



(12) 发明专利申请

(10) 申请公布号 CN 117238280 A

(43) 申请公布日 2023. 12. 15

(21) 申请号 202311386996.X

(22) 申请日 2023.10.24

(71) 申请人 京东科技信息技术有限公司

地址 101116 北京市大兴区北京经济技术
开发区科创十一街18号院2号楼6层
601

(72) 发明人 付立 范璐 陈蒙 何晓冬

(74) 专利代理机构 北京唯智勤实知识产权代理
事务所(普通合伙) 11557

专利代理师 姜悦

(51) Int. Cl.

G10L 15/06 (2013.01)

G10L 15/26 (2006.01)

权利要求书2页 说明书12页 附图3页

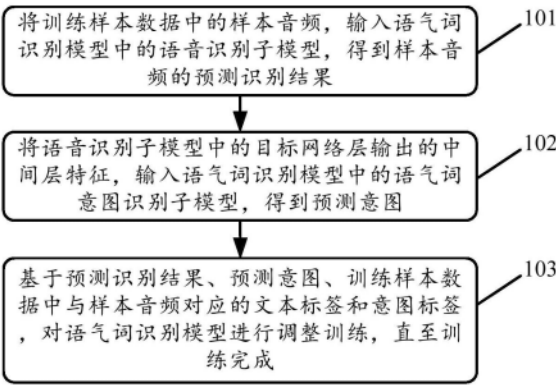
(54) 发明名称

语气词识别模型的生成方法和装置、音频数
据分析方法

(57) 摘要

本公开的实施例公开了语气词识别模型的生成方法和装置、音频数据分析方法。该生成方法的一具体实施方式包括：将训练样本数据中的样本音频，输入语气词识别模型中的语音识别子模型，得到样本音频的预测识别结果，其中，样本音频为包含有语气词的音频数据；将语音识别子模型中的目标网络层输出的中间层特征，输入语气词识别模型中的语气词意图识别子模型，得到预测意图；基于预测识别结果、预测意图、训练样本数据中与样本音频对应的文本标签和意图标签，对语气词识别模型进行调整训练，直至训练完成。该实施方式与语音识别技术有关，通过将语音识别子模型的中间层特征作为输入，来构建语气词意图识别子模型。有助于提高语气词意图识别的准确性。

100



1. 一种语气词识别模型的生成方法,包括:

将训练样本数据中的样本音频,输入语气词识别模型中的语音识别子模型,得到所述样本音频的预测识别结果,其中,所述样本音频为包含有语气词的音频数据;

将所述语音识别子模型中的目标网络层输出的中间层特征,输入所述语气词识别模型中的语气词意图识别子模型,得到预测意图,其中,所述目标网络层为除输入层和输出层以外的网络层;

基于所述预测识别结果、所述预测意图、所述训练样本数据中与所述样本音频对应的文本标签和意图标签,对所述语气词识别模型进行调整训练,直至训练完成。

2. 根据权利要求1所述的生成方法,其中,所述语音识别子模型采用语音识别ASR预训练模型;以及

所述语气词意图识别子模型还用于,对输入的中间层特征进行语气词识别,输出得到预测语气词。

3. 根据权利要求2所述的生成方法,其中,所述基于所述预测识别结果、所述预测意图、所述训练样本数据中与所述样本音频对应的文本标签和意图标签,对所述语气词识别模型进行调整训练,包括:

根据所述预测语气词、所述预测意图,以及与所述样本音频对应的文本标签和意图标签,对所述语气词意图识别子模型进行微调训练。

4. 根据权利要求3所述的生成方法,其中,所述对所述语气词意图识别子模型进行微调训练,包括:

利用交叉熵损失函数,确定损失值是否小于阈值;

响应于确定损失值不小于阈值,利用自适应梯度优化算法,调整所述语气词意图识别子模型中的相关参数,以继续模型训练。

5. 根据权利要求1所述的生成方法,其中,所述语气词意图识别子模型采用包括至少一个解码层的解码网络结构。

6. 根据权利要求1-5之一所述的生成方法,其中,所述训练样本数据通过以下方法得到:

获取表示语气词的音频数据;

根据所述音频数据所表示的语气词,进行文本标注以得到语气词文本标签;

根据所述音频数据所表示的意图,进行意图类别标注以得到意图标签;

