(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 113035200 A
(43)申请公布日 2021.06.25

(21)申请号 202110236299.0

(22)申请日 2021.03.03

(71)申请人 科大讯飞股份有限公司
地址 230088 安徽省合肥市高新开发区望江西路666号

(72)发明人 李锐 刘权 陈志刚

(74)专利代理机构 北京维澳专利代理有限公司 11252
代理人 常小溪 王立民

(51)Int.Cl.
G10L 15/26(2006.01)
G10L 15/183(2013.01)

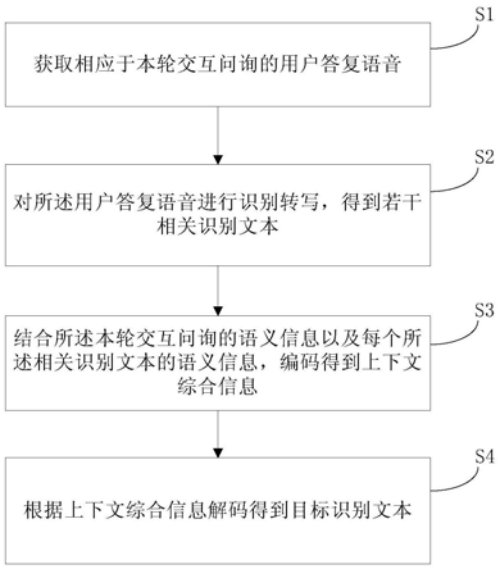
权利要求书2页 说明书11页 附图3页

(54)发明名称

基于人机交互场景的语音识别纠错方法、装置以及设备

(57)摘要

本发明公开了一种基于人机交互场景的语音识别纠错方法、装置以及设备，本发明的构思在于充分利用人机交互场景中多轮问答机制，将机器抛出的本轮问询内容与相应的用户答复内容经由语言识别处理获得的若干相关转写结果相结合，并从二者的语义层面进行深层挖掘，获得涉及本轮问询及若干答复语音的中间识别结果等上下文相关信息的综合表征，进而再对该综合表征进行解码，便可以精准、可靠地得到用户当前答复的正确识别文本。本发明的覆盖度、通用性可以得到显著提升，并且是对语音识别过程中的相关识别文本融入与真实交互场景息息相关的信息，因而实施复杂度也远低于单纯迁移语言模型进行纠错的现有方案，所以能够更易于被业内接受、认可及推广使用。



CN 113035200 A

1.一种基于人机交互场景的语音识别纠错方法,其特征在于,包括:

获取相应于本轮交互问询的用户答复语音;

对所述用户答复语音进行识别转写,得到若干相关识别文本;

结合所述本轮交互问询的语义信息以及每个所述相关识别文本的语义信息,编码得到上下文综合信息;

根据所述上下文综合信息解码得到目标识别文本。

2.根据权利要求1所述的基于人机交互场景的语音识别纠错方法,其特征在于,获得所述本轮交互问询的语义信息的方式包括:

预设若干种交互问询类型;

分别获取所述交互问询类型的第一表征信息以及所述本轮交互问询的第二表征信息;

融合所述第一表征信息以及所述第二表征信息,得到所述本轮交互问询的语义信息。

3.根据权利要求2所述的基于人机交互场景的语音识别纠错方法,其特征在于,所述获取所述交互问询类型的第一表征信息包括:

基于上一轮交互后的语义理解结果,确定所述本轮交互问询的问题内容;

从预设的多种交互问询类型中选出相应于当前问题内容的若干种特定类型;

将所述特定类型向量化后得到所述第一表征信息。

4.根据权利要求1所述的基于人机交互场景的语音识别纠错方法,其特征在于,获得所述相关识别文本的语义信息的方式包括:

获取针对所述本轮交互问询的历史交互信息;

分别获取所述历史交互信息的第三表征信息以及所述相关识别文本的第四表征信息;

融合所述第三表征信息以及所述第四表征信息,得到所述相关识别文本的语义信息。

5.根据权利要求4所述的基于人机交互场景的语音识别纠错方法,其特征在于,所述融合所述第三表征信息以及所述第四表征信息包括:

利用各所述历史交互信息的每个字向量与各所述相关识别文本的句子向量进行多维注意力计算。

6.根据权利要求1～5任一项所述的基于人机交互场景的语音识别纠错方法,其特征在于,所述得到若干相关识别文本包括:

按语音识别过程中解码路径的得分,得到所述相关识别文本。

7.一种基于人机交互场景的语音识别纠错装置,其特征在于,包括:

当前答复语音获取模块,用于获取相应于本轮交互问询的用户答复语音;

转写中间结果获取模块,用于对所述用户答复语音进行识别转写,得到若干相关识别文本;

编码模块,用于结合所述本轮交互问询的语义信息以及每个所述相关识别文本的语义信息,编码得到上下文综合信息;

解码模块,用于根据所述上下文综合信息解码得到目标识别文本。

8.一种电子设备,其特征在于,包括:

一个或多个处理器、存储器以及一个或多个计算机程序,其中所述一个或多个计算机程序被存储在所述存储器中,所述一个或多个计算机程序包括指令,当所述指令被所述电子设备执行时,使得所述电子设备执行权利要求1～6任一项所述的基于人机交互场景的语

音识别纠错方法。

9.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有计算机程序，当所述计算机程序在计算机上运行时，使得计算机执行权利要求1～6任一项所述的基于人机交互场景的语音识别纠错方法。

10.一种计算机程序产品，其特征在于，当所述计算机程序产品被计算机执行时，用于执行权利要求1～6任一项所述的基于人机交互场景的语音识别纠错方法。

基于人机交互场景的语音识别纠错方法、装置以及设备

技术领域

[0001] 本发明涉及人机交互领域，尤其涉及一种基于人机交互场景的语音识别纠错方法、装置以及设备。

背景技术

[0002] 语义理解(natural language understanding,NLU)作为自然语言处理领域中最重要的一环，被广泛应用于人机交互领域，例如但不限于对话系统、智能问答系统等。对于用户输入的一段自然语言文本，一个良好的NLU模块能准确判断出该句所表达的用户意图，然而，在真实的人机交互场景中，机器接收到的文本输入，全部是由用户语音经过语音识别(ASR)后所得。在这个过程中，由于个体发音方式、识别准确率、背景环境等因素影响，很可能导致进入NLU前的语音识别结果已经发生偏差，例如机器问用户“你最近去过动物园么?”，用户回答“去过”，但是被识别成了“吃过”，进而后续送入NLU处理时产生误差传递，且误差很可能存在叠加效应，最终导致用户体验不佳的负面效果。

