视觉地渲染以经由客户端设备的显示呈现给用户的自动化助理的实例的可视化表示。该方法还包括:基于由用户从多个不同的助理角色中指派给自动化助理实例的给定助理角色修改响应于口头话语的给定助理输出,以生成修改后的给定助理输出。修改后的给定助理输出包括:(i)与文本内容流不同的修改后的文本内容流;以及(ii)与视觉提示流不同的修改后的视觉提示流,该修改后的视觉提示流用于响应于口头话语控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示。该方法还包括:响应于接收到捕获客户端设备的用户的口头话语的音频数据流:使捕获与修改后的文本内容流相对应的合成语音的合成语音音频数据被可听地渲染以经由客户端设备的一个或多个扬声器呈现给用户;以及使修改后的视觉提示流用于控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示。

[0126] 本文所公开的技术的这些和其他实现方式可以可选地包括以下特征中的一个或多个特征。

[0127] 在一些实现方式中,该方法还可以包括:使与修改后的文本内容流相对应的合成语音的可听渲染和修改后的视觉提示流在控制客户端设备的显示和/或在控制自动化助理的实例的可视化表示中的利用同步,以呈现给用户。

[0128] 在那些实现方式的一些版本中,该方法还可以包括:在使捕获与修改后的文本内容流相对应的合成语音的合成语音音频数据被可听地渲染以呈现给用户之前:用一个或多个视觉提示时间戳注释修改后的文本内容流,这些视觉提示时间戳指示视觉提示流何时用

于控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示。使与修改后的文本内容流相对应的合成语音的可听渲染和视觉提示流在控制客户端设备的显示和/或在控制自动化助理的实例的可视化表示中的利用同步可以基于一个或多个视觉提示时间戳。

[0129] 在那些实现方式的一些其他版本中,一个或多个视觉提示时间戳可以至少包括开始视觉提示时间戳和停止视觉提示时间戳,该开始视觉提示时间戳指示包括在视觉提示流中的给定视觉提示何时将开始用于控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示,停止视觉提示时间戳指示包括在视觉提示流中的给定视觉提示何时将停止用于控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示。

[0130] 在一些实现方式中,基于处理音频数据流生成响应于口头话语的给定助理输出可以包括:使用自动语音辨识(ASR)模型处理捕获口头话语的音频数据流以生成ASR输出流;使用自然语言理解(NLU)模型处理ASR输出流以生成NLU输出流;以及至少基于NLU输出流确定响应于口头话语的给定助理输出。

[0131] 在那些实现方式的一些版本中,基于NLU输出流确定响应于口头话语的给定助理输出可以包括:使用大语言模型(LLM)处理ASR输出流和/或NLU输出流以确定包括在给定助理输出中的文本内容流和视觉提示流。

[0132] 在那些实现方式的附加或替代版本中,基于NLU输出流确定响应于口头话语的给定助理输出可以包括:使用先前基于口头话语的先前实例使用大语言模型(LLM)生成的LLM输出处理ASR输出流和/或NLU输出流,以确定包括在给定助理输出中的文本内容流和视觉提示流。

[0133] 在那些实现方式的附加或替代版本中,基于NLU输出流确定响应于口头话语的给定助理输出可以包括:基于ASR输出流和/或NLU输出流生成一个或多个结构化请求;向一个或多个第一方智能体和/或一个或多个第三方智能体传输一个或多个结构化请求;以及基于响应于一个或多个结构化请求而接收到的内容确定包括在给定助理输出中的文本内容流和视觉提示流。

[0134] 在一些实现方式中,基于指派给自动化助理的实例的给定助理角色修改响应于口头输出的给定助理输出以生成修改后的给定助理输出可以包括:从一个或多个数据库获得特定于指派给自动化助理的实例的给定角色的给定角色数据;以及处理文本内容流和视觉提示流以及特定于指派给自动化助理的实例的给定角色的角色数据以生成与文本内容流不同的修改后的文本内容流和与视觉提示流不同的修改后的视觉提示流。

[0135] 在那些实现方式的一些版本中,特定于指派给自动化助理的实例的给定角色的角色数据可以包括特定于指派给自动化助理的实例的给定角色的给定角色词元和/或特定于指派给自动化助理的实例的给定角色的给定嵌入。

[0136] 在那些实现方式的附加或替代版本中,处理文本内容流和视觉提示流以及特定于指派给自动化助理的实例的给定角色的角色数据以生成与文本内容流不同的修改后的文本内容流和与视觉提示流不同的修改后的视觉提示流可以包括:使用大语言模型(LLM)处理文本内容流和视觉提示流以及特定于指派给自动化助理的实例的给定角色的角色数据以生成与文本内容流不同的修改后的文本内容流和与视觉提示流不同的修改后的视觉提示流。

[0137] 在那些实现方式的附加或替代版本中,处理文本内容流和视觉提示流以及特定于

指派给自动化助理的实例的给定角色的角色数据以生成与文本内容流不同的修改后的文本内容流和与视觉提示流不同的修改后的视觉提示流可以包括:使用先前基于口头话语的先前实例使用大语言模型(LLM)生成的LLM输出处理文本内容流和视觉提示流以及特定于指派给自动化助理的实例的给定角色的角色数据以生成与文本内容流不同的修改后的文本内容流和与视觉提示流不同的修改后的视觉提示流。

[0138] 在一些实现方式中,指派给自动化助理的实例的给定角色可以与以下相关联:多个不同词汇中的第一词汇,用于修改文本内容流以生成修改后的文本内容流;多个不同的韵律属性集合中的第一韵律属性集合,用于生成捕获与修改后的文本内容流相对应的合成语音的合成语音音频数据,以被可听地渲染以呈现给用户;以及/或者多个不同的视觉提示集合中的第一视觉提示集合,用于修改视觉提示流以生成修改后的视觉提示流。

[0139] 在那些实现方式的一些版本中,与文本内容流不同的修改后的文本内容流可以使用第一词汇来修改,并且与视觉提示流不同的用于响应于口头话语控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示的修改后的视觉提示流可以使用第一视觉提示集合来修改。

[0140] 在那些实现方式的一些其他版本中,该方法还可以包括:使用文本转语音(TTS)模型并且基于第一韵律属性集合处理修改后的文本内容流以生成合成语音音频数据。

[0141] 在那些实现方式的附加或替代版本中,该方法还可以包括:接收捕获附加客户端设备的附加用户的附加口头话语的附加音频数据流,该附加音频数据流由附加客户端设备的一个或多个附加麦克风生成,该附加口头话语指向至少部分地在附加客户端设备处执行的自动化助理的附加实例,并且该附加口头话语与该口头话语相同;以及基于处理附加音频数据流生成响应于附加口头话语的给定助理输出。给定助理输出可以包括:(i)文本内容流;以及(ii)视觉提示流,该视觉提示流用于响应于附加口头话语控制附加客户端设备的附加显示和/或用于控制被视觉地渲染以经由附加客户端设备的附加显示呈现给附加用户的自动化助理的附加实例的附加可视化表示。该方法还可以包括:基于由附加用户从多个不同的助理角色中指派给自动化助理的附加实例的除给定助理角色之外的给定附加助理角色修改响应于附加口头话语的给定助理输出以生成修改后的给定附加助理输出。修改后的给定附加助理输出可以包括:(i)与文本内容流不同并且与修改后的文本内容流不同的修改后的附加文本内容流;以及(ii)与视觉提示流不同并且与修改后的视觉提示流不同的修改后的附加视觉提示流,该修改后的附加视觉提示流用于响应于附加口头话语控制附加客户端设备的附加显示和/或用于控制自动化助理的附加实例的附加可视化表示。该方法还可以包括:响应于接收到捕获附加客户端设备的附加用户的附加口头话语的附加音频数据流:使捕获与修改后的附加文本内容流相对应的附加合成语音的附加合成语音音频数据被可听地渲染以经由附加客户端设备的一个或多个附加扬声器呈现给附加用户;以及使修改后的附加视觉提示流用于控制附加客户端设备的附加显示和/或控制自动化助理的附加实例的附加可视化表示。

