(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116711006 A
(43)申请公布日 2023.09.05

(21)申请号 202180087499.2

(22)申请日 2021.09.15

(30)优先权数据
10-2021-0023992 2021.02.23 KR

(85)PCT国际申请进入国家阶段日
2023.06.26

(86)PCT国际申请的申请数据
PCT/KR2021/012596 2021.09.15

(87)PCT国际申请的公布数据
WO2022/181911 KO 2022.09.01

(71)申请人 三星电子株式会社
地址 韩国京畿道水原市灵通区三星路129号

(72)发明人 韩尉汎 宋雅诗 李在原

(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
专利代理师 付乐 陈颖慧

(51)Int.Cl.
G10L 17/14(2006.01)

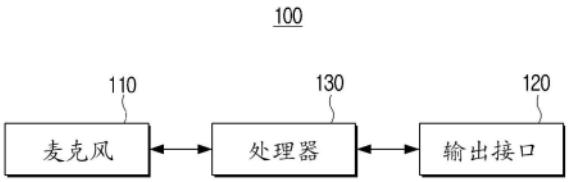
权利要求书2页 说明书13页 附图6页

(54)发明名称

电子装置及其控制方法

(57)摘要

电子装置包括:麦克风;输出接口;以及处理器,被配置为:基于由用户在电子装置上登记的词,从通过麦克风接收到的音频信号中检测说出登记词的说话者的语音;基于在电子装置上登记的登记说话者的语音信息,确定所检测到的语音是否是登记说话者的语音;以及基于确定出所检测到的语音是登记说话者的语音,控制输出接口输出语音通知,其中所述登记词提及用户。



CN 116711006 A

1.一种电子装置,包括:

麦克风;

输出接口;以及

处理器,被配置为:

基于由用户在所述电子装置上登记的词,从通过所述麦克风接收到的音频信号中检测说出登记词的说话者的语音;

基于在所述电子装置上登记的登记说话者的语音信息,

确定所检测到的语音是否是所述登记说话者的语音;以及

基于确定出所检测到的语音是所述登记说话者的语音,

控制所述输出接口输出语音通知,

其中,所述登记词提及所述用户。

2.根据权利要求1所述的装置,其中,所述语音通知指示所述登记说话者正在呼叫所述用户。

3.根据权利要求1所述的装置,其中,所述处理器还被配置为:

将所述登记词的语音输入到第一神经网络模型中,并且从所述第一神经网络模型获得第一输出值;

将使用所述麦克风获得的语音输入到所述第一神经网络模型,并且从所述第一神经网络模型获得第二输出值;以及

基于所述第一输出值和所述第二输出值,从通过所述麦克风接收到的音频信号中检测说出所述登记词的所述说话者的语音。

4.根据权利要求3所述的装置,其中,所述处理器还被配置为:基于输入到所述电子装置以登记所述词的所述用户的文本和语音中的至少一个来获得所登记词的语音。

5.根据权利要求1所述的装置,其中,所述处理器还被配置为:

将所述登记词的语音和所述登记说话者的语音信息输入到第二神经网络模型,并且基于所述登记说话者的语音风格来获得从所述登记词的语音转换的转换语音;以及

基于所述转换语音来确定所检测到的语音是否是所述登记说话者的语音。

6.根据权利要求5所述的装置,其中,所述处理器还被配置为:

将所述转换语音输入到第三神经网络模型中,从所述第三神经网络模型获得第三输出值;

将所检测到的语音输入到所述第三神经网络模型,并且从所述第三神经网络模型获得第四输出值;以及

基于所述第三输出值和所述第四输出值来确定所检测到的语音是否是所述登记说话者的语音。

7.根据权利要求5所述的装置,其中,所述登记说话者的语音信息包括由所述登记说话者说出的语音。

8.根据权利要求1所述的装置,其中,所述处理器还被配置为:基于确定出所检测到的语音不是所述登记说话者的语音,通过使用所检测到的语音来将所述说话者的语音信息存储在所述电子装置中。

9.一种用于控制包括麦克风的电子装置的方法,所述方法包括:

基于由用户在所述电子装置上登记的词,从通过所述麦克风接收到的音频信号中检测说出登记词的说话者的语音;

基于在所述电子装置上登记的登记说话者的语音信息,确定所检测到的语音是否是所述登记说话者的语音;以及

基于确定出所检测到的语音是所述登记说话者的语音,输出语音通知,

其中,所述登记词提及所述用户。

10.根据权利要求9所述的方法,其中,所述语音通知指示所述登记说话者正在呼叫所述用户。

11.根据权利要求9所述的方法,还包括:

将所述登记词的语音输入到第一神经网络模型中,并且从所述第一神经网络模型获得第一输出值;

将使用所述麦克风获得的语音输入到所述第一神经网络模型,并且从所述第一神经网络模型获得第二输出值;以及

基于所述第一输出值和所述第二输出值,从通过所述麦克风接收到的音频信号中识别说出所述登记词的所述说话者的语音。

12.根据权利要求11所述的方法,还包括:

基于输入到所述电子装置以登记所述词的所述用户的文本和语音中的至少一个来获得所登记词的语音。

13.根据权利要求9所述的方法,还包括:

将所述登记词的语音和所述登记说话者的语音信息输入到第二神经网络模型,并且基于所述登记说话者的语音风格来获得从所述登记词的语音转换的转换语音;以及

基于所述转换语音来确定所检测到的语音是否是所述登记说话者的语音。

14.根据权利要求13所述的方法,还包括:

将所述转换语音输入到第三神经网络模型中,从所述第三神经网络模型获得第三输出值;

将所检测到的语音输入到所述第三神经网络模型,并且从所述第三神经网络模型获得第四输出值;以及

基于所述第三输出值和所述第四输出值来确定所检测到的语音是否是所述登记说话者的语音。

15.根据权利要求13所述的方法,其中,所述登记说话者的语音信息包括由所述登记说话者说出的语音。

电子装置及其控制方法

技术领域

[0001] 本公开涉及一种电子装置及其控制方法，更具体地，涉及一种用于收集和提供周围语音信息的电子装置及其控制方法。

背景技术

[0002] 近年来，诸如无线耳机、无线头戴式耳机等的、经由无线通信输出从移动装置接收到的音频信号的电子装置已经被商用。

[0003] 然而，由于用户在佩戴和使用这种电子装置时难以听到用户周围的语音信息，因此会存在用户可能听不到正呼叫该用户的另一个人的语音的问题。

发明内容

[0004] [技术问题]

[0005] 提供了一种用于采集和输出周围语音信息的电子装置及其控制方法。

[0006] [技术方案]

[0007] 根据本公开的一个方面，一种电子装置包括麦克风、输出接口、以及处理器。所述处理器被配置为：基于由用户在所述电子装置上登记的词，从通过所述麦克风接收到的音频信号中检测说出登记词的说话者的语音；基于在所述电子装置上登记的登记说话者的语音信息，确定所检测到的语音是否是所述登记说话者的语音；以及基于确定出所检测到的语音是所述登记说话者的语音，控制所述输出接口输出语音通知。其中所述登记词提及所述用户。

[0008] 所述语音通知可表示示所述登记说话者正在呼叫所述用户。

[0009] 所述处理器还可被配置为：将所述登记词的语音输入到第一神经网络模型中，并且从所述第一神经网络模型获得第一输出值；将使用所述麦克风获得的语音输入到所述第一神经网络模型，并且从所述第一神经网络模型获得第二输出值；以及基于所述第一输出值和所述第二输出值，从通过所述麦克风接收到的音频信号中检测说出所述登记词的所述说话者的语音。