[0003] 因此，有必要在进入NLU前，对语音识别结果进行纠偏处理，现有的语音识别纠错技术，通常可以归为错别字词典、编辑距离、语言模型等三种主要方式。然而，构建错别字词典的人工成本较高，而且覆盖面较窄，仅适用于错别字有限的部分垂直领域;编辑距离采用类似字符串模糊匹配的方法，通过对照正确样本可以纠正部分常见错别字和语病，同样存在通用性不足的问题;2018年之后，在本技术领域中，预训练语言模型逐步得到重视，并且现阶段在本领域学术界和工业界也取得一定的效果，但是，经真实的人机交互场景的测试和应用后发现，由单纯将语言模型迁移用作语音识别纠错处理，其复杂度相对较高，难以形成具备规模的产品化部署，因而，单纯采用语言模型进行识别纠错处理的技术方案，在业内并未获得普遍认可的落地实践及市场。

发明内容

[0004] 鉴于上述，本发明旨在提供一种基于人机交互场景的语音识别纠错方法、装置以及设备，以及相应地提供了一种计算机可读存储介质和计算机程序产品，主要规避了现有的错别字词典、编辑距离、语言模型等现有纠错方案的弊端，而结合人机交互的场景特点以实现精准度高、通用性广、复杂度低的语音纠错处理。

[0005] 本发明采用的技术方案如下：

[0006] 第一方面，本发明提供了一种基于人机交互场景的语音识别纠错方法，其中包括：

[0007] 获取相应于本轮交互问询的用户答复语音；

[0008] 对所述用户答复语音进行识别转写，得到若干相关识别文本；

[0009] 结合所述本轮交互问询的语义信息以及每个所述相关识别文本的语义信息，编码得到上下文综合信息；

[0010] 根据所述上下文综合信息解码得到目标识别文本。

[0011] 在其中至少一种可能的实现方式中，获得所述本轮交互问询的语义信息的方式包

括：

[0012] 预设若干种交互问询类型；

[0013] 分别获取所述交互问询类型的第一表征信息以及所述本轮交互问询的第二表征信息；

[0014] 融合所述第一表征信息以及所述第二表征信息，得到所述本轮交互问询的语义信息。

[0015] 在其中至少一种可能的实现方式中，所述获取所述交互问询类型的第一表征信息包括：

[0016] 基于上一轮交互后的语义理解结果，确定所述本轮交互问询的问题内容；

[0017] 从预设的多种交互问询类型中选出相应于当前问题内容的若干种特定类型；

[0018] 将所述特定类型向量化后得到所述第一表征信息。

[0019] 在其中至少一种可能的实现方式中，获得所述相关识别文本的语义信息的方式包括：

[0020] 获取针对所述本轮交互问询的历史交互信息；

[0021] 分别获取所述历史交互信息的第三表征信息以及所述相关识别文本的第四表征信息；

[0022] 融合所述第三表征信息以及所述第四表征信息，得到所述相关识别文本的语义信息。

[0023] 在其中至少一种可能的实现方式中，所述融合所述第三表征信息以及所述第四表征信息包括：

[0024] 利用各所述历史交互信息的每个字向量与各所述相关识别文本的句子向量进行多维注意力计算。

[0025] 在其中至少一种可能的实现方式中，所述得到若干相关识别文本包括：

[0026] 按语音识别过程中解码路径的得分，得到所述相关识别文本。

[0027] 第二方面，本发明提供了一种基于人机交互场景的语音识别纠错装置，其中包括：

[0028] 当前答复语音获取模块，用于获取相应于本轮交互问询的用户答复语音；

[0029] 转写中间结果获取模块，用于对所述用户答复语音进行识别转写，得到若干相关识别文本；

[0030] 编码模块，用于结合所述本轮交互问询的语义信息以及每个所述相关识别文本的语义信息，编码得到上下文综合信息；

[0031] 解码模块，用于根据所述上下文综合信息解码得到目标识别文本。

[0032] 在其中至少一种可能的实现方式中，所述编码模块包括第一语义信息获取子模块，所述第一语义信息获取子模块具体包括：

[0033] 问询类型设定单元，用于预设若干种交互问询类型；

[0034] 表征信息第一获取单元，用于分别获取所述交互问询类型的第一表征信息以及所述本轮交互问询的第二表征信息；

[0035] 问询语义获取单元，用于融合所述第一表征信息以及所述第二表征信息，得到所述本轮交互问询的语义信息。

[0036] 在其中至少一种可能的实现方式中，所述表征信息第一获取单元包括问询类型信

息获取子单元，所述问询类型信息获取子单元具体包括：

[0037] 本轮问题确定组件，用于基于上一轮交互后的语义理解结果，确定所述本轮交互问询的问题内容；

[0038] 特定类型选择组件，用于从预设的多种交互问询类型中选出相应于当前问题内容的若干种特定类型；

[0039] 问询类型向量表征组件，用于将所述特定类型向量化后得到所述第一表征信息。

[0040] 在其中至少一种可能的实现方式中，所述编码模块包括第二语义信息获取子模块，所述第二语义信息获取子模块具体包括：

[0041] 历史交互获取单元，用于获取针对所述本轮交互问询的历史交互信息；

[0042] 表征信息第二获取单元，用于分别获取所述历史交互信息的第三表征信息以及所述相关识别文本的第四表征信息；

[0043] 答复语义获取单元，用于融合所述第三表征信息以及所述第四表征信息，得到所述相关识别文本的语义信息。

[0044] 在其中至少一种可能的实现方式中，所述答复语义获取单元包括特征融合组件，所述特征融合组件用于利用各所述历史交互信息的每个字向量与各所述相关识别文本的句子向量进行多维注意力计算。

[0045] 在其中至少一种可能的实现方式中，所述转写中间结果获取模块具体用于：按语音识别过程中解码路径的得分，得到所述相关识别文本。

[0046] 第三方面，本发明提供了一种电子设备，其中包括：

[0047] 一个或多个处理器、存储器以及一个或多个计算机程序，所述存储器可以采用非易失性存储介质，其中所述一个或多个计算机程序被存储在所述存储器中，所述一个或多个计算机程序包括指令，当所述指令被所述电子设备执行时，使得所述电子设备执行如第一方面或者第一方面的任一可能实现方式中的所述方法。