[0142] 在那些实现方式的一些版本中,指派给自动化助理的附加实例的给定附加角色可以与以下相关联:多个不同词汇中除第一词汇之外的第二词汇,用于修改附加文本内容流以生成修改后的附加文本内容流;多个不同的韵律属性集合中的除第一韵律属性集合之外的第二韵律属性集合,用于生成捕获与修改后的附加文本内容流相对应的附加合成语音的

附加合成语音音频数据,以被可听地渲染以呈现给附加用户;以及/或者多个不同的视觉提示集合中的除第一视觉提示集合之外的第二视觉提示集合,用于修改附加视觉提示流以生成修改后的附加视觉提示流。

[0143] 在一些实现方式中,该方法还可以包括:接收捕获客户端设备的附加用户的附加口头话语的附加音频数据流,该附加音频数据流由客户端设备的一个或多个附加麦克风生成,该附加口头话语指向至少部分地在客户端设备处执行的自动化助理的实例,并且该附加口头话语与该口头话语相同;基于处理附加音频数据流生成响应于附加口头话语的给定助理输出。给定助理输出可以包括:(i)文本内容流;以及(ii)视觉提示流,该视觉提示流用于响应于附加口头话语控制客户端设备的显示和/或用于控制被视觉地渲染以经由客户端设备的显示呈现给附加用户的自动化助理的实例的附加可视化表示。该方法还可以包括:基于由附加用户从多个不同的助理角色中指派给自动化助理的附加实例的除给定助理角色之外的给定附加助理角色修改响应于附加口头话语的给定助理输出以生成修改后的给定附加助理输出。修改后的给定附加助理输出可以包括:(i)与文本内容流不同并且与修改后的文本内容流不同的修改后的附加文本内容流;以及(ii)与视觉提示流不同并且与修改后的视觉提示流不同的修改后的附加视觉提示流,该修改后的附加视觉提示流用于响应于附加口头话语控制客户端设备的显示和/或用于控制自动化助理的实例的附加可视化表示。该方法还可以包括:响应于接收到捕获客户端设备的附加用户的附加口头话语的附加音频数据流:使捕获与修改后的附加文本内容流相对应的附加合成语音的附加合成语音音频数据被可听地渲染以经由客户端设备的一个或多个扬声器呈现给附加用户;以及使修改后的附加视觉提示流用于控制客户端设备的显示和/或用于控制自动化助理的实例的附加可视化表示。

[0144] 在一些实现方式中,客户端设备的用户可以在初始配置自动化助理的实例的自动化助理帐户的同时或在与自动化助理的实例的自动化助理应用的助理设置交互的同时将给定角色指派给自动化助理的实例。

[0145] 在一些实现方式中,修改后的视觉提示流可以用于响应于口头话语控制客户端设备的显示。

[0146] 在那些实现方式的一些版本中,修改后的视觉提示流还可以用于控制自动化助理的实例的可视化表示。

[0147] 在那些实现方式的附加或替代版本中,用于响应于口头话语控制客户端设备的显示的修改后的视觉提示流可以包括一个或多个显示动画,这些显示动画使客户端设备的显示在合成语音音频数据被可听地渲染以呈现给用户的同时被动态地适配。

[0148] 在一些实现方式中,修改后的视觉提示流可以用于控制自动化助理的实例的可视化表示。

[0149] 在那些实现方式的一些版本中,修改后的视觉提示流还可以用于响应于口头话语控制客户端设备的显示。

[0150] 在那些实现方式的附加或替代版本中,用于控制自动化助理的实例的可视化表示的修改后的视觉提示流可以包括在合成语音音频数据被可听地渲染以呈现给用户的同时由自动化助理的实例的可视化表示执行的一个或多个动画物理手势动作。

[0151] 在一些实现方式中,提供了一种由一个或多个处理器实现的方法,该方法包括:接

收捕获客户端设备的用户的口头话语的音频数据流,该音频数据流由客户端设备的一个或多个麦克风生成,并且该口头话语指向至少部分地在客户端设备处执行的自动化助理的实例;以及基于处理音频数据流并且使用给定大语言模型(LLM)生成给定助理输出,该给定助理输出响应于口头话语并且特定于从多个不同的角色中指派给自动化助理的实例的给定角色。给定助理输出包括:(i)特定于自动化助理的实例的给定角色的文本内容流;以及(ii)视觉提示流,该视觉提示流用于响应于口头话语控制客户端设备的显示和/或用于控制被视觉地渲染以经由客户端设备的显示呈现给用户的自动化助理的实例的可视化表示并且特定于指派给自动化助理的实例的给定角色。该方法还包括:响应于接收到捕获客户端设备的用户的口头话语的音频数据流:使捕获与文本内容流相对应的合成语音的合成语音音频数据被可听地渲染以经由客户端设备的一个或多个扬声器呈现给用户;以及使视觉提示流用于控制客户端设备的显示和/或控制自动化助理的实例的可视化表示。

[0152] 本文所公开的技术的这些和其他实现方式可以可选地包括以下特征中的一个或多个特征。

[0153] 在一些实现方式中,该方法还可以包括:从多个不同的角色中识别由用户指派给自动化助理的实例的给定角色;以及从多个不同的LLM中选择与由用户指派给自动化助理的给定角色相关联的给定LLM。

[0154] 在一些实现方式中,该方法还可以包括:从多个不同的角色中识别由用户指派给自动化助理的实例的给定角色;以及选择特定于指派给自动化助理的实例的给定角色的给定角色数据。给定角色数据可以在生成给定助理输出时使用给定LLM处理。

[0155] 在那些实现方式的一些版本中,特定于指派给自动化助理的实例的给定角色的角色数据可以包括特定于指派给自动化助理的实例的给定角色的给定角色词元和/或特定于指派给自动化助理的实例的给定角色的给定嵌入。

[0156] 在一些实现方式中,提供了一种由一个或多个处理器实现的方法,该方法包括:接收捕获客户端设备的用户的口头话语的音频数据流,该音频数据流由客户端设备的一个或多个麦克风生成,并且口头话语指向至少部分在客户端设备处执行的自动化助理的实例;以及基于处理音频数据流生成响应于口头话语的给定助理输出。给定助理输出包括:(i)文本内容流;以及(ii)视觉提示流,该视觉提示流用于响应于口头话语控制客户端设备的显示和/或用于控制被视觉地渲染以经由客户端设备的显示呈现给用户的自动化助理的实例的可视化表示。该方法还包括:基于由用户从多个不同的助理角色中指派给自动化助理实例的给定助理角色修改响应于口头话语的给定助理输出,以生成修改后的给定助理输出。修改后的给定助理输出包括:(i)文本内容流;以及(ii)与视觉提示流不同的修改后的视觉提示流,该修改后的视觉提示流用于响应于口头话语控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示。该方法还包括:响应于接收到捕获客户端设备的用户的口头话语的音频数据流:使捕获与文本内容流相对应的合成语音的合成语音音频数据被可听地渲染以经由客户端设备的一个或多个扬声器呈现给用户;以及使修改后的视觉提示流用于控制客户端设备的显示和/或控制自动化助理的实例的可视化表示。

[0157] 在一些实现方式中,提供了一种由一个或多个处理器实现的方法,该方法包括:从与自动化助理相关联的开发者并且针对能够指派给自动化助理的给定角色接收与将用于相对于文本内容流控制客户端设备的显示和/或用于相对于文本内容流控制自动化助理的

实例的可视化表示的一个或多个视觉提示相关联的开发者输入;至少基于开发者输入生成将用于进一步训练特定于多个不同的角色中的给定角色的给定大语言模型(LLM)的实例的给定角色训练实例,该给定LLM先前被训练为生成文本内容流;至少基于给定角色训练实例训练给定LLM的实例;以及使给定LLM的实例用于随后处理捕获指向被指派给定角色的自动化助理的实例的口头话语的音频数据。

