(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104781782 A
(43)申请公布日 2015.07.15

(21)申请号 201380057286.0
(22)申请日 2013.10.01
(30)优先权数据
2012-246118 2012.11.08 JP
(85)PCT国际申请进入国家阶段日
2015.04.30
(86)PCT国际申请的申请数据
PCT/JP2013/005859 2013.10.01
(87)PCT国际申请的公布数据
WO2014/073149 EN 2014.05.15
(71)申请人 索尼公司
地址 日本东京都
(72)发明人 大村淳己 河野道成 池田卓郎 冈田宪一
(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 王萍 陈炜
(51)Int.Cl.
G06F 3/16(2006.01)
G06F 3/01(2006.01)
G06N 3/00(2006.01)

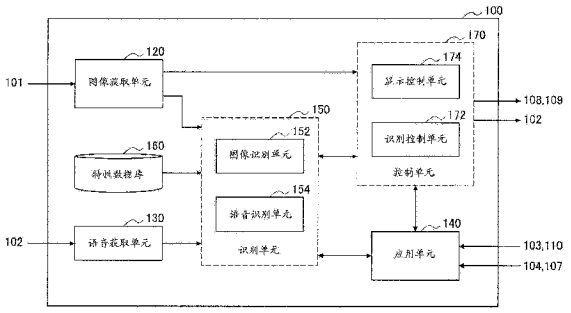
权利要求书2页 说明书19页 附图21页

(54)发明名称

信息处理设备、信息处理方法和程序

(57)摘要

一种信息处理包括：处理电路，被配置成生成用于控制显示装置的数据，以将与语音输入相关联的控制对象重叠在显示图像上，其中，显示图像是用户执行的姿势操作的反馈图像，并且显示图像是从摄像装置捕获图像得到的图像。



CN 104781782 A

1. 一种信息处理系统，包括：

处理电路，被配置成生成用于控制显示装置的数据，以在显示图像上重叠与语音输入相关联的控制对象，其中，所述显示图像是用户执行的姿势操作的反馈图像，并且所述显示图像是从摄像装置捕获图像获得的图像。

2. 根据权利要求 1 所述的信息处理系统，其中，

所述显示图像是所述用户的镜像图像。

3. 根据权利要求 1 所述的信息处理系统，其中，

所述处理电路被配置成基于所述控制对象与所述反馈图像中的用户的身体部位之间的显示位置关系而发起由语音输入触发的处理。

4. 根据权利要求 3 所述的信息处理系统，其中，

当所述显示位置关系包括所述控制对象在距所述反馈图像中的用户的身体部位预定距离内时，所述处理电路发起所述处理。

5. 根据权利要求 3 所述的信息处理系统，其中，

所述反馈图像中的所述用户的身体部位是所述用户的面部的至少一部分。

6. 根据权利要求 3 所述的信息处理系统，其中，

当所述显示位置关系包括所述反馈图像中的所述用户的身体部位在距所述控制对象的预定方向内时，所述处理电路发起所述处理。

7. 根据权利要求 3 所述的信息处理系统，其中，

所述处理电路被配置成控制所述显示装置以改变所述控制对象的图像属性，以指示所述处理电路发起了所述处理。

8. 根据权利要求 3 所述的信息处理系统，其中，

所述处理是语音识别处理。

9. 根据权利要求 1 所述的信息处理系统，其中，

所述处理电路被配置成控制所述显示装置以响应于所述用户执行的姿势操作而改变所述控制对象的显示位置。

10. 根据权利要求 1 所述的信息处理系统，其中，

所述处理电路被配置成控制所述显示装置以显示根据所述语音输入的检测状态而改变外观的指示符。

11. 根据权利要求 1 所述的信息处理系统，其中，

所述处理电路被配置成控制所述显示装置以与所述反馈图像同步地显示附加对象，所述附加对象与所述语音输入相关联并且与所述控制对象不同。

12. 根据权利要求 11 所述的信息处理系统，其中，

所述附加对象是基于所述语音输入生成的文本信息。

13. 根据权利要求 11 所述的信息处理系统，其中，

所述附加对象指示所述语音输入的音量水平。

14. 根据权利要求 8 所述的信息处理系统，其中，

所述处理电路被配置成基于所述语音识别而控制装置的功能。

15. 根据权利要求 14 所述的信息处理系统，其中，

所述装置被配置成控制内容的再现，并且所述处理电路被配置成控制所述显示装置以

同时显示所述反馈图像、所述控制对象和作为所述语音输入的对象的所述内容的图像。

16. 根据权利要求 1 所述的信息处理系统，还包括 ：

所述显示装置，其中，所述显示装置和所述处理电路是单个设备的部件。

17. 根据权利要求 3 所述的信息处理系统，其中，

所述镜像图像是所述用户的实际图像。

18. 根据权利要求 3 所述的信息处理系统，其中，

所述镜像图像是所述用户的体现。

19. 一种信息处理方法，包括 ：

利用处理电路生成用于控制显示装置的数据，以在显示图像上重叠与语音输入相关联的控制对象，其中，所述显示图像是用户执行的姿势操作的反馈图像，并且所述显示图像是从摄像装置捕获图像得到的图像。

20. 一种存储有计算机可读指令的非暂态计算机可读存储介质，所述计算机可读指令当由处理电路执行时执行信息处理方法，所述方法包括 ：

利用所述处理电路生成用于控制显示装置的数据，以在显示图像上重叠与语音输入相关联的控制对象，其中，所述显示图像是用户执行的姿势操作的反馈图像，并且所述显示图像是从摄像装置捕获图像得到的图像。

信息处理设备、信息处理方法和程序

技术领域

[0001] 本公开涉及一种信息处理设备、信息处理方法和程序。

[0002] 本公开包含与2012年11月8日向日本专利局提交的日本优先权专利申请JP 2012-246118中公开的主题相关的主题，其全部内容通过引用合并于此。

背景技术

[0003] 在过去，语音识别已被用作当输入到信息设施时辅助用户的技术。作为一个示例，JP 2012-58838公开了使用语音识别将用户产生的语音样本的内容转换为文本以及在用于多个用户之间的通信的屏幕上显示所获得的文本的技术。

[0004] 引用列表

[0005] 专利文献

[0006] PTL 1：JP 2012-58838A

发明内容

[0007] 技术问题

[0008] 然而，在许多情况下，在语音识别起作用并且语音输入有效的定时与用户产生用于语音识别的语音样本的定时之间存在偏差。如果这样的定时不匹配，则可能出现问题，诸如没有对期望的语音样本执行语音识别或者对不期望的语音样本执行语音识别。

[0009] 因此，期望提供一种新颖且改进的构架，其辅助用户以适当的定时产生用于语音识别的语音样本。

[0010] 问题的解决方案

[0011] 根据一个实施例，描述了一种信息处理系统，其包括处理电路，该处理电路被配置成生成用于控制显示装置的数据，以将与语音输入相关联的控制对象重叠在显示图像上，其中，显示图像是用户执行的姿势操作的反馈图像，并且显示图像是从摄像装置捕获图像获得的图像。

[0012] 根据另一实施例，一种信息处理方法包括：利用处理电路生成用于控制显示装置的数据，以将与语音输入相关联的控制对象重叠在显示图像上，其中，显示图像是用户执行的姿势操作的反馈图像，并且显示图像是从摄像装置捕获图像获得的图像。

[0013] 根据另一实施例，描述了一种存储有计算机可读指令的非暂态计算机可读存储介质，该计算机可读指令在被处理电路执行时执行信息处理方法，该方法包括：利用处理电路生成用于控制显示装置的数据，以将与语音输入相关联的控制对象重叠在显示图像上，其中，显示图像是用户执行的姿势操作的反馈图像，并且显示图像是从摄像装置捕获图像获得的图像。

[0014] 本发明的有利效果

[0015] 根据本公开的以上实施例，可以辅助用户以适当的定时产生用于语音识别的语音样本。

附图说明

[0016] 图 1 是用于说明根据本公开的第一实施例的信息处理设备的概况的图。

[0017] 图 2 是用于说明根据本公开的第二实施例的信息处理设备的概况的图。

[0018] 图 3 是示出根据第一实施例的信息处理设备的示例硬件配置的框图。

[0019] 图 4 是示出根据第一实施例的信息处理设备的逻辑功能的示例配置的框图。

[0020] 图 5 是用于说明图像识别的结果的一个示例的图。

[0021] 图 6 是用于说明图像识别的结果的另一示例的图。

[0022] 图 7 是用于说明用于控制语音识别的控制对象的第一示例的图。

[0023] 图 8 是用于说明用于控制语音识别的控制对象的第二示例的图。

[0024] 图 9 是用于说明用于激活语音输入的激活条件的第一示例的图。

[0025] 图 10 是用于说明用于激活语音输入的激活条件的第二示例的图。

[0026] 图 11 是用于说明语音识别结果的视觉反馈的一个示例的图。

[0027] 图 12 是用于说明表示语音样本的识别内容的附加显示对象的示例的第一图。

[0028] 图 13 是用于说明表示语音样本的识别内容的附加显示对象的示例的第二图。

[0029] 图 14 是用于说明辅助语音识别的附加显示对象的示例的图。

[0030] 图 15 是用于说明对麦克风的方向性的控制的示例的第一图。

[0031] 图 16 是用于说明对麦克风的方向性的控制的示例的第二图。

[0032] 图 17 是用于说明对麦克风的方向性的控制的示例的第三图。

[0033] 图 18 是用于说明输出图像的窗口布局的第一示例的图。

[0034] 图 19 是用于说明输出图像的窗口布局的第二示例的图。

[0035] 图 20 是用于说明第一控制场景的图。

[0036] 图 21 是用于说明第二控制场景的图。

[0037] 图 22 是用于说明第三控制场景的图。

[0038] 图 23 是用于说明第四控制场景的图。

[0039] 图 24 是示出根据第一实施例的处理流程示例的流程图的前半部。

[0040] 图 25 是示出根据第一实施例的处理流程示例的流程图的后半部。

[0041] 图 26 是示出根据第二实施例的信息处理设备的示例硬件配置的框图。

[0042] 图 27 是用于说明第二实施例中的控制场景的示例的图。

具体实施方式

[0043] 在下文中，将参照附图详细描述本公开的优选实施例。注意，在该说明书和附图中，具有基本上相同的功能和结构的结构元件以相同的附图标记来表示，并且省略对这些结构元件的重复说明。

[0044] 按以下示出的顺序来给出以下描述。

[0045] 1. 概况

[0046] 2. 第一实施例

[0047] 2-1. 示例硬件配置

[0048] 2-2. 示例功能配置

[0049] 2-3. 示例控制场景

[0050] 2-4. 示例处理流程

[0051] 3. 第二实施例

[0052] 4. 结论

[0053] <1. 概况 >

[0054] 首先，将参照图 1 和图 2 描述可以应用根据本公开的实施例的技术的信息处理设备的概况。根据本公开的实施例的技术可以应用于使用语音识别作为用户界面的一部分的多种设备和系统。作为示例，根据本公开的实施例的技术可以应用于诸如电视机设备、数字照相机或者数字摄像机的数字家庭设施。根据本公开的实施例的技术还可以应用于诸如 PC（个人计算机）、智能电话、PDA（个人数字助理）或游戏控制台的终端设备。根据本公开的实施例的技术也可以应用于诸如卡拉 OK 系统或娱乐设备的专用系统或设备。

[0055] 图 1 是用于说明根据本公开的第一实施例的信息处理设备 100 的概况的图。如图 1 所示，信息处理设备 100 是电视机设备。信息处理设备 100 包括摄像装置 101、麦克风 102 和显示器 108。摄像装置 101 拍摄正观看信息处理设备 100 的显示器 108 的用户的图像。麦克风 102 获得这样的用户产生的语音样本。显示器 108 显示信息处理设备 100 生成的图像。除内容图像之外，显示器 108 显示的图像可包括用户界面 (UI) 图像。在图 1 的示例中，用户 Ua 和 Ub 正观看显示器 108。UI 图像 W01 显示在显示器 108 上。UI 图像 W01 是使用由摄像装置 101 拍摄的拍摄图像而生成的，并且实现了可以是用户的实际图像或用户的体现 (avatar) 的所谓“镜像图像”显示。信息处理设备 100 具有语音识别功能。通过经由麦克风 102 对信息处理设备 100 进行语音输入，用户 Ua 和 Ub 能够操作信息处理设备 100 或者将信息输入到信息处理设备 100 中。

[0056] 图 2 是用于说明根据本公开的第二实施例的信息处理设备 200 的概况的图。如图 2 所示，信息处理设备 200 是平板 PC。信息处理设备 200 包括摄像装置 201、麦克风 202 和显示器 208。摄像装置 201 拍摄正观看信息处理设备 200 的显示器 208 的用户的图像。麦克风 202 获得这样的用户产生的语音样本。显示器 208 显示信息处理设备 200 生成的图像。除内容图像之外，显示器 208 显示的图像可包括用户界面 (UI) 图像。在图 2 的示例中，用户 Uc 正观看显示器 208。UI 图像 W02 显示在显示器 208 上。UI 图像 W02 是使用摄像装置 201 拍摄的拍摄图像而生成的并且实现了所谓的“镜像图像”显示。信息处理设备 200 具有语音识别功能。通过经由麦克风 202 对信息处理设备 200 进行语音输入，用户 Uc 能够操作信息处理设备 200 或者将信息输入到信息处理设备 200 中。

[0057] 对于这样的设备，在语音识别功能正工作并且语音输入有效时，不保证用户所说的任何事（即，不是每一个语音样本）都旨在被用于语音识别。还存在当语音输入无效时用户产生旨在用于语音识别的语音样本的可能性。这样的定时偏差可能为用户导致问题，诸如，不旨在用于语音识别的语音样本接受语音识别或者语音识别不成功。为此，信息处理设备 100 或 200 根据在以下章节详细描述的构架而辅助用户以适当的定时产生旨在用于语音识别的语音样本。

[0058] <2. 第一实施例 >

[0059] <2-1. 示例硬件配置 >

[0060] 图 3 是示出可在单个设备中实现的或者在多个单元的分布资源中实现的信息处

理设备 100 的示例硬件配置的框图。如图 3 所示，信息处理设备 100 包括摄像装置 101、麦克风 102、输入装置 103、通信接口 (I/F)104、存储器 105、调谐器 106、解码器 107、显示器 108、扬声器 109、远程控制 I/F 110、总线 111 和处理器 112（作为处理电路的一个示例，诸如 CPU）。

[0061] (1) 摄像装置

[0062] 摄像装置 101 包括诸如 CCD（电荷耦合器件）或 CMOS（互补金属氧化物半导体）的图像拍摄元件并且拍摄图像。摄像装置 101 拍摄的图像（构成视频的帧）被视为用于信息处理设备 100 的处理的输入图像。

[0063] (2) 麦克风

[0064] 麦克风 102 获得用户产生的语音样本并且生成语音信号。麦克风 102 生成的语音信号被视为旨在用于信息处理设备 100 的语音识别的输入语音。麦克风 102 可以是全方向麦克风或者具有固定的或可变的方向性的麦克风。在其它场景中，麦克风 102 具有可变的方向性并且使得其方向性被动态地控制。

[0065] (3) 输入装置

[0066] 输入装置 103 是用户用于直接操作信息处理设备 100 的装置。作为示例，输入装置 103 可包括布置在信息处理设备 100 的壳体上的按钮、开关、拨号盘等。在检测到用户输入时，输入装置 103 生成与所检测的用户输入对应的输入信号。

[0067] (4) 通信接口

[0068] 通信 I/F 104 用作信息处理设备 100 与其它设备之间的通信的媒介。通信 I/F 104 支持任意无线通信协议或者有线通信协议，并且建立与其它设备的通信连接。

[0069] (5) 存储器

[0070] 存储器 105 由诸如半导体存储器或硬盘驱动器的存储介质构成，并且存储用于信息处理设备 100 的处理的程序和数据以及内容数据。作为一个示例，存储器 105 存储的数据可包括用于稍后描述的图像识别和语音识别的特性数据。注意，本说明书中描述的程序和数据的一些或全部可不由存储器 105 来存储，而是替代地可从外部数据源（作为示例，数据服务器、网络存储装置或者外接存储器）来获取。

[0071] (6) 调谐器

[0072] 调谐器 106 从经由天线（未示出）接收的广播信号提取期望信道上的内容信号并进行解调。调谐器 106 然后将解调后的内容信号输出到解码器 107。

[0073] (7) 解码器

[0074] 解码器 107 根据从调谐器 106 输入的内容信号解码内容数据。解码器 107 可根据经由通信 I/F 104 接收的内容信号而解码内容数据。内容图像可基于由解码器 107 解码的内容数据来生成。

[0075] (8) 显示器

[0076] 显示器 108 具有由 LCD（液晶显示器）、OLED（有机发光二极管）、CRT（阴极射线管）等构成的屏幕，并且显示信息处理设备 100 生成的图像。作为示例，参照图 1 和图 2 描述的内容图像和 UI 图像可显示在显示器 108 的屏幕上。

[0077] (9) 扬声器

[0078] 扬声器 109 具有振动膜和诸如放大器的电路元件，并且基于信息处理设备 100 生

成的输出语音信号而输出音频。扬声器 109 的音量是可变的。

[0079] (10) 远程控制接口

[0080] 远程控制 I/F 110 是接收从用户使用的遥控器传送的远程控制信号（红外信号或其它无线信号）的接口。在检测到远程控制信号时，远程控制 I/F 110 生成与所检测的远程控制信号对应的输入信号。

[0081] (11) 总线

[0082] 总线 111 将摄像装置 101、麦克风 102、输入装置 103、通信 I/F 104、存储器 105、调谐器 106、解码器 107、显示器 108、扬声器 109、远程控制 I/F 110 和处理器 112 彼此连接。

[0083] (12) 处理器

[0084] 作为示例，处理器 112 可以是 CPU（中央处理单元）或 DSP（数字信号处理器）。通过执行存储在存储器 105 或其它存储介质中的程序，处理器 112 使得信息处理设备 100 以如稍后描述的各种方式起作用。

[0085] <2-2. 示例功能配置>

[0086] 图 4 是示出由图 3 所示的信息处理设备 100 的存储器 105 和处理器 112 实现的逻辑功能的示例配置的框图。如图 4 所示，信息处理设备 100 包括图像获取单元 120、语音获取单元 130、应用单元 140、识别单元 150、特性数据库 (DB)160 和控制单元 170。识别单元 150 包括图像识别单元 152 和语音识别单元 154。控制单元 170 包括识别控制单元 172 和显示控制单元 174。注意，图 4 所示的功能块的一些可由信息处理设备 100 外部的设备（诸如云计算环境中的设备）来实现。作为一个示例，取代由自身来执行以下描述的图像识别处理，图像识别单元 152 可使得这样的处理由外部图像识别功能来执行。以相同的方式，取代由自身来执行以下描述的语音识别功能，语音识别单元 154 可使得这样的处理由外部语音识别功能来执行。