基于所述音频数据、对应的语气词文本标签和意图标签,生成训练样本数据。

7. 根据权利要求6所述的生成方法,其中,所述基于所述音频数据、对应的语气词文本标签和意图标签,生成训练样本数据,包括:

对所述音频数据进行处理,得到设定属性的音频频谱特征;

将所述音频频谱特征、对应的语气词文本标签和意图标签,确定为标注数据;

按照设定比例,将所述标注数据划分为训练样本数据和测试样本数据。

8. 一种语气词识别模型的生成装置,包括:

音频识别单元,被配置成将训练样本数据中的样本音频,输入语气词识别模型中的语音识别子模型,得到所述样本音频的预测识别结果,其中,所述样本音频为包含有语气词的音频数据;

意图预测单元,被配置成将所述语音识别子模型中的目标网络层输出的中间层特征,输入所述语气词识别模型中的语气词意图识别子模型,得到预测意图,其中,所述目标网络层为除输入层和输出层以外的网络层;

模型调整单元,被配置成基于所述预测识别结果、所述预测意图、所述训练样本数据中与所述样本音频对应的文本标签和意图标签,对所述语气词识别模型进行调整训练,直至训练完成。

9.一种音频数据分析方法,包括:

将待分析的音频数据,输入预先训练完成的语气词识别模型中的语音识别子模型,预测得到所述音频数据的识别结果,其中,所述语气词识别模型采用如权利要求1-7之一所述的语气词识别模型的生成方法得到;

响应于确定所述识别结果表征为语气词,将所述语音识别子模型中目标网络层输出的中间层特征,输入所述语气词识别模型中的语气词意图识别子模型,预测得到所述语气词的意图;

根据所述语气词的意图,对所述音频数据进行处理。

10.根据权利要求9所述的音频数据分析方法,其中,所述根据所述语气词的意图,对所述音频数据进行处理,包括:

根据所述语气词的意图,生成与所述音频数据对应的回复信息。

11.一种电子设备,包括:

一个或多个处理器;

存储装置,其上存储有一个或多个程序,

当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如权利要求1-7或9-10中任一所述的方法。

12.一种计算机可读介质,其上存储有计算机程序,其中,所述计算机程序被处理器执行时,实现如权利要求1-7或9-10中任一所述的方法。

13.一种计算机程序产品,包括计算机程序,所述计算机程序在被处理器执行时,实现如权利要求1-7或9-10中任一所述的方法。

语气词识别模型的生成方法和装置、音频数据分析方法

技术领域

[0001] 本公开的实施例涉及语音识别技术领域,具体涉及语气词识别模型的生成方法和装置、音频数据分析方法。

背景技术

[0002] 近年来,随着人工智能(AI,Artificial Intelligence)技术的高速发展,采用AI代替人工的方式与客户进行电话沟通,已经广泛应用在多种业务场景中。在对话过程中,系统能否正确地识别与理解客户语音是保证用户体验的前提。通常情况下,AI对客户的语音理解流程分为两个级联的步骤:首先,通过自动语音识别(ASR, Automatic Speech Recognition)技术识别客户语音中的文本内容;其次,将该文本识别结果传递给下游的自然语言处理(NLP, Natural Language Processing)系统,从而识别客户想要表达的意图信息。

[0003] 然而,发明人发现,在实际的语音对话场景中,用户经常会使用口语语气词来表达不同的意图。在汉语表达习惯中,语气词虽然看似简单,但由于上下文语境和发音方式的细微差异,使得表达的意图大不相同。因此,系统对于口语语气词的识别与理解性能显得尤为重要。

[0004] 该背景技术部分中所公开的以上信息仅用于增强对本发明构思的背景的理解,并因此,其可包含并不形成本国的本领域普通技术人员已知的现有技术的信息。

发明内容

[0005] 本公开的内容部分用于以简要的形式介绍构思,这些构思将在后面的具体实施方式部分被详细描述。本公开的内容部分并不旨在标识要求保护的技术方案的关键特征或必要特征,也不旨在用于限制所要求的保护的技术方案的范围。

[0006] 本公开的一些实施例提出了语气词识别模型的生成方法、语气词识别模型的生成装置、音频数据分析方法、音频数据分析装置、电子设备、计算机可读介质和计算机程序产品,来解决以上背景技术部分提到的技术问题中的一项或多项。