[0158] 本文所公开的技术的这些和其他实现方式可以可选地包括以下特征中的一个或多个特征。

[0159] 在一些实现方式中,该方法还可以包括:从与自动化助理相关联的开发者并且针对能够指派给自动化助理的给定附加角色接收与将用于相对于附加文本内容流控制客户端设备的显示和/或用于相对于附加文本内容流控制自动化助理的附加实例的附加可视化表示的一个或多个附加视觉提示相关联的附加开发者输入;至少基于附加开发者输入生成将用于进一步训练特定于多个不同的角色中的给定附加角色的给定LLM的附加实例的给定附加角色训练实例;至少基于给定附加角色训练实例训练给定附加LLM的附加实例;以及使给定附加LLM的附加实例用于随后处理捕获指向被指派给定附加角色的自动化助理的附加实例的附加口头话语的附加音频数据。

[0160] 在一些实现方式中,开发者输入可以用一个或多个视觉提示时间戳注释文本内容流,这些视觉提示时间戳指示视觉提示流何时用于相对于文本内容流控制客户端设备的显示和/或用于相对于文本内容流控制自动化助理的实例的可视化表示。

[0161] 在那些实现方式的一些版本中,一个或多个视觉提示时间戳可以至少包括开始视觉提示时间戳和停止视觉提示时间戳,该开始视觉提示时间戳指示包括在视觉提示流中的给定视觉提示何时将开始用于相对于文本内容流控制客户端设备的显示和/或用于相对于文本内容流控制自动化助理的实例的可视化表示,该停止视觉提示时间戳指示包括在视觉提示流中的给定视觉提示何时将停止用于相对于文本内容流控制客户端设备的显示和/或用于相对于文本内容流控制自动化助理的实例的可视化表示。

[0162] 在一些实现方式中,开发者输入可以相对于文本内容流修改客户端设备的显示处的屏幕动画,和/或使自动化助理的实例的可视化表示相对于文本内容流执行一个或多个动画物理手势动作。

[0163] 在一些实现方式中,提供了一种由一个或多个处理器实现的方法,该方法包括:从在线多媒体储存库获得视频内容,该视频内容包括视频的可听内容的音频数据流和视频的视觉内容的视觉数据流;使用自动语音辨识模型处理视频的可听内容的音频数据流,以生成与在视频的可听内容的音频数据流中捕获的一个或多个口头话语相对应的文本内容流;使用一个或多个运动跟踪机器学习模型处理视频的视觉内容的视觉数据流以生成视觉提示流;基于处理音频数据流并且基于处理视频数据流生成将用于进一步训练特定于多个不同的角色中的给定角色的体现在视频内容中的给定大语言模型(LLM)的实例的给定角色训练数据实例;至少基于给定角色训练实例训练给定LLM的实例;以及使给定LLM的实例用于随后处理捕获指向被指派给定角色的自动化助理的实例的附加口头话语的附加音频数据。

[0164] 另外,一些实现方式包括一个或多个计算设备的一个或多个处理器(例如,中央处理单元(CPU)、图形处理单元(GPU)和/或张量处理单元(TPU)),其中一个或多个处理器能够操作以执行存储在相关联的存储器中的指令,并且其中这些指令被配置为使上述方法中的

任一种方法被执行。一些实现方式还包括一个或多个非暂时性计算机可读存储介质,其存储计算机指令,该计算机指令能够由一个或多个处理器执行来执行上述方法中的任何方法。一些实现方式进一步包括计算机程序产品,该计算机程序产品包括指令,该指令可由一个或多个处理器执行以执行上述方法中的任一者。

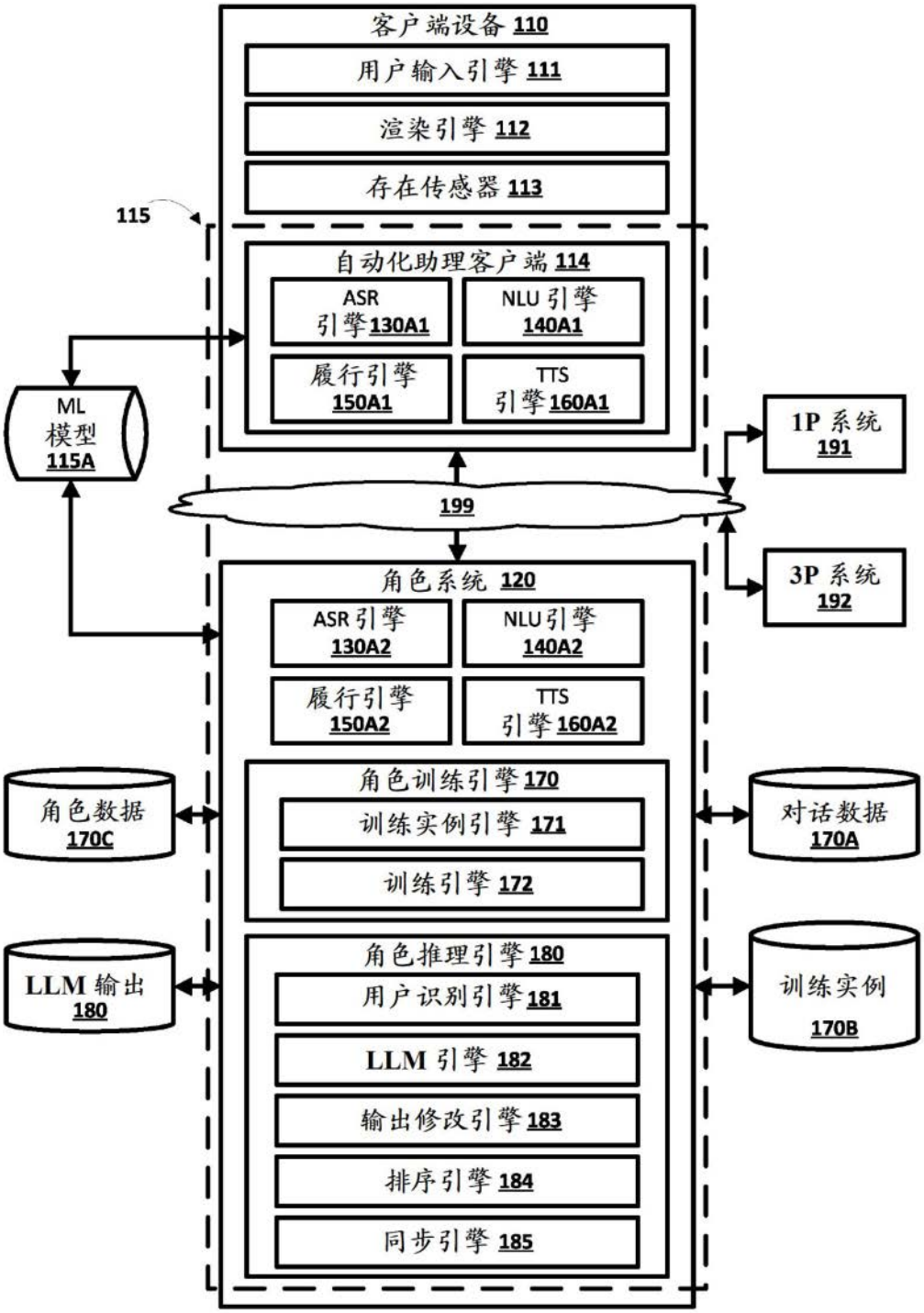

客户端设备 110
用户输入引擎 111
渲染引擎 112
存在传感器 113
115
自动化助理客户端 114
ASR 引擎130A1
NLU 引擎 140A1
履行引擎 150A1
TTS 引擎160A1
ML 模型 115A
1P 系统 191
199
3P 系统 192
角色系统 120
ASR引擎 130A2
NLU引擎 140A2
履行引擎 150A2
TTS 引擎160A2
角色训练引擎 170
训练实例引擎 171
训练引擎 172
角色数据 170C
对话数据 170A
角色推理引擎 180
用户识别引擎181
LLM 引擎 182
输出修改引擎183
排序引擎 184
同步引擎 185
LLM 输出 180
训练实例 170B