[0087] (1) 图像获取单元

[0088] 图像获取单元 120 获取摄像装置 101 拍摄的图像作为输入图像。输入图像通常是构成用户出现的视频的一系列帧中的单个帧。图像获取单元 120 然后将所获取的输入图像输出到识别单元 150 和控制单元 170。

[0089] (2) 语音获取单元

[0090] 语音获取单元 130 获取麦克风 102 生成的语音信号作为输入语音。语音获取单元 130 然后将所获取的输入语音输出到识别单元 150。

[0091] (3) 应用单元

[0092] 应用单元 140 执行信息处理设备 100 的各种应用功能。作为示例，可由应用单元 140 来执行电视节目再现功能、电子节目指南显示功能、记录设置功能、照片再现功能、视频再现功能、音乐再现功能和因特网浏览功能。应用单元 140 将经由应用功能生成的应用图像（可包括内容图像）和音频输出到控制单元 170。

[0093] 在本实施例中，应用单元 140 执行的应用功能的至少一部分与稍后描述的语音识别单元 154 一致地进行工作，并且接收来自用户的语音输入。作为一个示例，电视节目再现功能可根据语音识别单元 154 识别的语音命令而改变设置，诸如要再现的频道和音量。电子节目指南显示功能可根据语音识别单元 154 识别的语音命令而改变要显示的电子节目指南的频道或时段。照片再现功能可再现在语音识别单元 154 识别的指定日期拍摄的照

片。因特网浏览功能可使用语音识别单元 154 识别的关键词进行因特网搜索。

[0094] (4) 图像识别单元

[0095] 图像识别单元 152 识别出现在从图像获取单元 120 输入的输入图像中的用户的身体。作为一个示例，通过针对用户身体的特定部位使得从输入图像提取的图像特性值与特性 DB 160 预先存储的图像特性值进行匹配，识别这样的特定部位。作为示例，“特定部位”可包括用户的手、嘴和面部中的至少一个。

[0096] 图 5 是用于说明图像识别单元 152 的图像识别的结果的一个示例的图。如图 5 所示，用户 Ua 出现在输入图像 W03 中。用户 Ua 正面向摄像装置 101 并且抬起他的左手。通过匹配图像特征值或者使用其它已知方法，图像识别单元 152 能够识别输入图像 W03 中的手区域 A01、嘴区域 A02 和面部区域 A03。图像识别单元 152 然后将示出这样识别的区域的位置的位置数据输出到控制单元 170。

[0097] 作为一个示例，图像识别单元 152 可通过使得在输入图像内识别的面部区域的图像部分（面部图像）与特性 DB 160 预先存储的已知用户的面部图像数据进行匹配来标识用户。作为示例，图像识别单元 152 产生的用户标识结果可以用于对语音识别进行调整，对显示在 UI 图像中的菜单进行个性化，或者由应用单元 140 用于推荐内容。注意，用户的标识（即，个人识别）可基于输入语音而不是基于输入图像来执行。

[0098] 在本实施例中，图像识别单元 152 还可以识别出现在输入图像中的用户的姿势。注意，在本说明书中，表述“姿势”假设还包括不涉及用户身体的动态移动的所谓“姿态”（形式）。

[0099] 图 6 是用于说明图像识别单元 152 的图像识别的结果的另一示例的图。如图 6 所示，用户 Ua 和 Ub 出现在输入图像 W04 中。用户 Ua 正通过将他的右手的食指放在他的嘴上而做出姿势。图像识别单元 152 能够识别输入图像 W04 中的手区域 A04 并且还能够识别用户 Ua 做出的这样的姿势。用户 Ub 正通过用他的双手盖住他的嘴而做出姿势。图像识别单元 152 能够识别输入图像 W04 中的手区域 A05 并且还能够识别用户 Ub 做出的这样的姿势。在识别出用户的姿势时，图像识别单元 152 将示出所识别的姿势的类型的姿势数据输出到控制单元 170。

[0100] (5) 语音识别单元

[0101] 语音识别单元 154 基于从语音获取单元 130 输入的输入语音而对用户输入的语音样本执行语音识别。在本实施例中，从语音获取单元 130 到语音识别单元 154 的语音输入由识别控制单元 172 来激活或去激活。当语音输入有效时，语音识别单元 154 将输入语音转换为示出输入语音的内容的文本。如果正执行的应用接收到自由文本的输入，则语音识别单元 154 可将示出经过了语音识别的语音样本的内容的文本输出到应用单元 140。替选地，如果正执行的应用接收到语音命令的特定集合中的语音命令的输入，则语音识别单元 154 可将标识从用户的语音样本识别的语音命令的标识符输出到应用单元 140。当语音输入无效时，语音识别单元 154 不执行语音识别。

[0102] 语音识别单元 154 还可确定从语音获取单元 130 输入的输入语音的水平以及向控制单元 170 通知所确定的水平。稍后描述的识别控制单元 172 能够根据语音识别单元 154 指示的输入语音的水平而在屏幕上向用户给出各种反馈。

[0103] 如之前所述，在给定场景中，麦克风 102 具有可变方向性。在该情况下，麦克风 102

的方向性由稍后描述的识别控制单元 172 来设置。语音识别单元 154 然后使用麦克风 102 获取的语音信号，对位于与所设置的方向性对应的方向上的用户的语音样本执行语音识别。

[0104] (6) 特性 DB

[0105] 特性 DB 160 预先存储要用在图像识别单元 152 的图像识别中的图像特性数据和要用在语音识别单元 154 的语音识别中的语音特性数据。作为一个示例，图像特性数据可包括对于用户的特定部位（诸如手、嘴或面部）的已知图像特性值。图像特性数据还可包括对于每个用户的面部图像数据。图像特性数据还可包括定义要由图像识别单元 152 识别的姿势的姿势定义数据。语音特性数据可包括例如示出各个用户的说话特性的语音特性值。

[0106] (7) 识别控制单元

[0107] 识别控制单元 172 生成与说话相关并且要重叠在输入图像上的对象。识别控制单元 172 使用所生成的对象来控制语音识别单元 154 执行的语音识别。用于控制语音识别的这样的对象在下文中称为“控制对象”。控制对象可根据用户的操作而在屏幕上移动或者可显示在固定位置。

[0108] 图 7 是用于说明控制对象的第一示例的图。如图 7 所示，控制对象 IC1 重叠在输入图像 W05 上。控制对象 IC1 是类似手持麦克风的图标。作为一个示例，当启动了从用户接收语音输入的应用（下文中称为“语音兼容应用”）时，识别控制单元 172 使得控制对象 IC1 显示在屏幕上的指定显示位置处或者图像识别单元 152 识别的用户的身体附近。识别控制单元 172 然后根据用户的移动（例如，手区域的移动）而改变控制对象 IC1 的显示位置。识别控制单元 172 可根据用户的移动（例如，手区域的旋转）而改变控制对象 IC1 的朝向。当语音兼容应用结束时，控制对象 IC1 可从屏幕被删除或者被去激活并被移动到屏幕的边缘部分或默认显示位置。

[0109] 图 8 是用于说明控制对象的第二示例的图。如图 8 所示，控制对象 IC2 重叠在输入图像 W06 上。控制对象 IC2 是类似直立型麦克风的图标。作为一个示例，当启动了语音兼容应用时，识别控制单元 172 使得控制对象 IC2 显示在屏幕上的默认显示位置处。控制对象 IC2 的显示位置不移动。当语音兼容应用结束时，控制对象 IC2 可从屏幕被删除。

[0110] 注意，图 7 和图 8 所示的控制对象 IC1 和 IC2 仅是示例。作为示例，类似嘴或扩音器或文本标签的其它类型图标可用作控制对象。另外，取代控制对象的外观，控制对象的功能可与说话有关。

[0111] 在本实施例中，识别控制单元 172 基于控制对象与图像识别单元 152 识别的用户身体的特定部位之间在屏幕上的位置关系，控制语音识别单元 154 执行的语音识别。作为一个示例，如果基于这样的位置关系满足激活条件，则识别控制单元 172 激活到语音识别单元 154 的语音输入。如果不满足激活条件，则识别控制单元 172 不激活到语音识别单元 154 的语音输入。

[0112] 图 9 是用于说明用于激活语音输入的激活条件的第一示例的图。如图 9 所示，用户 Ua 出现在输入图像 W07a 和 W07b 中。图像识别单元 152 识别出现在输入图像中的用户的嘴区域和手区域。在该第一示例中，激活条件是用户的嘴与控制对象之间的距离短于距离阈值 D1 的条件。在图中，以嘴区域的中心点 G2 为中心并且其半径等于距离阈值 D1 的圆以虚线示出。识别控制单元 172 根据所识别的手区域 A01 的移动而在屏幕上移动控制对象

IC1。在图 9 的上部，由于用户的嘴与控制对象 IC1 之间的距离大于距离阈值 D1，因此语音输入无效。即，即使用户产生了语音样本（或者在附近产生了噪声），语音识别单元 154 也不会执行语音识别。因此，在这样的时间期间，防止了由于用户不期望的语音识别而导致的应用的意外操作。在图 9 的下部，作为用户移动其手的结果，用户的嘴与控制对象 IC1 之间的距离短于距离阈值 D1。为此，识别控制单元 172 确定满足激活条件并且激活语音输入。结果，用户产生的语音样本接受语音识别单元 154 的语音识别。注意，除了嘴之外的用户身体的部位与控制对象之间的距离可与上述距离阈值进行比较。

[0113] 图 10 是用于说明用于激活语音输入的激活条件的第二示例的图。如图 10 所示，用户 Ub 出现在输入图像 W08a 和 W08b 中。控制对象 IC2 也重叠在输入图像 W08a 和 W08b 上。图像识别单元 152 识别出现在输入图像中的用户的嘴区域 A06。在该第二示例中，激活条件是用户的嘴与控制对象之间的距离短于距离阈值 D2 的条件。在图中，以控制对象的中心点 G2 为中心并且其半径等于距离阈值 D2 的圆以虚线示出。在图 10 的上部，由于用户的嘴与控制对象 IC2 之间的距离大于距离阈值 D2，因此语音输入无效。即，即使用户产生语音样本（或者在附近产生了噪声），语音识别单元 154 也不会执行语音识别。因此，在这样的时间期间，防止了由于用户不期望的语音识别而导致的应用的意外操作。在图 10 的下部，作为用户移动的结果，用户的嘴与控制对象 IC2 之间的距离短于距离阈值 D2。为此，识别控制单元 172 确定满足激活条件并且激活语音输入。结果，用户产生的语音样本接受语音识别单元 154 的语音识别。

[0114] 注意，参照图 9 和图 10 描述的激活条件仅是示例。作为另一示例，可以将对与控制对象有关的特定姿势的检测（诸如触摸控制对象或者向上抬高控制对象）作为激活条件。

[0115] 一旦激活了语音输入，识别控制单元 172 使得维持语音输入的有效状态直到满足特定去激活条件为止。作为一个示例，去激活条件可以是上述激活条件的简单相反条件（例如，用户的嘴与控制对象之间的距离超过距离阈值）。替选地，去激活条件可以是图像识别单元 152 对用户的特定姿势的识别等。作为一个示例，用于去激活语音输入的姿势可以是用户用他的 / 她的食指触摸他的 / 她的嘴的姿势。去激活条件还可包括单个语音命令的成功识别或者从激活开始过去特定时间长度。

[0116] 在语音输入有效时，识别控制单元 172 还控制与语音识别单元 154 的语音识别有关的对用户的视觉反馈。

[0117] 作为一个示例，识别控制单元 172 通过改变控制对象的显示属性而向用户通知到语音识别单元 154 的语音输入已被激活。作为示例，识别控制单元 172 改变的控制对象的显示属性可包括颜色、亮度、透明度、大小、形状和纹理中的至少一个。在图 9 和图 10 的示例中，通过改变控制对象的纹理来将语音输入示出为有效或无效。

[0118] 作为另一示例，识别控制单元 172 向用户给出关于语音识别单元 153 指示的输入语音的水平的反馈。可通过改变控制对象的属性或者通过改变重叠了控制对象的 UI 图像的状态来给出关于输入语音的水平的反馈。图 11 是用于说明语音识别结果的视觉反馈的一个示例的图。如图 11 所示，效果 Fb1 被施加于重叠了控制对象 IC1 的 UI 图像 W09。效果 Fb1 表示波看起来从控制对象 IC1（可以是用户的嘴）发出的 UI 图像的状态。如果输入语音的水平下降到特定阈值以下，则可移除这样的效果 Fb1。通过提供这样的反馈，用户可以直观地了解用户他 / 她自身产生的语音样本是否正适当地被信息处理设备 100 检测。识别

控制单元 172 可根据超过上述特定阈值的输入语音的水平而改变控制对象的显示属性的变化水平或者输出图像的状态变化。作为一个示例，效果 Fb1 可被施加于其大小随着输入语音的水平增加而增加的图像区域。这样，用户可以直观地了解信息处理设备 100 针对用户他 / 她自身产生的语音样本所检测的水平。注意，识别控制单元 172 可改变效果 Fb1 的显示属性（例如，颜色），以便指示语音识别的状态以及错误的存在与否。将输入语音的水平与特定标准值进行比较的结果可被显示为 UI 图像 W09 中的文本。

[0119] 另外，作为一个示例，识别控制单元 172 可在出现在输入图像中的用户附近重叠包括表示语音识别单元 154 识别的语音样本的内容的文本的附加显示对象。图 12 和图 13 是用于说明表示语音样本的识别内容的附加显示对象的示例的图。如图 12 所示，控制对象 IC1 和附加对象 Fb2 重叠在 UI 图像 W10 上。附加对象 Fb2 是包括表示出现在 UI 图像 W10 中的用户 Ua 产生的语音样本的内容的文本的说话气泡。通过接收这样的反馈，用户可以立即了解信息处理设备 100 是否已正确地识别了用户产生的语音样本。如图 13 所示，附加对象 Fb2 包括随机字符串 Str1。当检测到超过特定阈值的水平的输入语音但是基于这样的输入语音的语音识别不成功时，随机字符串 Str1 可被插入到附加对象 Fb2 中。通过接收这样的反馈，用户可以立即了解虽然用户他 / 她自身产生的语音样本的水平足够但是语音识别不成功。可通过改变附加对象 Fb2 的显示属性来向用户通知语音识别不成功。注意，附加对象 Fb2 可取代随机字符串而包括空格。随机字符串或空格的长度可根据其语音识别不成功的说话的长度（时间长度）来确定。

[0120] 作为另一示例，识别控制单元 172 可重叠示出语音识别单元 154 检测的语音的水平以及用于有效地执行语音识别所需的语音水平的附加对象。用于有效地执行语音识别所需的语音水平可由存储器 105 预先存储或者可动态地计算以便根据环境中的噪声水平。图 14 是用于说明辅助语音识别的附加显示对象的示例的图。如图 14 所示，控制对象 IC1、附加对象 Fb2 和附加对象 Fb3 重叠在 UI 图像 W12 上。附加对象 Fb2 是包括表示说话的内容的文本的说话气泡。这里，作为由于用户语音的水平不足而导致语音识别不成功的结果，附加对象 Fb2 的背景颜色被改变为暗色。附加对象 Fb3 是指示语音水平的指示符。在附加对象 Fb3 的外侧以虚线绘制的圆的半径对应于有效地执行语音识别所需的语音水平。有色圆的半径对应于语音识别单元 154 指示的输入语音的水平。如果输入语音的水平增加，则有色圆增大。注意，附加对象 Fb3 不限于图 14 中的示例，并且可以是例如条形指示符。通过接收这样的反馈，用户可以直观地了解当用户产生的语音样本的水平不足时，他的 / 她的语音应该提高多少以使得语音识别能够成功。注意，识别控制单元 172 可改变附加对象 Fb3 的显示属性（例如，颜色）以指示语音识别的状态或者错误的存在与否。将输入语音的水平与特定标准值进行比较的结果可被显示为 UI 图像 W12 中的文本。

[0121] 如果麦克风 102 具有可变方向性，则识别控制单元 172 可通过使用控制对象设置麦克风 102 的方向性来改进语音识别的精度。作为一个示例，识别控制单元 172 可根据控制对象在屏幕上的位置，设置麦克风 102 的方向性。另外，识别控制单元 172 可根据控制对象在屏幕上的朝向来设置麦克风 102 的方向性。

[0122] 图 15 至图 17 是用于说明对麦克风的方向性的控制的示例的图。在图 15 的上部，控制对象 IC1 重叠在 UI 图像 W13 上。控制对象 IC1 的显示位置可根据用户 Ua 的手区域的移动而改变。在所示出的时间，控制对象 IC1 的显示位置在屏幕的中心略靠左侧。在图 15

的下部，示出了当从用户 Ua 的头上方的视点查看时信息处理设备 100 与用户 Ua 之间的真实空间中的位置关系。作为一个示例，识别控制单元 172 基于摄像装置 101 的视角和控制对象 IC1 的显示位置而以角度 R1 设置麦克风 102 的方向性。由于用户 Ua 存在于角度 R1 的方向上，因此，麦克风 102 变得可以以较高的质量获得用户 Ua 产生的语音样本。

[0123] 在图 16 的上部，控制对象 IC1 重叠在 UI 图像 W14 上。用户 Ua 和 Ub 也出现在 UI 图像 W14 中。在所示出的时间，与用户 Ua 相比，控制对象 IC1 的显示位置更靠近用户 Ub 的面部。在图 16 的下部，示出了当从用户 Ua 和 Ub 的头上方的视点查看时信息处理设备 100 与用户 Ua 和 Ub 之间的真实空间中的位置关系。作为一个示例，识别控制单元 172 基于摄像装置 101 的视角和控制对象 IC1 的显示位置而以角度 R2 设置麦克风 102 的方向性。由于用户 Ub 存在于角度 R2 的方向上，因此，麦克风 102 变得可以以较高的质量获得用户 Ub 产生的语音样本。

[0124] 在图 17 的上部，控制对象 IC1 重叠在 UI 图像 W15 上。控制对象 IC1 在屏幕上的朝向可以根据用户 Ua 的手区域的朝向而改变。用户 Ua 和 Ub 出现在 UI 图像 W15 中。在所示出的时间，控制对象 IC1 正由用户 Ua 操作并且正指向用户 Ub 的面部区域 A07 的方向。在图 17 的下部，示出了当从用户 Ua 和 Ub 的头上方的视点查看时信息处理设备 100 与用户 Ua 和 Ub 之间的真实空间中的位置关系。作为一个示例，识别控制单元 172 基于控制对象 IC1 的显示位置和朝向以及用户 Ub 的面部区域 A07 的位置而以角度 R3 设置麦克风 102 的方向性。由于用户 Ub 存在于角度 R3 的方向上，因此，麦克风 102 变得可以以较高的质量获得用户 Ub 产生的语音样本。