[0048] 第四方面，本发明提供了一种计算机可读存储介质，该计算机可读存储介质中存储有计算机程序，当其在计算机上运行时，使得计算机至少执行如第一方面或者第一方面的任一可能实现方式中的所述方法。

[0049] 第五方面，本发明还提供了一种计算机程序产品，当所述计算机程序产品被计算机执行时，用于至少执行第一方面或者第一方面的任一可能实现方式中的所述方法。

[0050] 在第五方面的至少一种可能的实现方式中，该产品涉及到的相关程序可以全部或者部分存储在与处理器封装在一起的存储器上，也可以部分或者全部存储在不与处理器封装在一起的存储介质上。

[0051] 本发明的构思在于，充分利用人机交互场景中多轮问答机制，将机器抛出的本轮问询内容与相应的用户答复内容经由语言识别处理获得的若干相关转写结果相结合，并从二者的语义层面进行深层挖掘，获得涉及本轮问询及若干答复语音的中间识别结果等上下文相关信息的综合表征，进而再对该综合表征进行解码，便可以精准、可靠地得到用户当前答复的正确识别文本，本发明提供的上述解决思路相对现有技术不依赖既定的正确样本或有限的词典，因而覆盖度、通用性可以得到显著提升，并且由于本发明构思并非单纯迁移语言模型对识别后的最终结果进行纠错，而是对语音识别过程中对应多条解码路径的相关识别文本，融入与真实交互场景息息相关的信息，因而从实施角度而言，复杂度也远低于单纯

以语言模型进行纠错的现有方案，所以能够更易于被业内接受、认可及推广使用。

[0052] 进一步地，在本发明的其他实施例中，对于本轮交互问询的语义挖掘，还考虑将本轮问询语句本身与预设的若干问询类型相互融合，从人机交互中的问询角度丰富语义信息。

[0053] 进一步地，在本发明的其他实施例中，对于用户答复内容的语义挖掘，还考虑将语音识别出的相关中间结果与针对相同问询的以往答复内容进行多维关联，从人机交互中的答复角度丰富语义信息。

[0054] 进一步地，在本发明的其他实施例中，对于获取包含上下文信息的综合表征，还考虑按得分排序机制获得的若干条相关识别结果各自与历史答复融合后的语义信息，进行整体融合，从而为获得正确的识别结果提供更全面的参考信息。

附图说明

[0055] 为使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明作进一步描述，其中：

[0056] 图1为本发明提供的基于人机交互场景的语音识别纠错方法的实施例的流程图；

[0057] 图2为本发明提供的纠错模型的实施例的处理过程示意图；

[0058] 图3为本发明提供的基于人机交互场景的语音识别纠错装置的实施例的示意图；

[0059] 图4为本发明提供的电子设备的实施例的示意图。

具体实施方式

[0060] 下面详细描述本发明的实施例，实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能解释为对本发明的限制。

[0061] 本发明提出了如下至少一种基于人机交互场景的语音识别纠错方法的实施例，如图1所示的，具体可以包括：

[0062] 步骤S1、获取相应于本轮交互问询的用户答复语音。

[0063] 人机交互场景最为普遍的形式即是问答机制，通过机器逐轮地抛出问题，用户对每轮问题进行回应，以此实现人与机器的对话。因此，本实施例提出结合人机交互该特性特点，对交互过程中每轮用户输入的语音进行识别纠错。在本步骤中，可以通过常规的拾音设备获取到用户针对当前轮问题的答复语音，这里还需指出，所称答复语音并不限定是陈述句这类回答、回复等形式，只要是响应本轮交互问询的用户输入语音，皆可以视为所述用户答复语音，例如机器提问“宝莲灯里面你最喜欢谁?”，假设用户未听清或未理解该问题，这时所述用户答复语音便可以是指“什么里面?”、“是指宝莲灯里的动画角色么?”这类疑问句，本实施例对此不作限定。

[0064] 此外，还需说明的是，本实施例是基于先问后答的形式设计的，并且主要针对的是相应于某个问题的答复语音，因而如果在用户答复前，不存在交互问询，可以预先设定一些主动式的交互问询内容，例如机器激活交互后输出“你想表达什么?”、“有什么我可以帮助您的?”、“请说出您的需求”等这类问题作为首轮交互的问询内容。

[0065] 步骤S2、对所述用户答复语音进行识别转写，得到若干相关识别文本。

[0066] 本实施例涉及的语音识别处理(ASR)与现有技术无异,因此不作过多赘述,而这里需要指出两点:

[0067] 其一、站位于对语音识别处理的一般理解,本步骤所称相关识别文本可以包含识别过程采用的语言模型解码后的若干非正确转写文本,也即是可以理解为是语音识别过程中的解码环节得到的若干个中间结果,因此,对于相关识别文本中是否存在正确识别结果则可以不作限定;尤其需指出的是,在本发明构思中,识别结果正确与否,并不是此阶段明确可知的,也即是只有当本实施例完整方案执行结束后得到目标识别结果,才将此目标识别结果视作最终的正确转写文本,换言之,本实施例的完整执行过程可以看作是对语音识别过程之中出现的多条解码路径结果的纠偏,而非对语音识别已然最终输出的唯一识别结果进行纠错。

[0068] 其二、如前所述,“相关识别文本”在本步骤可以看作是识别解码的中间产物,因而为了最终获得正确无误的转写文本,可以按语音识别过程中解码路径的得分,得到多条“相关识别文本”。例如但不限于从语音识别中WFST(weighted finaite-state transducer,加权有限状态转换器)解码后的多条路径中遴选出top-N条得分相对较高的路径结果作为所述相关识别文本(N-best),这里的N其取值则可以根据实际所需自定义,例如以N=4为例,发音为‘beijing’的相关识别文本按解码路径得分排序有:1.北京;2.背景;3.倍镜;4.背影。以N=4再结合本实施例的场景以及前文示例,当机器抛出本轮交互问询的内容(robot-query)“宝莲灯里面你最喜欢谁?”时,对用户输入的答复语音进行识别则可以得到4条相关识别文本(ASR-4-best结果):“我喜欢陈香”、“我喜欢沉香”、“我喜欢陈翔”、“我喜欢沉箱”。当然可以理解的是,其中“我喜欢沉香”是正确的识别文本,但处理至此步骤时并不知晓其是正确的,或者可以假设,ASR-4-best结果中并不包含“我喜欢沉香”,而可能是其他的相关识别结果,例如可以是“我喜欢晨翔”等。