图1

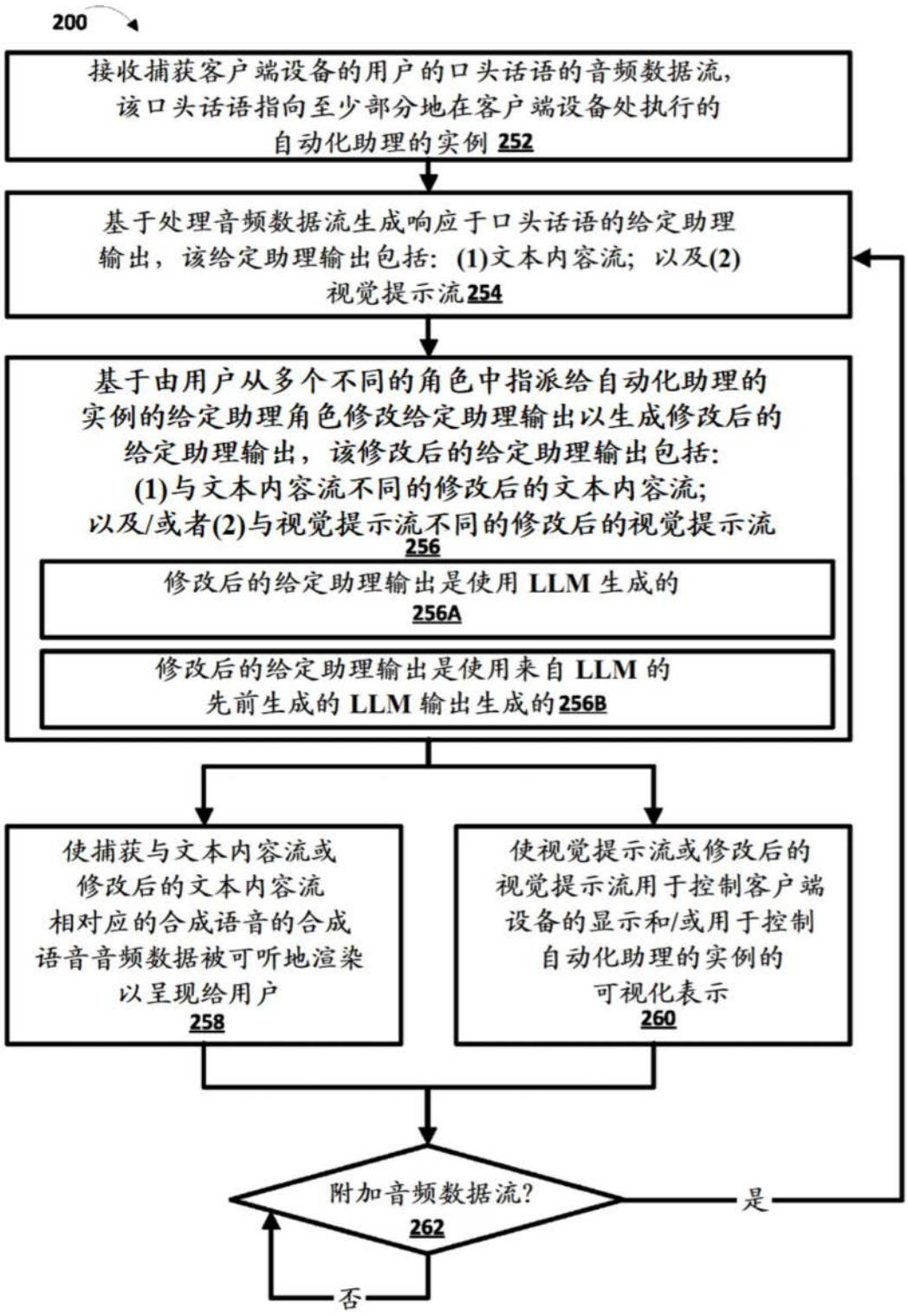

200
接收捕获客户端设备的用户的口头话语的音频数据流，该口头话语指向至少部分地在客户端设备处执行的自动化助理的实例 252
基于处理音频数据流生成响应于口头话语的给定助理输出，该给定助理输出包括：(1)文本内容流；以及(2)视觉提示流254
基于由用户从多个不同的角色中指派给自动化助理的实例的给定助理角色修改给定助理输出以生成修改后的给定助理输出，该修改后的给定助理输出包括：(1)与文本内容流不同的修改后的文本内容流；以及/或者(2)与视觉提示流不同的修改后的视觉提示流 256
修改后的给定助理输出是使用 LLM 生成的 256A
修改后的给定助理输出是使用来自 LLM 的先前生成的 LLM 输出生成的256B
使捕获与文本内容流或修改后的文本内容流相对应的合成语音的合成语音音频数据被可听地渲染以呈现给用户 258
使视觉提示流或修改后的视觉提示流用于控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示 260
附加音频数据流? 262
是
否