[0010] 所述处理器还可被配置为：基于输入到所述电子装置以登记所述词的所述用户的文本和语音中的至少一个来获得所登记词的语音。

[0011] 所述处理器还可被配置为：将所述登记词的语音和所述登记说话者的语音信息输入到第二神经网络模型，并且基于所述登记说话者的语音风格来获得从所述登记词的语音转换的转换语音；以及基于所述转换语音来确定所检测到的语音是否是所述登记说话者的语音。

[0012] 所述处理器还可被配置为：将所述转换语音输入到第三神经网络模型中，从所述第三神经网络模型获得第三输出值；将所检测到的语音输入到所述第三神经网络模型，并且从所述第三神经网络模型获得第四输出值；以及基于所述第三输出值和所述第四输出值来确定所检测到的语音是否是所述登记说话者的语音。

[0013] 所述登记说话者的语音信息可包括由所述登记说话者说出的语音。

[0014] 所述处理器还可被配置为:基于确定出所检测到的语音不是所述登记说话者的语音,通过使用所检测到的语音来将所述说话者的语音信息存储在所述电子装置中。

[0015] 根据本公开的一个方面,一种用于控制包括麦克风的电子装置的方法包括:基于由用户在所述电子装置上登记的词,从通过所述麦克风接收到的音频信号中检测说出登记词的说话者的语音;基于在所述电子装置上登记的登记说话者的语音信息,确定所检测到的语音是否是所述登记说话者的语音;以及基于确定出所检测到的语音是所述登记说话者的语音,输出语音通知,其中所述登记词提及所述用户。

[0016] 所述语音通知可表示示所述登记说话者正在呼叫所述用户。

[0017] 所述方法还可包括:将所述登记词的语音输入到第一神经网络模型中,并且从所述第一神经网络模型获得第一输出值;将使用所述麦克风获得的语音输入到所述第一神经网络模型,并且从所述第一神经网络模型获得第二输出值;以及

[0018] 基于所述第一输出值和所述第二输出值,从通过所述麦克风接收到的音频信号中识别说出所述登记词的所述说话者的语音。

[0019] 所述方法还可包括:基于输入到所述电子装置以登记所述词的所述用户的文本和语音中的至少一个来获得所登记词的语音。

[0020] 所述方法还可包括:将所述登记词的语音和所述登记说话者的语音信息输入到第二神经网络模型,并且基于所述登记说话者的语音风格来获得从所述登记词的语音转换的转换语音;以及基于所述转换语音来确定所检测到的语音是否是所述登记说话者的语音。

[0021] 所述方法还可包括:将所述转换语音输入到第三神经网络模型中,从所述第三神经网络模型获得第三输出值;将所检测到的语音输入到所述第三神经网络模型,并且从所述第三神经网络模型获得第四输出值;以及基于所述第三输出值和所述第四输出值来确定所检测到的语音是否是所述登记说话者的语音。

[0022] 所述登记说话者的语音信息可包括由所述登记说话者说出的语音。

[0023] [有益效果]

[0024] 根据本公开的各方面,当在所述电子装置上登记的说话者正在呼叫佩戴电子装置的用户时,可向用户提供对该说话者的语音通知。因此,可增强用户便利性。

附图说明

[0025] 图1是示出根据实施方式的电子装置的图;

[0026] 图2是示出根据实施方式的电子装置的配置的框图;

[0027] 图3是示出根据实施方式的电子装置的操作的流程图;

[0028] 图4a和图4b是示出根据实施方式的从电子装置输出的语音通知的示例的图;

[0029] 图5是示出根据实施方式的与外部电子装置相关联的电子装置的操作的图;

[0030] 图6A和图6B是示出根据实施方式在电子装置上显示的UI屏的示例的图;

[0031] 图7是示出根据实施方式的电子装置的其它配置的框图;以及

[0032] 图8是示出根据实施方式的用于控制电子装置的方法的流程图。

具体实施方式

[0033] 在描述本公开时，当确定详细描述可能会不必要地模糊本公开的要点时，可省略对相关技术或配置的详细描述。此外，下面的实施方式可以以各种形式改变，并且本公开技术思想的范围不限于下面的实施方式。提供实施方式以完成本公开，并且将本公开的技术思想充分传达给本领域技术人员。

[0034] 应当注意，在本公开中公开的技术不是将本公开的范围限制为特定实施方式，而是应当被解释为包括本公开实施方式的所有修改、等同和/或替换形式。关于附图的解释，类似的附图标记可用于类似的元件。

[0035] 在本公开中使用的表述“第一”、“第二”等可表示不论顺序和/或重要性的各种元件，并且可用于将一个元件与另一个元件区分开，并且不限制这些元件。

[0036] 在本公开中，诸如“A或B”、“A[和/或]B中的至少一个”、或“A[和/或]B中的一个或多个”的表述包括所列项的所有可能的组合。例如，“A或B”、“A和B中的至少一个”、或“A或B中的至少一个”包括(1)至少一个A、(2)至少一个B、或(3)至少一个A和至少一个B中的任一个。

[0037] 除非另有具体定义，单数表述可包括复数表述。应当理解，诸如“包括”或“组成”的术语在本文中用于表示特征、数字、步骤、操作、元件、部分或其组合的存在，并且不排除添加一个或多个其它特征、数字、步骤、操作、元件、部分或其组合的存在或可能性。

[0038] 如果描述了某一元件(例如，第一元件)与另一个元件(例如，第二元件)“可操作地或通信地联接”或“连接到”另一个元件(例如，第二元件)，应当理解的是，某一元件可直接或通过另一个元件(例如，第三元件)连接到另一个元件。另一方面，如果描述了某一元件(例如，第一元件)“直接联接到”或“直接连接到”另一元件(例如，第二元件)，则可理解的是，在某一元件与另一元件之间不存在元件(例如，第三元件)。

[0039] 此外，本公开中使用的表述“被配置为”可取决于情况与其它表述互换使用，例如“适于”、“具有……的能力”、“被设计成”、“适用于”、“被制造成”和“能够”。表述“被配置为(或被设置为)”并不总是指仅“被专门设计为”通过硬件实现。相反，在一些情况下，表述“被配置成……的装置”可意味着“能够”与另一装置或部件一起操作的装置。例如，短语“被配置(或设置)以执行A、B和C的处理器”可例如但不限于用于执行相应操作的专用处理器(例如，嵌入式处理器)或可通过执行存储在存储器中的至少一个软件程序来执行操作的通用处理器(例如，中央处理单元(CPU)或应用处理器)。

[0040] 图1是示出根据实施方式的电子装置的图。

[0041] 参考图1，根据本公开实施方式的电子装置100可输出音频信号。例如，电子装置100可输出存储在电子装置100中的音频信号，或者从以有线或无线方式连接的外部电子装置接收音频信号并输出所接收到的音频信号。

[0042] 在这种情况下，电子装置100可被实现为耳机或头戴式耳机，其在佩戴在用户的耳朵上或覆盖用户的耳朵时输出声音信息。

[0043] 同时，由于音频信号是从电子装置100输出的，因此佩戴电子装置100的用户10可能难以听到正在呼叫该用户10的另一用户20的语音。

[0044] 根据本公开实施方式的电子装置100可接收周围的音频信号，并且如果从所接收到的音频信号中识别出正在呼叫该用户的另一用户语音，则电子装置100可输出语音通知，

该语音通知用于通知另一用户正在呼叫该用户。

[0045] 因此，根据本公开的实施方式，用户可在佩戴电子装置100时识别另一用户正在呼叫该用户的情况，从而增强了便利性。

[0046] 图2是示出根据实施方式的电子装置的配置的框图。

[0047] 参考图2，电子装置100可包括麦克风110、输出接口120和处理器130。

[0048] 麦克风110可以是用于接收音频信号的配置。换句话说，麦克风110可接收周围的声音作为音频信号。在这种情况下，麦克风110可连续地接收音频信号。音频信号可包括每次关于特定频率、幅度、振动次数、波形等的信息，并且音频信号可以是模拟信号或数字信号的形式。