[0069] 步骤S3、结合所述本轮交互问询的语义信息以及每个所述相关识别文本的语义信息,编码得到上下文综合信息。

[0070] 本实施例提出的纠错框架优选采用的是自然语言处理中常见且成熟的encode-decode结构,而本步骤的作用即是编码(encode)过程,也即是将输入变量转换成特定的表征形式,具体来说,是将前述本轮交互问询的语义信息与前述相关识别文本的语义信息相结合,得到充分利用了交互特性的上下文综合信息。

[0071] 需要指出的是,这里的语义信息是指文本本身的深层知识,并不是指“语义理解”,本领域技术人员可以理解的是,人机交互场景下的语义理解通常是指发生于获得确定、准确的语音识别结果之后的处理环节,而本发明并不强调如何进行语义理解处理,而主要目标是为了得到每一轮交互中用户答复语音的正确识别文本,从这个角度而言,可以理解为是衔接于ASR处理的强化操作。基于此目的,本实施例提出分别从交互过程的问题层面以及答复层面挖掘出相关文本的语义信息,作为最终得到准确的用户答复转写文本的参考因素,以此避免语音识别过程可能产生的偏误。

[0072] 在实际操作中,获得本轮交互问询的语义信息、所述相关识别文本的语义信息以及将二者结合编码的方式,可以有多种选择。例如但不限于直接从本轮交互问询的问题语句中提取语义特征,并从前述各N-best结果中分别提取语义特征,而后再将提取到的各语义特征进行拼接或融合计算。对此,本发明在一些较佳实施例中分别对此三者的处理过程

进行了优化选择,下文将分别给出参考说明:

[0073] (一)获得本轮交互问询的语义信息

[0074] 较佳地,可以预先设定出若干种交互问询类型,简称robot-query-type,具体是指在人机交互场景中由机器所提问询内容的各种问题类型,在实际操作中可以预先设定用于交互问询的多种问题模板,并基于各问题模板中的内容确定出与其对应的一种或多种问题种类,例如但不限于如下示例(左侧为问题类型的标签):

[0075] Select——该问题属于选择型,并可以提供选项

[0076] Confirm——该问题属于一个是非类问题

[0077] Judge——该问题属于一个判断类问题

[0078] Why——该问题属于询问原因类问题

[0079] When——该问题属于询问时间的时间类问题

[0080] Where——该问题属于询问地点的地点类问题

[0081] How——该问题属于询问做法的怎样做问题

[0082] Open——该问题属于开放式问询,支持用户随意答复

[0083] 如此,接着便可以获取到所述交互问询类型的第一表征信息,例如但不限于将上述问题类型标签随机初始化一组200维的问题类型向量,作为每轮交互纠错中的固定参数,当然,更佳地还可以基于上一轮交互后的语义理解结果(这里的语义理解即是对准确的目标识别结果进行语义理解的过程),先确定出本轮交互问询的问题内容,再从预设的多种交互问询类型中,筛选出相应于本轮交互的当前问题内容的一种或多种特定类型,并对其向量化后得到所述第一表征信息,而不是采用固化的所有问题类型向量作为第一表征信息。这里需指出的是,对于所述特定类型的向量化处理依然可以发生在初始阶段,也即是只需匹配到当前文本内容的若干类型便可以得到相应的类型向量表征。

[0084] 除了类型标签的表征信息外,还可以获取本轮交互问询的第二表征信息,也即是得到机器当前抛出的问题的抽象表征。具体地,每轮对话中,机器抛出的问题文本(robot-query)可以但不限于先通过语言模型获取到该问题文本各个字符的300维大小的字向量(char-embedding),这里提及的语言模型其作用仅是对问题文本单元进行抽象信息提取,在实际操作中可以借鉴成熟的现有技术,此处不作赘余介绍。进一步地,可以将获取到的字向量序列送入提前初始化训练好的诸如BERT模型等,得到本轮交互问询的第二表征信息,也即是针对当前问题语句的抽象表达。

[0085] 接着,便可以融合所述第一表征信息以及所述第二表征信息,得到所述本轮交互问询的语义信息。第一表征信息与第二表征信息的融合方式可以有多种选择,例如可以依次将不同问题类型对应的各第一表征信息通过加权的方式分别与第二表征信息融合;也可以将不同问题类型对应的各第一表征信息先融合,再与第二表征信息计算相关性;在本发明的一些优选实施例中,是将不同问题类型对应的各第一表征信息与第二表征群信息进行注意力计算(可简称Q-Attention),得到本轮交互问询在encode端的最终表达,在此加强表达中,能够充分利用到问题类型的信息,进而为后续构建上下文综合表征提供精准且丰富的参考信息之一。

[0086] (二)获得相关识别文本的语义信息

[0087] 较佳地,可以获取针对所述本轮交互问询的历史交互信息,由于问题模板可以预

先构建的,因而抛出的问题模板的对象可以是面向预设范围的用户群体,例如某智能玩具的用户、某智能音箱的用户、某APP的用户等等,因而可以理解将某个应用设备的群体视为社区,而这里的历史交互信息可以是指某特定群体针对相同问询问题的社区回复(community-answer),在实际操作中,可以通过后台实时获取到各种各样的回复日志,并且这些已存在的历史回复是已然经过预处理且可以实现实时更新的,因而能够提供可靠的辅助参考信息。在具体选择社区回复时可以通过设定排序及阈值机制,从中挑选出若干条历史回复,例如可以取热度值排名前10的针对当前问题的历史交互信息,这里的热度值则可以结合用户点赞数和/或同类型回复的覆盖度等角度计算得来,进一步地还可以通过计算各历史回复信息之间的相关度并进行聚类和/或排序等,本发明对此不作限定。

[0088] 之后,便可以获取到上述历史交互信息的第三表征信息,具体实现方式可以参考前文提及的对本轮交互问询提取特征的方式,例如对每条社区回复进行char-embedding等,此处不作赘述。同理地,对所述相关识别文本提取第四表征信息,也可以先对各N-BEST结果进行char-embedding,更佳地,可以继续将字向量序列送入诸如BERT模型之中,从而获取到每条所述相关识别文本的语句表达。