[0125] 根据参照图 16 或图 17 描述的方法，当存在多个用户时，通过如作为真实的麦克风一样来使用控制对象 IC1，可以在用户之间传递说话以进行语音识别的权利。

[0126] 除了上述示例之外，可实现基于用户的姿势的多种用户界面。作为一个示例，识别控制单元 172 可根据用户用他的 / 她的手盖住他的 / 她的嘴的姿势的识别而取消语音识别单元 154 至此产生的语音识别结果。这样，当用户产生了具有错误内容的语音样本时或者当语音识别单元 154 错误地识别了语音样本的内容时，用户可以容易地重复语音输入。识别控制单元 172 还可根据预先定义的姿势的识别而增加或减小从扬声器 109 输出的音频的音量。

[0127] 识别控制单元 172 还可将分别表示至少一个语音命令候选的文本对象重叠在输入图像上。这样，甚至当用户预先不知道应用功能接收的语音命令时，用户也可以适当地给出所需语音命令。

[0128] (8) 显示控制单元

[0129] 显示控制单元 174 经由显示器 108 控制图像的显示。作为一个示例，显示控制单元 174 在显示器 108 上显示从应用单元 140 输入的应用图像。另外，如果启动了语音兼容应用，则显示控制单元 174 在显示器 108 上显示识别控制单元 172 生成的 UI 图像。显示控制单元 174 可仅在显示器 108 上显示 UI 图像或者可在显示器 108 上显示通过组合应用图像和 UI 图像而生成的单个输出图像。

[0130] 图 18 和图 19 示出了本实施例可使用的输出图像的窗口布局的示例。在这样的图中，显示器 108 显示了 UI 窗口 W_{UI} 和应用窗口 W_{APP}。UI 窗口 W_{UI} 显示识别控制单元 172 生成的 UI 图像。应用窗口 W_{APP} 显示从应用单元 140 输入的应用图像（例如，内容图像）。在图

18 的第一示例中，应用窗口 W_{APP} 在 UI 窗口 W_{UI} 的右下角处被组合。在图 19 的第二示例中，UI 窗口 W_{UI} 与应用窗口 W_{APP} 的一个部分混合。通过使用这样的窗口布局，作为一个示例，即使当用户手边没有遥控器时，用户也可以在观看内容图像的同时使用控制对象利用他的 / 她的语音来操作信息处理设备 100。

[0131] <2-3. 示例控制场景 >

[0132] 现在将参照图 20 至图 23 描述上述信息处理设备 100 可以执行的控制场景的示例。

[0133] (1) 第一场景

[0134] 图 20 是用于说明第一控制场景的图。在图 20 中，沿着时间轴示出了五个 UI 图像 ST11 至 ST15。

[0135] 用户 Ud 出现在 UI 图像 ST11 中，并且实现了镜像图像显示。

[0136] 下一 UI 图像 ST12 可例如在启动了语音兼容应用之后或者当用户做出了姿势（诸如抬起他的手）之后被显示。控制对象 IC1 重叠在 UI 图像 ST12 上。然而，此时，到语音识别单元 154 的语音输入尚未被激活。

[0137] 下一 UI 图像 ST13 可例如在用户 Ud 将控制对象 IC1 移动到他的嘴附近之后被显示。作为满足激活条件的结果，识别控制单元 172 激活到语音识别单元 154 的语音输入。在 UI 图像 ST13 中，控制对象 IC1 的显示属性改变以指示有效状态。

[0138] 下一 UI 图像 ST14 可在用户 Ud 正产生语音样本的同时被显示。在 UI 图像 ST14 中，控制对象 IC1 的显示属性继续指示有效状态。另外，效果 Fb1 被施加于 UI 图像 ST14，并且示出所识别的语音样本的内容的附加对象 Fb2 重叠在 UI 图像 ST14 上。

[0139] 下一 UI 图像 ST15 可在满足去激活条件时被显示。这里，假设用食指触摸嘴的姿势被定义为用于去激活语音输入的姿势。根据对这样的姿势的识别，识别控制单元 172 去激活到语音识别单元 154 的语音输入。控制对象 IC1 的显示位置返回到默认显示位置，并且控制对象 IC1 的显示属性被改变以指示无效状态。

[0140] (2) 第二场景

[0141] 图 21 是用于说明第二控制场景的图。在图 21 中，沿着时间轴示出了五个 UI 图像 ST21 至 ST25。

[0142] 用户 Ud 出现在 UI 图像 ST21 中。控制对象 IC1 也重叠在 UI 图像 ST21 上。然而，此时，到语音识别单元 154 的语音输入尚未被激活。

[0143] 下一 UI 图像 ST22 可例如在用户 Ud 将控制对象 IC1 移动到他的嘴附近之后被显示。作为满足激活条件的结果，识别控制单元 172 激活到语音识别单元 154 的语音输入。在 UI 图像 ST22 中，控制对象 IC1 的显示属性改变以指示有效状态。

[0144] 下一 UI 图像 ST23 可在用户 Ud 正产生语音样本的同时被显示。在 UI 图像 ST23 中，控制对象 IC1 的显示属性继续指示有效状态。在该第二场景中，在用户 Ud 正产生语音样本的同时，控制对象 IC1 的显示位置保持在用户 Ud 的嘴附近而与手移动无关。因此，如果用户输入诸如电子邮件消息的长文本作为语音样本，则可以在无需用户持续抬起他的手并且变累的情况下继续语音输入。

[0145] 在下一 UI 图像 ST24 中，用户 Ud 正做出用他的手盖住他的嘴的姿势。识别控制单元 172 根据对这样的姿势的识别而取消至此的语音识别结果。在第二控制场景中，此后维

持到语音识别单元 154 的语音输入的有效状态。

[0146] 在下一 UI 图像 ST25 中，用户 Ud 产生另一语音样本。结果，语音识别单元 154 适当地识别具有与用户 Ud 初始产生的语音样本的内容不同的内容的语音样本。

[0147] (3) 第三场景

[0148] 图 22 是用于说明第三控制场景的图。在图 22 中，沿着时间轴示出了三个 UI 图像 ST31 至 ST33。

[0149] 用户 Ud 出现在 UI 图像 ST31 中，并且实现了镜像图像显示。

[0150] 下一 UI 图像 ST32 可例如在用户做出了诸如举起他的手的姿势之后被显示。控制对象 IC2 重叠在 UI 图像 ST32 上。分别表示语音兼容应用接收的语音命令候选（命令 A 至命令 D）的四个文本对象也重叠在 UI 图像 ST32 上。

[0151] 在下一 UI 图像 ST33 中，作为用户 Ud 例如接近控制对象 IC12 附近的结果，激活语音输入。用户 Ud 然后产生语音样本以便读出命令 B，并且语音识别单元 154 适当地识别说出的命令 B。作为示例，语音命令候选可以是预先提供的以便使得用户远程控制信息处理设备 100 的至少一个命令。

[0152] 以此方式，在本实施例中，即使用户手边没有遥控器，用户也可以远程控制信息处理设备 100。作为示例，即使当遥控器丢失或者遥控器正由其他用户持有时，用户仍能够以期望的定时控制信息处理设备 100 而不会感觉到任何压力。注意，在显示 UI 图像 ST32 之后，根据对特定语音命令或姿势的识别，表示语音命令 A 至 D 的文本对象可用表示其它语音命令候选的文本对象来取代。

[0153] (4) 第四场景

[0154] 第四场景是不涉及控制对象的补充场景。图 23 是用于说明第四控制场景的图。在图 23 中，沿着时间轴示出了三个 UI 图像 ST41 至 ST43。

[0155] 用户 Ud 出现在 UI 图像 ST41 中，并且实现了镜像图像显示。

[0156] 在下一 UI 图像 ST42 中，用户 Ud 正做出用他的手罩住他的耳朵的姿势。识别控制单元 172 根据对这样的姿势的识别而增加从扬声器 109 输出的音频的音量。音量的增加可根据识别姿势的时间长度而改变。

[0157] 在下一 UI 图像 ST43 中，用户 Ud 做出用他的食指触摸他的嘴的姿势。识别控制单元 172 根据对这样的姿势的识别而减小从扬声器 109 输出的音频的音量。音量的减小可根据识别姿势的时间长度来改变。

[0158] 以此方式，在本实施例中，可基于用户姿势而实现各种用户界面。根据语音输入是否有效或者是否正执行语音兼容应用，同一类型的姿势可被解释为具有不同的含义。注意，可提供用于允许用户登记来源于用户的姿势的用户界面。作为一个示例，可登记用手推开（控制对象）的姿势，并且这样的姿势可被定义为用于激活 / 去激活语音输入的姿势。还可提供用于允许用户定制对于各个姿势的移动与对应于这样的姿势的处理之间的映射的用户界面。

[0159] <2-4. 示例处理流程 >

[0160] 图 24 和图 25 中的流程图示出了根据本实施例的信息处理设备 100 可执行的处理的流程的示例。针对构成摄像装置 101 拍摄的视频的一系列帧中的每个帧重复这里描述的处理。

[0161] 如图 24 所示，首先，图像获取单元 120 获取摄像装置 101 拍摄的图像作为输入图像（步骤 S100）。图像获取单元 120 然后将所获取的输入图像输出到识别单元 150 和控制单元 170。

[0162] 接下来，图像识别单元 152 识别出现在从图像获取单元 120 输入的输入图像中的用户的身体（步骤 S105）。例如，图像识别单元 152 识别输入图像中的用户的手区域和嘴区域，并且将示出这样识别的区域的位置的位置数据输出到控制单元 170。图像识别单元 152 可另外地识别预先定义的多个用户姿势。

[0163] 识别控制单元 172 接下来确定是否启动了语音兼容应用（步骤 S110）。如果尚未启动语音兼容应用，则跳过以下步骤 S115 至 S160 中的处理。如果启动了语音兼容应用（或者如果通过在步骤 S105 中识别的姿势启动了语音兼容应用），则处理进行到步骤 S115。

[0164] 在步骤 S115 中，识别控制单元 172 确定与说话有关的控制对象的显示位置和朝向（步骤 S115）。控制对象的显示位置可以是默认位置或者可移动以便跟踪图像识别单元 152 识别的用户手的移动。以相同方式，控制对象的朝向可以是默认朝向或者可旋转以便跟踪用户手的移动。

[0165] 此后，如果麦克风 102 具有可变方向性，则识别控制单元 172 根据在步骤 S115 中确定的控制对象的显示位置和朝向而设置麦克风 102 的方向性（步骤 S120）。

[0166] 接下来，识别控制单元 172 将具有在步骤 S115 中确定的显示位置和朝向的控制对象重叠在显示输入图像的镜像图像的 UI 图像上（步骤 S125）。这里，控制对象的显示属性可被设置为指示语音输入尚未被激活的值。

[0167] 返回到图 25，识别控制单元 172 接下来根据先前描述的激活条件和去激活条件来确定语音输入是否有效（步骤 S130）。作为一个示例，当用户的嘴区域与控制对象之间的距离低于距离阈值时，确定满足激活条件。如果没有确定语音输入有效，则跳过以下步骤 S135 至 S160 中的处理。如果确定语音输入有效，则处理进行到步骤 S135。

[0168] 在步骤 S135 中，识别控制单元 172 根据需要激活到语音识别单元 154 的语音输入，并且将控制对象的显示属性设置为指示语音输入已被激活的值（步骤 S135）。

[0169] 接下来，语音获取单元 130 将从麦克风 102 获取的输入语音输出到语音识别单元 154（步骤 S140）。

[0170] 此后，语音识别单元 154 基于从语音获取单元 130 输入的输入语音而对用户的语音样本执行语音识别（步骤 S145）。语音识别单元 154 然后将语音识别的结果输出到应用单元 140 和识别控制单元 172。

[0171] 接下来，识别控制单元 172 将关于从语音识别单元 154 输入的语音识别结果的反馈合并到 UI 图像中（步骤 S150）。作为一个示例，识别控制单元 172 将图 11 所示的效果 Fb1 施加于 UI 图像。识别控制单元 172 还可将图 12 至图 14 所示的附加对象 Fb2 或 Fb3 重叠在 UI 图像上。

[0172] 此后，识别控制单元 172 确定语音识别是否成功（步骤 S155）。如果语音识别不成功，则跳过以下步骤 S160 中的处理。如果语音识别成功，则处理进行到步骤 S160。

[0173] 在步骤 S160 中，应用单元 140 基于语音识别结果而执行应用处理（步骤 S160）。例如，应用单元 140 可执行与所识别的语音命令对应的处理。应用单元 140 还可接收示出所识别的语音样本的内容的文本作为输入信息。

[0174] 接下来，显示控制单元 174 在显示器 108 上显示包括 UI 图像的输出图像（步骤 S165）。这里显示的输出图像可仅包括 UI 图像或者可包括 UI 图像和应用图像两者。此后，处理返回到图 24 中的步骤 S100。

[0175] 注意，至此已主要描述了仅将一个控制对象重叠在 UI 图像上的示例。然而，本公开不限于这样的示例，并且多个控制对象可重叠在 UI 图像上。作为一个示例，当多个用户出现在输入图像中时，如果针对各个用户重叠分别的控制对象，则各个用户可以以期望的定时输入语音命令而不需要在用户之间传递控制对象的操作。

[0176] <3. 第二实施例 >

[0177] 如之前所述，根据本公开的实施例的技术不限于电视机设备，并且可以应用于各种类型的设备。为此，现在将描述根据本公开的实施例的技术应用于包括消息应用的信息处理设备 200 的示例作为第二实施例。如参照图 2 所描述的，信息处理设备 200 是平板 PC。

[0178] (1) 示例硬件配置

[0179] 图 26 是示出信息处理设备 200 的示例硬件配置的框图。如图 26 所示，信息处理设备 200 包括摄像装置 201、麦克风 202、输入装置 203、通信 I/F 204、存储器 205、显示器 208、扬声器 209、总线 211 和处理器 212。

[0180] 摄像装置 201 包括诸如 CCD 或 CMOS 的图像拍摄元件并且拍摄图像。摄像装置 201 拍摄的图像（构成视频的帧）被视为用于信息处理设备 200 的处理的输入图像。

[0181] 麦克风 202 获得用户产生的语音样本并且生成语音信号。麦克风 202 生成的语音信号被视为旨在用于信息处理设备 200 的语音识别的输入语音。

[0182] 输入装置 203 是用户用于直接操作信息处理设备 200 或将信息输入到信息处理设备 200 的装置。作为一个示例，输入装置 103 可包括触摸板、按钮、开关等。在检测到用户输入时，输入装置 203 生成与所检测的用户输入对应的输入信号。

[0183] 通信 I/F 204 用作信息处理设备 200 与其它设备之间的通信的媒介。通信 I/F 204 支持任意无线通信协议或有线通信协议，并且建立与其它设备的通信连接。

[0184] 存储器 205 由诸如半导体存储器或硬盘驱动器的存储介质构成，并且存储用于信息处理设备 200 的处理的程序和数据以及内容数据。注意，程序和数据的一些或全部可不由存储器 205 来存储，而是替代地可从外部数据源（作为示例，数据服务器、网络存储装置或外接存储器）来获取。

[0185] 显示器 208 具有由 LCD、OLED 等构成的屏幕并且显示信息处理设备 200 生成的图像。作为一个示例，与在第一实施例中描述的 UI 图像相同的 UI 图像可显示在显示器 208 的屏幕上。

[0186] 扬声器 209 具有振动膜和诸如放大器的电路元件，并且基于信息处理设备 200 生成的输出音频信号而输出音频。扬声器 209 的音量是可变的。

[0187] 总线 211 将摄像装置 201、麦克风 202、输入装置 203、通信 I/F 204、存储器 205、显示器 208、扬声器 209 和处理器 212 彼此连接。

[0188] 作为示例，处理器 112 可以是 CPU 或 DSP。通过执行存储在存储器 205 或其它存储介质中的程序，以与根据第一实施例的信息处理设备 100 的处理器 112 相同的方式，处理器 212 使得信息处理设备 200 以各种方式起作用。除了应用功能中的差别之外，信息处理设备 200 的存储器 205 和处理器 212 实现的逻辑功能的配置可与图 4 所示的信息处理设备

100 的配置相同。

[0189] (2) 示例控制场景

[0190] 图 27 是用于说明第二实施例中的控制场景的示例的图。在图 27 中，沿着时间轴示出了四个 UI 图像 ST51 至 ST54。在该场景中，各个输出图像由在顶部的消息应用的应用图像和底部的 UI 图像构成。

[0191] 在输出图像 ST51 中，应用图像包括消息输入框。消息尚未被输入到消息输入框中。用户 Ud 出现在 UI 图像中并且实现了镜像图像显示。

[0192] 下一输出图像 ST52 可例如在用户做出诸如抬起他的手的姿势之后被显示。在输出图像 ST52 中，控制对象 IC1 重叠在 UI 图像上。然而，此时，语音输入尚未被激活。

[0193] 下一输出图像 ST53 可例如在用户 Ud 将控制对象 IC1 移动到他的嘴附近之后被显示。语音输入被激活，并且控制对象 IC1 的显示属性被改变以指示有效状态。用户产生的语音样本的内容被输入到消息输入框中。

[0194] 下一输出图像 ST54 可例如在用户 Ud 将控制对象 IC1 移动远离他的嘴附近之后被显示。语音输入被去激活，并且控制对象 IC1 的显示属性改变以便示出无效状态。即使用户在该状态下产生语音样本，这样的语音样本的内容也不会被输入到消息输入框中。因此，通过仅做出移动他的 / 她的手的简单操作，用户可以切换语音输入的状态，并且仅包括用户希望在消息中输入的语音样本的内容。

[0195] <4. 结论 >

[0196] 至此参照图 1 至图 27 详细描述了本公开的实施例。根据先前描述的实施例，重叠在输入图像上所显示的控制对象用于控制信息设施执行的语音识别。因此，通过使用屏幕上的控制对象的状态作为向导，用户能够确定用于语音识别的合适定时。

[0197] 另外，根据以上实施例，基于控制对象与在输入图像中识别的用户身体的特定部位之间的位置关系来控制语音识别。因此，通过移动显示在屏幕上的他 / 她自己的身体，用户能够处置与语音识别有关的各种功能。

[0198] 另外，根据上述实施例，可基于用户的嘴与控制对象之间的距离来激活用于语音识别的语音输入。控制对象还可根据用户手的移动来在屏幕上移动。因此，通过移动控制对象或者朝向控制对象相反地移动他 / 她自己，用户能够容易地以期望的定时仅对期望的语音输入执行语音识别。由于在这样的时间要执行的所需移动类似当处置真实麦克风时的移动，因此这样的构架使得可以实现对用户直观的用户界面。

