(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111583909 B
(45)授权公告日 2024.04.12

(21)申请号 202010418728.1

(22)申请日 2020.05.18

(65)同一申请的已公布的文献号
申请公布号 CN 111583909 A

(43)申请公布日 2020.08.25

(73)专利权人 科大讯飞股份有限公司
地址 230088 安徽省合肥市高新区望江西路666号

(72)发明人 熊世富 刘聪 魏思 刘庆峰 高建清 潘嘉

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
专利代理师 杨华

(51)Int.Cl.
G10L 15/02(2006.01)
G10L 15/06(2013.01)
G10L 15/22(2006.01)
G10L 15/26(2006.01)
G10L 25/03(2013.01)

(56)对比文件
CN 110956959 A,2020.04.03
CN 102592595 A,2012.07.18
CN 110879839 A,2020.03.13
CN 109215662 A,2019.01.15
CN 108228565 A,2018.06.29
CN 108984529 A,2018.12.11
CN 108899030 A,2018.11.27
CN 109559752 A,2019.04.02
CN 110047467 A,2019.07.23
CN 102968987 A,2013.03.13
CN 111105799 A,2020.05.05
CN 108831456 A,2018.11.16
CN 103310790 A,2013.09.18
US 4520499 A,1985.05.28
CN 109523991 A,2019.03.26
CN 110517692 A,2019.11.29
CN 105719649 A,2016.06.29
CN 111009237 A,2020.04.14

审查员 李哲

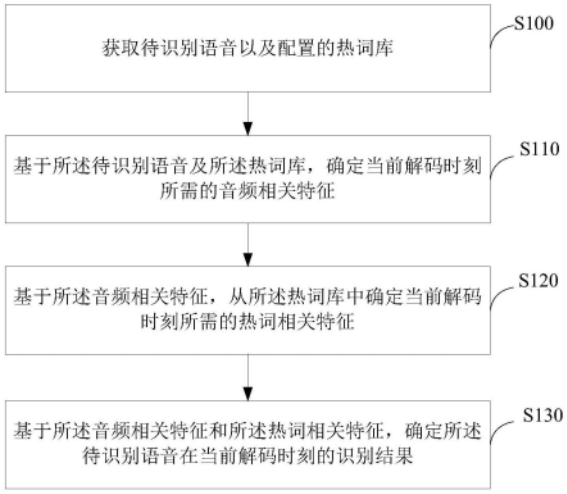
权利要求书3页 说明书18页 附图4页

(54)发明名称

一种语音识别方法、装置、设备及存储介质

(57)摘要

本申请提供了一种语音识别方法、装置、设备及存储介质，本申请配置有热词库，在对待识别语音进行识别过程，基于待识别语音及热词库，确定当前解码时刻所需的音频相关特征，由于音频相关特征确定过程利用了热词信息，如果当前解码时刻的语音片段中包含某个热词，则确定的音频相关特征中能够包含该热词对应的完整音频信息，进一步基于该音频相关特征从热词库中确定当前解码时刻所需的热词相关特征，热词相关特征能够准确表示当前解码时刻的语音片段是否包含热词以及具体包含哪个热词，最终基于音频相关特征和热词相关特征，确定待识别语音在当前解码时刻的识别结果，该识别结果对热词的识别更加准确。



CN 111583909 B

1.一种语音识别方法,其特征在于,包括:

获取待识别语音以及配置的热词库;

基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征,所述音频相关特征包含有潜在热词的完整音频信息;

基于所述音频相关特征,从所述热词库中确定当前解码时刻所需的热词相关特征;

基于所述音频相关特征和所述热词相关特征,确定所述待识别语音在当前解码时刻的识别结果;

所述基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征,包括:

获取当前解码时刻之前的已解码结果信息;

基于所述已解码结果信息及所热词库,从所述待识别语音中确定当前解码时刻所需的音频相关特征。

2.根据权利要求1所述的方法,其特征在于,所述基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征;基于所述音频相关特征,从所述热词库中确定当前解码时刻所需的热词相关特征;基于所述音频相关特征和所述热词相关特征,确定所述待识别语音在当前解码时刻的识别结果的过程,包括:

利用预先训练的语音识别模型处理所述待识别语音及所述热词库,得到语音识别模型输出的待识别语音的识别结果,其中:

所述语音识别模型具备接收并处理待识别语音及热词库,以输出待识别语音的识别结果的能力。

3.根据权利要求2所述的方法,其特征在于,所述语音识别模型包括音频编码器模块、热词编码器模块、联合注意力模块、解码器模块及分类器模块;

所述音频编码器模块对所述待识别语音进行编码,得到音频编码结果;

所述热词编码器模块对所述热词库中各热词进行编码,得到热词编码结果;

所述联合注意力模块接收并处理所述音频编码结果和所述热词编码结果,得到当前解码时刻所需的拼接特征,所述拼接特征包括音频相关特征和热词相关特征;

所述解码器模块接收并处理所述当前解码时刻所需的拼接特征,得到解码器模块当前解码时刻的输出特征;

所述分类器模块利用解码器模块当前解码时刻的输出特征,确定待识别语音在当前解码时刻的识别结果。

4.根据权利要求3所述的方法,其特征在于,所述联合注意力模块包括:

第一注意力模型和第二注意力模型;

所述第一注意力模型基于解码器模块在当前解码时刻输出的表示已解码结果信息的状态向量,以及所述热词编码结果,从所述音频编码结果中确定当前解码时刻所需的音频相关特征;

所述第二注意力模型基于所述音频相关特征,从所述热词编码结果中确定当前解码时刻所需的热词相关特征;

由所述音频相关特征和所述热词相关特征组合成当前解码时刻所需的拼接特征。

5.根据权利要求4所述的方法,其特征在于,所述第一注意力模型基于解码器模块在当

前解码时刻输出的表示已解码结果信息的状态向量,以及所述热词编码结果,从所述音频编码结果中确定当前解码时刻所需的音频相关特征,包括:

将所述状态向量、所述热词编码结果作为第一注意力模型的输入,由所述第一注意力模型从所述音频编码结果中确定当前解码时刻所需的音频相关特征。

6.根据权利要求4所述的方法,其特征在于,所述第二注意力模型基于所述音频相关特征,从所述热词编码结果中确定当前解码时刻所需的热词相关特征,包括:

将所述音频相关特征作为第二注意力模型的输入,由所述第二注意力模型从所述热词编码结果中确定当前解码时刻所需的热词相关特征。

7.根据权利要求3所述的方法,其特征在于,所述分类器模块的分类节点包括固定的常用字符节点和可动态扩展的热词节点;

分类器模块利用解码器模块当前解码时刻的输出特征,确定待识别语音在当前解码时刻的识别结果,包括:

分类器模块利用解码器模块在当前解码时刻的输出特征,确定各所述常用字符节点的概率得分和各所述热词节点的概率得分;

根据各所述常用字符节点的概率得分和各所述热词节点的概率得分,确定待识别语音在当前解码时刻的识别结果。

8.根据权利要求7所述的方法,其特征在于,所述可动态扩展的热词节点,与所述热词库中的热词一一对应。

9.根据权利要求1-8任一项所述的方法,其特征在于,所述获取待识别语音以及配置的热词库,包括:

获取待识别语音,并确定所述待识别语音的会话场景;

获取与所述会话场景相关的热词库。

10.根据权利要求1-8任一项所述的方法,其特征在于,所述获取待识别语音以及配置的热词库,包括:

获取人机交互场景下,用户所产出的语音,作为待识别语音;

获取预先配置的在人机交互场景下,由用户语音操控指令中的操作关键词组成的热词库。

11.根据权利要求10所述的方法,其特征在于,还包括:

基于所述待识别语音的识别结果,确定与所述识别结果相匹配的交互响应,并输出该交互响应。

12.一种语音识别装置,其特征在于,包括:

数据获取单元,用于获取待识别语音以及配置的热词库;

音频相关特征获取单元,用于基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征,所述音频相关特征包含有潜在热词的完整音频信息;所述基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征,包括:获取当前解码时刻之前的已解码结果信息;基于所述已解码结果信息及所热词库,从所述待识别语音中确定当前解码时刻所需的音频相关特征;

热词相关特征获取单元,用于基于所述音频相关特征,从所述热词库中确定当前解码时刻所需的热词相关特征;

识别结果获取单元,用于基于所述音频相关特征和所述热词相关特征,确定所述待识别语音在当前解码时刻的识别结果。

13.一种语音识别设备,其特征在于,包括:存储器和处理器;

所述存储器,用于存储程序;

所述处理器,用于执行所述程序,实现如权利要求1~11中任一项所述的语音识别方法的各个步骤。

14.一种可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时,实现如权利要求1~11中任一项所述的语音识别方法的各个步骤。

一种语音识别方法、装置、设备及存储介质

技术领域

[0001] 本申请涉及语音识别技术领域,尤其涉及一种语音识别方法、装置、设备及存储介质。

背景技术

[0002] 语音识别即对输入的语音数据进行识别,已得到语音对应的识别文本内容。随着深度学习序列建模的发展,端到端建模方法成为语音识别领域的研究热点。

[0003] 如图1示例的现有基于注意力机制的端到端语音识别框架,能够对输入语音进行编码,并基于注意力机制对编码音频进行处理,经过解码、分类得到输入语音对应的识别文本。这种语音识别方法对训练数据的需求量很大,导致训练后的模型出现过度自信(over-confidence)的问题,表现在模型上就是计算出的后验概率得分很尖锐,也即对于高频词的识别效果很好,得分很高;但是对于低频词的识别效果很差,得分很低。对于一些热词如,专业名词、专业术语、日常社会活动中产生的实时热点词汇,相对于模型来说就属于低频词,对于此类热词模型的识别效果很差。