[0007] 第一方面,本公开的一些实施例提供了一种语气词识别模型的生成方法,包括:将训练样本数据中的样本音频,输入语气词识别模型中的语音识别子模型,得到样本音频的预测识别结果,其中,样本音频为包含有语气词的音频数据;将语音识别子模型中的目标网络层输出的中间层特征,输入语气词识别模型中的语气词意图识别子模型,得到预测意图,其中,目标网络层为除输入层和输出层以外的网络层;基于预测识别结果、预测意图、训练样本数据中与样本音频对应的文本标签和意图标签,对语气词识别模型进行调整训练,直至训练完成。

[0008] 在一些实施例中,语音识别子模型采用语音识别ASR预训练模型;以及语气词意图识别子模型还用于,对输入的中间层特征进行语气词识别,输出得到预测语气词。

[0009] 在一些实施例中,基于预测识别结果、预测意图、训练样本数据中与样本音频对应

的文本标签和意图标签,对语气词识别模型进行调整训练,包括:根据预测语气词、预测意图,以及与样本音频对应的文本标签和意图标签,对语气词意图识别子模型进行微调训练。

[0010] 在一些实施例中,对语气词意图识别子模型进行微调训练,包括:利用交叉熵损失函数,确定损失值是否小于阈值;响应于确定损失值不小于阈值,利用自适应梯度优化算法,调整语气词意图识别子模型中的相关参数,以继续模型训练。

[0011] 在一些实施例中,语气词意图识别子模型采用包括至少一个解码层的解码网络结构。

[0012] 在一些实施例中,训练样本数据通过以下方法得到:获取表示语气词的音频数据;根据音频数据所表示的语气词,进行文本标注以得到语气词文本标签;根据音频数据所表示的意图,进行意图类别标注以得到意图标签;基于音频数据、对应的语气词文本标签和意图标签,生成训练样本数据。

[0013] 在一些实施例中,基于音频数据、对应的语气词文本标签和意图标签,生成训练样本数据,包括:对音频数据进行处理,得到设定属性的音频频谱特征;将音频频谱特征、对应的语气词文本标签和意图标签,确定为标注数据;按照设定比例,将标注数据划分为训练样本数据和测试样本数据。

[0014] 第二方面,本公开的一些实施例提供了一种语气词识别模型的生成装置,包括:音频识别单元,被配置成将训练样本数据中的样本音频,输入语气词识别模型中的语音识别子模型,得到样本音频的预测识别结果,其中,样本音频为包含有语气词的音频数据;意图预测单元,被配置成将语音识别子模型中的目标网络层输出的中间层特征,输入语气词识别模型中的语气词意图识别子模型,得到预测意图,其中,目标网络层为除输入层和输出层以外的网络层;模型调整单元,被配置成基于预测识别结果、预测意图、训练样本数据中与样本音频对应的文本标签和意图标签,对语气词识别模型进行调整训练,直至训练完成。

[0015] 在一些实施例中,语音识别子模型采用语音识别ASR预训练模型;以及语气词意图识别子模型还用于,对输入的中间层特征进行语气词识别,输出得到预测语气词。

[0016] 在一些实施例中,模型调整单元进一步被配置成根据预测语气词、预测意图,以及与样本音频对应的文本标签和意图标签,对语气词意图识别子模型进行微调训练。

[0017] 在一些实施例中,模型调整单元进一步被配置成利用交叉熵损失函数,确定损失值是否小于阈值;响应于确定损失值不小于阈值,利用自适应梯度优化算法,调整语气词意图识别子模型中的相关参数,以继续模型训练。

[0018] 在一些实施例中,语气词意图识别子模型采用包括至少一个解码层的解码网络结构。

[0019] 在一些实施例中,该生成装置还包括训练样本数据生成单元,被配置成获取表示语气词的音频数据;根据音频数据所表示的语气词,进行文本标注以得到语气词文本标签;根据音频数据所表示的意图,进行意图类别标注以得到意图标签;基于音频数据、对应的语气词文本标签和意图标签,生成训练样本数据。

[0020] 在一些实施例中,训练样本数据生成单元进一步被配置成对音频数据进行处理,得到设定属性的音频频谱特征;将音频频谱特征、对应的语气词文本标签和意图标签,确定为标注数据;按照设定比例,将标注数据划分为训练样本数据和测试样本数据。