[0199] 另外，根据上述实施例，经由控制对象的显示属性的改变向用户通知语音输入是否已被激活。因此，用户可以仅通过仅注意屏幕上的控制对象而以适当的定时说话。

[0200] 注意，通常使用软件来实现被描述为本公开的实施例的各种设备执行的一系列处理。作为一个示例，由实现这样的一系列处理的软件构成的程序预先存储在设置在这样的设备内部或外部的存储介质（非暂态介质）中。作为一个示例，在执行期间，这样的程序然后被写入到 RAM（随机存取存储器）中并且由诸如 CPU 的处理器来执行。

[0201] 尽管以上参照附图详细描述了本公开的优选实施例，但是本公开的技术范围不限于此。本领域技术人员应理解，在所附权利要求或其等同方案的范围内，可根据设计要求和其它因素进行各种修改、组合、子组合和变更。

[0202] 另外，本技术还可如下配置。

[0203] (1) 一种信息处理系统，包括：

[0204] 处理电路，被配置成生成用于控制显示装置的数据，以在显示图像上重叠与语音输入相关联的控制对象，其中，所述显示图像是用户执行的姿势操作的反馈图像，并且所述显示图像是从摄像装置捕获图像获得的图像。

[0205] (2) 根据 (1) 所述的信息处理系统，其中，

[0206] 所述显示图像是所述用户的镜像图像。

[0207] (3) 根据 (1) 所述的信息处理系统，其中，

[0208] 所述处理电路被配置成基于所述控制对象与所述反馈图像中的用户的身体部位之间的显示位置关系而发起由语音输入触发的处理。

[0209] (4) 根据 (3) 所述的信息处理系统，其中，

[0210] 当所述显示位置关系包括所述控制对象在距所述反馈图像中的用户的身体部位预定距离内时，所述处理电路发起所述处理。

[0211] (5) 根据 (3) 所述的信息处理系统，其中，

[0212] 所述反馈图像中的所述用户的身体部位是所述用户的面部的至少一部分。

[0213] (6) 根据 (3) 所述的信息处理系统，其中，

[0214] 当所述显示位置关系包括所述反馈图像中的所述用户的身体部位在距所述控制对象的预定方向内时，所述处理电路发起所述处理。

[0215] (7) 根据 (3) 所述的信息处理系统，其中，

[0216] 所述处理电路被配置成控制所述显示装置以改变所述控制对象的图像属性，以指示所述处理电路发起了所述处理。

[0217] (8) 根据 (3) 所述的信息处理系统，其中，

[0218] 所述处理是语音识别处理。

[0219] (9) 根据 (1) 所述的信息处理系统，其中，

[0220] 所述处理电路被配置成控制所述显示装置以响应于所述用户执行的姿势操作而改变所述控制对象的显示位置。

[0221] (10) 根据 (1) 所述的信息处理系统，其中，

[0222] 所述处理电路被配置成控制所述显示装置以显示根据所述语音输入的检测状态而改变外观的指示符。

[0223] (11) 根据 (1) 所述的信息处理系统，其中，

[0224] 所述处理电路被配置成控制所述显示装置以与所述反馈图像同步地显示附加对象，所述附加对象与所述语音输入相关联并且与所述控制对象不同。

[0225] (12) 根据 (11) 所述的信息处理系统，其中，

[0226] 所述附加对象是基于所述语音输入生成的文本信息。

[0227] (13) 根据 (11) 所述的信息处理系统，其中，

[0228] 所述附加对象指示所述语音输入的音量水平。

[0229] (14) 根据 (8) 所述的信息处理系统，其中，

[0230] 所述处理电路被配置成基于所述语音识别而控制装置的功能。

[0231] (15) 根据 (14) 所述的信息处理系统，其中，

[0232] 所述装置被配置成控制内容的再现，并且所述处理电路被配置成控制所述显示装

置以同时显示所述反馈图像、所述控制对象和作为所述语音输入的对象的所述内容的图像。

[0233] (16) 根据 (1) 所述的信息处理系统，还包括：

[0234] 所述显示装置，其中，所述显示装置和所述处理电路是单个设备的部件。

[0235] (17) 根据 (3) 所述的信息处理系统，其中，

[0236] 所述镜像图像是所述用户的实际图像。

[0237] (18) 根据 (3) 所述的信息处理系统，其中，

[0238] 所述镜像图像是所述用户的体现。

[0239] (19) 一种信息处理方法，包括：

[0240] 利用处理电路生成用于控制显示装置的数据，以在显示图像上重叠与语音输入相关联的控制对象，其中，所述显示图像是用户执行的姿势操作的反馈图像，并且所述显示图像是从摄像装置捕获图像得到的图像。

[0241] (20) 一种存储有计算机可读指令的非暂态计算机可读存储介质，所述计算机可读指令当由处理电路执行时执行信息处理方法，所述方法包括：

[0242] 利用所述处理电路生成用于控制显示装置的数据，以在显示图像上重叠与语音输入相关联的控制对象，其中，所述显示图像是用户执行的姿势操作的反馈图像，并且所述显示图像是从摄像装置捕获图像得到的图像。

[0243] 另外，本技术还可如下配置。

[0244] (1) 一种信息处理设备，包括：

[0245] 图像获取单元，获取输入图像；以及

[0246] 控制单元，在屏幕上显示重叠在所述输入图像上的与说话有关的对象，

[0247] 其中，所述控制单元使用所述对象控制对用户的语音样本执行的语音识别。

[0248] (2) 根据 (1) 所述的信息处理设备，还包括：

[0249] 图像识别单元，识别出现在所述输入图像中的用户的身体，

[0250] 其中，所述控制单元基于所述对象与所述图像识别单元识别的所述用户的身体的特定部位之间在所述屏幕上的位置关系，控制所述语音识别。

[0251] (3) 根据 (2) 所述的信息处理设备，

[0252] 其中，所述特定部位包括用户的嘴，以及

[0253] 其中，所述控制单元基于所述对象与所述用户的嘴之间的距离而激活用于所述语音识别的语音输入。

[0254] (4) 根据 (3) 所述的信息处理设备，

[0255] 其中，所述特定部位包括用户的手，以及

[0256] 其中，所述控制单元根据所述用户的手的移动而在所述屏幕上移动所述对象。

[0257] (5) 根据 (3) 或 (4) 所述的信息处理设备，

[0258] 其中，所述控制单元根据出现在所述输入图像中的用户的姿势而去激活用于所述语音识别的语音输入。

[0259] (6) 根据 (1) 至 (5) 中任一项所述的信息处理设备，

[0260] 其中，所述控制单元通过改变所述对象的显示属性而向所述用户通知用于所述语音识别的语音输入是否被激活。

[0261] (7) 根据 (1) 至 (6) 中任一项所述的信息处理设备，

[0262] 其中，所述控制单元通过改变所述对象的显示属性和重叠了所述对象的输出图像的状态中的任意一个，向所述用户通知在所述语音识别期间是否检测到语音样本。

[0263] (8) 根据 (7) 所述的信息处理设备，

[0264] 其中，所述控制单元根据在所述语音识别期间检测的语音样本的水平，改变所述对象的显示属性的变化水平或者所述输出图像的状态。

[0265] (9) 根据 (1) 至 (8) 中任一项所述的信息处理设备，

[0266] 其中，对具有可变方向性的麦克风获取的语音信号执行所述语音识别。

[0267] (10) 根据 (9) 所述的信息处理设备，

[0268] 其中，所述控制单元根据所述用户的移动而改变所述对象的位置，以及

[0269] 其中，所述麦克风的方向性是根据所述对象的位置而设置的。

[0270] (11) 根据 (9) 或 (10) 所述的信息处理设备，

[0271] 其中，所述控制单元根据所述用户的移动而改变所述对象的朝向，以及

[0272] 其中，所述麦克风的方向性是根据所述对象的朝向设置的。

[0273] (12) 根据 (1) 至 (11) 中任一项所述的信息处理设备，

[0274] 其中，所述控制单元还将第一附加对象重叠在出现在所述输入图像中的所述用户附近，所述第一附加对象包括表示通过所述语音识别而识别的语音样本的内容的文本。

[0275] (13) 根据 (12) 所述的信息处理设备，

[0276] 其中，当所述语音识别失败时，所述控制单元能够被操作以通过改变所述第一附加对象的显示属性以及将特殊字符串插入到所述文本中之一而向所述用户通知所述语音识别失败。

[0277] (14) 根据 (1) 至 (13) 中任一项所述的信息处理设备，

[0278] 其中，所述控制单元还将第二附加对象重叠在所述输入图像上，所述第二附加对象指示在所述语音识别期间检测的语音样本的水平和有效地执行所述语音识别所需的语音水平。

[0279] (15) 根据 (1) 至 (14) 中任一项所述的信息处理设备，

[0280] 其中，所述控制单元还将分别表示至少一个语音命令的候选的文本对象重叠在所述输入图像上。

[0281] (16) 根据 (15) 所述的信息处理设备，

[0282] 其中，所述信息处理设备是电视机设备，以及

[0283] 其中，所述语音命令是所述用户发出的用于远程控制所述信息处理设备的命令。

[0284] (17) 根据 (1) 至 (16) 中任一项所述的信息处理设备，

[0285] 其中，所述对象是类似麦克风的图标。

[0286] (18) 一种由信息处理设备执行的信息处理方法，所述信息处理方法包括：

[0287] 获取输入图像；以及

[0288] 在屏幕上显示重叠在所述输入图像上的与说话有关的对象，使用所述对象控制对用户的语音样本执行的语音识别。

[0289] (19) 一种用于使得控制信息处理设备的计算机用作以下单元的程序：

[0290] 图像获取单元，获取输入图像；以及

[0291] 控制单元，在屏幕上显示重叠在所述输入图像上的与说话有关的对象，

[0292] 其中，所述控制单元使用所述对象控制对用户的语音样本执行的语音识别。

[0293] 附图标记列表

[0294] 100，200 信息处理设备

[0295] 120 图像获取单元

[0296] 152 图像识别单元

[0297] 154 语音识别单元

[0298] 172 识别控制单元

[0299] 174 显示控制单元

[0300] IC1，IC2 控制对象

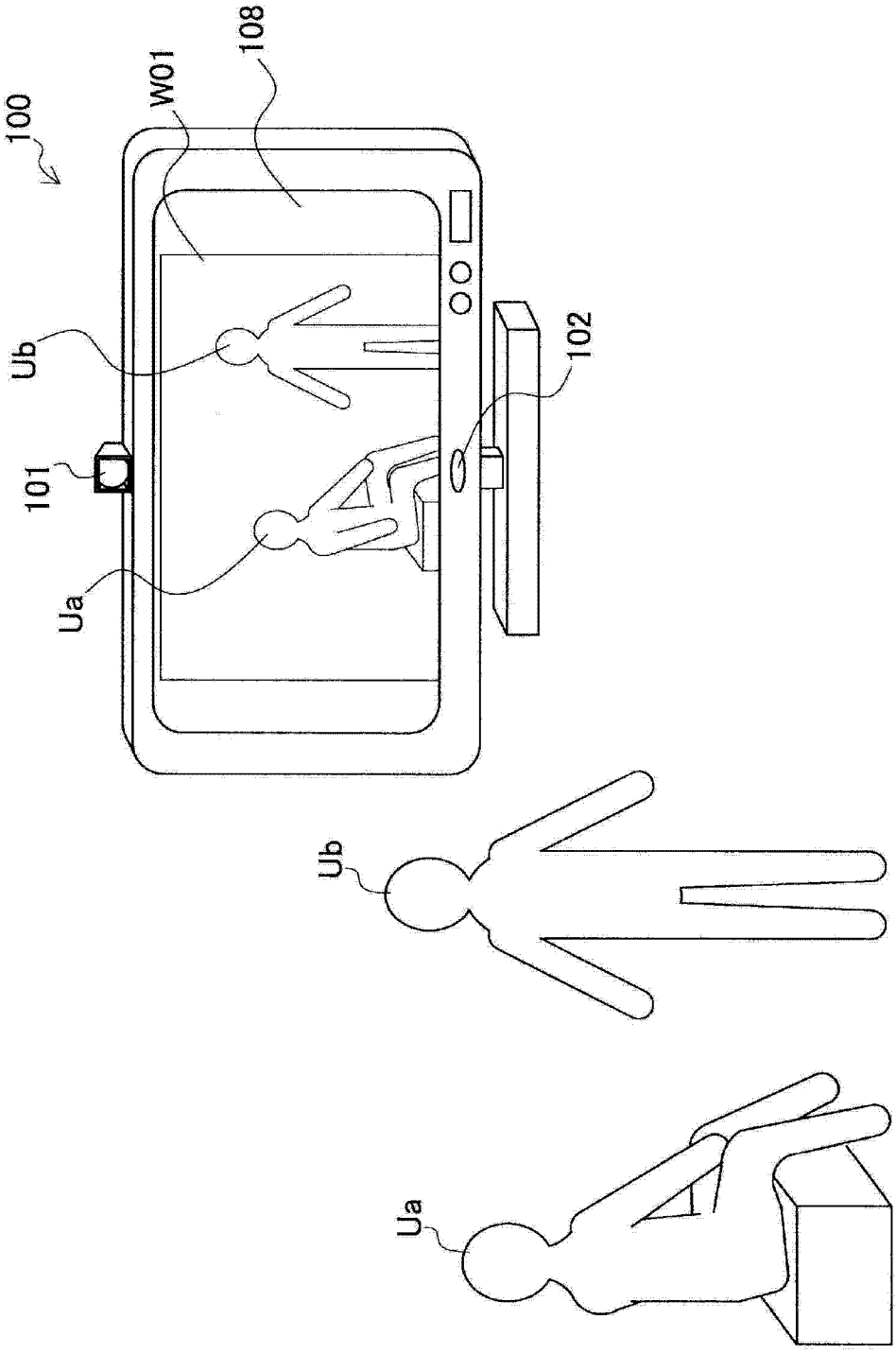


图 1

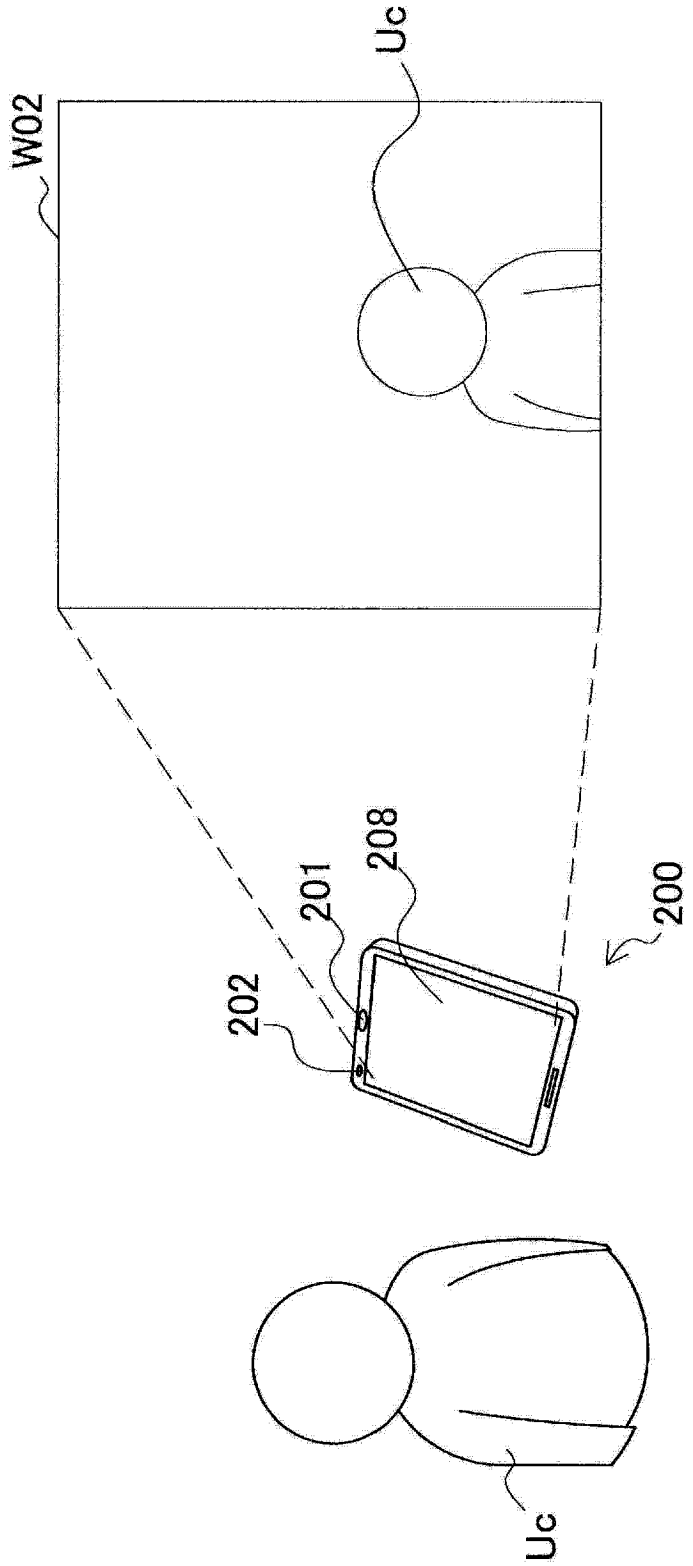


图 2

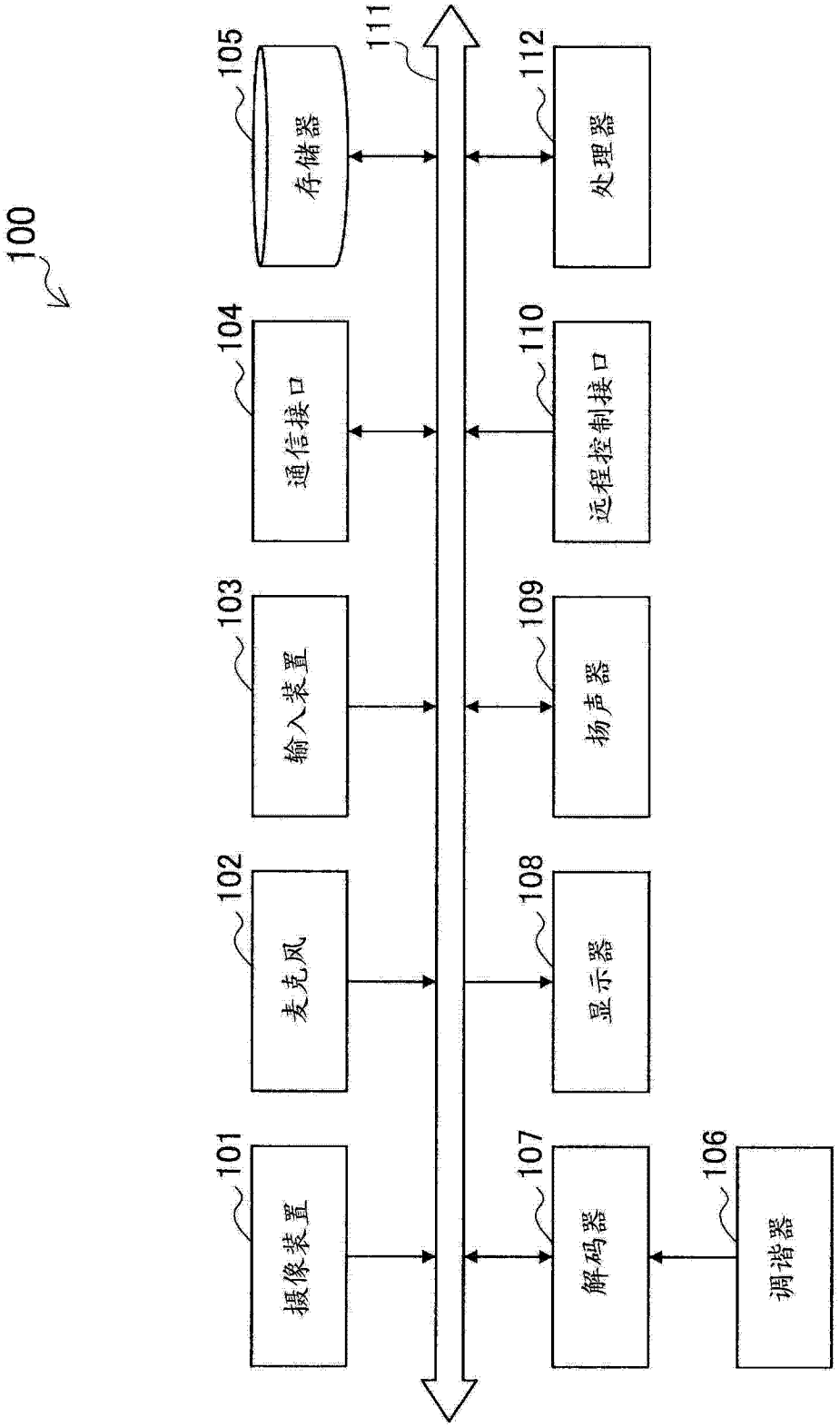

100
111
105
存储器
112
处理器
104
通信接口
110
远程控制接口
103
输入装置
109
扬声器
102
麦克风
108
显示器
101
摄像装置
107
解码器
106
调谐器