发明内容

[0004] 有鉴于此,本申请提供了一种语音识别方法、装置、设备及存储介质,用以解决现有语音识别方案对热词识别效果差的问题,其技术方案如下:

[0005] 一种语音识别方法,包括:

[0006] 获取待识别语音以及配置的热词库;

[0007] 基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征;

[0008] 基于所述音频相关特征,从所述热词库中确定当前解码时刻所需的热词相关特征;

[0009] 基于所述音频相关特征和所述热词相关特征,确定所述待识别语音在当前解码时刻的识别结果。

[0010] 优选地,所述基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征,包括:

[0011] 获取当前解码时刻之前的已解码结果信息;

[0012] 基于所述已解码结果信息及所热词库,从所述待识别语音中确定当前解码时刻所需的音频相关特征。

[0013] 优选地,所述基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征;基于所述音频相关特征,从所述热词库中确定当前解码时刻所需的热词相关特征;基于所述音频相关特征和所述热词相关特征,确定所述待识别语音在当前解码时刻的识别结果的过程,包括:

[0014] 利用预先训练的语音识别模型处理所述待识别语音及所述热词库,得到语音识别模型输出的待识别语音的识别结果,其中:

[0015] 所述语音识别模型具备接收并处理待识别语音及热词库,以输出待识别语音的识别结果的能力。

[0016] 优选地,所述语音识别模型包括音频编码器模块、热词编码器模块、联合注意力模块、解码器模块及分类器模块;

[0017] 所述音频编码器模块对所述待识别语音进行编码,得到音频编码结果;所述热词编码器模块对所述热词库中各热词进行编码,得到热词编码结果;

[0018] 所述联合注意力模块接收并处理所述音频编码结果和所述热词编码结果,得到当前解码时刻所需的拼接特征,所述拼接特征包括音频相关特征和热词相关特征;

[0019] 所述解码器模块接收并处理所述当前解码时刻所需的拼接特征,得到解码器模块当前解码时刻的输出特征;

[0020] 所述分类器模块利用解码器模块当前解码时刻的输出特征,确定待识别语音在当前解码时刻的识别结果。

[0021] 优选地,所述联合注意力模块包括:

[0022] 第一注意力模型和第二注意力模型;

[0023] 所述第一注意力模型基于解码器模块在当前解码时刻输出的表示已解码结果信息的状态向量,以及所述热词编码结果,从所述音频编码结果中确定当前解码时刻所需的音频相关特征;

[0024] 所述第二注意力模型基于所述音频相关特征,从所述热词编码结果中确定当前解码时刻所需的热词相关特征;

[0025] 由所述音频相关特征和所述热词相关特征组合成当前解码时刻所需的拼接特征。

[0026] 优选地,所述第一注意力模型基于解码器模块在当前解码时刻输出的表示已解码结果信息的状态向量,以及所述热词编码结果,从所述音频编码结果中确定当前解码时刻所需的音频相关特征,包括:

[0027] 将所述状态向量、所述热词编码结果作为第一注意力模型的输入,由所述第一注意力模型从所述音频编码结果中确定当前解码时刻所需的音频相关特征。

[0028] 优选地,所述第二注意力模型基于所述音频相关特征,从所述热词编码结果中确定当前解码时刻所需的热词相关特征,包括:

[0029] 将所述音频相关特征作为第二注意力模型的输入,由所述第二注意力模型从所述热词编码结果中确定当前解码时刻所需的热词相关特征。

[0030] 优选地,所述分类器模块的分类节点包括固定的常用字符节点和可动态扩展的热词节点;

[0031] 分类器模块利用解码器模块当前解码时刻的输出特征,确定待识别语音在当前解码时刻的识别结果,包括:

[0032] 分类器模块利用解码器模块在当前解码时刻的输出特征,确定各所述常用字符节点的概率得分和各所述热词节点的概率得分;

[0033] 根据各所述常用字符节点的概率得分和各所述热词节点的概率得分,确定待识别语音在当前解码时刻的识别结果。

[0034] 优选地,所述可动态扩展的热词节点,与所述热词库中的热词一一对应。

[0035] 优选地,所述获取待识别语音以及配置的热词库,包括:

[0036] 获取待识别语音,并确定所述待识别语音的会话场景;

[0037] 获取与所述会话场景相关的热词库。

[0038] 优选地,所述获取待识别语音以及配置的热词库,包括:

[0039] 获取人机交互场景下,用户所产出的语音,作为待识别语音;

[0040] 获取预先配置的在人机交互场景下,由用户语音操控指令中的操作关键词组成的热词库。

[0041] 优选地,还包括:

[0042] 基于所述待识别语音的识别结果,确定与所述识别结果相匹配的交互响应,并输出该交互响应。

[0043] 一种语音识别装置,包括:

[0044] 数据获取单元,用于获取待识别语音以及配置的热词库;

[0045] 音频相关特征获取单元,用于基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征;

[0046] 热词相关特征获取单元,用于基于所述音频相关特征,从所述热词库中确定当前解码时刻所需的热词相关特征;

[0047] 识别结果获取单元,用于基于所述音频相关特征和所述热词相关特征,确定所述待识别语音在当前解码时刻的识别结果。

[0048] 一种语音识别设备,包括:存储器和处理器;

[0049] 所述存储器,用于存储程序;

[0050] 所述处理器,用于执行所述程序,实现如上所述的语音识别方法的各个步骤。

[0051] 一种可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时,实现如上的语音识别方法的各个步骤。

[0052] 经由上述方案可知,本申请提供的语音识别方法,配置有热词库,也即待识别语音中可能存在的热词,进而在对待识别语音进行识别过程,基于待识别语音及热词库,确定当前解码时刻所需的音频相关特征,由于音频相关特征确定过程利用了热词信息,如果当前解码时刻的语音片段中包含某个热词,则确定的音频相关特征中能够包含该热词对应的完整音频信息而非局部信息,进一步基于该音频相关特征从热词库中确定当前解码时刻所需的热词相关特征,由于音频相关特征中能够包含热词对应的完整的音频信息,因此确定的热词相关特征能够准确表示当前解码时刻的语音片段是否包含热词以及具体包含哪个热词,最终基于音频相关特征和热词相关特征,确定待识别语音在当前解码时刻的识别结果,该识别结果对热词的识别更加准确。

附图说明

[0053] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据提供的附图获得其他的附图。

[0054] 图1示例的现有基于注意力机制的端到端语音识别框架;

[0055] 图2示例了一种改进的基于注意力机制的端到端语音识别框架;

[0056] 图3为本申请实施例提供的一种语音识别方法流程图；

[0057] 图4为本申请实施例提供的另一种改进的基于注意力机制的端到端语音识别框架；

[0058] 图5为本申请实施例示例的一种一层双向长短时记忆层的热词编码器对热词的编码示意图；

[0059] 图6为本申请实施例提供的一种语音识别装置结构示意图；

[0060] 图7为本申请实施例提供的电子设备的结构示意图。

具体实施方式

[0061] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0062] 为了解决现有语音识别方法对热词识别效果差的问题，本案发明人进行了研究，首先想到的就是对热词得分进行激励，也即对于语音识别模型输出的各候选识别字符中，属于热词的候选识别字符的得分进行激励，以达到提高热词识别率的目的。

[0063] 但是，进一步研究发现，对于端到端的语音识别模型，其对于热词这种低频词的得分过低，导致热词在解码过中很容易被裁减掉，恶劣的情况下甚至没有被激励的机会，从而无法真正实现对热词提升识别度的目的。

[0064] 为此，发明人进一步提出了一种在模型层面提高热词得分概率的方案，通过修改语音识别模型的结构来实现。修改后的语音识别模型的示意框架如图2所示。

[0065] 相比于现有语音识别模型，增加了热词编码器模块Bias encoder，其能够对热词进行编码。进一步，利用解码器Decoder的状态信息通过注意力机制分别对音频编码特征和热词编码特征进行操作，得到解码所需的音频相关特征和热词相关特征。进而基于音频相关特征和热词相关特征进行解码、分类得到输入语音对应的识别文本。

[0066] 这种方案由于从模型结构层面即考虑了热词，因而相比于直接对模型输出的热词得分进行激励的方式，效果更好。

[0067] 但是，经过发明人深入研究发现，不同热词的长度可能不一样，要准确地确定音频中是否包含热词、包含哪个热词，不同热词所需的信息是不一样的。而解码器的状态信息只包含已解码结果的历史文本和历史音频信息，单纯用一个只包含历史信息的状态信息作为注意力机制的查询项，对音频编码特征执行注意力操作所得到的音频相关特征并不一定是完整的，同时对热词编码特征执行注意力操作所得到的热词相关特征也不一定是准确的，进而导致最终的热词识别准确度也不是特别高。

[0068] 为此，发明人进一步提出了另一种改进方案，来解决上述问题。接下来，对本案提出的语音识别方法进行详细介绍。

[0069] 可以理解的是，本案的语音识别方法可以应用于任何需要进行语音识别的场景下。语音识别方法可以基于电子设备实现，如通过手机、翻译机、电脑、服务器等具备数据处理能力的设备。