[0049] 输出接口120可以是用于输出音频信号的配置。例如，输出接口120可使用电信号来移动音圈，并且随着音圈的移动来振动附接到音圈的振动膜以再现音频信号。

[0050] 处理器130可控制电子装置100的总体操作。为此，处理器130可以是诸如中央处理单元(CPU)或应用处理器(AP)的通用处理器，诸如图形处理单元(GPU)或视觉处理单元(VPU)的图形专用处理器，或诸如神经处理单元(NPU)的人工智能专用处理器等。此外，处理器130可包括加载至少一个指令或模块的易失性存储器。

[0051] 在下文中，将参考图3更详细地描述本公开的操作。

[0052] 首先，在操作S310，处理器130可经由输出接口120输出音频信号。

[0053] 具体地，处理器130可经由输出接口120输出存储在电子装置100的存储器140(参见图7)中的音频信号，或者从有线或无线方式连接的外部电子装置接收音频信号、并且经由输出接口120输出所接收到的音频信号。

[0054] 例如，外部电子装置可从提供音乐内容流服务的服务器接收音乐内容，并且将所接收到的音乐内容发送到电子装置100。在这种情况下，处理器130可控制输出接口120，以输出从电子装置200接收到的音乐内容。

[0055] 此外，基于用户在电子装置100上登记的词(或关键字)，处理器130可从经由麦克风110接收到的音频信号中识别说出登记词的说话者的语音。在实施方式中，术语“说话者”可表示说出特定词的说话者，这可意味着说出或讲出特定词的人。换句话说，在操作S320，处理器130可从麦克风110接收到的语音中识别登记词。

[0056] 为此，处理器130可从经由麦克风110接收到的音频信号中检测语音。例如，处理器130可在经由麦克风110连续地接收到的音频信号中检测出电平超过预定电平的部分中的音频信号作为语音信号。这是为了通过在音频信号中识别出某段中的音频信号作为语音信号并且仅针对该语音信号进行处理，从而减少数据处理的目标数以减少操作量。同时，音频信号的电平可以以分贝(dB)、电压或能量为单位。然而，这仅仅是一个示例，并且处理器130可通过使用各种公知方法来从经由麦克风110接收到的音频信号中检测语音。

[0057] 处理器130可从检测到的语音中识别说出登记词的说话者的语音。

[0058] 在本文中，词可在电子装置100上预先登记。换句话说，关于词的信息可预先存储在存储器140中。

[0059] 在这种情况下，处理器130可基于用户输入来登记词。具体地，处理器130可将与根据用户输入而接收到的词有关的信息存储在存储器140中，以用于登记该词。

[0060] 例如，用户可通过使用语音来登记词。在这种情况下，在登记词的过程中，当经由

麦克风110接收到用户说出的语音时，处理器130可将所接收到的语音数据存储在存储器140中。

[0061] 在另一示例中，用户可使用显示在外部电子装置上的虚拟键盘将词输入到外部电子装置。在这种情况下，外部电子装置可将输入词的文本数据发送到电子装置100。在登记词的过程中，如果从外部电子装置接收到文本形式的词，则处理器130可将所接收到的文本数据存储在存储器140中。

[0062] 同时，登记词可包括提及用户的词。例如，该词可包括该用户通常被另一个人呼叫的词，诸如用户的姓名、职位等。

[0063] 同时，处理器130可通过使用第一神经网络模型来识别说出登记词的说话者的语音。

[0064] 具体地，处理器130可将登记词的语音和经由麦克风110接收到的语音分别输入到第一神经网络模型，并且识别说出登记词的说话者的语音。

[0065] 在本文中，第一神经网络模型可以是经训练以对语音进行分类的分类模型。例如，第一神经网络模型可通过使用训练数据集来训练，该训练数据集包括多个语音(例如，语音数据)和针对每个语音的标签。在这种情况下，针对每个语音的标签可以是由每个语音表示的词。此外，当处理器130向第一神经网络模型输入语音(即，语音数据)时，第一神经网络模型可输出输入语音所属的标签作为结果值。

[0066] 首先，处理器130可将登记词的语音输入到第一神经网络模型，并且从第一神经网络模型获得输出值(第一输出值)。

[0067] 为此，处理器130可获得登记词的语音。在本文中，登记词的语音可以是以语音形式表示登记词的语音。

[0068] 具体地，处理器130可基于输入到电子装置100以用于登记该词的用户文本和语音中的至少一个来获得登记词的语音。

[0069] 例如，如果用户已经使用语音登记了词，则存储器140可存储通过说出该登记词而获得的用户语音。在这种情况下，处理器130可通过使用存储在存储器140中的语音数据来获得登记词的语音。

[0070] 在另一示例中，如果用户通过使用显示在外部电子装置上的虚拟键盘登记了词，则存储器140可存储登记词的文本数据。在这种情况下，处理器130可通过使用文本到语音(TTS)模块，将文本数据转换为语音数据以获得登记词的语音。

[0071] 因此，处理器130可将登记词的语音输入到第一神经网络模型，以从第一神经网络模型获得输出值(第一输出值)。

[0072] 在本文中，输出值可以不是第一神经网络模型的最终输出值，而可以是从配置第一神经网络模型的多个层中的一个层获得的输出值(即，向量值)。

[0073] 具体地，第一神经网络模型可包括卷积层、池化层、完全连接层等。在这种情况下，处理器130可将登记词的语音输入到第一神经网络模型，并且从位于第一神经网络模型后部的完全连接层获得来自一个层(例如，嵌入层)的输出值。

[0074] 此外，处理器130可将经由麦克风110接收到的语音输入到第一神经网络模型，以从第一神经网络模型获得输出值(第二输出值)。

[0075] 在本文中，输出值可以不是第一神经网络模型的最终输出值，而可以是从配置第

一神经网络模型的多个层中的一个层获得的输出值(即,向量值)。

[0076] 具体地,第一神经网络模型可包括卷积层、池化层、完全连接层等。在这种情况下,处理器130可将经由麦克风110接收到的语音输入到第一神经网络模型,并且从位于第一神经网络模型后部的完全连接层获得来自一个层(例如,嵌入层)的输出值。

[0077] 此外,处理器130可基于从第一神经网络模型获得的输出值,从经由麦克风110接收到的音频信号中识别说出登记词的说话者的语音。

[0078] 具体地,如果通过将登记词的语音输入到第一神经网络模型而获得的输出值(第一输出值)与通过将经由麦克风110接收到的语音输入到第一神经网络模型而获得的输出值(第二输出值)之间的差等于或小于预定值,则处理器130可识别出经由麦克风110接收到的语音识别是说出登记词的说话者的语音。此外,如果通过将登记词的语音输入到第一神经网络模型而获得的输出值(第一输出值)与通过将经由麦克风110接收到的语音输入到第一神经网络模型而获得的输出值(第二输出值)之间的差大于预定值,则处理器130可确定出经由麦克风110接收到的语音不是说出登记词的说话者的语音。

[0079] 例如,在本文中,输出值之间的差值可通过内积运算来计算内部输出值(即,向量值)之间的距离来获得。

[0080] 如上所述,处理器130可针对经由麦克风110接收到的语音执行上述处理,以在经由麦克风110接收到的语音中识别说出登记词的说话者的语音。

[0081] 然后,处理器130可基于在电子装置100上登记的说话者的语音信息,识别所识别的语音是否是登记说话者的语音。换句话说,在操作S330,处理器130可识别说出登记词的说话者是否是登记说话者。

[0082] 处理器130可将登记词的语音和登记说话者的语音信息输入到第二神经网络模型,以基于登记说话者的语音风格来获得由登记词的语音转换的语音。

[0083] 在本文中,登记说话者的语音信息可包括由登记说话者说出的语音。换句话说,存储器140可预先存储登记说话者的语音数据。例如,在佩戴电子装置100的用户与说话者之间的对话期间,处理器130可经由麦克风110接收说话者的语音,并将所接收到的语音存储在存储器140中。