[0021] 第三方面,本公开的一些实施例提供了一种音频数据分析方法,包括:将待分析的

音频数据,输入预先训练完成的语气词识别模型中的语音识别子模型,预测得到音频数据的识别结果,其中,语气词识别模型采用上述第一方面中,任一实现方式所描述的语气词识别模型的生成方法得到;响应于确定识别结果表征为语气词,将语音识别子模型中目标网络层输出的中间层特征,输入语气词识别模型中的语气词意图识别子模型,预测得到语气词的意图;根据语气词的意图,对音频数据进行处理。

[0022] 在一些实施例中,根据语气词的意图,对音频数据进行处理,包括:根据语气词的意图,生成与音频数据对应的回复信息。

[0023] 第四方面,本公开的一些实施例提供了一种音频数据分析装置,包括:识别单元,被配置成将待分析的音频数据,输入预先训练完成的语气词识别模型中的语音识别子模型,预测得到音频数据的识别结果,其中,语气词识别模型采用上述第一方面中,任一实现方式所描述的语气词识别模型的生成方法得到;意图识别单元,被配置成响应于确定识别结果表征为语气词,将语音识别子模型中目标网络层输出的中间层特征,输入语气词识别模型中的语气词意图识别子模型,预测得到语气词的意图;处理单元,被配置成根据语气词的意图,对音频数据进行处理。

[0024] 在一些实施例中,处理单元进一步被配置成根据语气词的意图,生成与音频数据对应的回复信息。

[0025] 第五方面,本公开的一些实施例提供了一种电子设备,包括:一个或多个处理器;存储装置,其上存储有一个或多个程序,当一个或多个程序被一个或多个处理器执行,使得一个或多个处理器实现上述第一方面或第三方面中任一实现方式所描述的方法。

[0026] 第六方面,本公开的一些实施例提供了一种计算机可读介质,其上存储有计算机程序,其中,计算机程序被处理器执行时实现上述第一方面或第三方面中任一实现方式所描述的方法。

[0027] 第七方面,本公开的一些实施例提供了一种计算机程序产品,包括计算机程序,计算机程序在被处理器执行时实现上述第一方面或第三方面中任一实现方式所描述的方法。

[0028] 本公开的上述各个实施例具有如下有益效果:本公开的一些实施例的语气词识别模型的生成方法,可以提高模型识别理解语气词意图的准确性。具体来说,在相关语气词识别与理解技术中,通常是将ASR模型识别得到的文本数据,输入NLP模型以识别理解用户意图。然而,对于同一语气词,发音不同会导致所要表达的意思截然不同。因此,即使用户发出不同发音的“嗯”均被正确识别为嗯,但输出的文本结果往往丢失了语气词的语调或语义信息,这就导致下游NLP模型无法准确地判断用户的真实意图。

[0029] 基于此,本公开一些实施例中的语气词识别模型的生成方法,可以在语音识别子模型对样本音频进行识别分析时,将其目标网络层输出的中间层特征作为输入数据,从而输入到语气词意图识别子模型中,以识别样本音频中语气词所表达的意图。与传统文本输入数据相比,中间层特征中不仅包含了音频表达的文本信息,往往还包含了语调或语义等信息。这样可以提高意图预测的准确性。进而可以促进人工智能与用户之间的对话沟通更加顺畅,改善用户体验,从而提升使用人工智能对话场景的用户量。

附图说明

[0030] 结合附图并参考以下具体实施方式,本公开各实施例的上述和其他特征、优点及

方面将变得更加明显。贯穿附图中,相同或相似的附图标记表示相同或相似的元素。应当理解附图是示意性的,元件和元素不一定按照比例绘制。

- [0031] 图1是本公开的语气词识别模型的生成方法的一些实施例的流程图;
- [0032] 图2A是本公开的语气词识别模型的一些实施例的结构示意图;
- [0033] 图2B是本公开的语气词识别模型的生成装置的一些实施例的结构示意图;
- [0034] 图3是本公开的音频数据分析方法的一些实施例的流程图;
- [0035] 图4A是本公开的语气词识别模型的一些运行逻辑示意图;
- [0036] 图4B是本公开的音频数据分析装置的一些实施例的结构示意图;
- [0037] 图5是适于用来实现本公开的一些实施例的电子设备的结构示意图。