[0089] 接着,便可以融合所述第三表征信息以及所述第四表征信息,得到所述相关识别文本的语义信息。第三表征信息与第四表征信息的融合方式同样可以有多种选择,例如可以将不同的历史交互信息对应的各第三表征信息通过加权的方式分别与各第四表征信息融合;也可以以语句为单位,将若干条社区回复语句与若干条所述相关识别文本进行句子级别的相关性计算;在本发明的一些优选实施例中,则是利用各所述历史交互信息的每个字向量,与各所述相关识别文本的句子向量进行多维注意力(Multi-Dim Att)计算,也即是利用每一条社区回复中的每一个字符或字词向量分别与N-BEST中各个识别文本经BERT预编码后的结果进行信息融合,进而可以得到信息量丰富的相关识别文本的语义信息。

[0090] (三)将本轮交互问询的语义信息与相关识别文本的语义信息结合

[0091] 由前述(一)(二)分别从问询及答复两个角度,获得了充足的语义信息,进而可以将(一)(二)得到的语义信息整合为所述上下文综合信息。传统的获得上下文表征(context vector)的方式是在编码层将输入句子压缩成固定长度的向量,理论上该定长向量可完整表达输入句子,再通过后续的解码层即可将context vector内的信息进行转换并输出。在本发明提出的一些实施例中,则不再是输入单句转化为定长向量,而是可以将前述包含有丰富信息的本轮交互问询的语义表征与相关识别文本的语义表征进行拼接,进一步地,由于通常存在多条所述相关识别文本,因而在进行拼接操作之前,可以先对前述(二)中得到的每一条相关识别文本的语义信息再作一次融合(例如但不限于注意力计算),之后再与(一)得到本轮交互问询的语义信息拼接,进而得到最终的上下文综合信息。

[0092] 接续前文,回到图1,步骤S4、根据上下文综合信息解码得到目标识别文本。

[0093] 获得上下文综合信息之后便可以在解码端(decode)按常规解码方式进行逐序解码,形成正确字符序列,进而得到目标识别文本。由前文所述,本发明的目的是对用户输入的答复语音进行识别纠错,因而这里所述的解码过程可以看做是ASR解码的延伸,即ASR解码得到若干个中间识别结果后再次通过前述encode-decode过程得到最终的精准转写文本,也因此在实际操作中可以将前述实施构思以纠错模型体现,该纠错模型拼接在ASR后端并引入其他条件因子(例如前述本轮交互问询、预设问题类型、历史答复信息等,还可以进

一步考虑引入本次多轮交互场景下的在先交互内容),从而由该纠错模型输出精准的、用于后续语义理解的目标识别文本。

[0094] 为了便于理解本发明上述实施例及其优选方案,结合图2所示的encode-decode纠错模型架构的示例再作如下说明,其中涉及的数量及具体内容皆不是对本发明技术方案的限定:

[0095] 在某轮人机交互中,当前机器问询采用的问题模板是“宝莲灯里面你最喜欢谁?”也即是在encode端输入有本轮交互问询(robot-query),并接收到用户语音“woxihuanchenxiang”解码后得到的四个较佳的相关识别文本(ASR-4-best)“我喜欢陈香”、“我喜欢沉香”、“我喜欢陈翔”、“我喜欢沉箱”,并同时输入有按既定策略选择的或预先设定的三种交互问询类型(robot-query-type):select、confirm、open,以及按照既定挑选策略选出的四条可能来自其他用户的历史交互信息也即是回应当前问题模板的社区答复(community-answer)“我没看过这个啊,是电影吗”、“我喜欢里面的二郎神”、“我觉得沉香最可爱”、“我喜欢小玉,我觉得她很好看”。接着将三个robot-query-type分别进行抽象表征,将robot-query先表达为字向量再经由BERT编码为语句抽象表征,同样地,将ASR-4-best的四个中间转写文本分别表达为字向量再经由BERT编码为语句抽象表征,以及将4条community-answer分别以字向量表达。然后由robot-query-type的抽象表征与robot-query的语句级表征进行注意力计算(Q-Attention),由community-answer的字向量表达与ASR-4-best的语句级表征进行多维注意力计算(Multi-Dim-Att),多维注意力计算之后将ASR-4-best的四个信息加强表征再一次进行融合(Attention)。然后将问题层面最终的语义表征与答复层面最终的语义表征进行context处理得到上下文综合表征C_i,最后将C_i送入decode端按序解码出文本序列“我喜欢沉香”。这里还需补充说明,目标识别文本“我喜欢沉香”与encode端输入的相关识别文本之一“我喜欢沉香”没有筛选关系,即,本发明提供的方案不存在从输入的若干相关识别文本之中筛选出其中之一作为目标识别文本的思想。

[0096] 综上所述,本发明的构思在于,充分利用人机交互场景中多轮问答机制,将机器抛出的本轮问询内容与相应的用户答复内容经由语言识别处理获得的若干相关转写结果相结合,并从二者的语义层面进行深层挖掘,获得涉及本轮问询及若干答复语音的中间识别结果等上下文相关信息的综合表征,进而再对该综合表征进行解码,便可以精准、可靠地得到用户当前答复的正确识别文本,本发明提供的上述解决思路相对现有技术不依赖既定的正确样本或有限的词典,因而覆盖度、通用性可以得到显著提升,并且由于本发明构思并非单纯迁移语言模型对识别后的最终结果进行纠错,而是对语音识别过程中对应多条解码路径的相关识别文本,融入与真实交互场景息息相关的信息,因而从实施角度而言,复杂度也远低于单纯以语言模型进行纠错的现有方案,所以能够更易于被业内接受、认可及推广使用。

[0097] 相应于上述各实施例及优选方案,本发明还提供了一种基于人机交互场景的语音识别纠错装置的实施例,如图3所示,具体可以包括如下部件:

[0098] 当前答复语音获取模块1,用于获取相应于本轮交互问询的用户答复语音;

[0099] 转写中间结果获取模块2,用于对所述用户答复语音进行识别转写,得到若干相关识别文本;

[0100] 编码模块3,用于结合所述本轮交互问询的语义信息以及每个所述相关识别文本

的语义信息,编码得到上下文综合信息;

[0101] 解码模块4,用于根据所述上下文综合信息解码得到目标识别文本。

[0102] 在其中至少一种可能的实现方式中,所述编码模块包括第一语义信息获取子模块,所述第一语义信息获取子模块具体包括:

[0103] 问询类型设定单元,用于预设若干种交互问询类型;

[0104] 表征信息第一获取单元,用于分别获取所述交互问询类型的第一表征信息以及所述本轮交互问询的第二表征信息;

[0105] 问询语义获取单元,用于融合所述第一表征信息以及所述第二表征信息,得到所述本轮交互问询的语义信息。

[0106] 在其中至少一种可能的实现方式中,所述表征信息第一获取单元包括问询类型信息获取子单元,所述问询类型信息获取子单元具体包括:

[0107] 本轮问题确定组件,用于基于上一轮交互后的语义理解结果,确定所述本轮交互问询的问题内容;

[0108] 特定类型选择组件,用于从预设的多种交互问询类型中选出相应于当前问题内容的若干种特定类型;

[0109] 问询类型向量表征组件,用于将所述特定类型向量化后得到所述第一表征信息。

[0110] 在其中至少一种可能的实现方式中,所述编码模块包括第二语义信息获取子模块,所述第二语义信息获取子模块具体包括:

[0111] 历史交互获取单元,用于获取针对所述本轮交互问询的历史交互信息;

[0112] 表征信息第二获取单元,用于分别获取所述历史交互信息的第三表征信息以及所述相关识别文本的第四表征信息;

[0113] 答复语义获取单元,用于融合所述第三表征信息以及所述第四表征信息,得到所述相关识别文本的语义信息。

[0114] 在其中至少一种可能的实现方式中,所述答复语义获取单元包括特征融合组件,所述特征融合组件用于利用各所述历史交互信息的每个字向量与各所述相关识别文本的句子向量进行多维注意力计算。

[0115] 在其中至少一种可能的实现方式中,所述转写中间结果获取模块具体用于:按语音识别过程中解码路径的得分,得到所述相关识别文本。

[0116] 应理解以上图3所示的基于人机交互场景的语音识别纠错装置可中各个部件的划分仅仅是一种逻辑功能的划分,实际实现时可以全部或部分集成到一个物理实体上,也可以物理上分开。且这些部件可以全部以软件通过处理元件调用的形式实现;也可以全部以硬件的形式实现;还可以部分部件以软件通过处理元件调用的形式实现,部分部件通过硬件的形式实现。例如,某个上述模块可以为单独设立的处理元件,也可以集成在电子设备的某一个芯片中实现。其它部件的实现与之类似。此外这些部件全部或部分可以集成在一起,也可以独立实现。在实现过程中,上述方法的各步骤或以上各个部件可以通过处理器元件中的硬件的集成逻辑电路或者软件形式的指令完成。

[0117] 例如,以上这些部件可以是被配置成实施以上方法的一个或多个集成电路,例如:一个或多个特定集成电路(Application Specific Integrated Circuit;以下简称:ASIC),或,一个或多个微处理器(Digital Singnal Processor;以下简称:DSP),或,一个或

者多个现场可编程门阵列(Field Programmable Gate Array;以下简称:FPGA)等。再如,这些部件可以集成在一起,以片上系统(System-On-a-Chip;以下简称:SOC)的形式实现。

[0118] 综合上述各实施例及其优选方案,本领域技术人员可以理解的是,在实际操作中,本发明所涉及的技术构思可适用于多种实施方式,本发明以下述载体作为示意性说明:

[0119] (1)一种电子设备。该设备具体可以包括:一个或多个处理器、存储器以及一个或多个计算机程序,其中所述一个或多个计算机程序被存储在所述存储器中,所述一个或多个计算机程序包括指令,当所述指令被所述电子设备执行时,使得所述电子设备执行前述实施例或者等效实施方式的步骤/功能。

[0120] 图4为本发明提供的电子设备的实施例的结构示意图,该设备具体可以为与计算机相关且用于人与计算机交互的电子设备,例如但不限于各类交互终端、智能玩具、智能家居、导航系统、便携电子产品等。

[0121] 具体如图4所示,电子设备900包括处理器910和存储器930。其中,处理器910和存储器930之间可以通过内部连接通路互相通信,传递控制和/或数据信号,该存储器930用于存储计算机程序,该处理器910用于从该存储器930中调用并运行该计算机程序。上述处理器910可以和存储器930可以合成一个处理装置,更常见的是彼此独立的部件,处理器910用于执行存储器930中存储的程序代码来实现上述功能。具体实现时,该存储器930也可以集成在处理器910中,或者,独立于处理器910。

[0122] 除此之外,为了使得电子设备900的功能更加完善,该设备900还可以包括输入单元960、显示单元970、音频电路980、摄像头990和传感器901等中的一个或多个,所述音频电路还可以包括扬声器982、麦克风984等。其中,显示单元970可以包括显示屏。

[0123] 进一步地,上述设备900还可以包括电源950,用于给该设备900中的各种器件或电路提供电能。

[0124] 应理解,该设备900中的各个部件的操作和/或功能,具体可参见前文中关于方法、系统等实施例的描述,为避免重复,此处适当省略详细描述。

[0125] 应理解,图4所示的电子设备900中的处理器910可以是片上系统SOC,该处理器910中可以包括中央处理器(Central Processing Unit;以下简称:CPU),还可以进一步包括其他类型的处理器,例如:图像处理器(Graphics Processing Unit;以下简称:GPU)等,具体在下文中再作介绍。

[0126] 总之,处理器910内部的各部分处理器或处理单元可以共同配合实现之前的方法流程,且各部分处理器或处理单元相应的软件程序可存储在存储器930中。

[0127] (2)一种可读存储介质,在可读存储介质上存储有计算机程序或上述装置,当计算机程序或上述装置被执行时,使得计算机执行前述实施例或等效实施方式的步骤/功能。

[0128] 在本发明所提供的几个实施例中,任一功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明的某些技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以如下所述软件产品的形式体现出来。

[0129] (3)一种计算机程序产品(该产品可以包括上述装置),该计算机程序产品在终端设备上运行时,使终端设备执行前述实施例或等效实施方式的基于人机交互场景的语音识别纠错方法。

[0130] 通过以上的实施方式的描述可知,本领域的技术人员可以清楚地了解到上述实施方法中的全部或部分步骤可借助软件加必需的通用硬件平台的方式来实现。基于这样的理解,上述计算机程序产品可以包括但不限于是指APP;接续前文,上述设备/终端可以是一台计算机设备,并且,该计算机设备的硬件结构还可以具体包括:至少一个处理器,至少一个通信接口,至少一个存储器和至少一个通信总线;处理器、通信接口、存储器均可以通过通信总线完成相互间的通信。其中,处理器可能是一个中央处理器CPU、DSP、微控制器或数字信号处理器,还可包括GPU、嵌入式神经网络处理器(Neural-network Process Units;以下简称:NPU)和图像信号处理器(Image Signal Processing;以下简称:ISP),该处理器还可包括特定集成电路ASIC,或者是被配置成实施本发明实施例的一个或多个集成电路等,此外,处理器可以具有操作一个或多个软件程序的功能,软件程序可以存储在存储器等存储介质中;而前述的存储器/存储介质可以包括:非易失性存储器(non-volatile memory),例如非可移动磁盘、U盘、移动硬盘、光盘等,以及只读存储器(Read-Only Memory;以下简称:ROM)、随机存取存储器(Random Access Memory;以下简称:RAM)等。