[0070] 接下来，结合附图3示例的流程，对本案的语音识别方法进行介绍，详细包括如下

步骤：

[0071] 步骤S100、获取待识别语音以及配置的热词库。

[0072] 具体的，对于本次语音识别任务所需要识别的语音作为待识别语音。在语音识别之前，可以获取配置的热词库，热词库中存储有多个热词。可以理解的是，热词库可以是由与语音识别任务相关的热词组成，如将待识别语音中所有可能存在的热词，如专业术语等组成热词库。

[0073] 此外，还可以直接调用已有的一些热词库作为本实施例中配置的热词库。步骤S110、基于所述待识别语音及所述热词库，确定当前解码时刻所需的音频相关特征。

[0074] 具体的，为了提高语音识别对热词的识别度，当待解码字符为一个潜在的热词时，需要获取该潜在的热词的完整音频信息。因此，本步骤中，为了确保得到的当前解码时刻所需的音频相关特征中包含有潜在热词的完整音频信息，将热词库考虑进来，也即让热词库参与到确定音频相关特征的计算过程中，起到检测当前解码时刻的待解码字符是否为热词的功能。

[0075] 最终得到的音频相关特征能够包含当前待解码字符的完整音频信息。

[0076] 步骤S120、基于所述音频相关特征，从所述热词库中确定当前解码时刻所需的热词相关特征。

[0077] 上一步骤中已经确定了当前解码时刻所需的音频相关特征，因此可以基于该音频相关特征，从热词库中确定出当前解码时刻所需的热词相关特征，该热词相关特征表示了当前解码时刻所可能出现的热词内容。

[0078] 可以理解的是，由于音频相关特征能够包含当前待解码字符的完整音频信息，基于此从热词库中确定的当前解码时刻所需的热词相关特征，其更加能够适应热词长度不一的情况。

[0079] 步骤S130、基于所述音频相关特征和所述热词相关特征，确定所述待识别语音在当前解码时刻的识别结果。

[0080] 在得到当前解码时刻所需的音频相关特征和热词相关特征之后，可以基于二者对当前时刻的待解码字符进行解码识别，从而确定待识别语音在当前解码时刻的识别结果。

[0081] 本申请实施例提供的语音识别方法，配置有热词库，也即待识别语音中可能存在的热词，进而在对待识别语音进行识别过程，基于待识别语音及热词库，确定当前解码时刻所需的音频相关特征，由于音频相关特征确定过程利用了热词信息，如果当前解码时刻的语音片段中包含某个热词，则确定的音频相关特征中能够包含该热词对应的完整音频信息而非局部信息，进一步基于该音频相关特征从热词库中确定当前解码时刻所需的热词相关特征，由于音频相关特征中能够包含热词对应的完整的音频信息，因此确定的热词相关特征能够准确表示当前解码时刻的语音片段是否包含热词以及具体包含哪个热词，最终基于音频相关特征和热词相关特征，确定待识别语音在当前解码时刻的识别结果，该识别结果对热词的识别更加准确。

[0082] 本申请实施例介绍了上述步骤S100，获取待识别语音以及配置的热词库的实现方式。

[0083] 可选的，在获取到待识别语音后，可以确定待识别语音的会话场景。进一步，可以获取与该会话场景相关的热词库，作为本案中配置的热词库。

[0084] 可以理解的是,不同的会话场景下产出的待识别语音中,包含的热词也可能不同,因此本申请可以预先确定与各会话场景相对应的热词库,进而在确定了待识别语音的会话场景后,获取对应的热词库。

[0085] 另一种可选的方式,当本申请方案应用于人机交互场景下对语音进行识别时:

[0086] 可以理解的是,在人机交互场景下,用户与机器进行交互,会涉及到用户语音操控指令,也即用户向机器下发语音操控指令,以实现预定的目的。例如,用户语音控制智能电视实现切换台、增减音量等相关操作,再比如,用户语音控制智能机器人实现播放歌曲、查询天气、执行预定动作等。

[0087] 在此基础上,若要使得机器能够正确响应用户,则需要机器能够准确识别语音语音操控指令。为此,本申请可以将用户语音操控指令中的操作关键词组成热词库。

[0088] 基于此,本申请实施例可以获取人机交互场景下,用户所产出的语音,作为待识别语音,并获取预先配置的在人机交互场景下,由用户语音操控指令中的操作关键词组成的热词库。

[0089] 在上述基础上,按照本申请方案确定了待识别语音的识别结果后,可以基于该识别结果,确定与该识别结果相匹配的交互响应,并输出该交互响应。

[0090] 按照本实施例介绍的方案,能够准确识别人机交互过程中用户的操控指令,从而使得机器能够基于正确的识别结果做出匹配的交互响应。

[0091] 本申请的另一个实施例中,对上述步骤S110,基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征。

[0092] 具体的,待识别语音的各帧语音之间存在上下文关系,为了确定当前解码时刻所需的音频相关特征,本实施例中可以先获取当前解码时刻之前的已解码结果信息。已解码结果信息可以包含已解码字符的文本信息、音频信息。

[0093] 此外,为了提高语音识别对热词的识别度,当待解码字符为一个潜在的热词时,需要获取该潜在的热词的完整音频信息。因此,本步骤中,为了确保得到的当前解码时刻所需的音频相关特征中包含有潜在热词的完整音频信息,将热词库考虑进来,也即让热词库参与到确定音频相关特征的计算过程中,起到检测当前解码时刻的待解码字符是否为热词的功能。在此基础上,后续可以基于该音频相关特征,从热词库中确定当前解码时刻所需的热词相关特征,其更加能够适应热词长度不一的情况。

[0094] 由上可知,步骤S110确定音频相关特征的过程可以包括:

[0095] S1、获取当前解码时刻之前的已解码结果信息。

[0096] S2、基于已解码结果信息及热词库,从待识别语音中确定当前解码时刻所需的音频相关特征。

[0097] 具体的,本实施例中可以先确定待识别语音的音频特征,其音频特征可以选用滤波器组FilterBank特征、梅尔频率倒谱系数MFCC特征、感知线性预测PLP特征等。进一步的,基于已解码结果信息及热词库,从待识别语音的音频特征中确定当前解码时刻所需的音频相关特征。

[0098] 其中,音频相关特征携带有当前解码时刻的待解码字符的完整音频信息。在此基础上,才能够为后续准确识别热词提供足够的音频相关特征。

[0099] 在本申请的另一个实施例中,介绍了上述实施例中步骤S110-S130,确定当前解码

时刻所需的音频相关特征、热词相关特征，并基于此确定待识别语音在当前解码时刻的识别结果的一种可选实现方式。

[0100] 具体的，可以通过语音识别模型来实现。

[0101] 当然，本实施例中提供的语音识别模型和传统的语音识别模型不同，本实施例中的语音识别模型被配置为，具备接收并处理待识别语音及热词库，以输出待识别语音的识别结果的能力。

[0102] 具体的，语音识别模型可以具备基于当前解码时刻之前的已解码结果信息及热词库，从待识别语音中确定当前解码时刻所需的音频相关特征，并基于音频相关特征从热词库中确定当前解码时刻所需的热词相关特征，基于所述音频相关特征和所述热词相关特征，确定待识别语音在当前解码时刻的识别结果的能力。

[0103] 在此基础上，本实施例中可以将利用语音识别模型处理前述步骤S100中获取的待识别语音及热词库，则语音识别模型即可输出待识别语音的识别结果。

[0104] 具体的，可以将待识别语音的音频特征及热词库输入语音识别模型，得到模型输出的待识别语音的识别结果。

[0105] 接下来，结合图4，对语音识别模型的框架进行介绍。

[0106] 语音识别模型可以包括音频编码器模块、热词编码器模块、联合注意力模块、解码器模块及分类器模块。由各个模块配合实现对接收的热词库中各热词和待识别语音的音频特征进行处理，并最终输出识别结果的过程。接下来分别对各个模块进行介绍。

[0107] 为了便于说明，定义待识别语音的音频特征为$X=[x_1,x_2,...,x_k]$，其中，x_k表示第k帧音频特征向量，k为待识别语音的总语音帧数目。

[0108] 1、音频编码器模块：

[0109] 音频编码器模块对待识别语音进行编码，得到音频编码结果。

[0110] 具体的，音频编码器模块可以对待识别语音的音频特征X进行编码，得到编码后由每帧语音的音频特征向量组成的音频特征向量序列。

[0111] 其中，编码后得到的音频特征向量序列表示为：

[0112] $$H^x=\left[h_1^x,h_2^x,...,h_k^x\right]。$$

[0113] 其中，h_k^x表示第k帧音频特征向量，h_k^x对应x_k经过音频编码器模块编码后的结果。

[0114] 音频编码器模块可以包含为一层或多层编码层，编码层可以采用单向或双向长短时记忆神经网络中长短时记忆层或卷积神经网络的卷积层，具体采用哪种结构可以根据应用需求确定。如对于有实时性要求的语音识别可以使用3～5层的单向长短时记忆层，对于没有实时性要求的语音识别可以使用3～5层的双向长短时记忆层。其中，实时性要求指的是边说边识别，而不是等语音说完了才一次性出识别结果。