具体实施方式

[0038] 下面将参照附图更详细地描述本公开的实施例。虽然附图中显示了本公开的某些实施例,然而应当理解的是,本公开可以通过各种形式来实现,而且不应该被解释为限于这里阐述的实施例。相反,提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是,本公开的附图及实施例仅用于示例性作用,并非用于限制本公开的保护范围。

[0039] 另外还需要说明的是,为了便于描述,附图中仅示出了与有关发明相关的部分。在不冲突的情况下,本公开中的实施例及实施例中的特征可以相互组合。

[0040] 需要注意,本公开中提及的“第一”、“第二”等概念仅用于对不同的装置、模块或单元进行区分,并非用于限定这些装置、模块或单元所执行的功能的顺序或者相互依存关系。

[0041] 需要注意,本公开中提及的“一个”、“多个”的修饰是示意性而非限制性的,本领域技术人员应当理解,除非在上下文另有明确指出,否则应该理解为“一个或多个”。

[0042] 下面将参考附图并结合实施例来详细说明本公开。

[0043] 图1示出了根据本公开的语气词识别模型的生成方法的一些实施例的流程100。该生成方法包括以下步骤:

[0044] 步骤101,将训练样本数据中的样本音频,输入语气词识别模型中的语音识别子模型,得到样本音频的预测识别结果。

[0045] 在一些实施例中,语气词识别模型的生成方法的执行主体(例如服务器),可以通过有线连接方式或无线连接方式,来获取或接收训练样本数据。其中,训练样本数据中可以包括样本音频、样本音频的文本标签和意图标签。这里的样本音频的格式并不限制,能够被执行主体读取识别即可。其中,样本音频通常为包含有语气词的音频数据。如包含语气词的一段话的录音,或者仅包含或表示语气词的音频等。这里的语气词可以根据实际情况进行设置,通常为口语语气词,如嗯、啊、哦等等。而文本标签可以是表征样本音频内容的文本数据。如整个样本音频的文本内容,或者样本音频中特定音频部分的文本内容等。例如,本申请主要用于语气词识别,因此这里的文本标签也可以是表示语气词的文本数据。另外,意图标签可以是表征样本音频中语气词所要表达的意图信息。

[0046] 在这里,执行主体在接收到训练样本数据后,可以将其中的样本音频输入语气词识别模型中的语音识别子模型,从而得到样本音频的预测识别结果。其中,语音识别子模型可以用于对输入的音频数据进行识别,以输出文本识别结果。这里的语音识别子模型通常为采用自动语音识别技术(ASR, Automatic Speech Recognition)的模型。ASR一般是一种

将人的语音转换为文本的技术。

[0047] 可以理解的是,随着人工智能技术的高速发展,采用人工智能代替人工的方式与客户进行电话沟通,已经广泛应用在各个行业。在智能通话场景中普遍采用ASR模型。说明该模型技术已经较为成熟完善。因此,为了提高模型的生成效率,这里的语音识别子模型可以采用ASR预训练模型。即采用预先训练好的ASR模型中的参数,来构建本实施例中语音识别子模型。这样在保证模型识别准确度的同时,可以减少模型的训练次数,从而有助于提高模型的生成效率。另外,采用预训练模型也有助于降低对训练数据的数据量要求。

[0048] 步骤102,将语音识别子模型中的目标网络层输出的中间层特征,输入语气词识别模型中的语气词意图识别子模型,得到预测意图。

[0049] 在一些实施例中,在语音识别子模型对样本音频进行处理的过程中,执行主体还可以对语音识别子模型中,目标网络层输出的中间层特征进行缓存。其中,目标网络层通常为除输入层和输出层以外的网络层。并且,执行主体可以将该中间层特征,输入语气词识别模型中的语气词意图识别子模型,从而得到预测意图。其中,语气词意图识别子模型可以用于对输入数据中的语气词进行意图识别。

[0050] 在这里,目标网络层具体为哪层可以根据实际情况进行设置。作为示例,可以采用第 $K/2$ 层的输出作为中间层特征。 K 为语音识别子模型的总层数。需要说明的是,采用 $K/2$ (采用上取整算法)层能够兼顾效率和性能。如果这个值太大,网络层太多,速度会变慢。如果这个值太小,用的网络层少,虽然可以提升速度,但是因为使用的输入参数太少,无法发挥ASR模型的作用,一般会导致性能下降。