[0084] 同时,第二神经网络模型可以是经训练以根据目标语音的风格来转换语音的模型。例如,第二神经网络模型可接收语音和目标语音的输入,通过使用输入目标语音的特征(例如,音调、语调、语速、口音等)来转换输入语音而使得输入语音具有目标语音的风格,并输出转换的语音。如上所述,第二神经网络模型可执行语音转换(或语音模拟)。在这种情况下,例如,第二神经网络模型可包括编码器、解码器等,并且可被实现为基于生成对抗网络(GAN)的各种模型。

[0085] 因此,处理器130可将登记词的语音和登记说话者的语音输入到第二神经网络模型,以从第二神经网络模型获得貌似登记说话者说出登记词的语音。

[0086] 处理器130可基于从第二神经网络模型获得的语音(即,转换语音),识别说出登记词的说话者的语音是否是登记说话者的语音。

[0087] 在这种情况下,处理器130可通过使用第三神经网络模型来识别说出登记词的说话者的语音是否是登记说话者的语音。

[0088] 具体地,处理器130可将转换语音和说出登记词的说话者的语音中分别输入到第

三神经网络模型,以识别说出登记词的说话者的语音是否是登记说话者的语音。

[0089] 在本文中,第三神经网络模型可以是经训练以对语音进行分类的分类模型。例如,第三神经网络模型可通过使用训练数据集来训练,该训练数据集包括多个语音和针对每个语音的标签。在这种情况下,针对每个语音的标签可以是说出每个语音的人。当输入语音时,第三神经网络模型可输出输入语音所属的标签作为结果值。

[0090] 首先,处理器130可将经转换语音输入到第三神经网络模型,以从第三神经网络模型获得输出值(第三输出值)。

[0091] 在本文中,输出值可以不是第三神经网络模型的最终输出值,而可以是从配置第三神经网络模型的多个层中的一个层获得的输出值(即,向量值)。

[0092] 具体地,第三神经网络模型可包括卷积层、池化层、完全连接层等。在这种情况下,处理器130可将经转换语音输入到第三神经网络模型,并且从位于第三神经网络模型后部的完全连接层获得来自一个层(例如,嵌入层)的输出值。

[0093] 此外,处理器130可将说出登记词的说话者的语音输入到第三神经网络模型,以从第三神经网络模型获得输出值(第四输出值)。

[0094] 在本文中,输出值可以不是第三神经网络模型的最终输出值,而可以是从配置第三神经网络模型的多个层中的一个层获得的输出值(即,向量值)。

[0095] 具体地,第三神经网络模型可包括卷积层、池化层、完全连接层等。在这种情况下,处理器130可将说出登记词的说话者的语音输入到第三神经网络模型,并且从位于第三神经网络模型的后部的完全连接层获得来自一个层(例如,嵌入层)的输出值。

[0096] 处理器130可基于从第三神经网络模型获得的输出值,识别说出登记词的说话者的语音是否是登记说话者的语音。

[0097] 具体地,如果通过将经转换语音输入到第三神经网络模型而获得的输出值(第三输出值)与通过将说话者的语音输入到第三神经网络模型而获得的输出值(第四输出值)之间的差等于或小于预定值,则处理器130可识别出说出登记词的说话者的语音是登记说话者的语音。此外,如果通过将经转换语音输入到第三神经网络模型而获得的输出值(第三输出值)与通过将说话者的语音输入到第三神经网络模型而获得的输出值(第四输出值)之间的差大于预定值,则处理器130可识别出说出登记词的说话者的语音不是登记说话者的语音。

[0098] 例如,在本文中,输出值之间的差值可通过内积运算来计算输出值(即,向量值)之间的距离来获得。

[0099] 通过上述方法,在操作S340,处理器130可识别说出登记词的说话者的语音是否是登记说话者的语音。

[0100] 当识别出说出登记词的说话者的语音是登记说话者的语音时(在操作S340为“是”),在操作S350,处理器130可控制输出接口120输出语音通知。

[0101] 在本文中,语音通知可包括通知登记说话者正在呼叫用户的语音通知。

[0102] 例如,假设通过麦克风110接收到的、说出登记词的说话者的语音是登记说话者AAA的语音。在这种情况下,参考图4a,处理器130可经由输出接口120输出语音通知“AAA呼叫用户”。

[0103] 在这种情况下,如果经由输出接口120输出声音,则处理器130可停止正在经由输

出接口120输出的声音输出，并经由输出接口120输出语音通知。在经由输出接口120输出语音通知之后，处理器130可再次输出声音或维持停止声音输出的状态。

[0104] 当识别出说出登记词的说话者的语音不是登记说话者的语音时(在操作S340为“否”)，在操作S360，处理器130可控制输出接口120输出语音通知。

[0105] 在这种情况下，可不指定正在呼叫用户的说话者，因此可输出与在说出登记词的说话者的语音是登记说话者的语音的情况下的语音通知不同的语音通知。

[0106] 具体地，语音通知可包括通知某人正在呼叫用户的语音通知。例如，参考图4b，处理器130可经由输出接口120输出语音通知“某人呼叫用户”。

[0107] 在这种情况下，如果正在经由输出接口120输出声音，则处理器130可停止经由输出接口120输出的声音输出，并经由输出接口120输出语音通知。在经由输出接口120输出语音通知之后，处理器130可再次输出声音或维持停止声音输出的状态。

[0108] 此外，当识别出说出登记词的说话者的语音不是登记说话者的语音时，处理器130可在操作中使用经由麦克风110接收到的语音来登记说出登记词的说话者。换句话说，处理器130可将说话者的语音信息存储在电子装置100中。

[0109] 具体地，当说话者呼叫用户时，用户随后可与说话者进行对话。对话可在用户与呼叫该用户的说话者之间进行，或者在多个人与其他人之间进行。

[0110] 在这种情况下，处理器130可从经由麦克风110接收到的音频信号中检测多个语音。

[0111] 当一对一地执行对话时，从音频信号检测到的多个语音可包括用户的语音和呼叫该用户的说话者的语音。此外，当在多个人之间进行对话时，从音频信号中检测到的多个语音可包括用户的语音、呼叫该用户的说话者的语音、以及其他人的语音。

[0112] 在这种情况下，处理器130可将从音频信号中检测到的多个语音聚类为多个组。

[0113] 具体地，处理器130可从多个语音中的每一个获得特征向量，并且基于所获得的特征向量将多个语音聚类为多个组。例如，处理器130可将多个语音聚类成多个组，使得距离等于或小于预定值的特征向量的语音属于同一组。在这种情况下，由于每个人都具有唯一的语音特征，因此当根据表示语音特征的特征向量划分多个语音时，可由说出语音的每个人来将多个语音划分成多个组。

[0114] 处理器130可经由输出接口120输出包括在多个组中的每一个中的至少一个语音。

[0115] 在这种情况下，处理器130可以以除了电子装置100的用户所属组之外的剩余组中的包括相对多语音的组的顺序，经由输出接口120输出包括在每个组中的至少一个语音。

[0116] 为此，存储器140可存储关于电子装置100的用户的语音特性(例如，特征向量)的信息。

[0117] 因此，处理器130可在多个组中确定出包括与存储在存储器140中的语音特征最相似的至少一个语音的组，作为包括该用户语音的组。在本文中，最相似的语音特征可意味着特征向量之间的距离最短。换句话说，处理器130可在多个组中确定出包括与存储在存储器140中的特征向量距离最短的组，作为包括该用户语音的组。

[0118] 当经由麦克风110接收到用于选择剩余组中的一个的用户语音、以及指示与说出包括在所选择的组中的至少一个语音的说话者有关的信息的用户语音时，处理器130可将包括在所选择的组中的至少一个语音存储在存储器140中，并且执行对该说话者的登记。