图 3

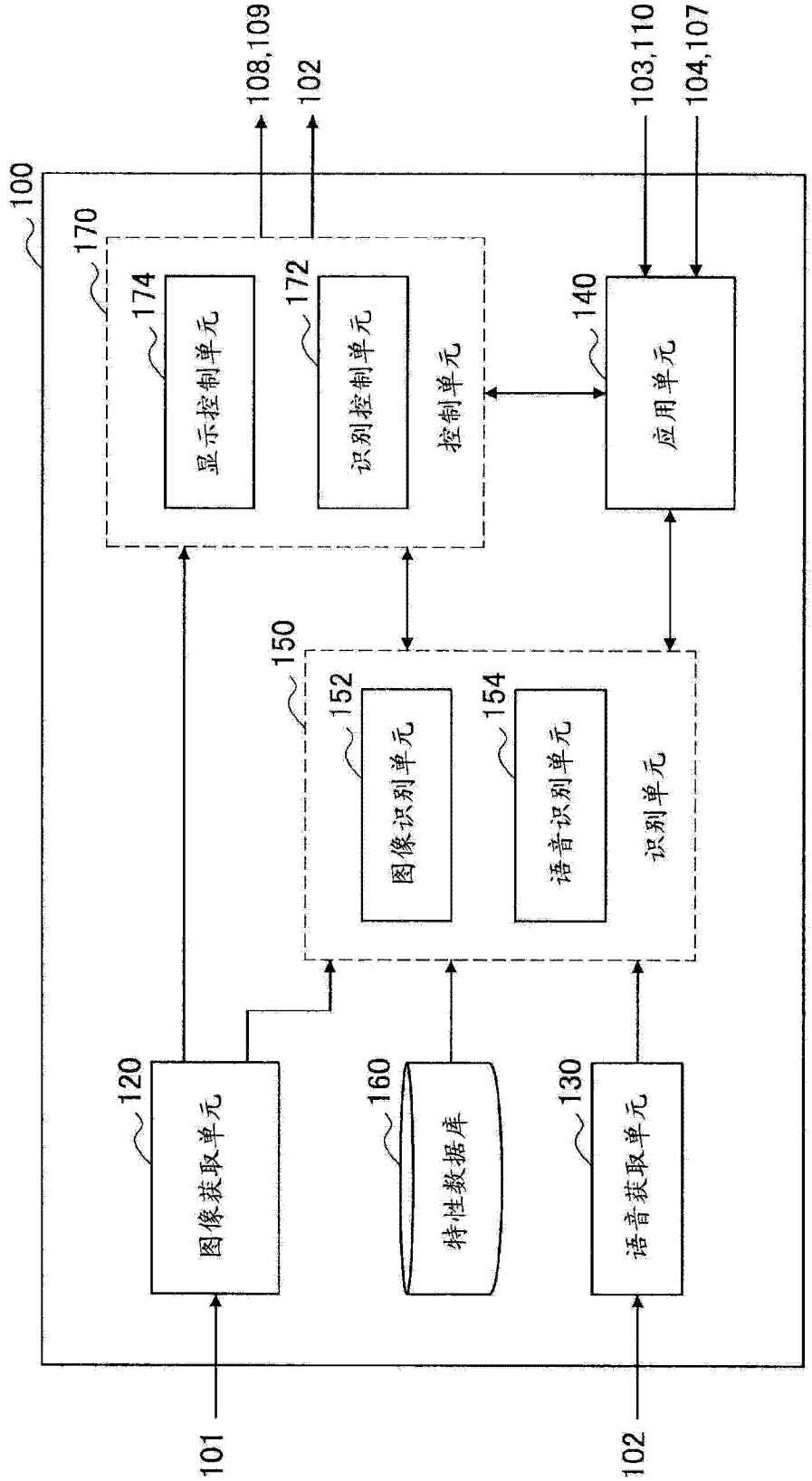

100
101
102
102
103,110
104,107
108,109
120
图像获取单元
130
语音获取单元
140
应用单元
150
识别单元
152
图像识别单元
154
语音识别单元
160
特性数据库
170
控制单元
172
识别控制单元
174
显示控制单元