图2

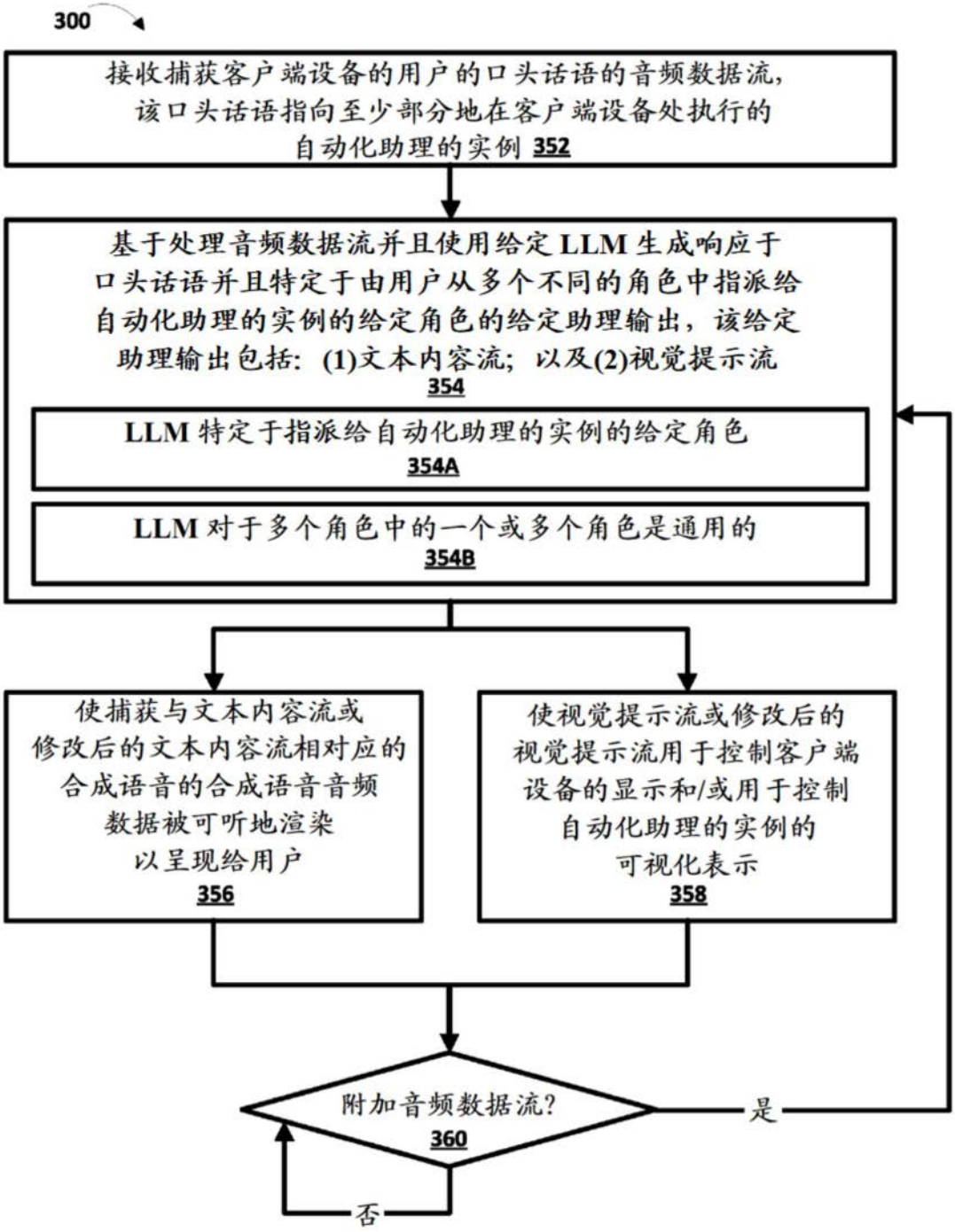

300
接收捕获客户端设备的用户的口头话语的音频数据流，该口头话语指向至少部分地在客户端设备处执行的自动化助理的实例 352
基于处理音频数据流并且使用给定 LLM 生成响应于口头话语并且特定于由用户从多个不同的角色中指派给自动化助理的实例的给定角色的给定助理输出，该给定助理输出包括：(1)文本内容流；以及(2)视觉提示流 354
LLM 特定于指派给自动化助理的实例的给定角色 354A
LLM 对于多个角色中的一个或多个角色是通用的 354B
使捕获与文本内容流或修改后的文本内容流相对应的合成语音的合成语音音频数据被可听地渲染以呈现给用户 356
使视觉提示流或修改后的视觉提示流用于控制客户端设备的显示和/或用于控制自动化助理的实例的可视化表示 358
附加音频数据流? 360
是
否

图3

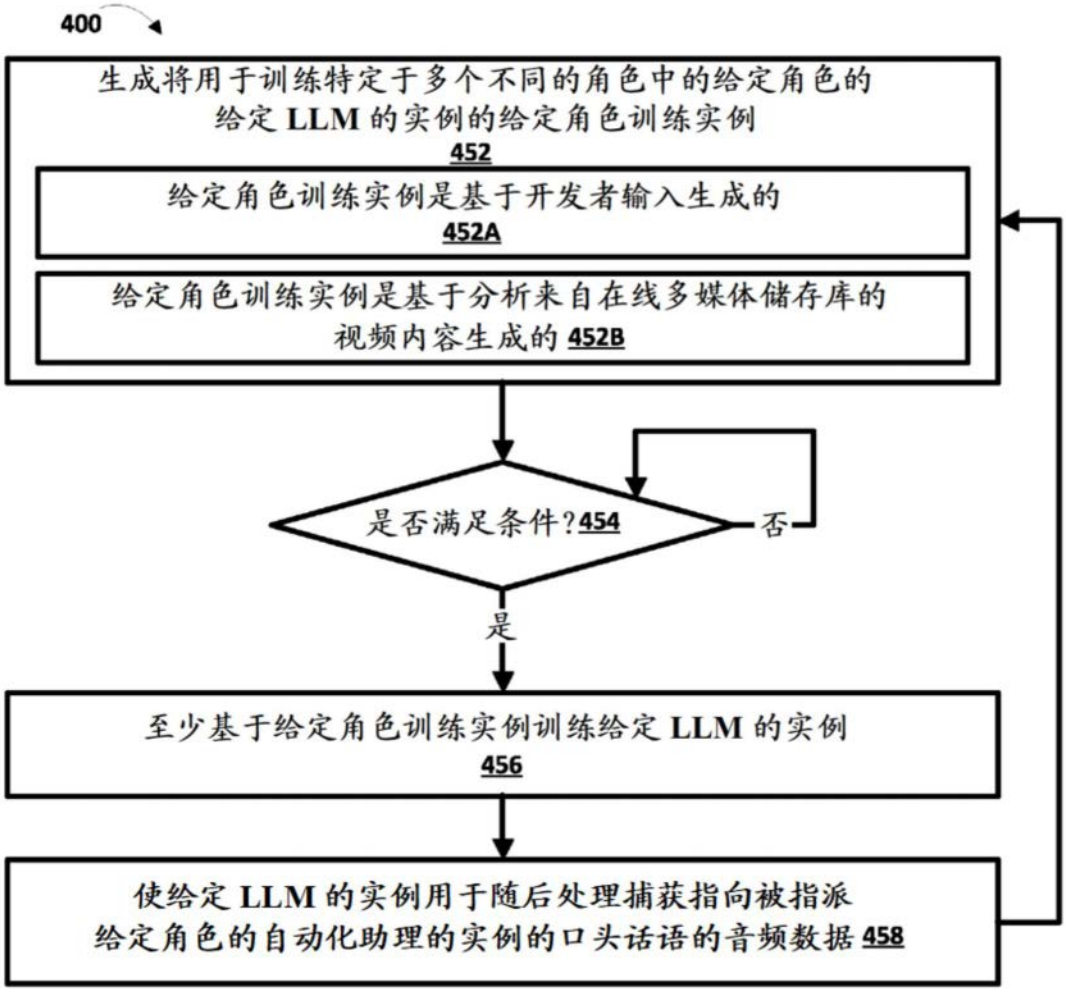

400
生成将用于训练特定于多个不同的角色中的给定角色的给定 LLM 的实例的给定角色训练实例
452
给定角色训练实例是基于开发者输入生成的
452A
给定角色训练实例是基于分析来自在线多媒体储存库的视频内容生成的 452B
是否满足条件?454
否
是
至少基于给定角色训练实例训练给定 LLM 的实例
456
使给定 LLM 的实例用于随后处理捕获指向被指派给定角色的自动化助理的实例的口头话语的音频数据 458