[0119] 在本文中，与说话者有关的信息可包括提及说话者的词。例如，与说话者有关的信息可包括用户通常被另一个人呼叫的词，诸如用户的姓名、职位等。

[0120] 为此，处理器130可对用户语音执行语音识别。

[0121] 在本文中，语音识别可通过自动语音识别(ASR)模块和自然语言理解(NLU)模块来执行。

[0122] ASR模块可通过使用语言模型和声学模型将检测到的语音信号转换为词或音素序列的文本(字符串)。语言模型可以是对词或音素序列分配概率的模型，并且声学模型可以是表示语音信号与该语音信号的文本之间关系的模型。这些模型可基于概率和统计数据或人工神经网络来配置。

[0123] NLU模块可通过对经转换文本使用诸如形态分析、句法分析、语义分析等各种分析方法来识别配置文本的词或句子的含义，并且基于所识别的含义来掌握语音的意图。

[0124] 例如，如果经由麦克风110接收到用户的语音“将刚输出的语音登记为BBB”，则处理器130可在接收到用户的语音所属之前，识别经由输出接口120向其输出语音的组，将包括在所识别的组中的至少一个语音存储在存储器140中作为BBB的语音信息，并登记BBB。因此，BBB可包括在登记说话者中。

[0125] 同时，处理器130可在登记说话者的过程中经由输出接口120输出各种引导语音。

[0126] 例如，处理器130可经由输出接口120输出用于引导说话者的语音是登记说话者所必需的语音指令，例如，语音“您需要与您的伙伴进行充分对话来进行说话者登记”。此外，处理器130可经由输出接口120输出用于引导关于说话者登记进程的信息的语音指令。

[0127] 通过上述方法，处理器130可登记新的说话者，然后识别说出登记词的说话者的语音是否是登记说话者的语音。

[0128] 同时，在上述示例中，描述了通过使用经由麦克风110接收到的语音来登记说话者，然而这仅仅是示例，并且电子装置100可通过使用在电话会话期间获得的语音来登记说话者。

[0129] 具体地，当用户使用电子装置100通过连接到电子装置100的诸如智能手机等外部装置以电话来与另一用户进行对话时，处理器130可将从外部电子装置接收到的另一用户的语音存储在存储器140中。

[0130] 在登记说话者的过程期间，处理器130可经由输出接口120输出存储在存储器140中的其他用户的语音，并且当经由麦克风110接收到用于选择输出语音中的至少一个的用户语音时，处理器130可将所选择的语音存储在存储器140中以执行对说话者的登记。

[0131] 图5是示出根据实施方式的与外部电子装置相关联的电子装置的操作的图。

[0132] 参考图5，电子装置100可与电子装置200通信。为此，电子装置100还可包括通信接口150，例如如图7所示。通信接口150可表示通过有线通信方法或无线通信方法与电子装置200通信的元件。通信接口150可向电子装置200发送数据或从电子装置200接收数据。例如，电子装置200可从提供音乐内容流服务的服务器接收音乐内容，并且将所接收到的音乐内容发送到电子装置100。在这种情况下，处理器130可控制输出接口120输出从电子装置200接收到的音乐内容。

[0133] 根据本公开的实施方式，电子装置100和电子装置200可彼此相关联。换句话说，上述操作可被划分，并且电子装置100和电子装置200可执行这些操作。电子装置200可被实现

为诸如服务器装置或用户智能手机等的各种装置。

[0134] 在示例中，在电子装置200上可执行以下操作中的至少一个：使用第一神经网络模型从经由麦克风110接收到的音频信号中识别说出登记词的说话者的语音的操作，使用第二神经网络模型生成转换语音的操作，以及识别说出登记词的说话者的语音是否是登记说话者的语音的操作。

[0135] 为此，电子装置200可预先存储第一神经网络模型至第三神经网络模型中的至少一个。在这种情况下，电子装置100可将经由麦克风110接收到的音频信号或说出登记词的说话者的语音发送到电子装置200。此外，电子装置200可向电子装置100发送指示如下项的信息：从音频信号中识别出的说出登记词的说话者的语音、由语音转换产生的语音、以及说出登记词的说话者的语音是否是登记说话者的语音。

[0136] 同时，在上述实施方式中，描述了电子装置100被实现为耳机或头戴式耳机，但是这仅仅是示例，并且电子装置100可被实现为诸如智能手机、平板个人计算机(PC)等装置。

[0137] 在这种情况下，电子装置100可通过与诸如耳机或头戴式耳机的外部电子装置进行通信来执行上述操作。

[0138] 具体地，电子装置100可识别出外部电子装置接收到的音频信号是来自外部电子装置的说出登记词的说话者的语音，并且识别出说出登记词的说话者是否是登记说话者。此外，电子装置100可根据说出登记词的说话者是否是登记说话者，向外部电子装置发送语音通知。因此，佩戴耳机或头戴式耳机的用户可听到语音通知。

[0139] 同时，当识别出说出登记词的说话者不是登记说话者时，电子装置100可在电子装置100的显示器上显示用于登记说话者的用户接口(UI)屏。

[0140] 例如，电子装置100可从外部电子装置接收由外部电子装置所接收到的音频信号，并且将包括在所接收到的音频信号中的多个语音聚合为多个组。电子装置100可将包括在多个组中的至少一个语音发送到外部电子装置。外部电子装置可输出从电子装置100接收到的语音。

[0141] 在这种情况下，参考图6A，电子装置100可显示用于接收组选择的UI屏610。参考图6B，当通过UI屏610选择出一个组时，电子装置100可显示用于接收与包括在所选择的组中的语音的说话者有关的信息输入的UI屏620。因此，当通过UI屏620输入与说话者有关的信息时，处理器130可将包括在所选择的组中的至少一个语音存储在存储器140中，以执行对说话者的登记。

[0142] 图7是示出根据实施方式的电子装置的其它配置的框图。

[0143] 参考图7，除了麦克风110、输出接口120和处理器130之外，根据本公开实施方式的电子装置100还可包括存储器140、通信接口150、传感器160、输入接口170、电源180等。然而，上述配置仅仅是一个示例。在执行本公开时，可将新的组成元件添加到上述配置中，或者可省略一些组成元件。

[0144] 存储器140可以是用于存储操作系统(OS)和各种数据的元件，其中OS用于控制电子装置100的组成元件的总体操作，以及各种数据与电子装置100的组成元件相关。

[0145] 为此，存储器140可被配置为临时或永久地存储数据或信息的硬件。例如，存储器140可被实现为非易失性存储器、易失性存储器、闪存、硬盘驱动器(HDD)或固态驱动器(SSD)、RAM、ROM等中的至少一种硬件。

[0146] 同时,存储器140可存储进行电子装置100的操作的各种数据。

[0147] 例如,存储器140可存储用于登记词的数据,登记说话者的语音数据,神经网络模型,以及诸如TTS模块、ASR模块、NLU模块等的各种模块。

[0148] 通信接口150可根据各种类型的通信方法与各种类型的外部装置通信,以发送和接收各种类型的数据。通信接口150可包括用于执行各种类型无线通信的电路中的至少一个、以及以太网模块、USB模块、高清多媒体接口(HDMI)、显示接口(DisplayPort)、D-subminimal(D-SUB)、数字可视接口(DVI)、迅雷(Thunderbolt)、以及执行有线通信的部件,其中各种类型的无线通信可以是诸如蓝牙模块(蓝牙或蓝牙低能量方法)、Wi-Fi模块(Wi-Fi方法)、无线通信模块(蜂窝方法,诸如3G、4G或5G)、近场通信(NFC)模块(NFC方法)、红外模块(红外方法)、zigbee模块(zigbee方法)、超宽带模块(UWB方法)、超声模块(超声方法)等。