[0115] 本实施例中可以选择使用5层单向长短时记忆层对输入的音频特征$X=[x_1,x_2,...,x_k]$进行处理，输出一组编码后的音频特征向量序列$H^x=\left[h_1^x,h_2^x,...,h_k^x\right]$。

[0116] 2、热词编码器模块：

[0117] 热词编码器模块对热词库中各热词进行编码，得到热词编码结果。

[0118] 具体的，热词编码器模块可以对热词库中每个热词进行独立编码，得到由各热词独立编码后的各热词特征向量组成的热词特征向量序列。

[0119] 定义热词库中共有N+1个热词：

[0120] $Z=[z_0,z_1,...,z_N]$

[0121] 其中，z_N为第N个热词。其中，z_0是一个特殊的热词“<no-bias>”，其表示不存在热词，当解码过程选中的热词为z_0时，表示当前解码时刻的待解码字符不是任何一个热词，用于处理正在识别的语音片段不是热词的情况。

[0122] 热词的总个数为N+1，则热词编码器对每个热词独立编码，得到的热词特征向量序列表示为：

[0123] $$H^z=\left[h_0^z,h_1^z,...,h_N^z\right]$$

[0124] 其中，h_N^z为第N个热词经过热词编码器模块独立编码之后的热词特征向量。

[0125] 可以理解的是，不同热词包含的字符数可能不一样，比如定义“中科大”和“科大讯飞”都是热词，则包含的字符数分别为3和4。

[0126] 为了便于模型处理，本实施例中可以将变长的热词，统一编码成相同维度的向量。具体的，热词编码器模块可以按照设定的维度，将各个热词分别独立编码成相同维度的热词特征向量。

[0127] 热词编码器模块可以包含为一层或多层编码层，编码层可以采用单向或双向长短时记忆神经网络中长短时记忆层或卷积神经网络的卷积层。通常来讲，能同时看到左右全部信息的双向长短时记忆层对热词的编码效果好于单向长短时记忆层，如选择使用一层双向长短时记忆层，以热词“科大讯飞”为例，该热词由“科”、“大”、“讯”、“飞”四个字组成，一层双向长短时记忆层的热词编码器模块对它编码的过程如图5所示：

[0128] 图5中的左边为双向长短时记忆层的正向部分，右边为反向部分，将正向和反向最后一步的输出向量h_f^z和h_b^z进行拼接，得到的向量h^z即为热词的编码向量表示。

[0129] 3、联合注意力模块：

[0130] 联合注意力模块接收并处理音频编码结果和热词编码结果，得到当前解码时刻所需的拼接特征，该拼接特征包括音频相关特征和热词相关特征。

[0131] 本实施例中介绍了联合注意力模块的一种可选架构，如图4所示，联合注意力模块可以包括：

[0132] 第一注意力模型和第二注意力模型。其中：

[0133] 第一注意力模型可以基于解码器模块在当前解码时刻输出的表示已解码结果信息的状态向量，以及热词编码结果，从音频编码结果中确定当前解码时刻所需的音频相关特征。

[0134] 具体的，可以将状态向量、热词编码结果作为第一注意力模型的输入，由第一注意力模型从音频编码结果中确定当前解码时刻所需的音频相关特征。

[0135] 第二注意力模型可以基于音频相关特征，从热词编码结果中确定当前解码时刻所需的热词相关特征。

[0136] 具体的，可以将音频相关特征作为第二注意力模型的输入，由第二注意力模型从热词编码结果中确定当前解码时刻所需的热词相关特征。

[0137] 最后，由所述音频相关特征和所述热词相关特征组合成当前解码时刻所需的拼接特征。

[0138] 由上可知,解码器模块在当前解码时刻输出的状态向量能够表示已解码结果信息,因此可以基于该状态向量和热词编码结果,对音频编码结果进行注意力机制操作,确定当前解码时刻所需的音频相关特征。也即,本实施例中第一注意力模型使用了音频、热词联合注意力机制,让热词参与到音频相关特征的计算中,由于利用了热词信息,如果当前解码时刻的待解码字符为某个热词,则音频相关特征可以抽取到该热词对应的完整音频信息。

[0139] 进一步,再利用音频相关特征对热词编码结果进行注意力机制操作,确定当前解码时刻所需的热词相关特征,因为音频相关特征包含了热词的完整音频信息,以此得到的热词相关特征也更加准确。

[0140] 其中,注意力机制使用一个向量作为查询项(query),对一组特征向量序列进行注意力机制操作,选出与查询项最匹配的特征向量作为输出,具体为:将查询项与特征向量序列中每个特征向量计算一个匹配系数,然后将这些匹配系数与对应的特征向量相乘并求和,得到一个新的特征向量即为与查询项最匹配的特征向量。

[0141] 定义当前时刻为t时刻,解码器模块在t时刻输出的状态向量为d_t,则第一注意力模型基于状态向量d_t和热词特征向量序列H^z,从音频特征向量序列H^x中确定当前解码时刻所需的音频相关特征c_t^x。第二注意力模型以c_t^x为查询项,对热词特征向量序列H^z执行注意力机制操作,确定当前解码时刻所需的热词相关特征c_t^z。

[0142] 接下来,对第一注意力模型的实施方式进行详细说明:

[0143] 首先,第一注意力模型分别以热词特征向量序列H^z中的每一热词特征向量h_i^z与状态向量d_t的组合为查询项,对音频特征向量序列H^x中每一个音频特征向量h_j^x进行注意力机制操作,得到匹配系数矩阵E^t,所述匹配系数矩阵E^t中包含任一热词与任一帧语音的匹配度e_{ij}^t,e_{ij}^t表示第i个热词与第j帧语音的匹配度,也就是第j帧语音是第i个热词的可能性。

[0144] 其中,e_{ij}^t的计算过程参照下式:

[0145] $$e_{ij}^t = < W_d * d_t + W_z * h_i^z, W_x * h_j^x >$$

[0146] 其中,W_d、W_z、W_x为模型参数,$W_d \in R^{D \times D_d}$,$W_z \in R^{D \times D_z}$,$W_x \in R^{D \times D_x}$,D_d、D_z、D_x分别表示向量d_t、h_i^z、h_j^x的维度,三个矩阵的行数相同,均为D,操作符<.,.>表示求向量内积。

[0147] 元素e_{ij}^t组成热词与语音帧的匹配系数矩阵E^t,$E^t \in R^{K \times (N+1)}$。其中$e_{ij}^t$表示$E^t$中第i行第j列的元素,$E^t$中的列向量表示某个热词与音频特征向量序列间的匹配度,E^t中的行向量表示某帧音频特征向量与热词特征向量序列的匹配度。

[0148] 进一步,第一注意力模型根据上述匹配系数矩阵E^t,从音频特征向量序列H^x中确定当前解码时刻所需的音频相关特征c_t^x。

[0149] 具体的,该过程可以包括如下步骤:

[0150] S1、根据匹配系数矩阵E^t,确定每个热词作为当前解码时刻的待解码字符的概率w^t。

[0151] E^t中第i行第j列的元素表示第j帧音频是第i个热词的可能性,那么将E^t中的每一行进行softmax归一化,然后再把所有行向量相加求平均,得到一个N+1维的行向量,记为:

[0152] $w^t=\left[w_0^t,w_1^t,...w_i^t,...w_N^t\right]$

[0153] 其中，w_i^t表示当前解码时刻t所要解码的字符为第i个热词的可能性。也即，确定出当前解码时刻t，语音中最有可能出现的是哪个热词。

[0154] S2、根据匹配系数矩阵E^t及每个热词作为当前解码时刻的待解码字符的概率w^t，确定每一帧语音作为当前解码时刻所需的语音内容的概率a^t。

[0155] 具体的，将E^t中的每一列进行softmax归一化，得到列向量归一化后的矩阵A^t，然后将w^t中的元素作为矩阵A^t中列向量的加权系数，将矩阵A^t中所有的列向量加权求和，得到一个K维的行向量，记为：

[0156] $a^t=\left[a_1^t,a_2^t,...,a_j^t,...a_K^t\right]$

[0157] 其中，a_j^t表示第j帧音频特征是当前解码时刻t解码所需的语音内容的可能性。

[0158] S3、以每一帧语音作为当前解码时刻所需的语音内容的概率a^t作为加权系数，将音频特征向量序列H^x中各帧语音的音频特征向量加权求和，得到当前解码时刻所需的音频相关特征c_t^x。

[0159] 具体的，以a^t中的元素作为音频特征向量序列$H^x=\left[h_1^x,h_2^x,...,h_k^x\right]$中对应位置音频特征向量的加权系数，将音频特征向量加权求和，得到音频相关特征向量c_t^x。

[0160] 再进一步的，对第二注意力模型的实施方式进行详细说明：

[0161] 第二注意力模型根据上述音频相关特征c_t^x，从热词特征向量序列H^z中确定当前解码时刻所需的热词相关特征c_t^z。

[0162] 具体的，该过程可以包括如下步骤：

[0163] S1、第二注意力模型以音频相关特征c_t^x作为查询项，对热词特征向量序列H^z进行注意力机制操作，得到热词匹配系数向量b^t，热词匹配系数向量b^t中包含每一热词作为当前解码时刻的待解码字符的概率。b^t记为：