图4

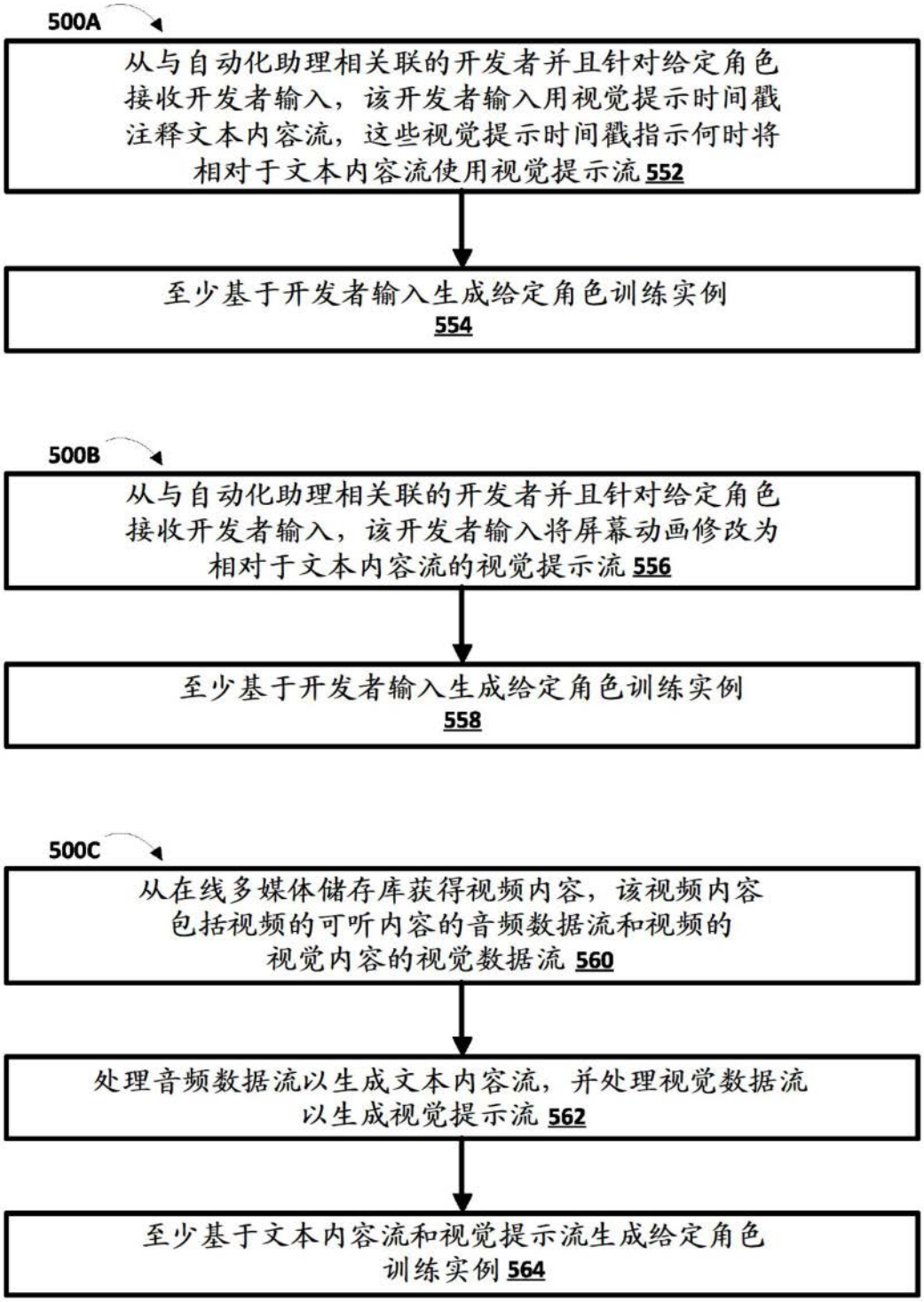

500A
从与自动化助理相关联的开发者并且针对给定角色接收开发者输入，该开发者输入用视觉提示时间戳注释文本内容流，这些视觉提示时间戳指示何时将相对于文本内容流使用视觉提示流 552
至少基于开发者输入生成给定角色训练实例 554
500B
从与自动化助理相关联的开发者并且针对给定角色接收开发者输入，该开发者输入将屏幕动画修改为相对于文本内容流的视觉提示流 556
至少基于开发者输入生成给定角色训练实例 558
500C
从在线多媒体储存库获得视频内容，该视频内容包括视频的可听内容的音频数据流和视频的视觉内容的视觉数据流 560
处理音频数据流以生成文本内容流，并处理视觉数据流以生成视觉提示流 562
至少基于文本内容流和视觉提示流生成给定角色训练实例 564

图5

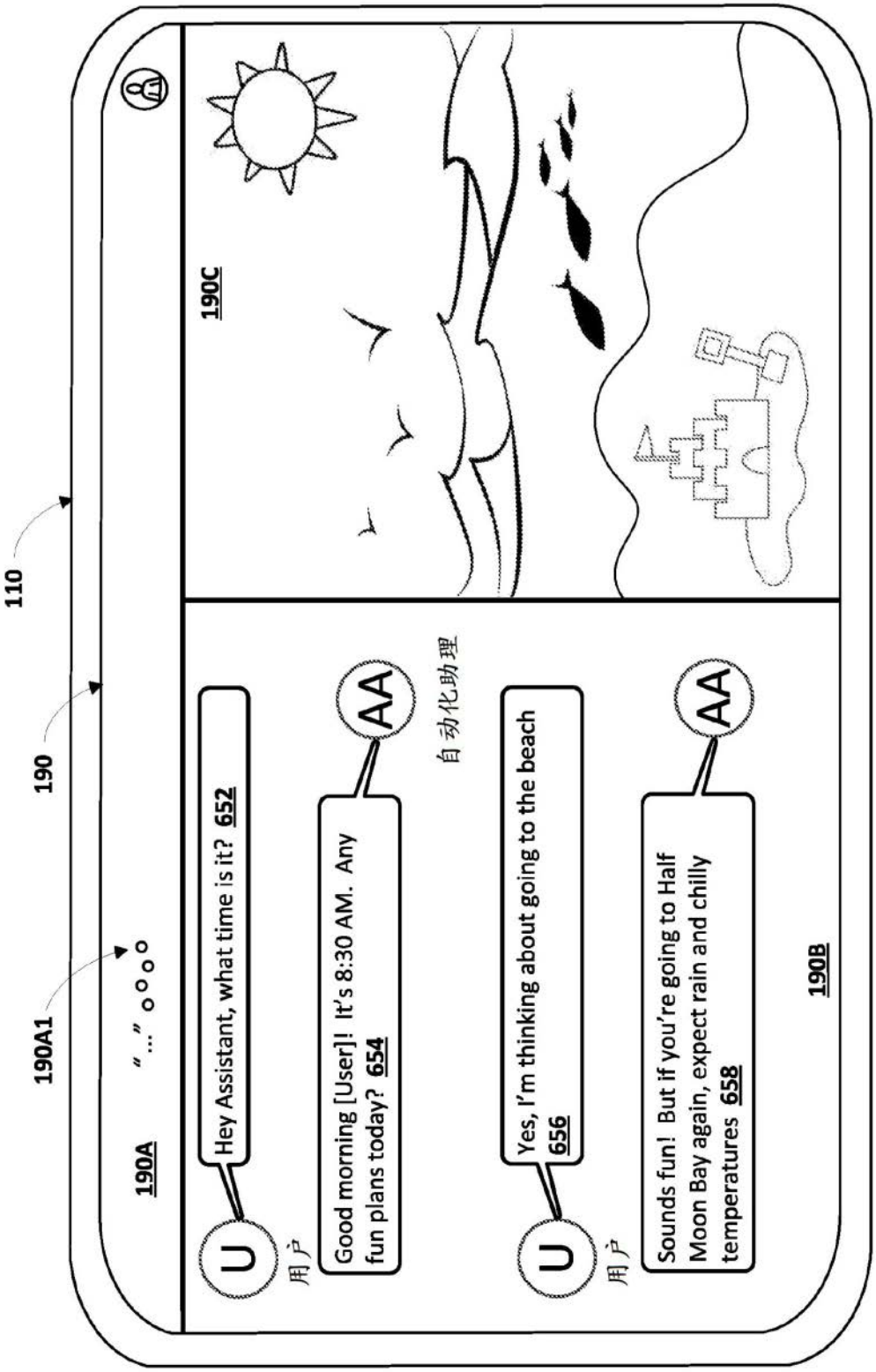

110
190
190A1
190A
"..."
190C
Hey Assistant, what time is it? 652
用户
Good morning [User]! It's 8:30 AM. Any fun plans today? 654
AA
自动化助理
Yes, I'm thinking about going to the beach 656
用户
Sounds fun! But if you're going to Half Moon Bay again, expect rain and chilly temperatures 658
AA
190B