[0149] 传感器160可被实现为各种传感器,诸如运动传感器。例如,运动传感器可检测电子装置100的移动距离、移动方向、倾斜等。为此,运动传感器可被实现为加速度传感器、陀螺仪传感器、电磁传感器等。然而,传感器160的实施方式仅仅是实施方式,并且传感器160可无任何限制地实施为各种类型的传感器。

[0150] 输入接口170可接收各种用户命令,并且将用户命令传送到处理器130。换句话说,处理器130可识别经由输入接口170从用户输入的用户命令。用户命令可通过各种方法来实现,诸如用户的触摸输入(触摸面板)、按压按键或按钮的输入、由用户说出语音的输入等。

[0151] 电源180可针对电子装置100的每个组成元件供电或停止供电。电源180可包括用于供电的电池,并且可根据有线充电方法或无线充电方法来对电池进行充电。

[0152] 图8是示出根据实施方式的用于控制电子装置的方法的流程图。

[0153] 首先,在操作S810,基于用户在电子装置100上登记的词,从经由麦克风接收到的音频信号中识别出说出登记词的说话者的语音。登记词可包括关于用户的词。

[0154] 在操作S820,基于在电子装置100上登记的说话者的语音信息,可识别出所识别的语音是否是登记说话者的语音。

[0155] 在操作S830,当识别出所识别的语音是登记说话者的语音时,输出语音通知。该语音通知可包括用于通知登记说话者正在呼叫用户的语音通知。

[0156] 同时,在操作S810中,可将登记词的语音输入到第一神经网络模型,以从第一神经网络模型获得输出值(第一输出值);可将经由麦克风接收到的语音输入到第一神经网络模型,以从第一神经网络模型获得输出值(第二输出值);并且可基于所获得的输出值,从经由麦克风接收到的音频信号中识别说出登记词的说话者的语音。

[0157] 同时,可基于输入到电子装置以用于登记该词的用户文本和语音中的至少一个来获得登记词的语音。

[0158] 在操作S820,可将登记词的语音和登记说话者的语音信息输入到第二神经网络模型,以基于登记说话者的语音风格来获得从登记词的语音转换的语音;并且可基于所获得的语音识别出所识别的语音是否是登记说话者的语音。

[0159] 在操作S820中,可将所转换的语音输入到第三神经网络模型,以从第三神经网络模型获得输出值(第三输出值);可将所识别的语音输入到第三神经网络模型,以从第三神经网络模型获得输出值(第四输出值),并且可基于所获得的输出值,识别出所识别的语音

是否是登记说话者的语音。

[0160] 同时,登记说话者的语音信息可包括由登记说话者说出的语音。

[0161] 当识别出所识别的语音不是登记说话者的语音时,可通过使用经由麦克风接收到的说话者的语音来将该说话者的语音信息存储在电子装置100中。

[0162] 同时,上面已经详细描述了从经由麦克风接收到的音频信号中识别说话者的语音、识别所识别的语音是否是登记说话者的语音、以及提供语音通知的方法。

[0163] 如上所述,根据本公开的各种实施方式,当接收到登记词的语音(即,预定关键词)时,电子装置可识别说出预定关键词的说话者是否是登记说话者。换句话说,在考虑用户说出的词的情况下执行说话者识别,即,仅在说出特定关键词的情况下执行说话者识别,因此,说话者识别率可相对优异。

[0164] 此外,当说出预定关键词的说话者是登记说话者时,电子装置可提供语音通知,用于向用户通知登记说话者正在呼叫用户。因此,用户可在佩戴电子装置的同时识别某人正在呼叫该用户的情况,从而增强了用户便利性。

[0165] 同时,可通过存储器和处理器来执行与上述神经网络模型相关的功能。处理器可由一个或多个处理器构成。一个或多个处理器可以是诸如CPU、AP等的通用处理器,诸如GPU、VPU等的图形专用处理器,或者诸如NPU等的人工智能专用处理器。该一个或多个处理器可执行控制以根据存储在非易失性存储器和易失性存储器中的预定动作规则或人工智能模型来处理输入数据。预定的动作规则或人工智能模型是通过训练来形成。

[0166] 通过本文的训练形成的可例如意味着通过将学习算法应用于多个训练数据来形成用于期望特征的预定动作规则或人工智能模型。这种训练可在根据本公开的演示人工智能的装置中执行或由单独的服务器和/或系统执行。

[0167] 人工智能模型可包括多个神经网络层。每个层具有多个权重值,并通过前一个层的处理结果以及多个权重值之间的处理来执行对该层的处理。神经网络的示例可包括卷积神经网络(CNN)、深神经网络(DNN)、递归神经网络(RNN)、受限Boltzmann机器(RBM)、深度信念网络(DBN)、双向递归深神经网络(BRDNN)和深Q网络,但是除非另有说明,本公开的神经网络不限于上述示例。

[0168] 学习算法可以是通过使用多个训练数据来训练预定目标机器(例如,机器人)以允许预定目标装置自行确定或预测的方法。学习算法的示例包括监督学习、无监督学习、半监督学习或强化学习,但是除非另有说明,本公开的学习算法不限于上述示例。

[0169] 机器可读存储介质可以以非暂时性存储介质的形式提供。在本文中,“非暂时性”存储介质是有形的,并且可不包括信号(例如,电磁波),并且该术语不区分数据被半永久地存储在存储介质中或暂时地存储在存储介质中。例如,“非暂时性存储介质”可包括临时存储数据的缓存。

[0170] 根据实施方式,根据在本公开中所公开的各种实施方式的方法可在计算机程序产品中提供。计算机程序产品可在买卖方双方之间作为可市购产品进行交易。计算机程序产品可以以机器可读存储介质(例如,光盘只读存储器(CD-ROM))的形式分发、或通过应用商城(例如,PlayStore™)在线分发(例如,下载或上传)、或直接在两个用户设备(例如,智能手机)之间分发。对于在线分发的情况,计算机程序产品的至少一部分(例如,可下载应用)可至少临时存储或临时生成在机器可读存储介质中,诸如制造商服务器的存储器、应用商城

的服务器、或中继服务器。

[0171] 根据上述各种实施方式的每个元件(例如,模块或程序)可包括单个实体或多个实体,并且在各种实施方式中可省略上述子元件中的一些子元件、或者还可进一步包括其它子元件。可替换地或另外,一些元件(例如,模块或程序)可被集成到一个实体中以执行在集成之前由每个相应元件执行的相同或类似的功能。

[0172] 根据各种实施方式,由模块、程序或其它元件执行的操作可顺序地、并行地、重复地或启发式地执行,或者至少一些操作可以以不同的顺序执行、被省略,或者可添加不同的操作。

[0173] 在本公开中,术语“单元”或“模块”可包括由硬件、软件或固件实现的单元,并且可与术语(例如,逻辑、逻辑块、部件或电路)互换使用。“单元”或“模块”可以是一体成型的部件,或者是执行一个或多个功能的部件中的最小单元或一部分。例如,模块可被实现为专用集成电路(ASIC)。

[0174] 本公开的各种实施方式可被实现为软件,该软件包括存储在机器(例如,计算机)可读存储介质中的指令。机器是调用存储在存储介质中的指令并且根据所调用的指令来操作的装置,并且可包括根据所公开的实施方式的电子装置(例如,电子装置100)。

[0175] 在由处理器执行指令的情况下,处理器可直接或者在处理器的控制下使用其它元件执行对应于指令的功能。该指令可包括由编译器产生的代码或由解析器执行的代码。尽管已经示出和描述了本公开的优选实施方式,但是本公开不限于上述具体实施方式,并且显而易见的是,本公开所属技术领域的普通技术人员可在不脱离如所附权利要求书所要求的本公开主旨的情况下对其进行各种修改。此外,这些修改不应独立于本公开的技术思想或前景来解释。