[0131] 本发明实施例中,“至少一个”是指一个或者多个,“多个”是指两个或两个以上。“和/或”,描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示单独存在A、同时存在A和B、单独存在B的情况。其中A,B可以是单数或者复数。字符“/”一般表示前后关联对象是一种“或”的关系。“以下至少一项”及其类似表达,是指的这些项中的任意组合,包括单项或复数项的任意组合。例如,a,b和c中的至少一项可以表示:a,b,c,a和b,a和c,b和c或a和b和c,其中a,b,c可以是单个,也可以是多个。

[0132] 本领域技术人员可以意识到,本说明书中公开的实施例中描述的各模块、单元及方法步骤,能够以电子硬件、计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。本领域技术人员可以对每个特定的应用来使用不同方式来实现所描述的功能,但是这种实现不应认为超出本发明的范围。

[0133] 以及,其中作为分离部件说明的模块、单元等可以是或者也可以不是物理上分开的,即可以位于一个地方,或者也可以分布到多个地方,例如系统网络的节点上。具体可根据实际的需要选择其中的部分或者全部模块、单元来实现上述实施例方案的目的。本领域技术人员在不付出创造性劳动的情况下,即可以理解并实施。

[0134] 以上依据图式所示的实施例详细说明了本发明的构造、特征及作用效果,但以上仅为本发明的较佳实施例,需要言明的是,上述实施例及其优选方式所涉及的技术特征,本领域技术人员可以在不脱离、不改变本发明的设计思路以及技术效果的前提下,合理地组合搭配成多种等效方案;因此,本发明不以图面所示限定实施范围,凡是依照本发明的构想所作的改变,或修改为等同变化的等效实施例,仍未超出说明书与图示所涵盖的精神时,均应在本发明的保护范围内。

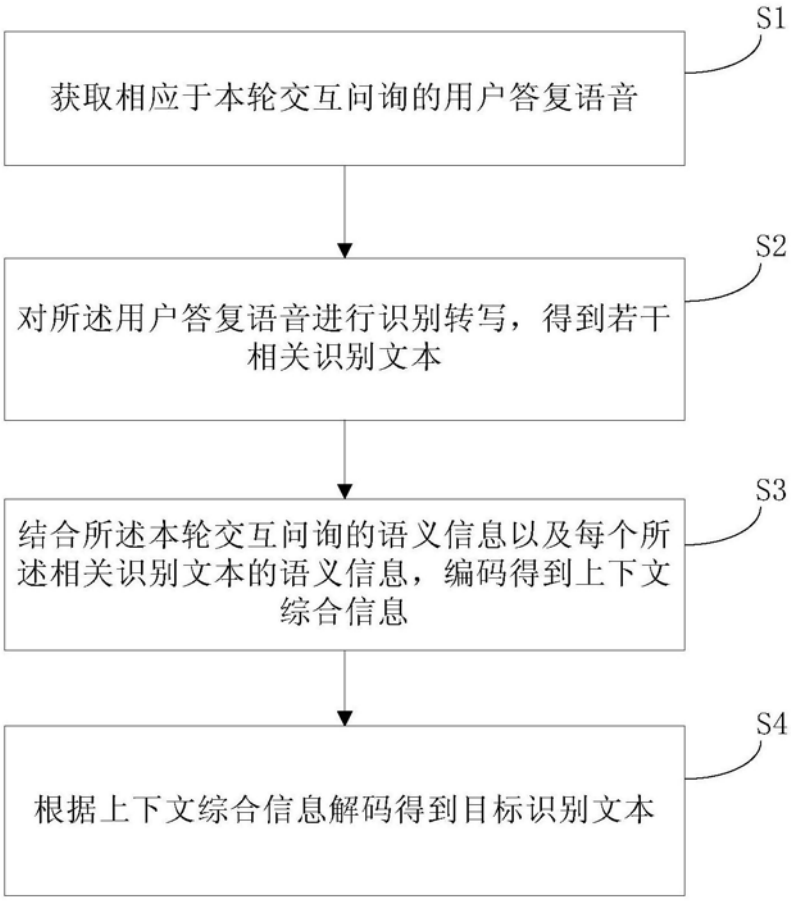

S1
获取相应于本轮交互问询的用户答复语音
S2
对所述用户答复语音进行识别转写，得到若干相关识别文本
S3
结合所述本轮交互问询的语义信息以及每个所述相关识别文本的语义信息，编码得到上下文综合信息
S4
根据上下文综合信息解码得到目标识别文本