图6A

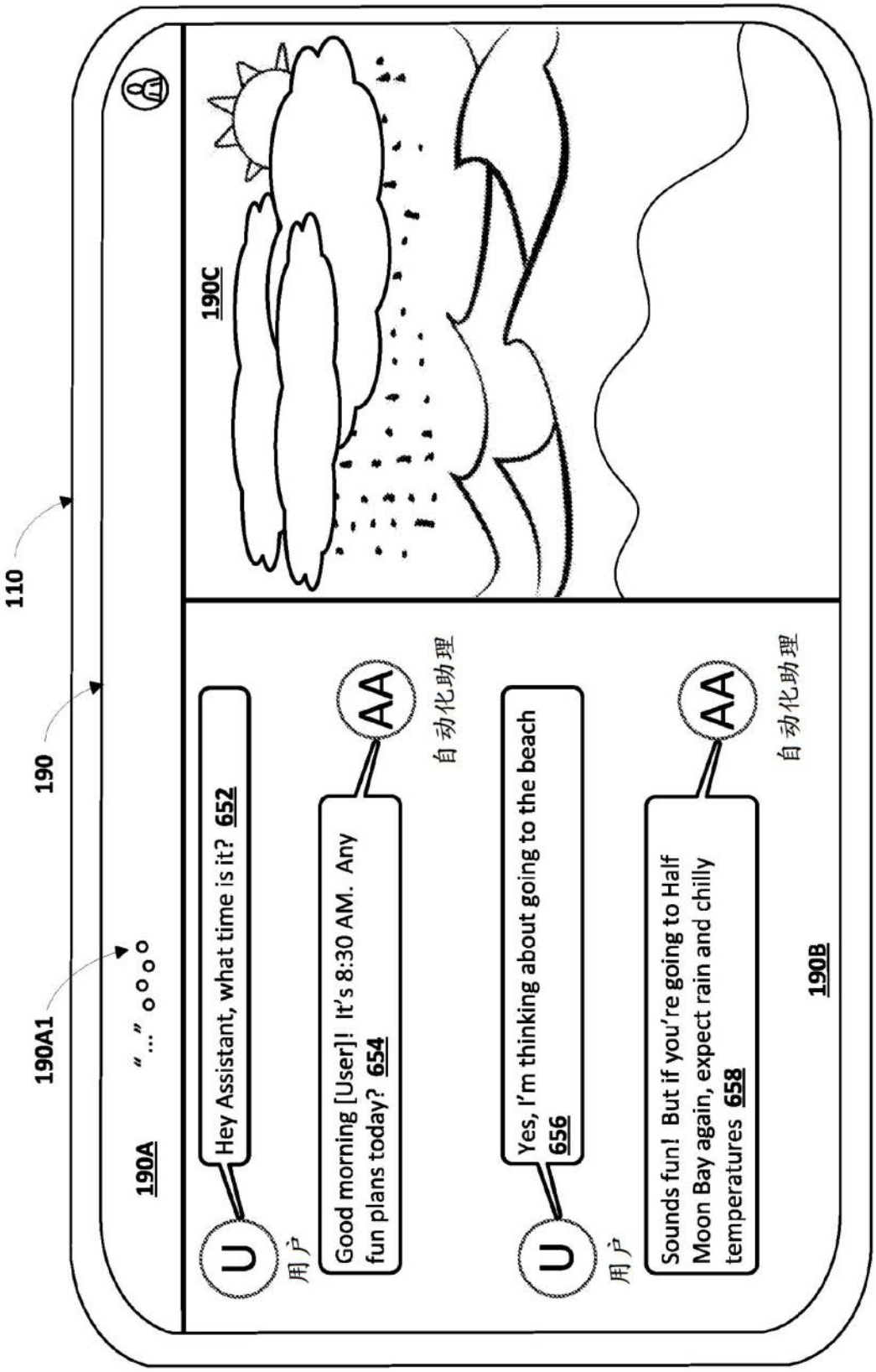

110
190
190A1
190A
190C
190B
Hey Assistant, what time is it? 652
用户
Good morning [User]! It's 8:30 AM. Any fun plans today? 654
AA
自动化助理
Yes, I'm thinking about going to the beach 656
用户
Sounds fun! But if you're going to Half Moon Bay again, expect rain and chilly temperatures 658
AA
自动化助理

图6B

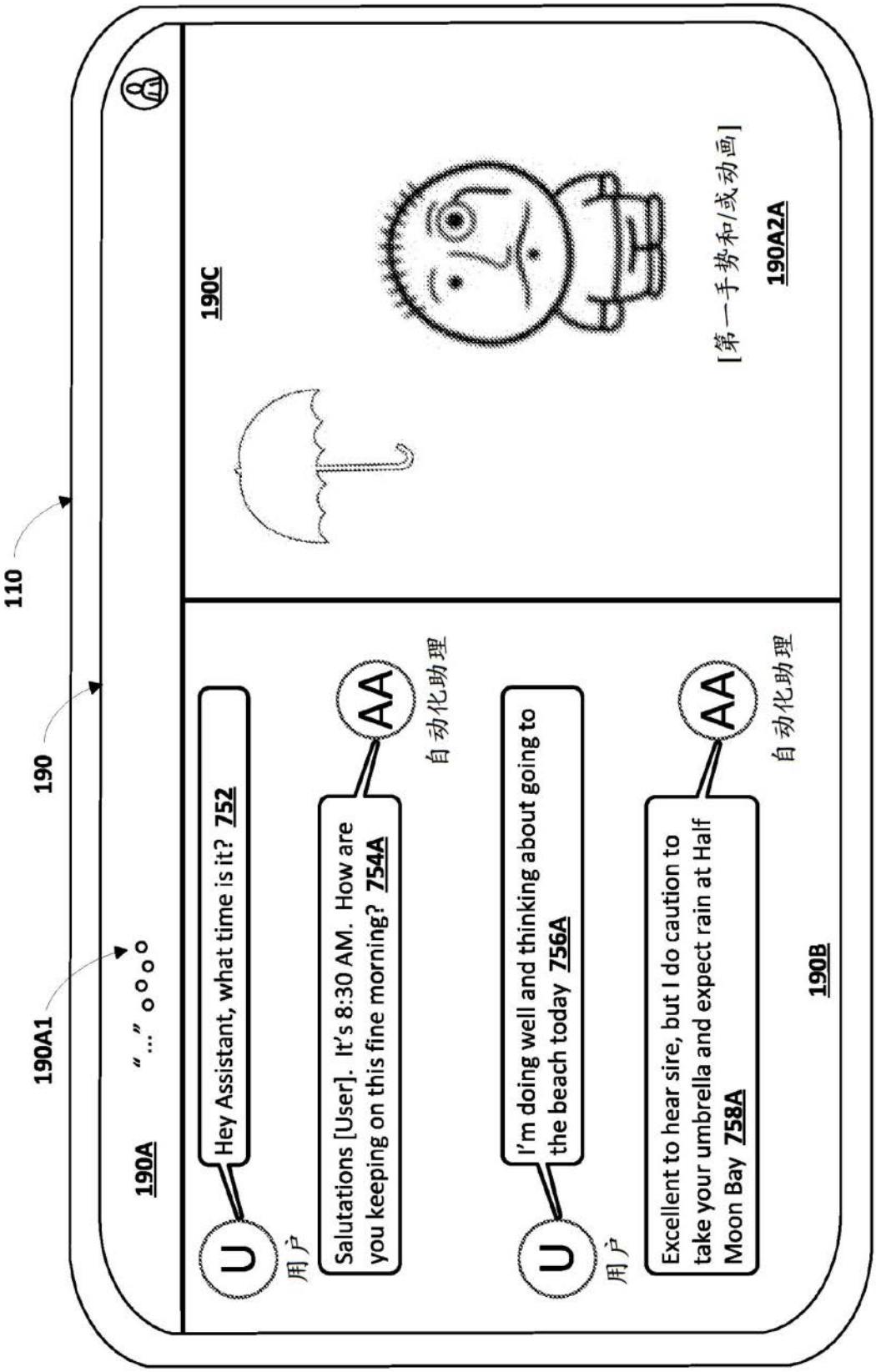

110
190
190A1
190A
"..."
Hey Assistant, what time is it? 752
U
用户
Salutations [User]. It's 8:30 AM. How are you keeping on this fine morning? 754A
AA
自动化助理
I'm doing well and thinking about going to the beach today 756A
U
用户
Excellent to hear sire, but I do caution to take your umbrella and expect rain at Half Moon Bay 758A
AA
自动化助理
190B
190C
[第一手势和/或动画]
190A2A