图 4

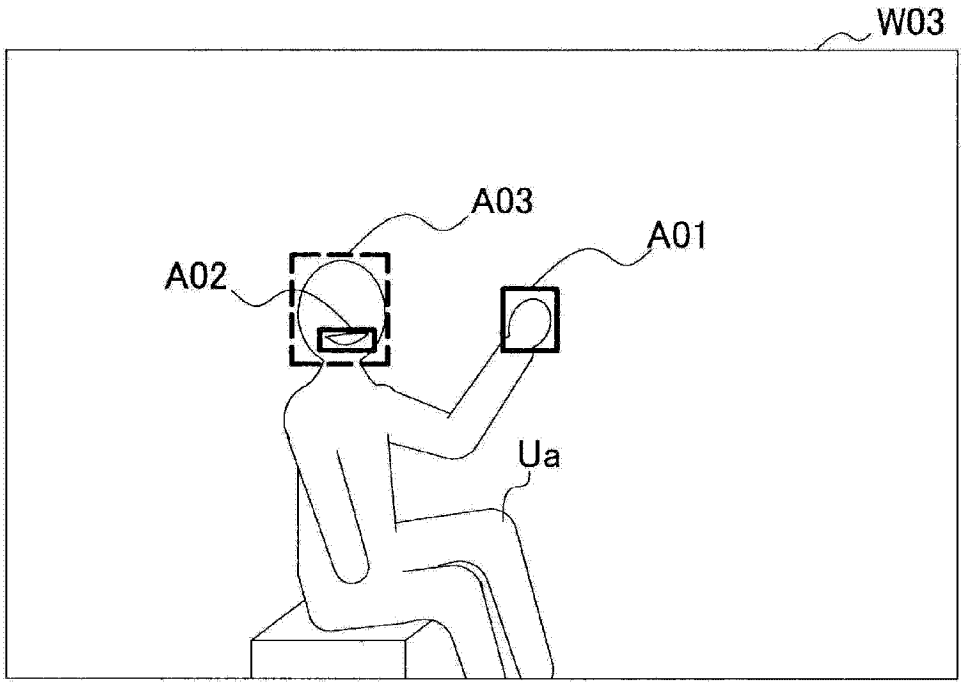


图 5

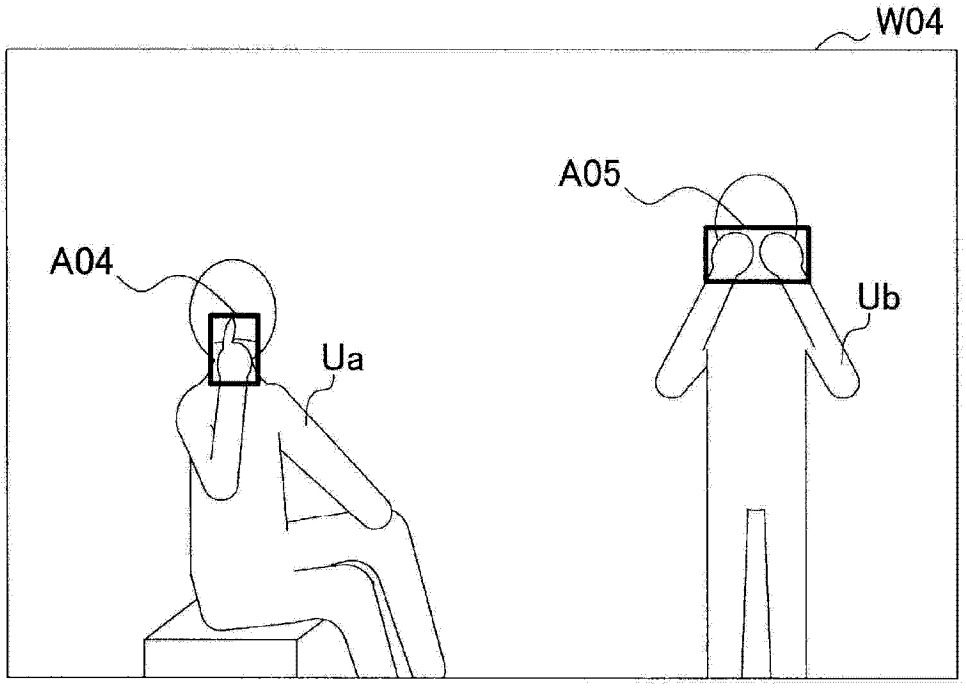


图 6

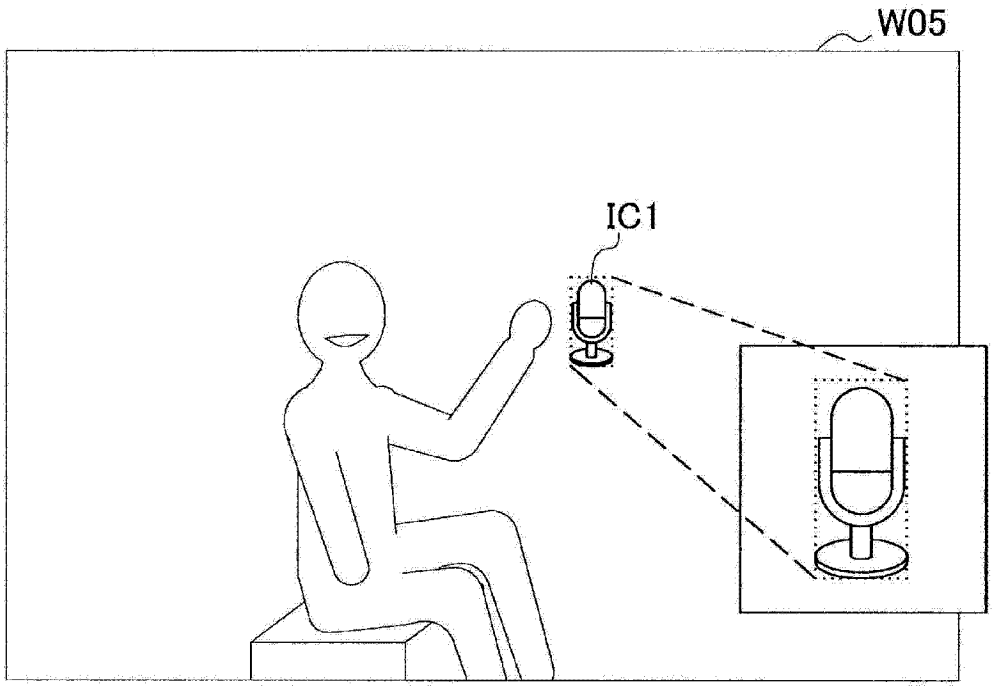


图 7

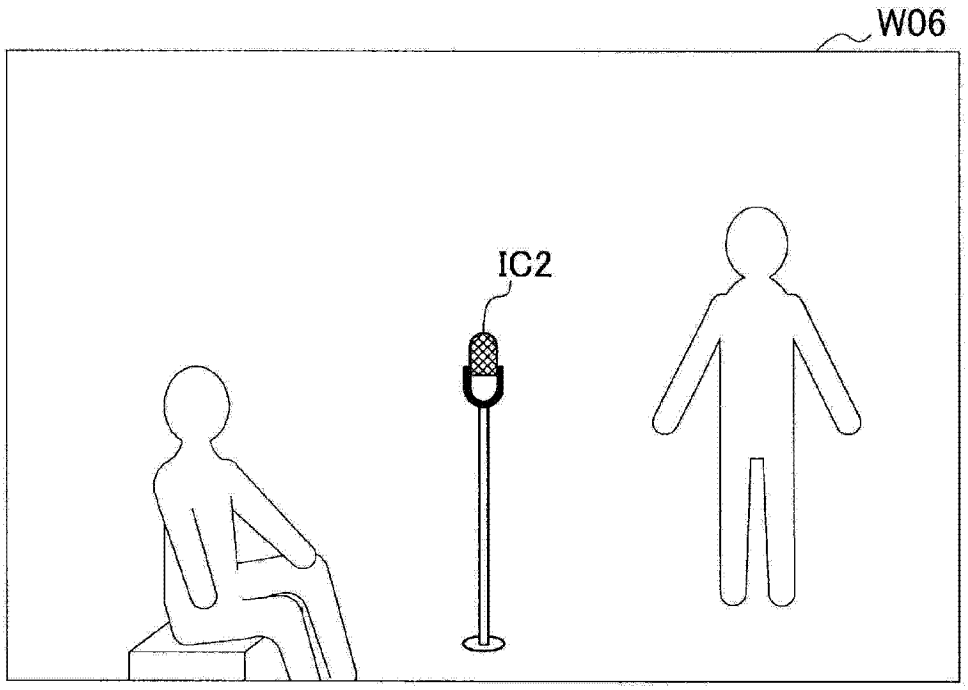


图 8

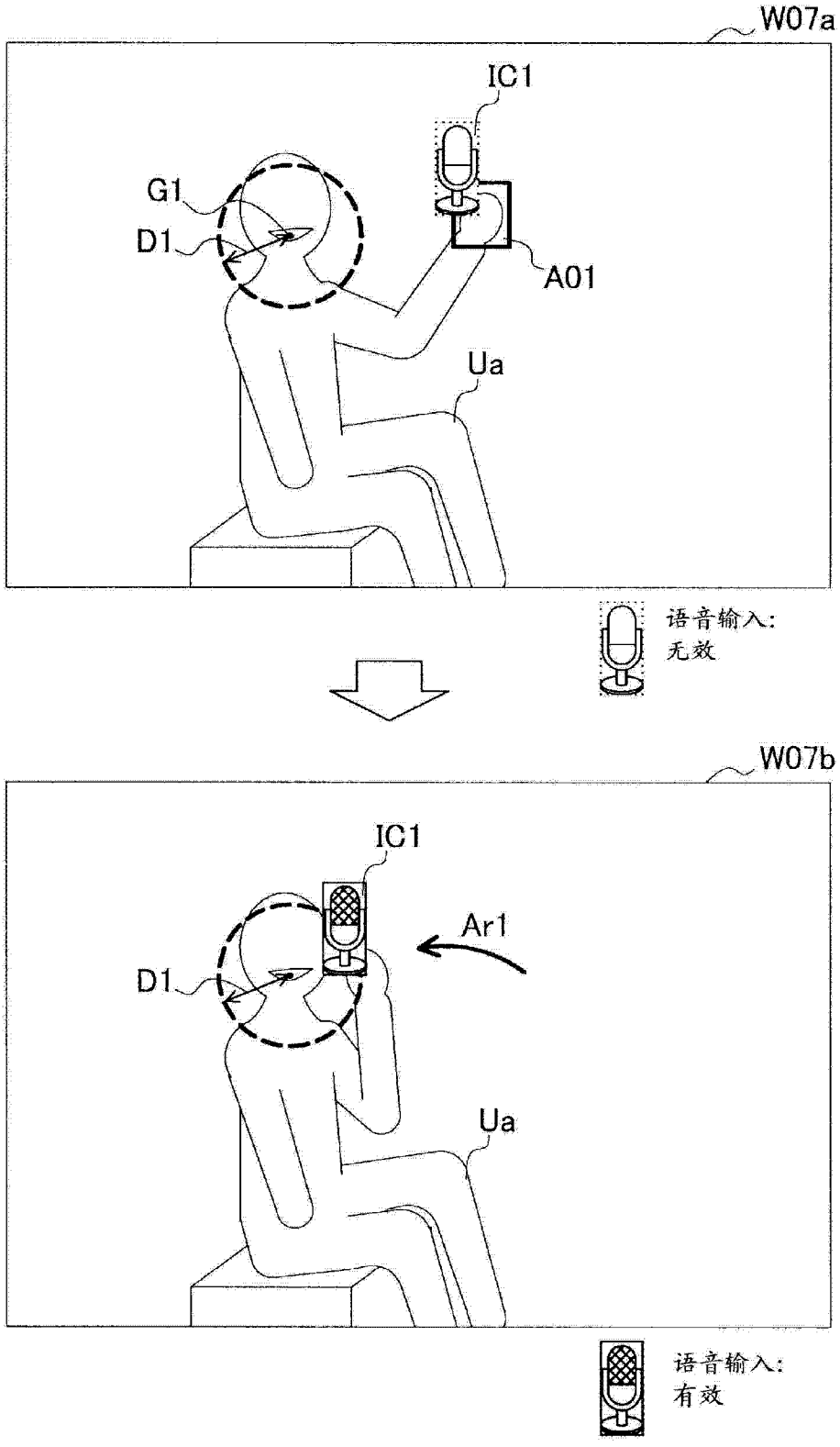


图 9

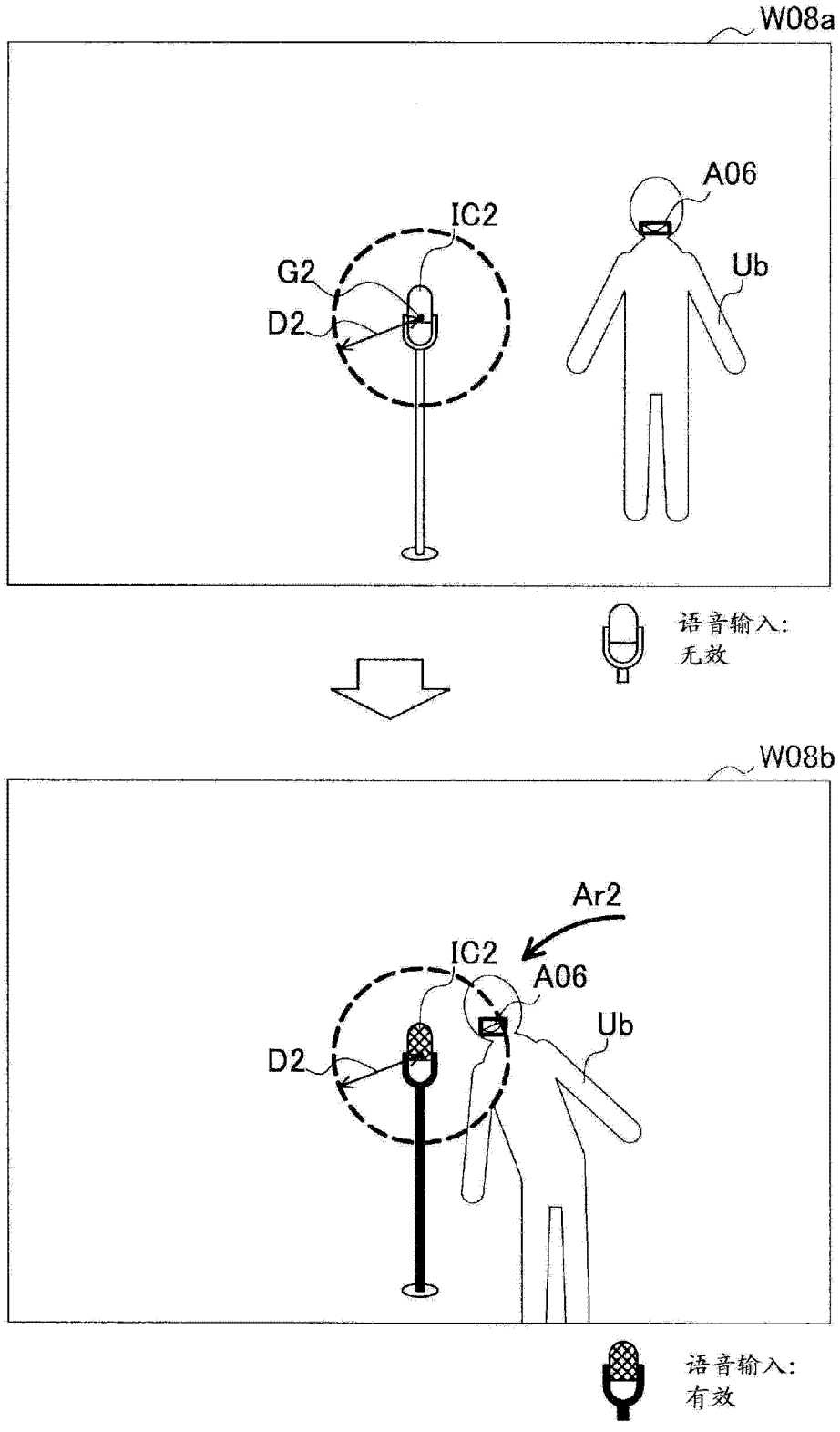


图 10

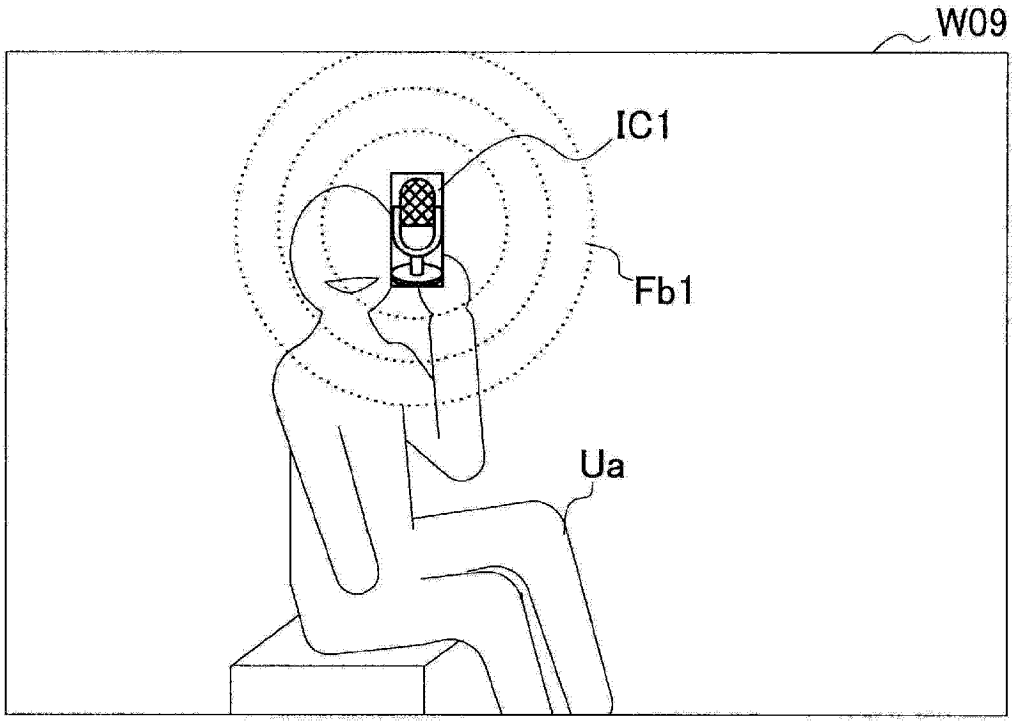


图 11

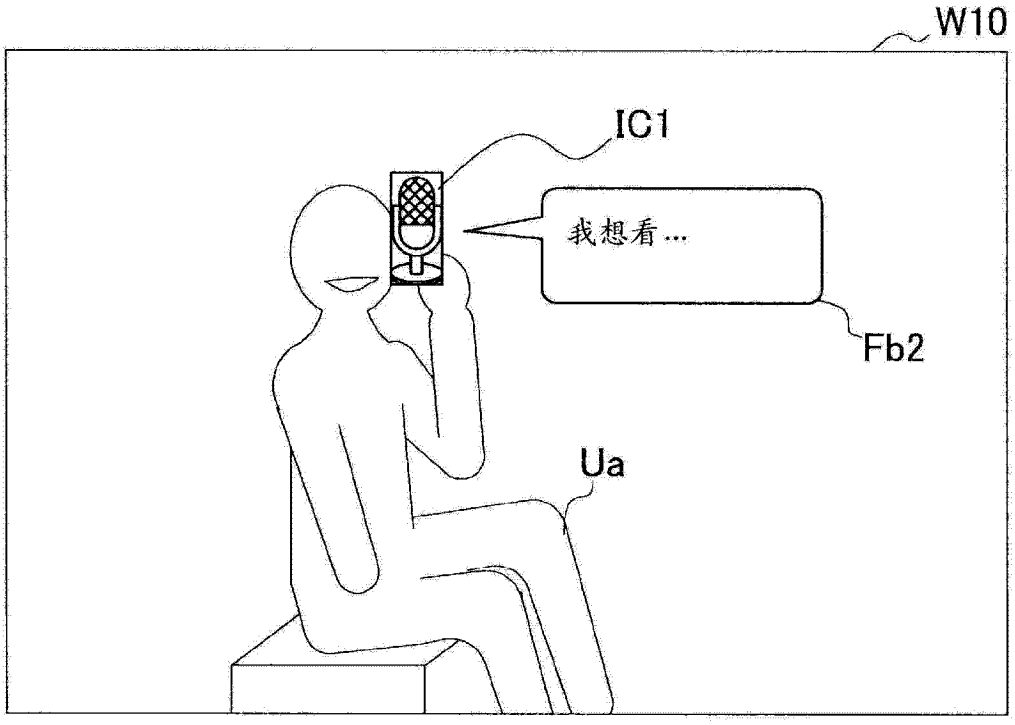


图 12

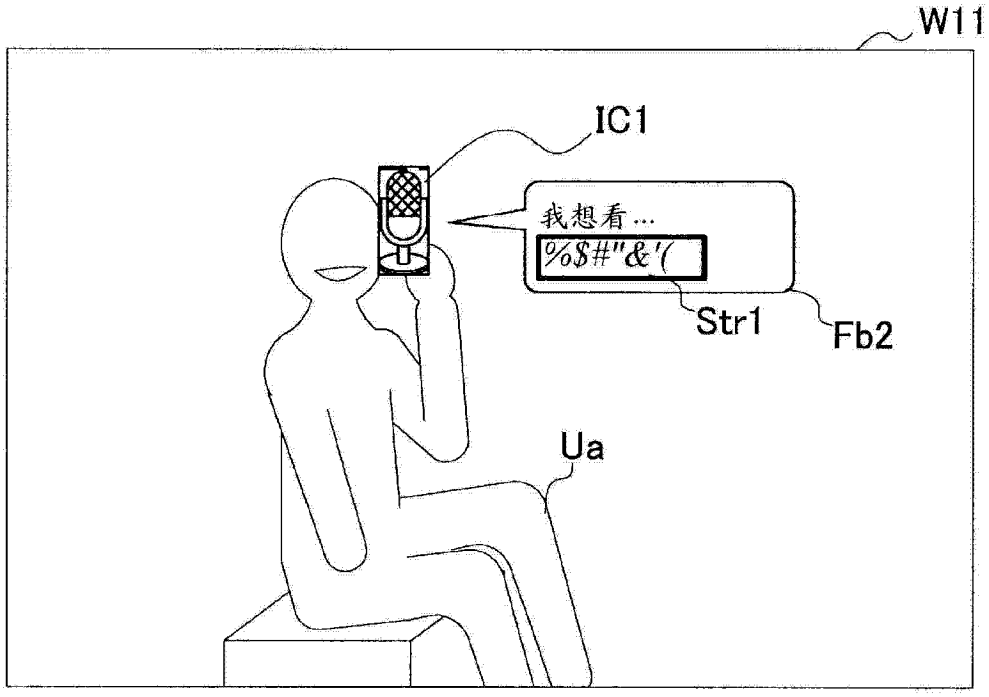


图 13

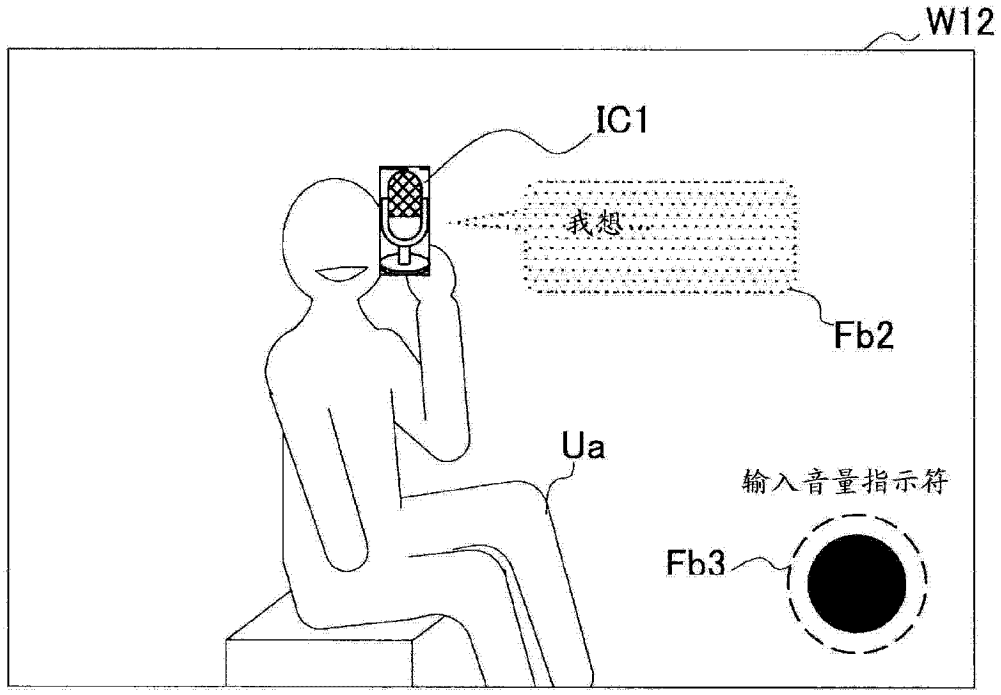


图 14

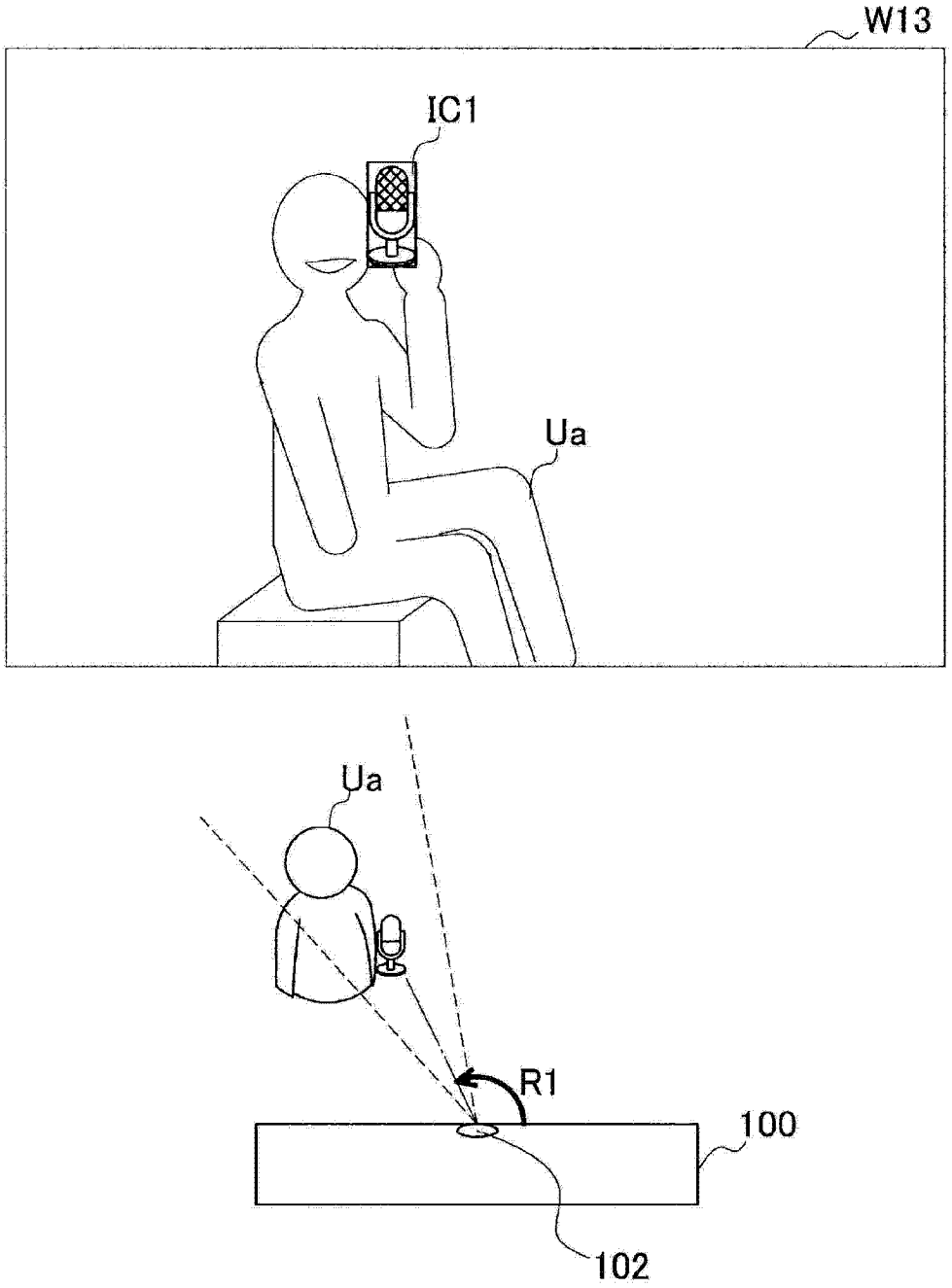