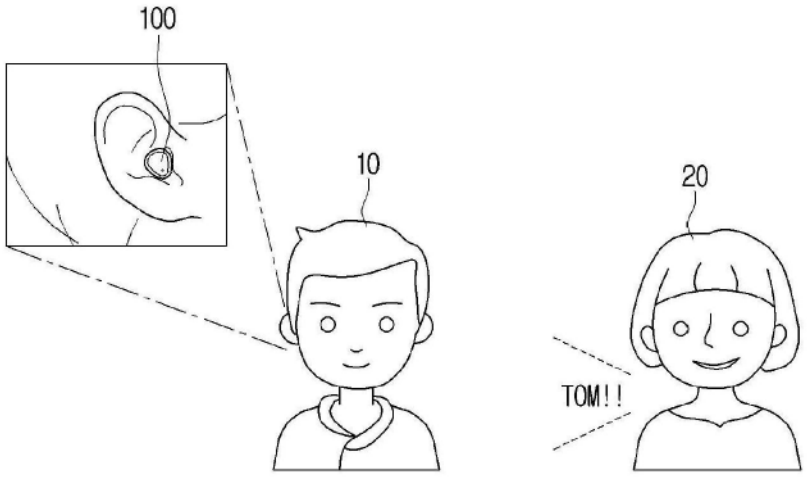


图1

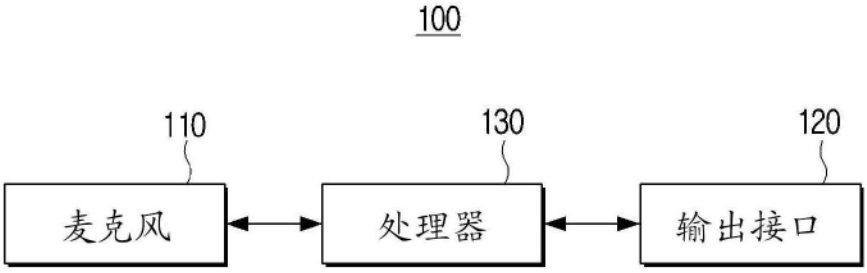


图2

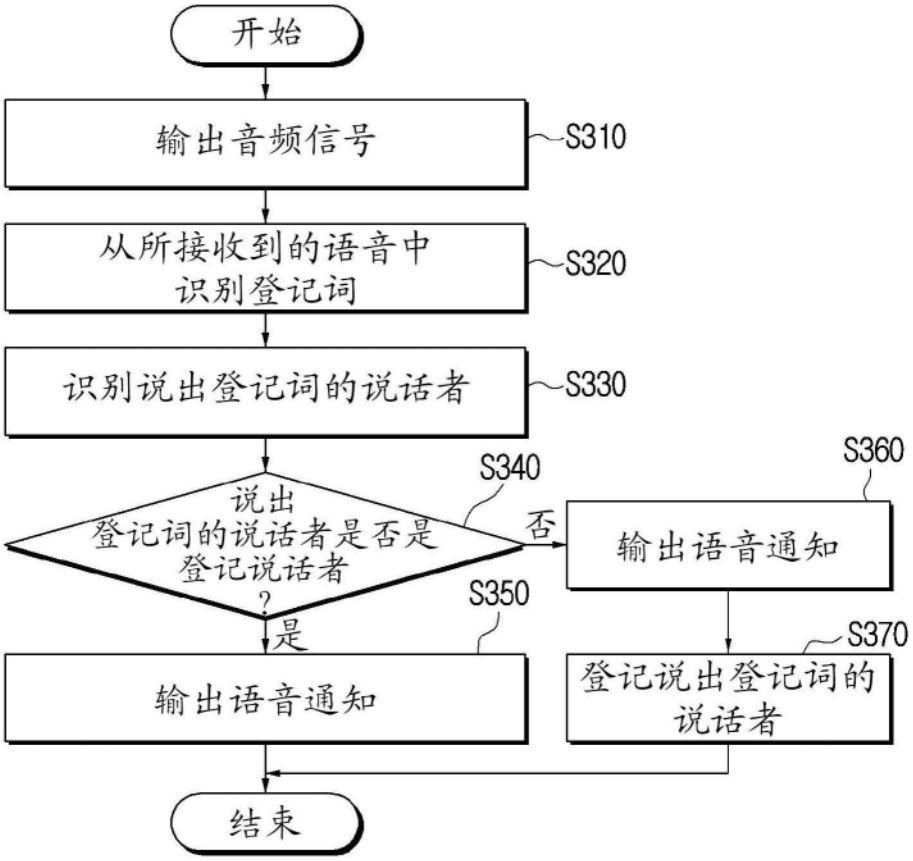

开始
输出音频信号
S310
从所接收到的语音中
识别登记词
S320
识别说出登记词的说话者
S330
S340
说出
登记词的说话者是否是
登记说话者
?
否
S360
输出语音通知
是
S350
输出语音通知
S370
登记说出登记词的
说话者
结束