图7A

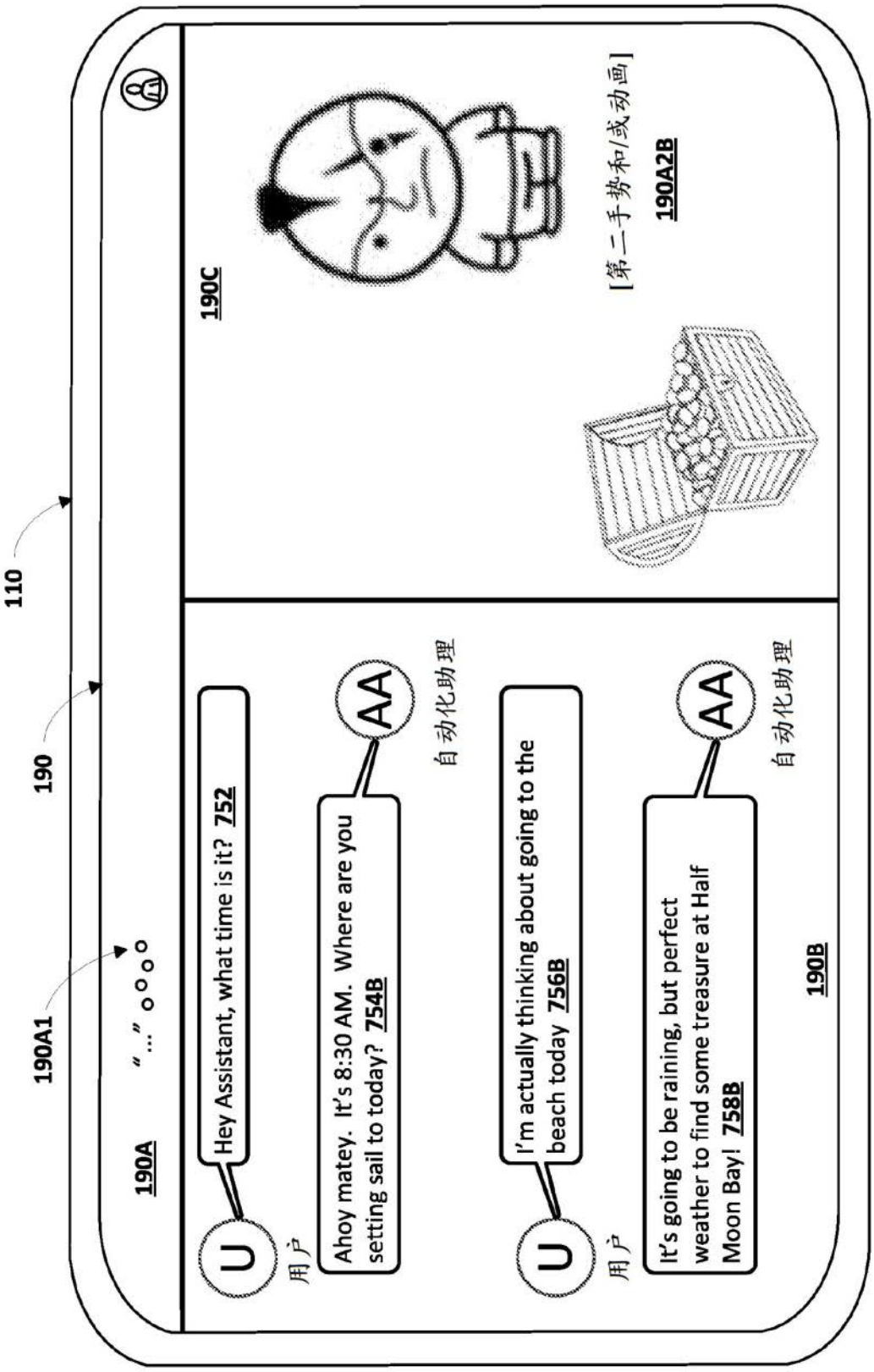

110
190
190A1
190A
"..."
190C
[第二手势和/或动画]
190A2B
U
用户
Hey Assistant, what time is it? 752
AA
自动化助理
Ahoy matey. It's 8:30 AM. Where are you setting sail to today? 754B
U
用户
I'm actually thinking about going to the beach today 756B
AA
自动化助理
It's going to be raining, but perfect weather to find some treasure at Half Moon Bay! 758B
190B

图7B

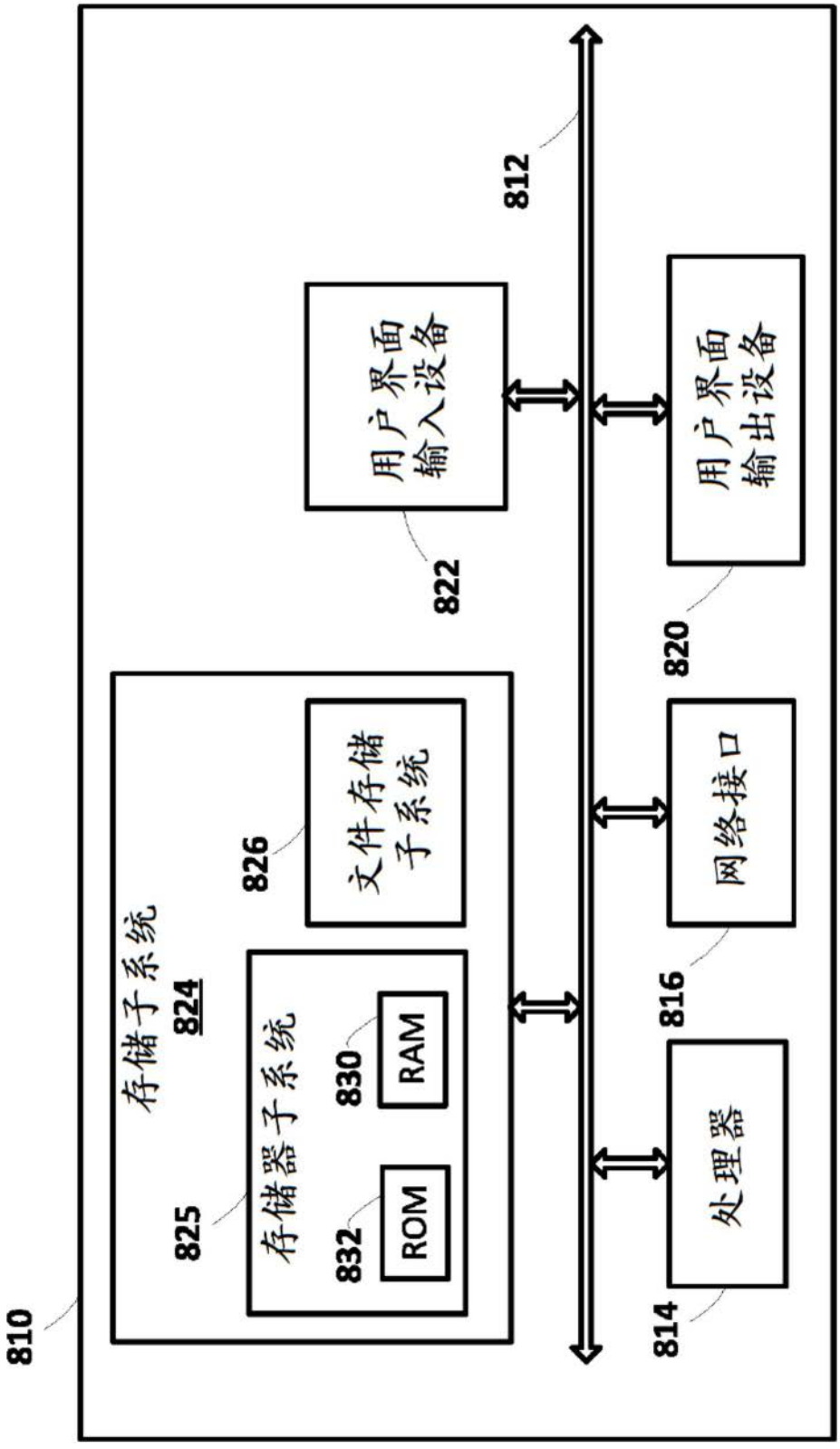

810
存储子系统
824
825
存储器子系统
830
RAM
832
ROM
826
文件存储子系统
822
用户界面输入设备
812
820
用户界面输出设备
816
网络接口
814
处理器

图8