图 15

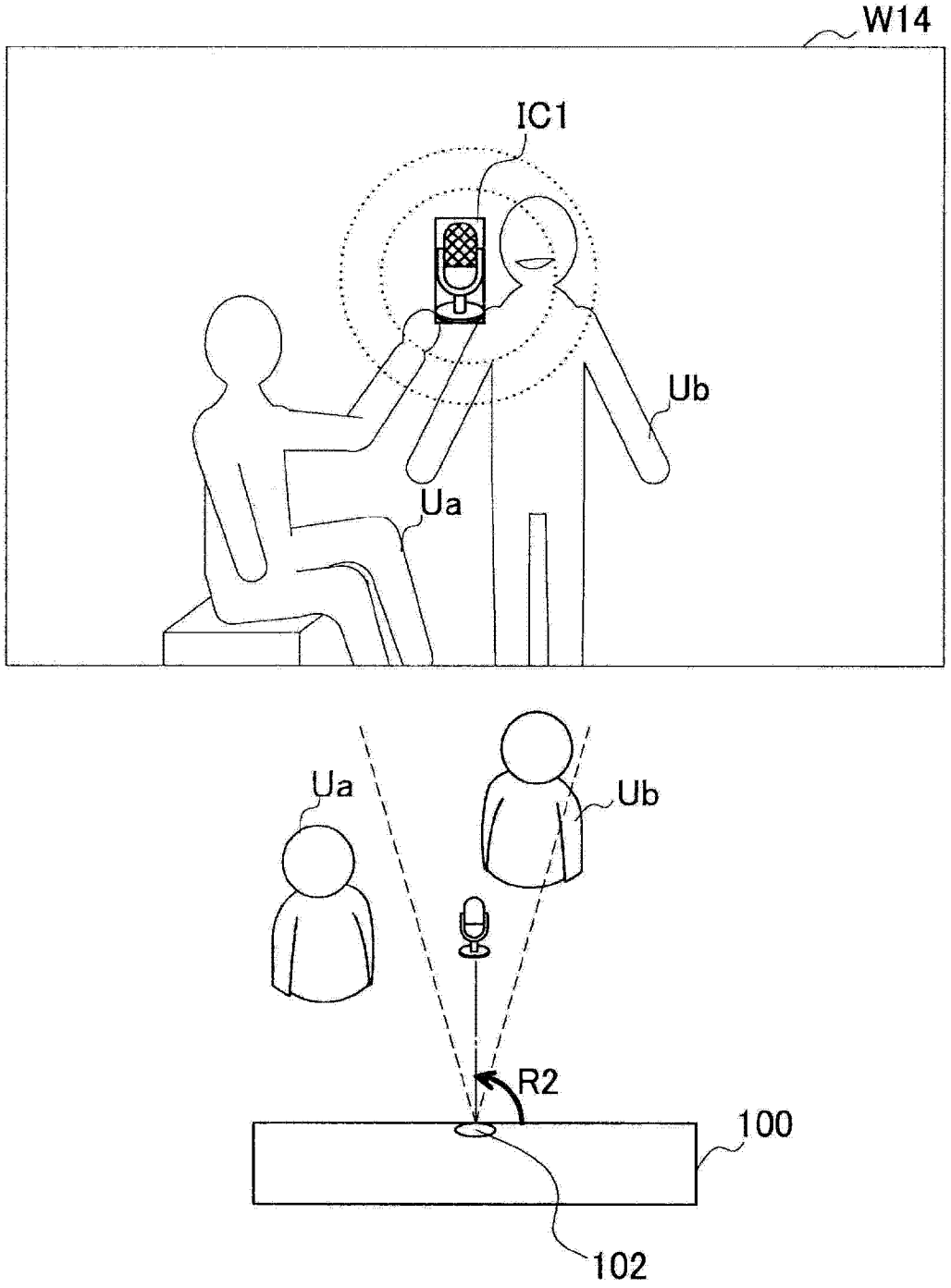


图 16

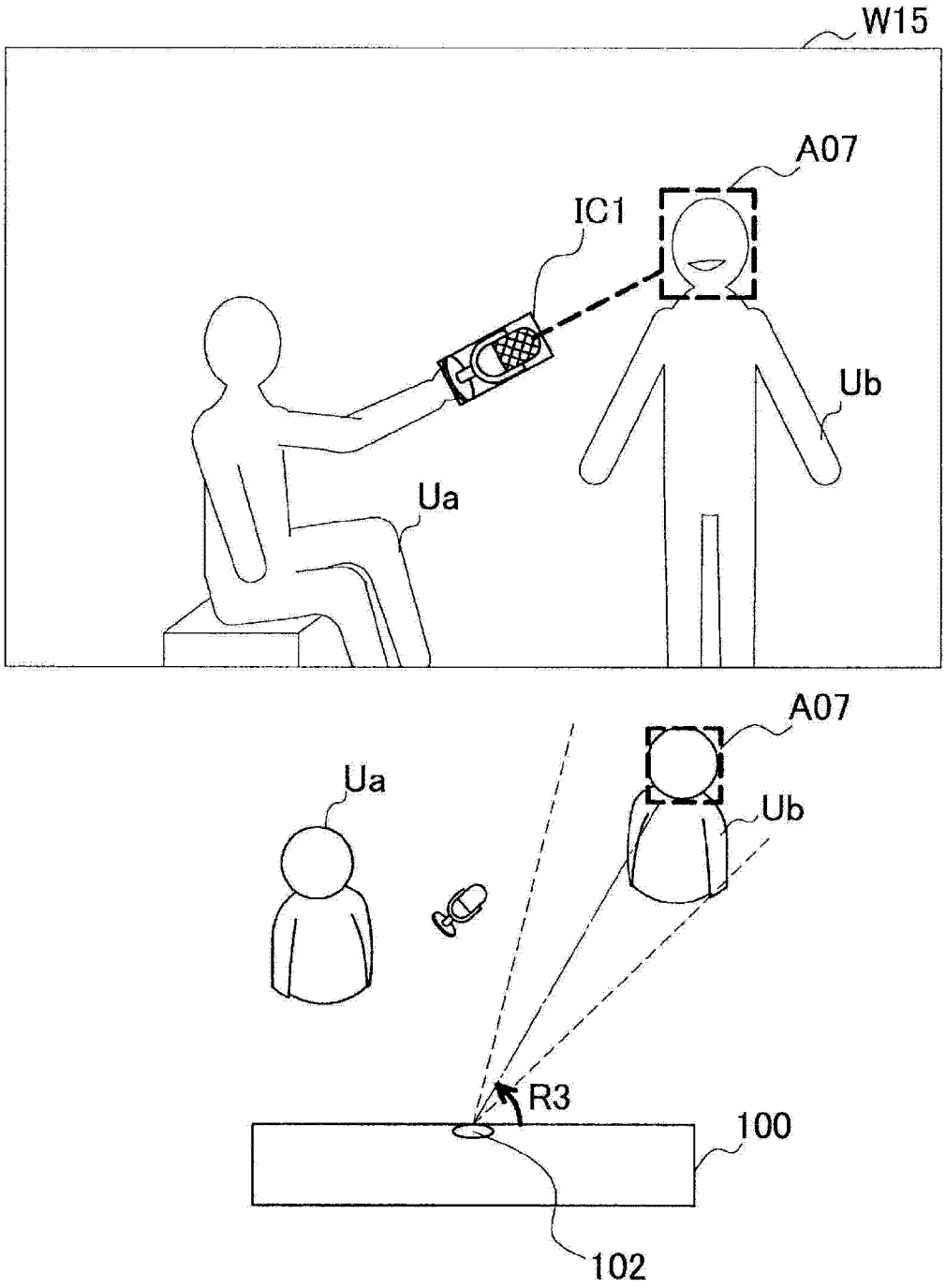


图 17

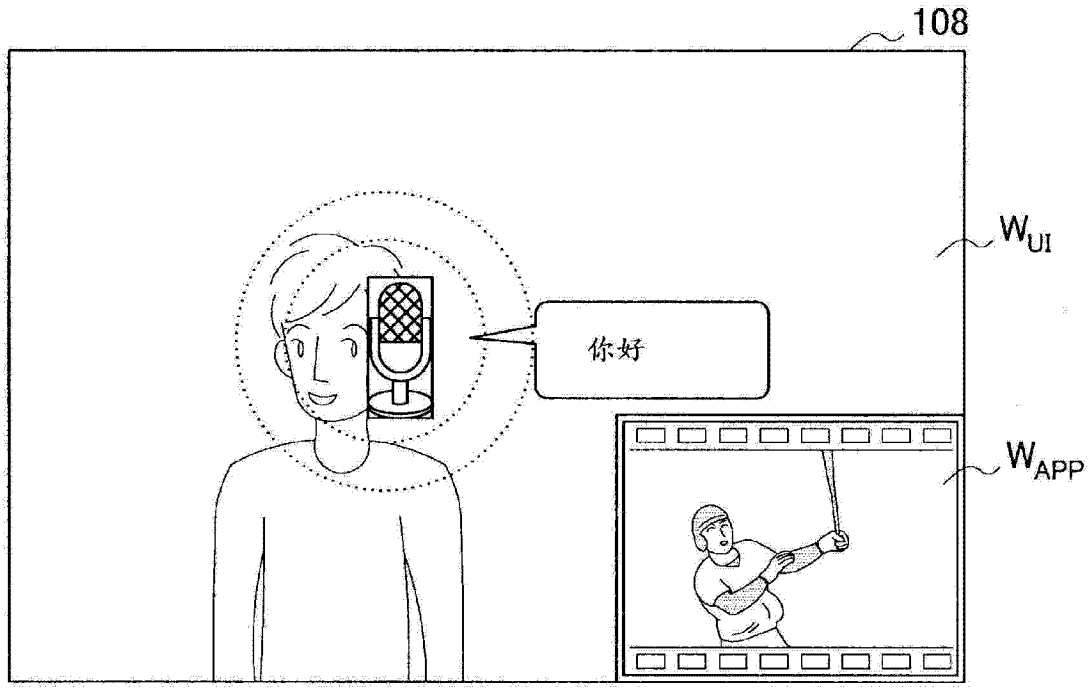


图 18

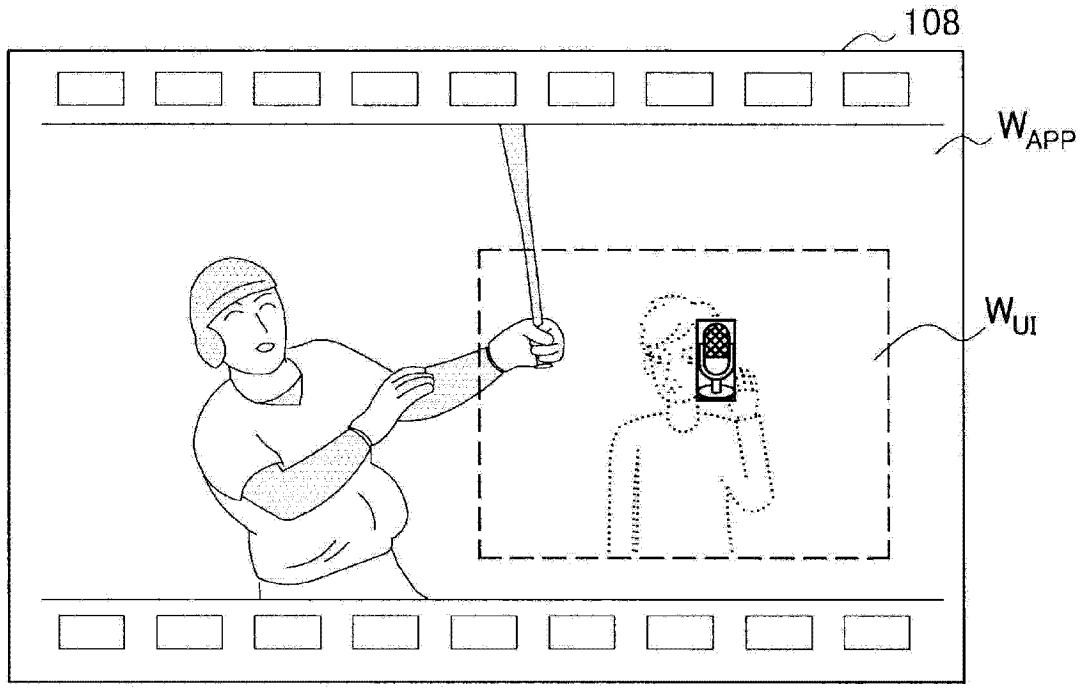


图 19

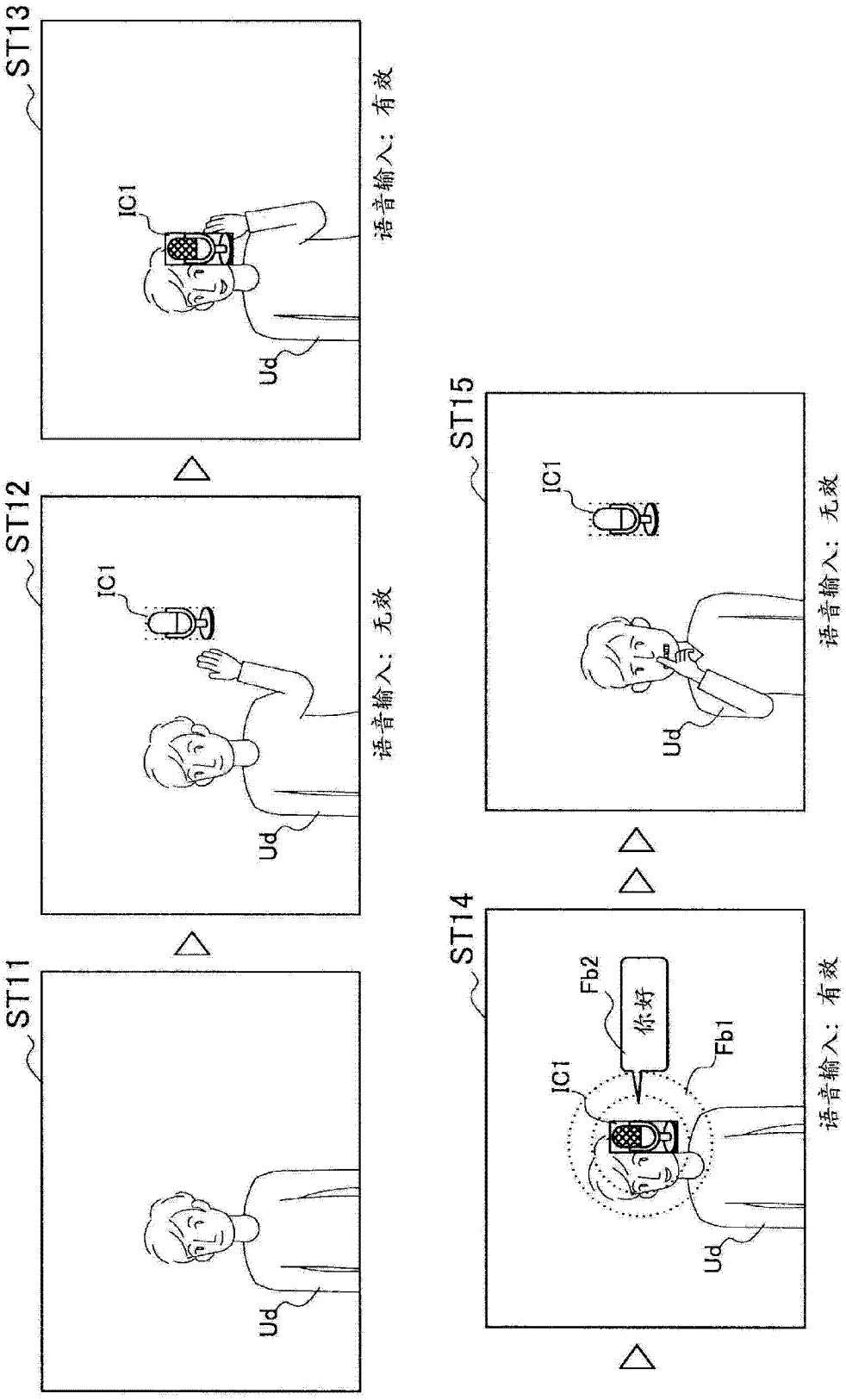

ST11
Ud
ST12
IC1
Ud
语音输入：无效
ST13
IC1
Ud
语音输入：有效
ST14
Fb2
你好
Fb1
IC1
Ud
语音输入：有效
ST15
IC1
Ud
语音输入：无效

图 20

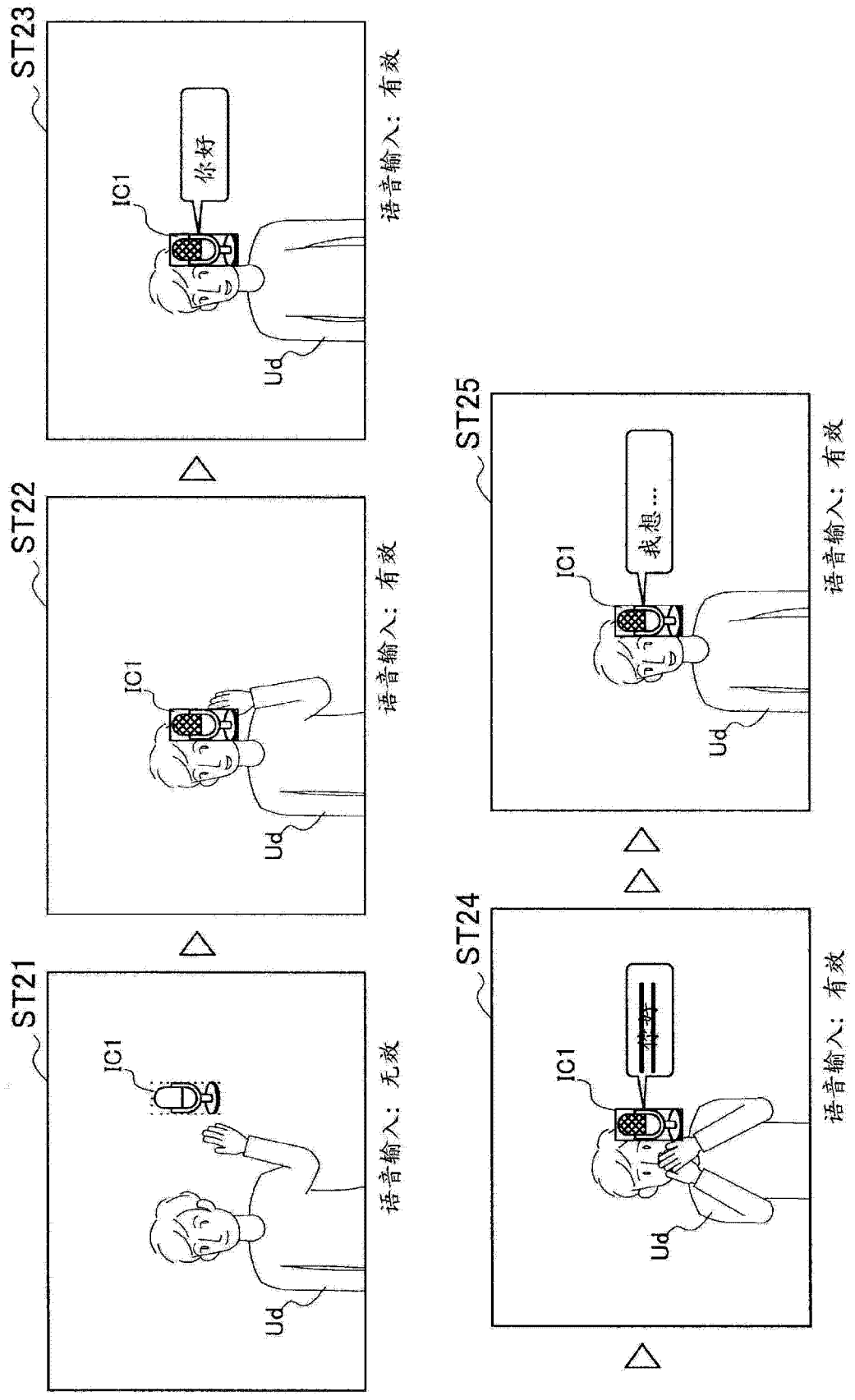

ST21
IC1
Ud
语音输入：无效
ST22
IC1
Ud
语音输入：有效
ST23
IC1
你好
Ud
语音输入：有效
ST24
IC1
Ud
语音输入：有效
ST25
IC1
我想...
Ud
语音输入：有效