[0164] $b^t=\left[b_0^t,b_1^t,...,b_i^t,...b_N^t\right]$

[0165] 其中，b_i^t表示第i个热词作为当前解码时刻的解码字符的概率。

[0166] 具体的，将c_t^x与每一个热词特征向量通过一个小型神经网络计算得到一个匹配系数，然后将这些匹配系数进行softmax归一化，得到b_i^t。

[0167] S2、以每一热词作为当前解码时刻的待解码字符的概率b_i^t作为加权系数，将热词特征向量序列H^z中每个热词的热词特征向量加权求和，得到当前解码时刻所需的热词相关特征c_t^z。

[0168] 由于c_t^x包含了潜在的热词的完整音频信息，而非热词的局部信息，基于此确定的热词相关特征c_t^z也更加准确。

[0169] c_t^x和c_t^z确定之后需要进行拼接，得到当前解码时刻所需的拼接特征c_t，将拼接特征c_t送入解码器模块。

[0170] 进一步的，还可以将上述确定的当前解码时刻的待解码字符的概率b^t送入分类器

模块供对热词分类使用。

[0171] 4、解码器模块：

[0172] 解码器模块接收并处理联合注意力模块输出的当前解码时刻所需的拼接特征，得到解码器模块当前解码时刻的输出特征。

[0173] 具体的，解码器模块可以利用当前解码时刻t的前一解码时刻t-1所需的拼接特征c_{t-1}及前一解码时刻t-1的识别结果字符，计算得到当前解码时刻t的状态向量d_t。

[0174] 其中，d_t有两个作用，其一是发送给联合注意力模块，以供联合注意力模块执行上述实施例介绍的操作过程，计算得到当前解码时刻的c_t。

[0175] 其二，解码器模块利用当前解码时刻的状态向量d_t和当前解码时刻所需的拼接特征c_t，计算得到解码器模块在当前解码时刻的输出特征h_t^d。

[0176] 需要说明的是，解码器模块可以包含多个神经网络层，本申请中可以选用两层单向长短时记忆层。在解码当前时刻t的待解码字符时，第一层长短时记忆层以t-1时刻的识别结果字符和注意力模块输出的拼接特征c_{t-1}为输入，计算得到解码器模块当前解码时刻的状态向量d_t。解码器模块将d_t与c_t作为第二层长短时记忆层的输入，计算得到解码器模块的输出特征h_t^d。

[0177] 5、分类器模块：

[0178] 分类器模块利用解码器模块当前解码时刻的输出特征，确定待识别语音在当前解码时刻的识别结果。

[0179] 具体的，分类器模块可以利用解码器模块在当前解码时刻的输出特征h_t^d，确定待识别语音在当前解码时刻的识别结果。

[0180] 由上可知，输出特征h_t^d是根据解码器模块的状态向量d_t和当前解码时刻所需的拼接特征c_t联合确定的，而拼接特征c_t中的c_t^x包含了潜在的热词的完整音频信息，而非热词的局部信息，基于此确定的热词相关特征c_t^z也更加准确，由此可以确定最终得到的输出特征h_t^d也会更加准确，进而基于此确定的识别结果也更加准确，并且能够提高热词的识别准确率。

[0181] 本申请的一个实施例中，提供了分类器模块的两种实现方式，一种是采用现有的常规静态分类器，该静态分类器中分类节点的数目始终维持不变，包含有常见的字符。分类器模块可以根据输出特征h_t^d来确定各个分类节点字符的得分概率，进而组合成最终的识别结果。

[0182] 但是，这种常规静态分类器将热词表示为常用字符的组合，对热词进行逐字符解码，很容易导致非热词片段的热词误触发。例如，对于待识别语音内容为“这个模型训飞了”的语音数据，假设热词为“科大讯飞”，则使用静态分类器的识别结果可能是“这个模型讯飞了”。因为“训飞”和热词“科大讯飞”中的“讯飞”二字发音一样，由于静态分类器对热词进行逐字解码、逐字激励，每个单字都存在被激励的可能，很可能将语音片段中与热词存在部分发音匹配的内容错误得激励为热词的一部分，也就是将“训飞”中的“训”错误识别为热词“科大讯飞”中的“讯”。

[0183] 为此，本申请提供了一种分类器模块的新结构，分类器模块的分类节点既包含固

定的常用字符节点,还包含可动态扩展的热词节点,从而实现直接对热词进行整词识别的目的,不需要像现有技术那样将热词拆分,单个字符逐字符进行识别、激励。仍以上述示例的例子说明,对于语音数据“这个模型训飞了”,由于“训飞”只是与热词“科大讯飞”中部分字符的读音相同,其与热词“科大讯飞”这个整词的匹配程度远远不过,因此不会发生整个热词误识别的问题。而当语音数据中包含某个热词时,按照本实施例的分类器模块,由于分类节点中包含了热词这个整词,因此可以直接识别出热词这个整词,很好的提升了热词识别效果。

[0184] 本实施例的分类器模块中的热词节点个数可以根据场景而动态调整,如当前场景对应的热词库中中N个热词,则可以设置相同个数的N个热词节点。以中文语音识别为例,以汉字为建模单元,假设常用汉字个数为V个,则分类器模块的固定常用字符节点的个数为V,当热词库中共存在N个热词时,则分类器模块的热词节点的个数可以为N,也即分类器模块的所有分类节点的数目为V+N。

[0185] 基于上述这种新型结构的分类器模块,分类器模块进行语音识别的过程可以包括:

[0186] 分类器模块利用解码器模块在当前解码时刻的输出特征h_t^d,确定各常用字符节点的概率得分和各热词节点的概率得分;进而确定最终的识别结果。

[0187] 一种可选的方式下,分类器模块可以利用解码器模块在当前解码时刻的输出特征h_t^d,分别确定各常用字符节点的概率得分,以及确定各热词节点的概率得分。

[0188] 另一种可选的方式下,分类器模块可以利用解码器模块在当前解码时刻的输出特征h_t^d,确定各常用字符节点的概率得分。进一步,利用前述实施例介绍的热词匹配系数向量b^t,确定各热词节点的概率得分。

[0189] 可以理解的是,对于分类器模块中的固定常用字符节点,其概率得分可以通过静态分类器来确定。具体的,静态分类器利用解码器模块在当前解码时刻的输出特征h_t^d确定各常用字符节点的概率得分。

[0190] 静态分类器输出一个V维的概率分布,记为:$P_v(y_t) = soft\max(W * h_t^d)$

[0191] 其中,y_t表示当前解码时刻t的待解码字符,矩阵W为静态分类器的模型参数,假设解码器模块的输出特征h_t^d的维度为M,则W为一个V*M的矩阵,$P_v(y_t)$中的元素表示对应常用字符节点的常用字符的概率得分。

[0192] 对于分类器模块中的动态可扩展热词节点,其概率得分可以通过热词分类器来确定。具体的,热词分类器可以利用热词匹配系数向量b^t来确定各热词节点的概率得分。

[0193] 前述已经介绍过程,热词匹配系数向量b^t中包含每一热词作为当前解码时刻的待解码字符的概率,因此可以用该概率,作为对应热词节点的概率得分。

[0194] $$b^t = \left[b_0^t, b_1^t, ..., b_i^t, ...b_N^t\right]$$

[0195] 其中,b_i^t表示第i个热词作为当前解码时刻的解码字符的概率,可以将此作为第i个热词节点的概率得分。第0个热词为“<no-bias>”表示“不是热词”,当i等于0时,b_0^t表示解码结果“不是热词”的概率得分。

[0196] 在确定了常用字符节点和热词节点的概率得分之后,可以根据两种类型节点的概率得分,确定待识别语音在当前解码时刻的识别结果。

[0197] 可以理解的是,由于同时存在静态分类器和热词分类器两个分类器,因此分类器模块还可以增加一个判断器,用于决定到底使用哪个分类器的结果作为最终结果,该判断器输出一个标量的概率值P_b^t,表示当前解码时刻t使用热词分类器/静态分类器的结果作为最终输出结果的概率得分。

[0198] 以P_b^t表示当前解码时刻t使用热词分类器的结果作为最终输出结果的概率得分为例进行说明,P_b^t可以表示为:

[0199] $$P_b^t = sig\,\mathrm{mo}id(< w_b, h_t^d >)$$

[0200] 其中,w_b为模型参数,是一个与h_t^d维度相同的权重向量,sigmoid为神经网络激活函数。

[0201] 则判断器可以根据两个分类器输出的概率得分,确定待识别语音在当前解码时刻的识别结果,具体可以包括:

[0202] 对于N个热词中的第i个热词节点(i的取值范围为[1,N]),在静态分类器输出的概率分布中它的得分为0,在热词分类器中它的概率得分为b_i^t,因此最终它的概率得分为$P_b^t * b_i^t$;对于V个常用字符y_t,在静态分类器输出的概率分布中它的得分为$P_v(y_t)$,在热词分类器中它的概率得分为$P_v(y_t) * b_0^t$,因此最终它的概率得分为$(1-P_b^t) * P_v(y_t) + P_b^t * P_v(y_t) * b_0^t$。

[0203] 在本申请的又一个实施例中,介绍了一种上述语音识别模型的训练方式。

[0204] 由于本申请提出的语音识别模型需要具备支持任意热词识别的能力,这就说明在模型训练中不能限定热词。因此本申请可以从训练数据的文本标注中随机挑选标注片段作为热词,参与整个模型训练。具体流程可以包括:

[0205] S1、获取标注有识别文本的语音训练数据。

[0206] 其中,语音训练数据的文本标注序列可以表示为:

[0207] $Y=[y_0, y_1, \ldots y_t, \ldots y_T]$

[0208] 其中,y_t表示文本标注序列中第t个字符,T+1为识别文本的总字符数目。其中,y_0为句子开始符“<s>”,y_T为句子结束符“</s>”。

[0209] 以中文语音识别为例,并用单个汉字作为建模单元。假设某句话的文本内容为“欢迎来到科大讯飞”,共有8个汉字,加上句子开始符和结束符,文本标注序列总共有10个字符,则文本标注序列表示为:

[0210] Y=[<s>,欢,迎,来,到,科,大,讯,飞,</s>]。

[0211] S2、获取所述语音训练数据的音频特征。

[0212] 其中,音频特征可以选用滤波器组FilterBank特征、梅尔频率倒谱系数MFCC特征、感知线性预测PLP特征等。

[0213] S3、从所述语音训练数据的标注文本中随机挑选标注片段作为训练热词。

[0214] 具体的,本申请可以预先设置P和N两个参数,P为某句训练数据是否挑选训练热词的概率,N为挑选的训练热词的最大字数。则任意一句训练数据被挑选出选取训练热词的概率为P,从该句训练数据的文本标注序列中,挑选最多连续N个字符作为训练热词。以“欢迎

来到科大讯飞”为例,从该句话中挑选训练热词的标注对比如下表所示:

[0215]

原始标注	<s>	欢	迎	来	到	科	大	讯	飞	</s>
第一种标注	<s>	欢	迎	来	到	科大讯飞	<bias>	</s>		
第二种标注	<s>	欢	迎	来	到	科大	讯	飞	<bias>	</s>

[0216] 其中,上述表格中第一种标注为:当“科大讯飞”被选为训练热词后的标注;第二种标注为:当“科大”“被选为训练热词后的标注。

[0217] 由上可知,当原始标注中“科”、“大”、“讯”、“飞”被挑选为训练热词,则需要把这四个字合并为整词“科大讯飞”,并且它的后面添加特殊标记符“<bias>”。“<bias>”的作用是引入训练错误,强迫模型训练时更新训练热词相关的模型参数,比如热词编码器模块。当“科大讯飞”或者“科大”被选为训练热词后,需要将它加入这次模型更新的训练热词列表中,作为热词编码器模块的输入和分类器模块的训练热词分类节点。每次模型更新时训练热词挑选工作独立进行,初始时刻训练热词列表为空。

[0218] S4、利用所述训练热词、所述音频特征及语音训练数据的识别文本,训练语音识别模型。

[0219] 具体的,以训练热词和音频特征作为训练样本输入,以语音训练数据的识别文本作为样本标签,训练语音识别模型。

[0220] 本申请实施例还提供了一种语音识别装置,下面对本申请实施例提供的语音识别装置进行描述,下文描述的语音识别装置与上文描述的语音识别方法可相互对应参照。

[0221] 请参阅图6,示出了本申请实施例提供的语音识别装置的结构示意图,该语音识别装置可以包括:

[0222] 数据获取单元11,用于获取待识别语音以及配置的热词库;

[0223] 音频相关特征获取单元12,用于基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征;

[0224] 热词相关特征获取单元13,用于基于所述音频相关特征,从所述热词库中确定当前解码时刻所需的热词相关特征;

[0225] 识别结果获取单元14,用于基于所述音频相关特征和所述热词相关特征,确定所述待识别语音在当前解码时刻的识别结果。

[0226] 可选的,上述音频相关特征获取单元可以包括:

[0227] 第一音频相关特征获取子单元,用于获取当前解码时刻之前的已解码结果信息;

[0228] 第二音频相关特征获取子单元,用于基于所述已解码结果信息及所热词库,从所述待识别语音中确定当前解码时刻所需的音频相关特征。

[0229] 可选的,上述音频相关特征获取单元、热词相关特征获取单元及识别结果获取单元的实现过程可以通过语音识别模型实现,具体的,利用预先训练的语音识别模型处理所

述待识别语音及所述热词库,得到语音识别模型输出的待识别语音的识别结果,其中:

[0230] 所述语音识别模型具备接收并处理待识别语音及热词库,以输出待识别语音的识别结果的能力。

[0231] 具体的,语音识别模型可以具备基于当前解码时刻之前的已解码结果信息及热词库,从音频特征中确定当前解码时刻所需的音频相关特征,并基于音频相关特征从热词库中确定当前解码时刻所需的热词相关特征,基于所述音频相关特征和所述热词相关特征,确定待识别语音在当前解码时刻的识别结果的能力。

[0232] 可选的,语音识别模型可以包括音频编码器模块、热词编码器模块、联合注意力模块、解码器模块及分类器模块;

[0233] 其中,所述音频编码器模块对所述待识别语音进行编码,得到音频编码结果。

[0234] 具体的,由所述音频编码器模块对所述音频特征进行编码,得到由每帧语音的音频特征向量组成的音频特征向量序列。

[0235] 所述热词编码器模块对所述热词库中各热词进行编码,得到热词编码结果。

[0236] 具体的,由所述热词编码器模块对每个所述热词进行独立编码,得到由各热词独立编码后的各热词特征向量组成的热词特征向量序列。

[0237] 所述联合注意力模块接收并处理所述音频编码结果和所述热词编码结果,得到当前解码时刻所需的拼接特征,所述拼接特征包括音频相关特征和热词相关特征。

[0238] 所述解码器模块接收并处理所述当前解码时刻所需的拼接特征,得到解码器模块当前解码时刻的输出特征。

[0239] 所述分类器模块利用解码器模块当前解码时刻的输出特征,确定待识别语音在当前解码时刻的识别结果。

[0240] 其中,可选的,所述联合注意力模块可以包括:

[0241] 第一注意力模型和第二注意力模型。

[0242] 所述第一注意力模型基于解码器模块在当前解码时刻输出的表示已解码结果信息的状态向量,以及所述热词编码结果,从所述音频编码结果中确定当前解码时刻所需的音频相关特征。

[0243] 具体的,可以将所述状态向量、所述热词编码结果作为第一注意力模型的输入,由所述第一注意力模型从所述音频编码结果中确定当前解码时刻所需的音频相关特征。

[0244] 所述第二注意力模型基于所述音频相关特征,从所述热词编码结果中确定当前解码时刻所需的热词相关特征。

[0245] 具体的,可以将所述音频相关特征作为第二注意力模型的输入,由所述第二注意力模型从所述热词编码结果中确定当前解码时刻所需的热词相关特征。

[0246] 由所述音频相关特征和所述热词相关特征组合成当前解码时刻所需的拼接特征。

[0247] 可选的,上述热词编码器模块对每个所述热词进行独立编码的过程,可以包括:

[0248] 所述热词编码器模块按照设定的维度,将各个所述热词分别独立编码成相同维度的热词特征向量。

[0249] 可选的,上述第一注意力模型基于解码器模块在当前解码时刻输出的表示已解码结果信息的状态向量及热词特征向量序列,从所述音频特征向量序列中确定当前解码时刻所需的音频相关特征的过程,可以包括:

[0250] 第一注意力模型分别以所述热词特征向量序列中的每一热词特征向量与所述状态向量的组合为查询项,对所述音频特征向量序列进行注意力机制操作,得到匹配系数矩阵,所述匹配系数矩阵中包含任一热词与任一帧语音的匹配度;

[0251] 根据所述匹配系数矩阵,从所述音频特征向量序列中确定当前解码时刻所需的音频相关特征。

[0252] 可选的,上述第一注意力模型根据所述匹配系数矩阵,从所述音频特征向量序列中确定当前解码时刻所需的音频相关特征的过程,可以包括:

[0253] 根据所述匹配系数矩阵,确定每个热词作为当前解码时刻的待解码字符的概率;

[0254] 根据所述匹配系数矩阵及每个热词作为当前解码时刻的待解码字符的概率,确定每一帧语音作为当前解码时刻所需的语音内容的概率;

[0255] 以每一帧语音作为当前解码时刻所需的语音内容的概率作为加权系数,将所述音频特征向量序列中各帧语音的音频特征向量加权求和,得到当前解码时刻所需的音频相关特征。

[0256] 可选的,上述第二注意力模型基于音频相关特征从热词特征向量序列中确定当前解码时刻所需的热词相关特征的过程,可以包括:

[0257] 第二注意力模型以所述音频相关特征作为查询项,对所述热词特征向量序列进行注意力机制操作,得到热词匹配系数向量,所述热词匹配系数向量中包含每一热词作为当前解码时刻的待解码字符的概率;

[0258] 以每一热词作为当前解码时刻的待解码字符的概率作为加权系数,将所述热词特征向量序列中每个热词的热词特征向量加权求和,得到当前解码时刻所需的热词相关特征。

[0259] 可选的,上述联合注意力模块还可以将所述热词匹配系数向量发送至所述分类器模块;则所述分类器模块具体为,利用所述解码器模块在当前解码时刻的输出特征及所述热词匹配系数向量,确定待识别语音在当前解码时刻的识别结果。