图1

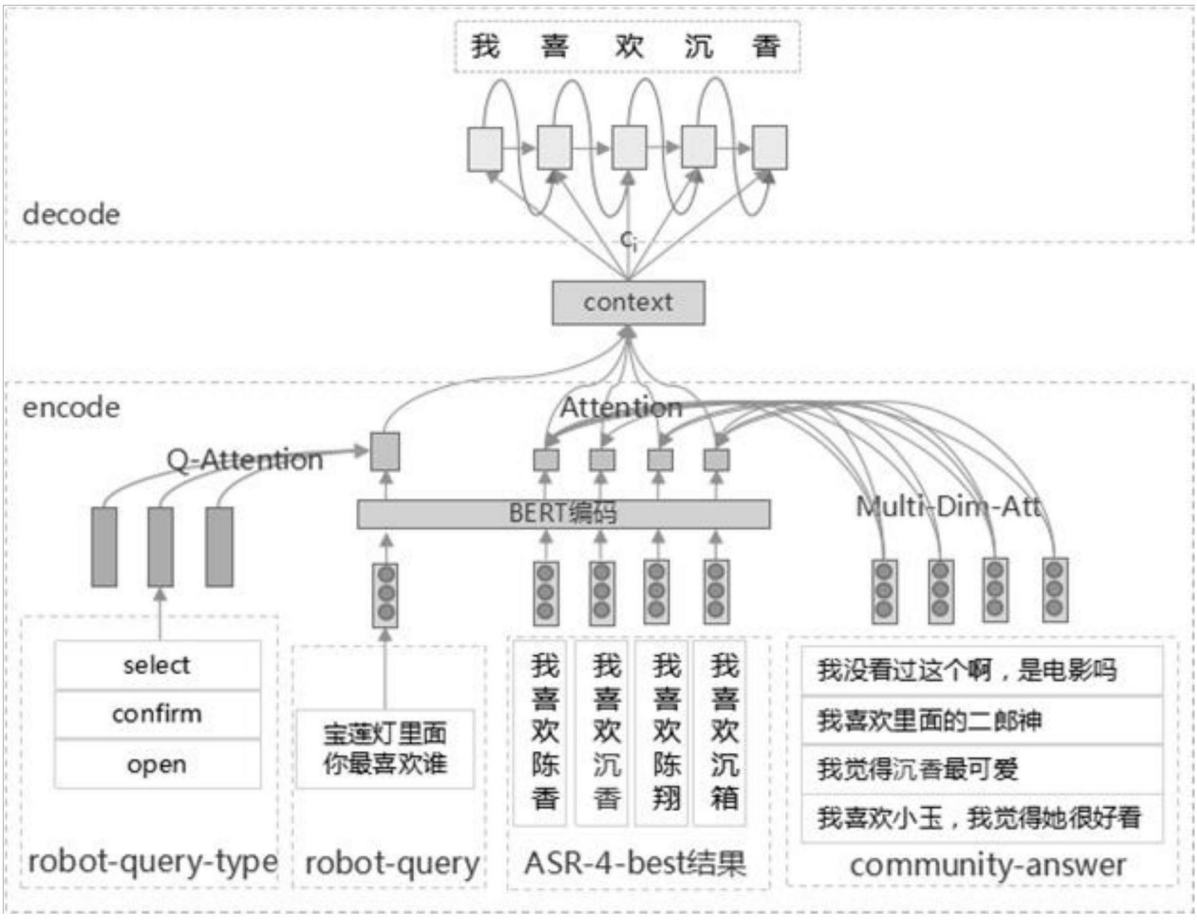

我 喜 欢 沉 香
decode
c_i
context
encode
Attention
Q-Attention
BERT编码
Multi-Dim-Att
select
confirm
open
宝莲灯里面
你最喜欢谁
我喜欢陈香
我喜欢沉香
我喜欢陈翔
我喜欢沉箱
我没看过这个啊，是电影吗
我喜欢里面的二郎神
我觉得沉香最可爱
我喜欢小玉，我觉得她很好看
robot-query-type
robot-query
ASR-4-best结果
community-answer

图2

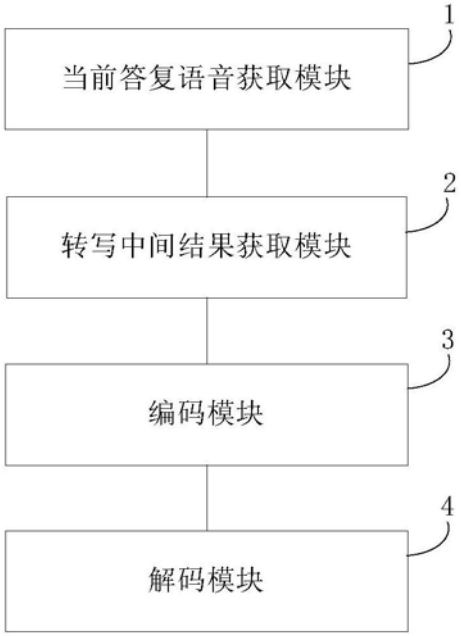

1
当前答复语音获取模块
2
转写中间结果获取模块
3
编码模块
4
解码模块

图3

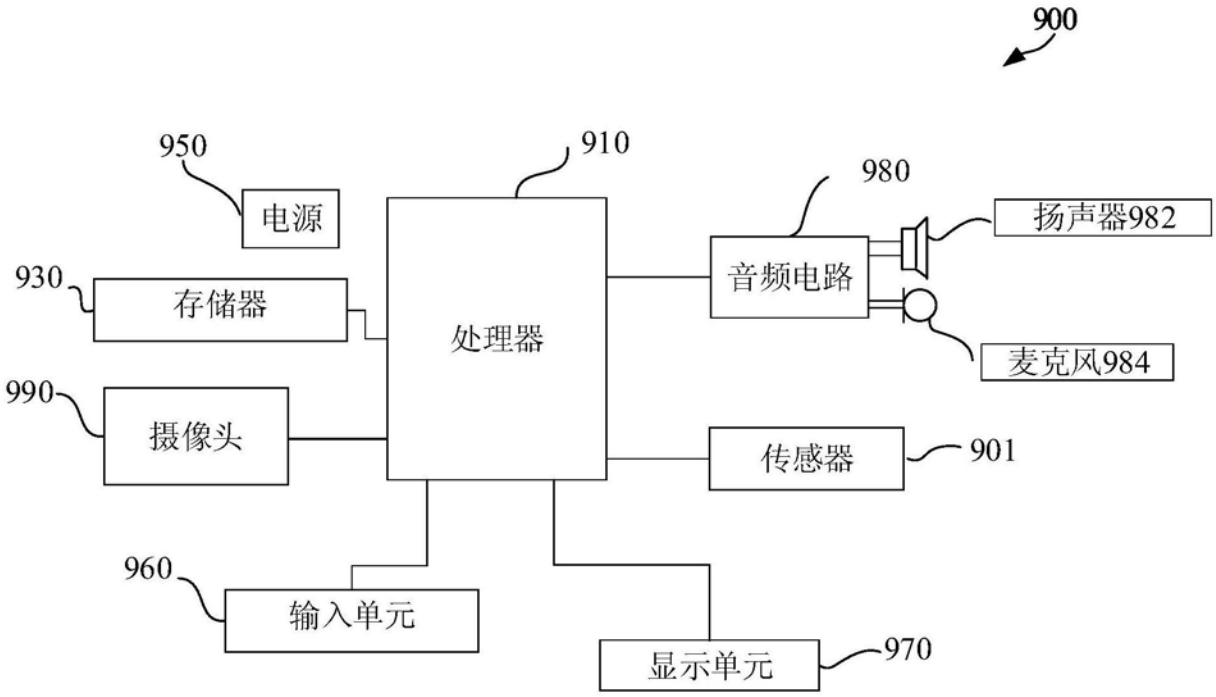

900
950
电源
910
980
扬声器982
930
存储器
处理器
音频电路
麦克风984
990
摄像头
传感器
901
960
输入单元
显示单元
970

图4