[0051] 在一些实施例中,上述语气词意图识别子模型可以采用各种机器学习模型结构。作为示例,语气词意图识别子模型可以采用自然语言处理(NLP,Natural Language Processing)模型结构。由于NLP模型(即语气词意图识别子模型)的输入为ASR模型的中间层特征,因此此时的NLP模型可以去除传统的输入层,或者减少对输入数据的编码层,而可以直接对中间层特征进行分析处理。这样可以简化语气词意图识别子模型的结构和处理过程,有利于进一步提高模型的处理效率。

[0052] 可选地,语气词意图识别子模型也可以采用包括至少一个解码层的解码网络结构。如为了兼顾性能和推理效率,语气词意图识别子模型可以采用仅1层Transformer的解码网络结构,以进行语气词的文本和意图识别。又例如,也可以采用基于注意力机制(attention)+长短期记忆网络(Long Short-Term Memory,LSTM)的网络结构。

[0053] 需要说明的是,在相关语气词识别与理解技术中,通常是将ASR模型识别得到的文本数据,输入NLP模型以识别理解用户意图。尽管NLP模型已经在文本理解的准确率上获得了巨大的提升。然而,由于在对话场景中,同一个语气词可能因为发音和上下文语境的差异,表达的客户意图存在很大的不同。例如,客户发出“嗯”这个发音,既可能表达肯定,也可能表达迟疑,还可能无特定意图的应和。由于ASR模型通常以汉字为建模单元,无法识别语调等更丰富的信息。因此,当仅以ASR模型正确识别的语气词文本作为NLP模型的输入时,通常依然难以准确地判断用户的真实意图。

[0054] 在本公开实施例中,采用中间层特征作为语气词意图识别子模型的输入数据。与传统文本输入数据相比,中间层特征中不仅包含了音频表达的文本信息,往往还包含了语调或语义等信息。这样可以提高意图预测的准确性。

[0055] 步骤103,基于预测识别结果、预测意图、训练样本数据中与样本音频对应的文本标签和意图标签,对语气词识别模型进行调整训练,直至训练完成。

[0056] 在一些实施例中,基于上述的预测识别结果、预测意图,以及训练样本数据中与样本音频对应的文本标签和意图标签,执行主体可以调整语气词识别模型中的相关参数,以继续模型训练直至训练完成。作为示例,执行主体可以根据预测识别结果与对应的文本标签的比较结果,来进行语音识别子模型的调整训练。同时,可以根据预测意图与对应的意图标签的比较结果,来进行语气词意图识别子模型进行调整训练。

[0057] 需要说明的是,在相关技术中,为了提高传统ASR和NLP级联系统对语气词的识别与理解效果,通常采用一种直接的方法就是利用大量的语气词标注数据,来分别训练智能对话系统中的ASR和NLP模型。然而,分别对ASR和NLP两个模型进行调优,训练过程较为复杂。且由于现有模型往往具有不可解释性,调优后可能出现过拟合,导致会对非语气词的识别和理解效果造成未知影响。

[0058] 在一些实施例中,若语音识别子模型为预训练模型,则在模型优化过程中可以固定该子模型。另外,上述语气词意图识别子模型还可以用于,对输入的中间层特征进行语气词识别,输出得到预测语气词。此时,可以根据预测语气词、预测意图,以及与样本音频对应的文本标签(语气词标签)和意图标签,对语气词意图识别子模型进行微调训练。这样在保证语气词识别模型的整体识别准确度的同时,可以减少语音识别子模型出现过拟合的情况。并且仅对语气词意图识别子模型微调,还可以简化模型整体的训练过程,提高模型整体的优化效率。

[0059] 具体地,执行主体可以利用交叉熵损失函数,确定损失值是否小于阈值。以及在确定损失值不小于阈值的情况下,可以利用自适应梯度(ADAM, Adaptive Moment Estimation)优化算法,来调整语气词意图识别子模型中的相关参数,以继续模型训练直至模型收敛。也就是说,在模型参数更新过程中,梯度反向传播可以采用ADAM优化算法。其中,ADAM可以采用热身策略,其学习率峰值可以设为0.001,热身策略步长可以设为1000步。