[0260] 可选的,上述分类器模块的分类节点可以包括固定的常用字符节点和可动态扩展的热词节点。基于此,

[0261] 分类器模块可以利用解码器模块在当前解码时刻的输出特征,确定各所述常用字符节点的概率得分和各所述热词节点的概率得分;根据各所述常用字符节点的概率得分和各所述热词节点的概率得分,确定待识别语音在当前解码时刻的识别结果。

[0262] 具体的,分类器模块可以利用解码器模块在当前解码时刻的输出特征,确定各所述常用字符节点的概率得分;

[0263] 分类器模块利用所述热词匹配系数向量,确定各所述热词节点的概率得分;

[0264] 根据各所述常用字符节点的概率得分和各所述热词节点的概率得分,确定待识别语音在当前解码时刻的识别结果。

[0265] 可选的,本申请的装置还可以包括模型训练单元,用于:

[0266] 获取标注有识别文本的语音训练数据;

[0267] 获取所述语音训练数据的音频特征;

[0268] 从所述语音训练数据的标注文本中随机挑选标注片段作为训练热词;

[0269] 利用所述训练热词、所述音频特征及语音训练数据的识别文本,训练语音识别模

型。

[0270] 可选的,上述数据获取单元获取待识别语音的音频特征的过程,可以包括:

[0271] 获取待识别语音的以下任意一项音频特征:

[0272] 滤波器组FilterBank特征、梅尔频率倒谱系数MFCC特征、感知线性预测PLP特征。

[0273] 本申请实施例还提供了一种电子设备,请参阅图7,示出了该电子设备的结构示意图,该电子设备可以包括:至少一个处理器1001,至少一个通信接口1002,至少一个存储器1003和至少一个通信总线1004;

[0274] 在本申请实施例中,处理器1001、通信接口1002、存储器1003、通信总线1004的数量为至少一个,且处理器1001、通信接口1002、存储器1003通过通信总线1004完成相互间的通信;

[0275] 处理器1001可能是一个中央处理器CPU,或者是特定集成电路ASIC(Application Specific Integrated Circuit),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;

[0276] 存储器1003可能包含高速RAM存储器,也可能还包括非易失性存储器(non-volatile memory)等,例如至少一个磁盘存储器;

[0277] 其中,存储器存储有程序,处理器可调用存储器存储的程序,所述程序用于:

[0278] 获取待识别语音以及配置的热词库;

[0279] 基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征;

[0280] 基于所述音频相关特征,从所述热词库中确定当前解码时刻所需的热词相关特征;

[0281] 基于所述音频相关特征和所述热词相关特征,确定所述待识别语音在当前解码时刻的识别结果。

[0282] 可选的,所述程序的细化功能和扩展功能可参照上文描述。

[0283] 本申请实施例还提供一种可读存储介质,该可读存储介质可存储有适于处理器执行的程序,所述程序用于:

[0284] 获取待识别语音以及配置的热词库;

[0285] 基于所述待识别语音及所述热词库,确定当前解码时刻所需的音频相关特征;

[0286] 基于所述音频相关特征,从所述热词库中确定当前解码时刻所需的热词相关特征;

[0287] 基于所述音频相关特征和所述热词相关特征,确定所述待识别语音在当前解码时刻的识别结果。

[0288] 最后,还需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0289] 本说明书中各个实施例采用递进的方式描述,每个实施例重点说明的都是与其他

实施例的不同之处,各个实施例之间可以根据需要进行组合,且相同相似部分互相参见即可。

[0290] 对所公开的实施例的上述说明,使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下,在其它实施例中实现。因此,本发明将不会被限制于本文所示的这些实施例,而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

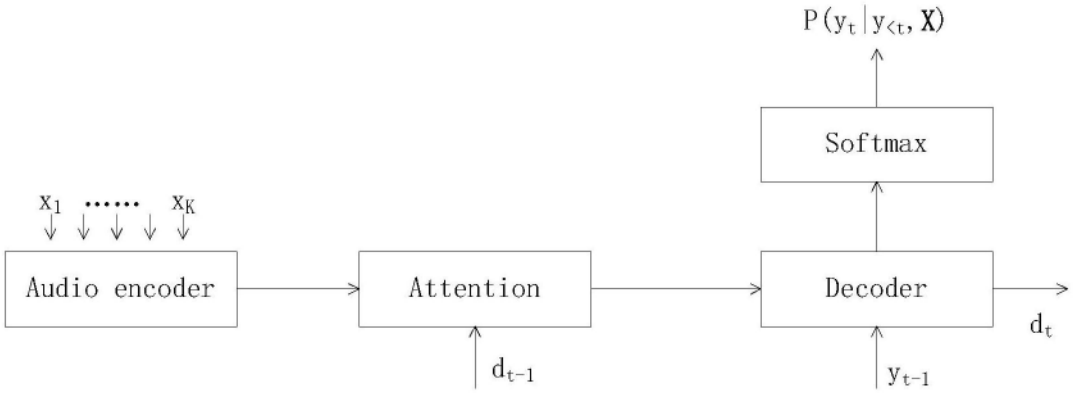

P(y_t|y_<t, X)
Softmax
x_1 ..... x_K
Audio encoder
Attention
Decoder
d_t
d_t-1
y_t-1

图1

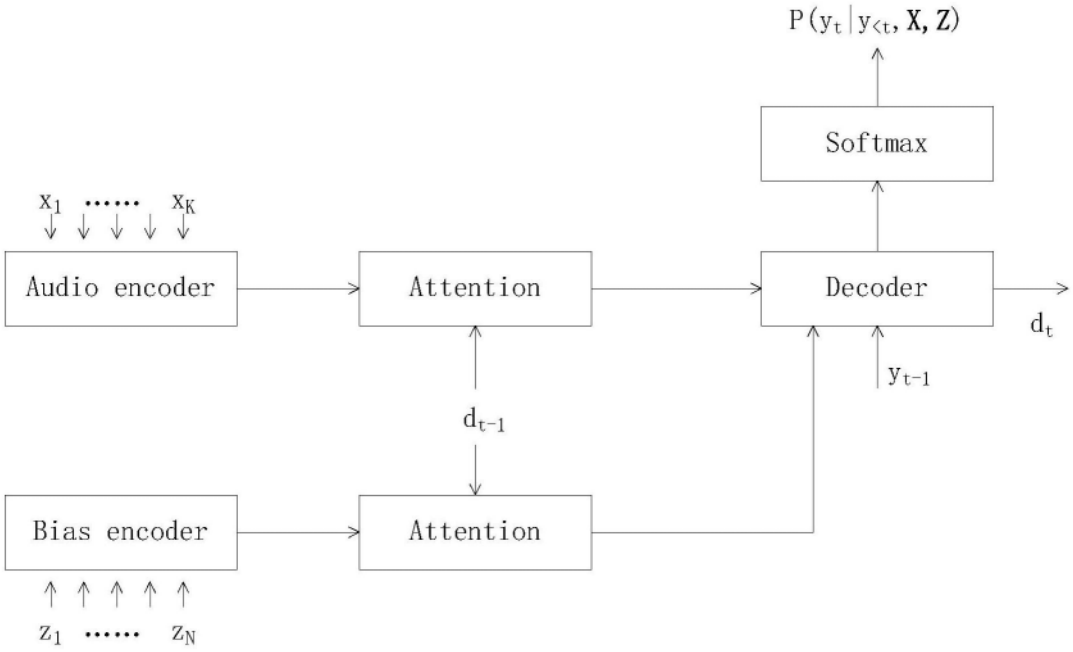

P(y_t|y_<t, X, Z)
Softmax
x_1 ..... x_K
Audio encoder
Attention
Decoder
d_t
y_t-1
d_t-1
Bias encoder
Attention
z_1 ..... z_N

图2

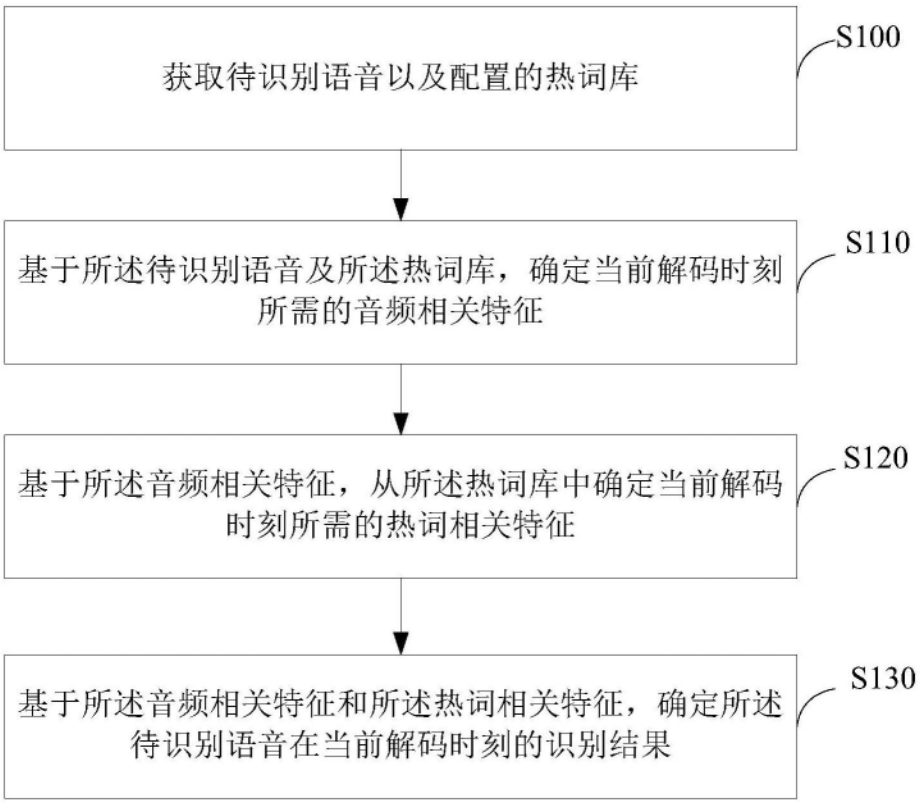

获取待识别语音以及配置的热词库
S100
基于所述待识别语音及所述热词库，确定当前解码时刻所需的音频相关特征
S110
基于所述音频相关特征，从所述热词库中确定当前解码时刻所需的热词相关特征
S120
基于所述音频相关特征和所述热词相关特征，确定所述待识别语音在当前解码时刻的识别结果
S130