图 21

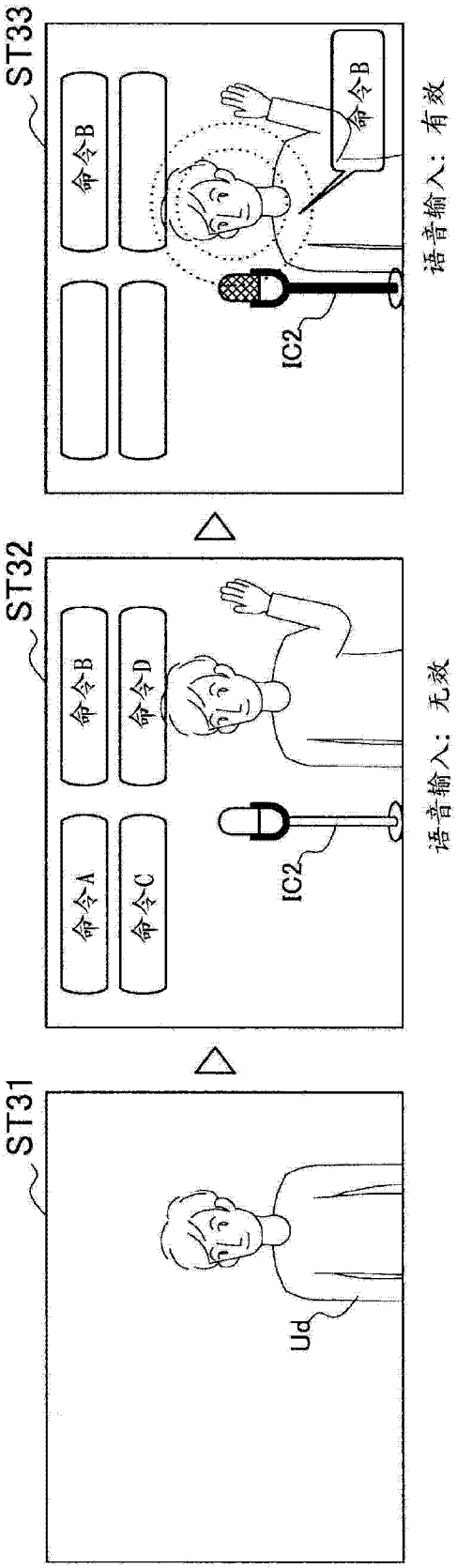

ST31
Ud
ST32
命令A
命令C
命令B
命令D
IC2
语音输入：无效
ST33
命令B
命令B
IC2
语音输入：有效

图 22

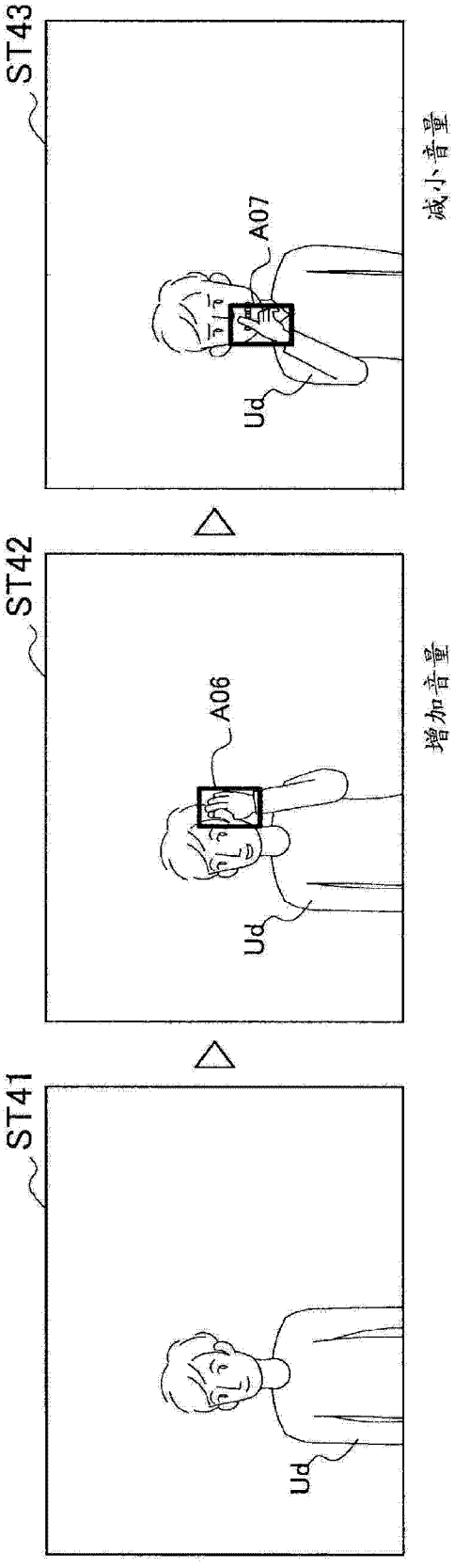

ST41
Ud
ST42
A06
Ud
增加音量
ST43
A07
Ud
减小音量

图 23

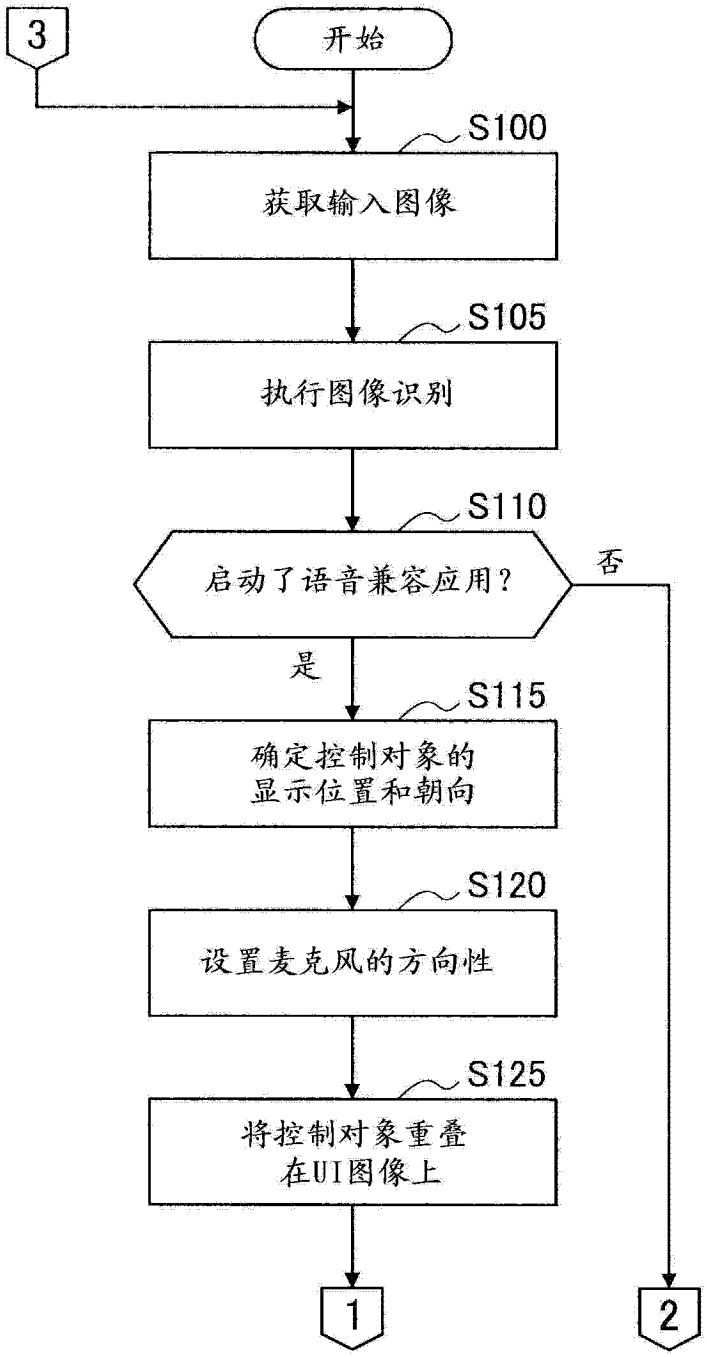

3
开始
S100
获取输入图像
S105
执行图像识别
S110
启动了语音兼容应用？
否
是
S115
确定控制对象的
显示位置和朝向
S120
设置麦克风的方向性
S125
将控制对象重叠
在UI图像上
1
2

图 24

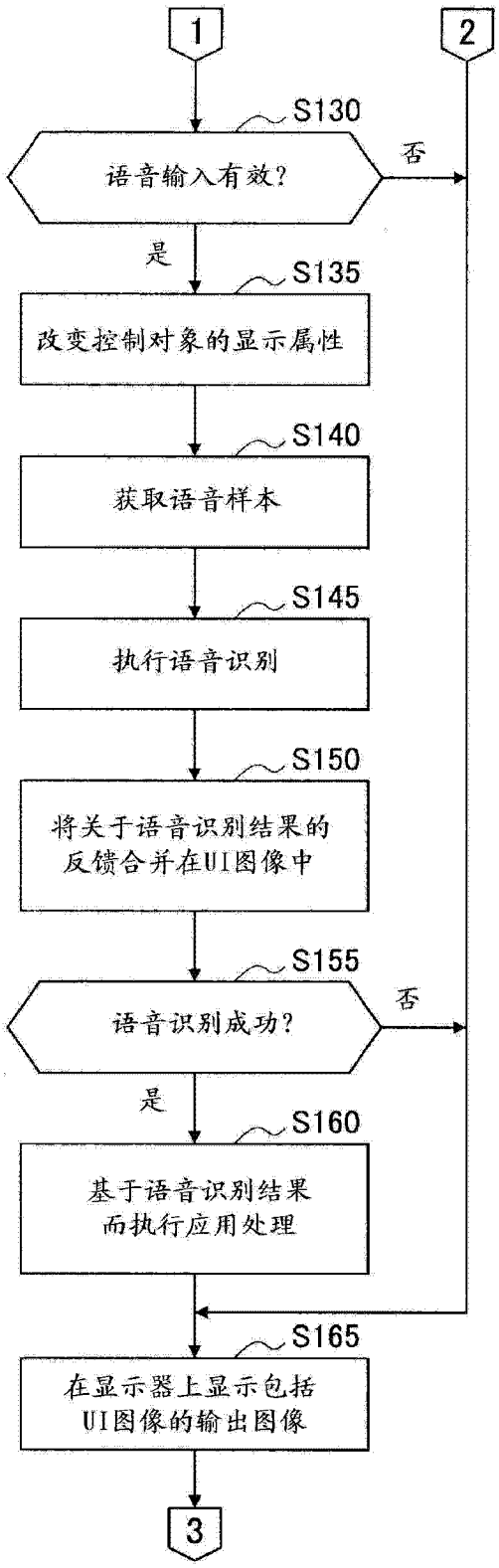

1
2
S130
语音输入有效?
否
是
S135
改变控制对象的显示属性
S140
获取语音样本
S145
执行语音识别
S150
将关于语音识别结果的
反馈合并在UI图像中
S155
语音识别成功?
否
是
S160
基于语音识别结果
而执行应用处理
S165
在显示器上显示包括
UI图像的输出图像
3

图 25

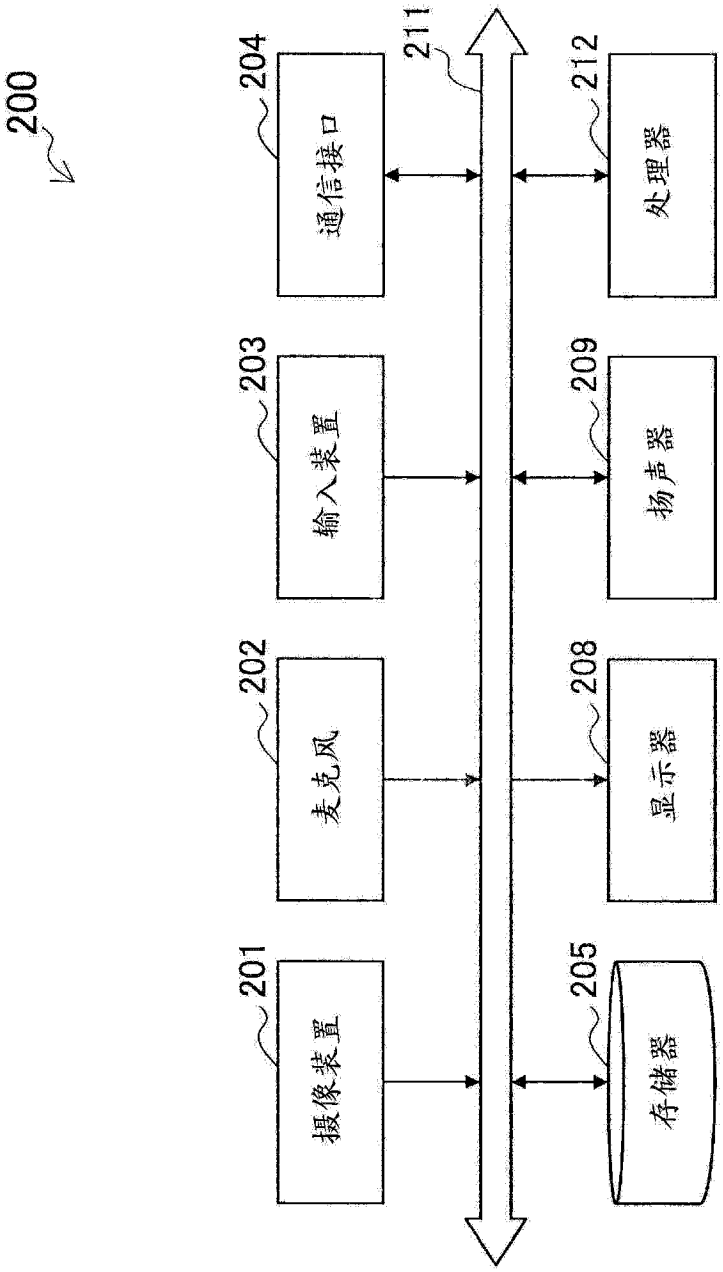

200
211
204
通信接口
212
处理器
203
输入装置
209
扬声器
202
麦克风
208
显示器
201
摄像装置
205
存储器

图 26

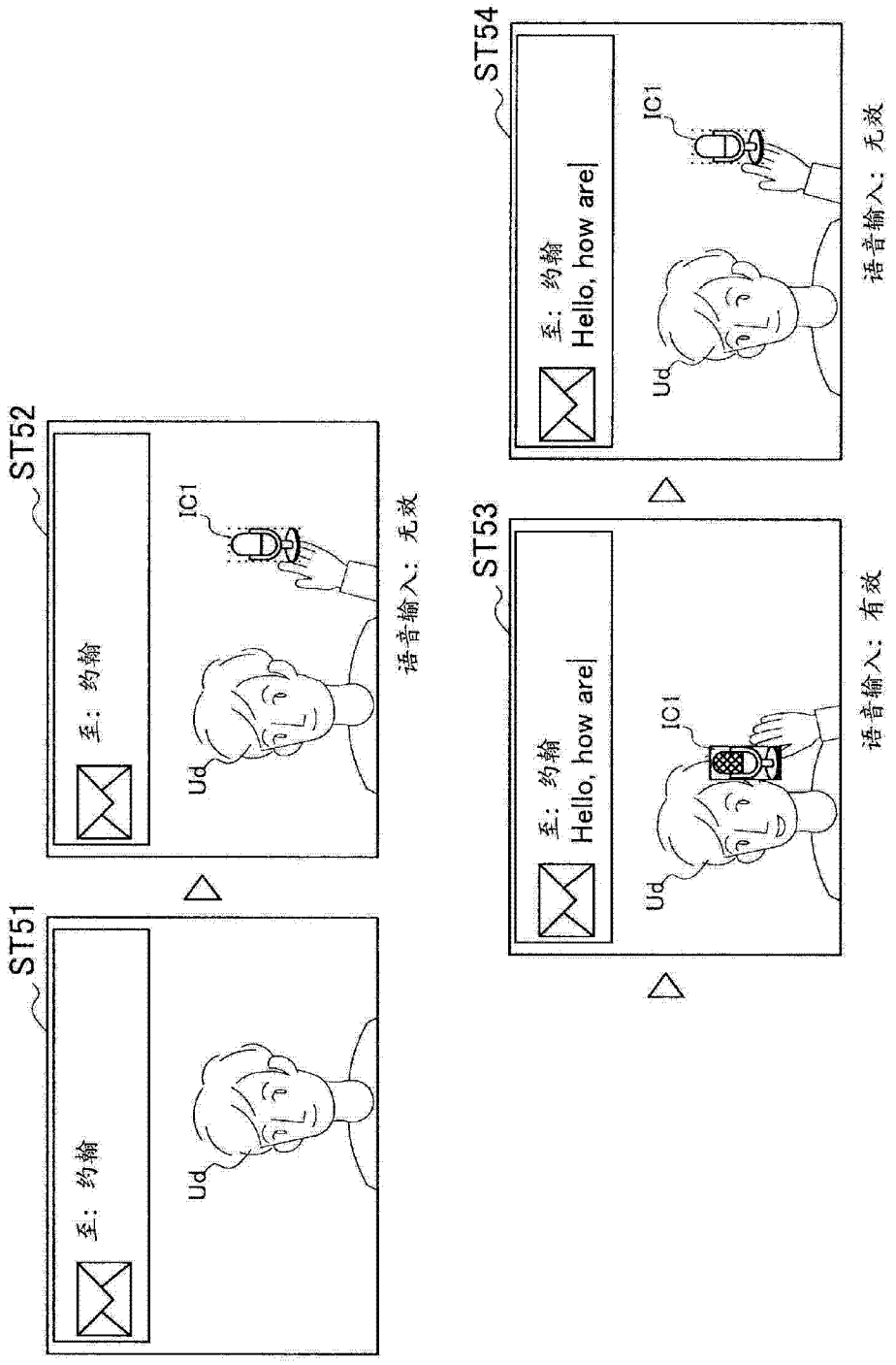

ST51
至：约翰
Ud
ST52
至：约翰
IC1
Ud
语音输入：无效
ST53
至：约翰
Hello, how are|
IC1
Ud
语音输入：有效
ST54
至：约翰
Hello, how are|
IC1
Ud
语音输入：无效

图 27