图3

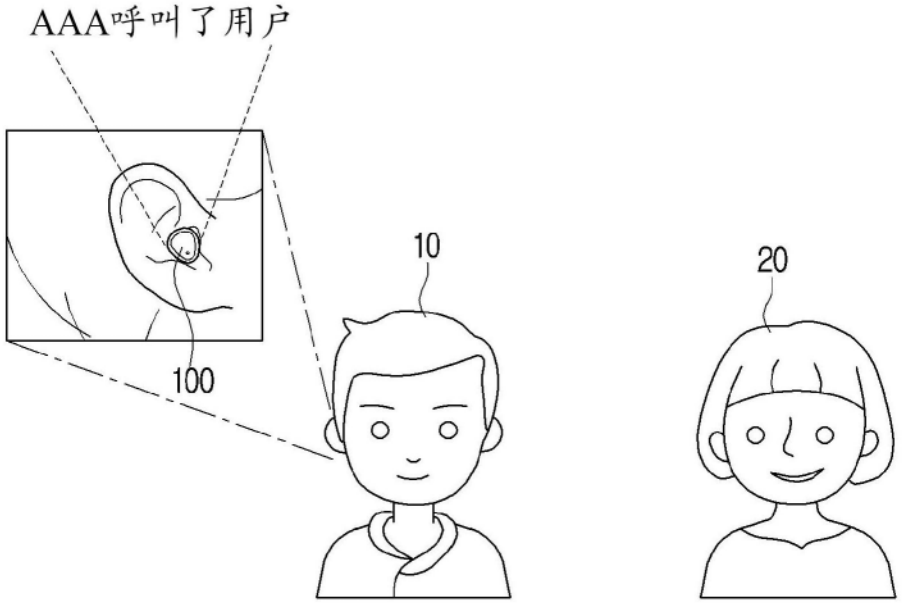

AAA呼叫了用户
10
20
100

图4a

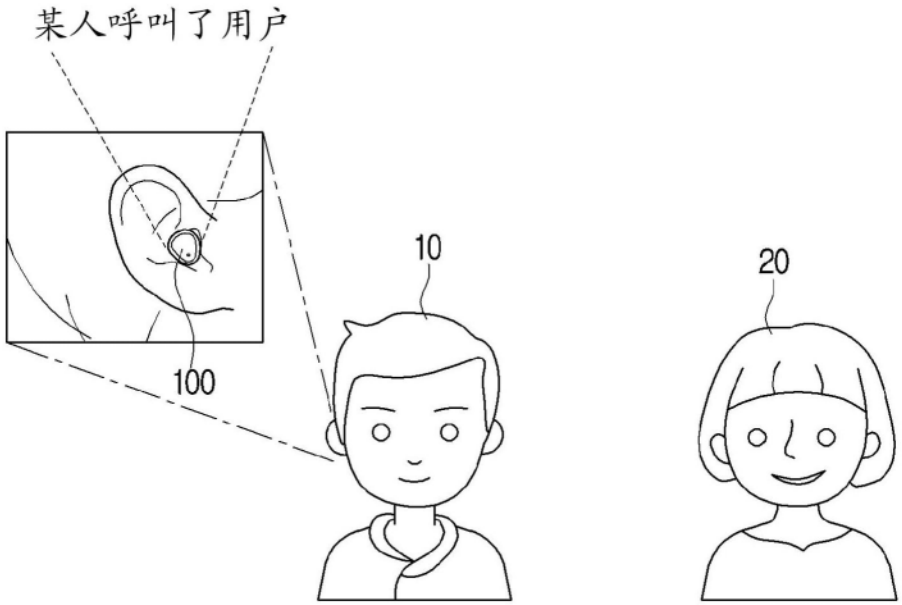


图4b

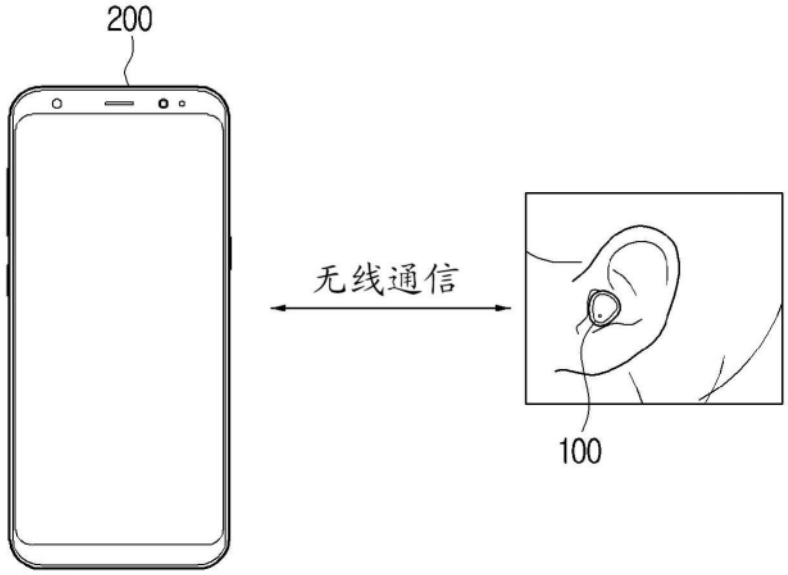


图5

图6a

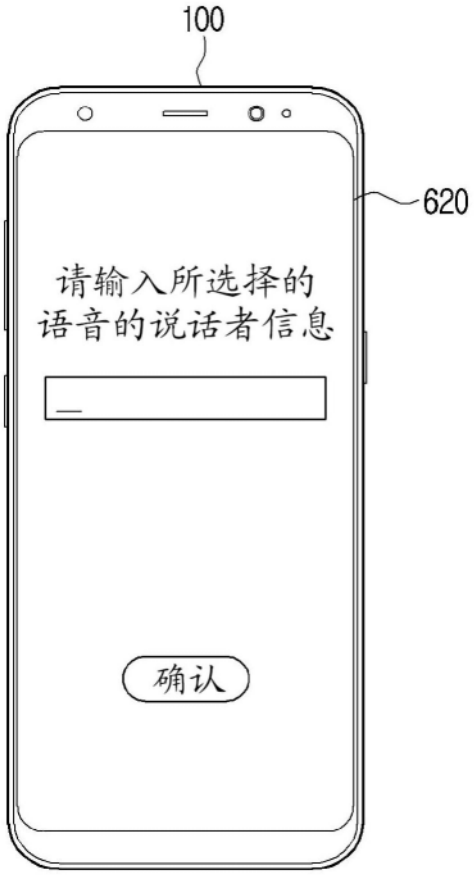

100
620
请输入所选择的
语音的说话者信息
确认

图6b

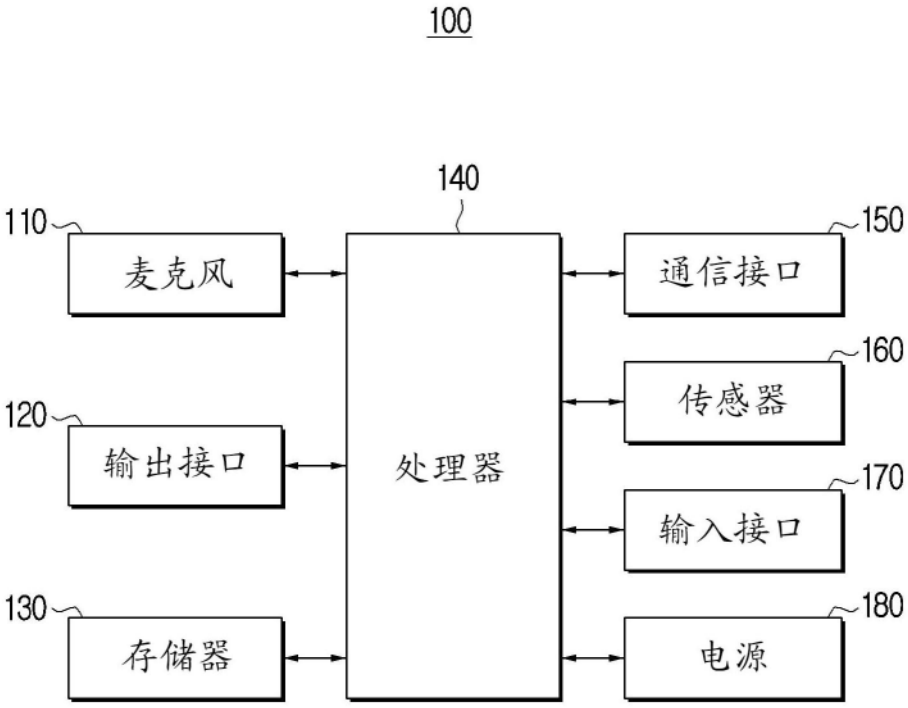

100
140
110
麦克风
120
输出接口
130
存储器
处理器
150
通信接口
160
传感器
170
输入接口
180
电源

图7

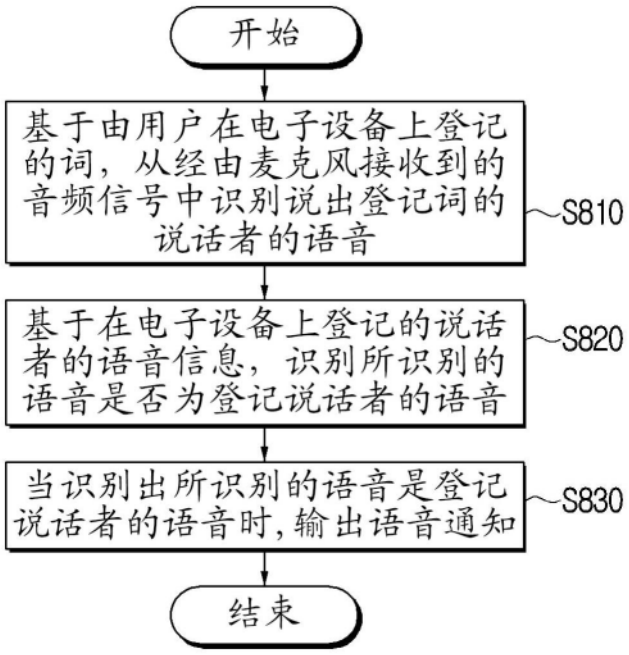

开始
基于由用户在电子设备上登记的词，从经由麦克风接收到的音频信号中识别说出登记词的说话者的语音
S810
基于在电子设备上登记的说话者的语音信息，识别所识别的语音是否为登记说话者的语音
S820
当识别出所识别的语音是登记说话者的语音时，输出语音通知
S830
结束

图8