图3

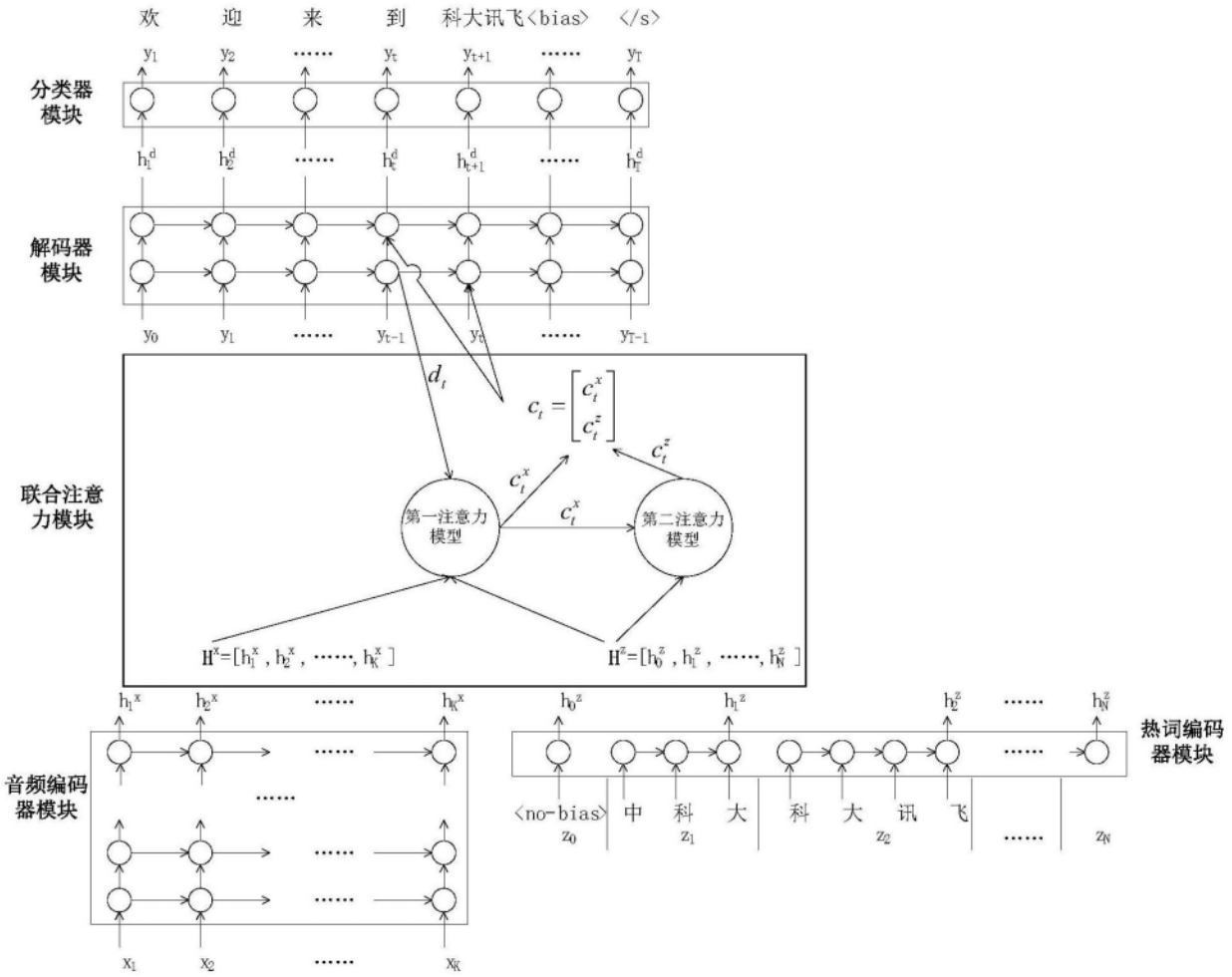

欢 迎 来 到 科大讯飞<bias> </s>
分类器模块
解码器模块
联合注意力模块
第一注意力模型
第二注意力模型
音频编码器模块
热词编码器模块
<no-bias> 中 科 大 科 大 讯 飞

图4

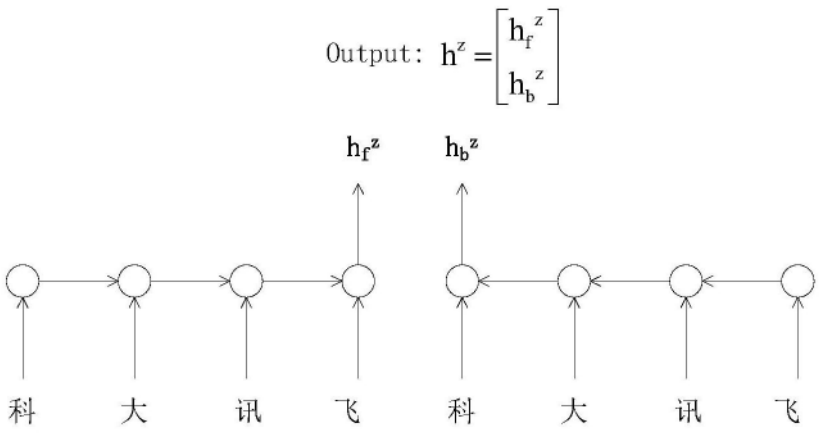

Output: $h^z = \begin{bmatrix} h_f^z \\ h_b^z \end{bmatrix}$
h_f^z
h_b^z
科 大 讯 飞 科 大 讯 飞
Input:“科大讯飞”

图5

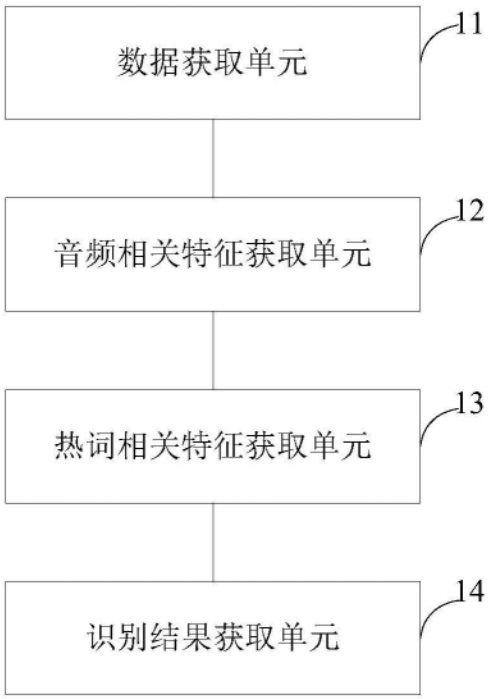

数据获取单元
11
音频相关特征获取单元
12
热词相关特征获取单元
13
识别结果获取单元
14

图6

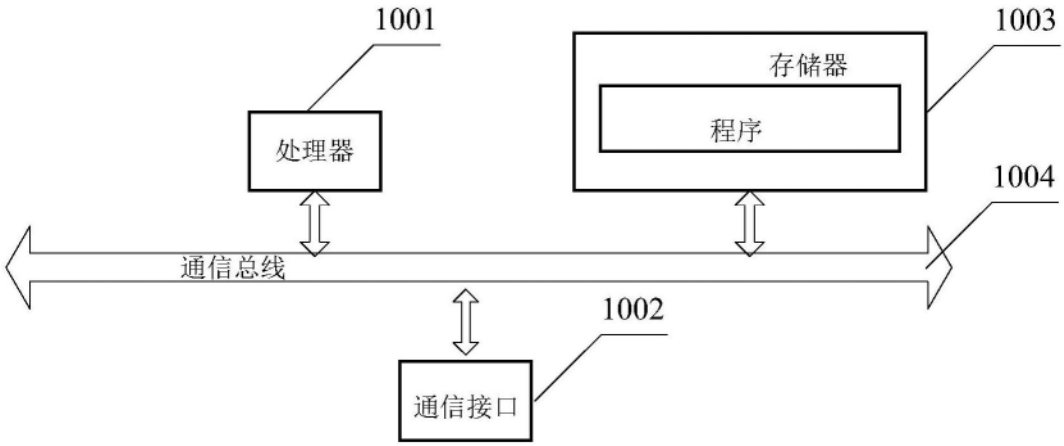

1001
处理器
1003
存储器
程序
1004
通信总线
1002
通信接口

图7