[0060] 此外,传统模型的优化往往需要大量的人工标注数据,这就会导致优化成本较高。为了解决这个问题,本公开实施例中在ASR预训练模型的基础上,采用模型网络中间层特征作为输入,构建基于注意力机制的语气词识别与意图分类任务网络。从而能够减少模型训练对标注数据量的依赖,从而降低数据标注成本。

[0061] 通过以上描述可知,本公开一些实施例中的语气词识别模型的生成方法,可以在语音识别子模型对样本音频进行识别分析时,将其目标网络层输出的中间层特征作为输入数据,从而输入到语气词意图识别子模型中,以识别样本音频中语气词所表达的意图。与传统文本输入数据相比,中间层特征中不仅包含了音频表达的文本信息,往往还包含了语调或语义等信息。这样可以提高意图预测的准确性。

[0062] 作为示例,如图2A所示的语气词识别模型。图中阴影部分即ASR预训练模型为语音识别子模型。其可分为浅层特征提取网络和深层网络。在这里,对于给定的样本音频输入数据 x_i ,一方面,经过ASR浅层网络,可以计算获得中间层特征 e_i 。然后,通过ASR深层网络,可以计算获得ASR识别结果。另一方面,以中间层特征 e_i 作为输入,通过语气词解码网络(即为语气词意图识别子模型),可以输出文本和意图的识别结果,即

[0063] $p(y_i | x_i) = \prod_u f(y_{i,u} | e_i, y_{i,1:u-1})$;

[0064] 其中, $y_{i,u}$ 为标注的 y_i 的第 u 个元素 (含开始字符 < sos > 和结束字符 < eos >); $y_{i,1:u-1}$ 为标注的 y_i 的前 $u-1$ 个元素; f 表示语气词解码网络; CI 中的 C 表示文本标签, I 表示意图标签。

[0065] 在一些实施例中, 本公开实施例的训练样本数据可以通过以下方法得到: 首先, 获取表示语气词的音频数据; 接着, 根据音频数据所表示的语气词, 进行文本标注以得到语气词文本标签; 然后, 根据音频数据所表示的意图, 进行意图类别标注以得到意图标签; 之后, 基于音频数据、对应的语气词文本标签和意图标签, 生成训练样本数据。

[0066] 可以理解的是, 本公开实施例主要针对语气词识别进行优化。因此在这里, 为了提高语气词音频数据的收集效率, 可以收集线上的 ASR 模型识别结果为语气词的音频, 对其进行人工标注, 获得文本和意图类别标注。此外, 也可以通过采样的方式, 针对漏识别的语气词音频样例, 如客户说了语气词, 但是 ASR 模型识别为非语气词, 作为补充数据。

[0067] 进一步地, 执行主体还可以对音频数据进行处理, 得到设定属性的音频频谱特征。接着, 将音频频谱特征、对应的语气词文本标签和意图标签, 确定为标注数据。之后, 按照设定比例, 将标注数据划分为训练样本数据和测试样本数据。

[0068] 通过以上方式可以得到标注数据集, 记为 $D = \{x_i, p_i | i \in [1, N]\}$ 。其中, N 为训练样本数据集中的样本数量; x_i 对应数据集中的第 i 个训练样本数据的音频特征矩阵; y_i 为音频特征 x_i 对应的文本-意图序列标注结果。上述设定属性的音频频谱特征可以根据实际情况进行设置。如可以为 80 维梅尔谱特征, 其中每一帧的时长为 25 毫秒, 步长为 10 毫秒。而设定比例同样不限制, 如可以为 9:1, 即 90% 的训练样本数据, 10% 的测试样本数据。测试样本数据可以用于对训练完成的模型进行测试。其中, 文本标注 C 可以分为“嗯”, “啊”, “哦”, “呃”, “喂”, “um”, “ang” 共 7 类语气词和 1 个空标注, 共 8 个类别。意图标注 I 可以分为“肯定”, “迟疑”, “无意图” 共 3 个类别。

[0069] 下面参考图 2B, 作为对上述图 1 所示语气词识别模型的生成方法的实现, 本公开提供了一种语气词识别模型的生成装置的一些实施例。这些生成装置实施例与图 1 所示的那些生成方法实施例相对应。该生成分析装置具体可以应用于各种电子设备中。

[0070] 如图 2B 所示, 一些实施例的语气词识别模型的生成装置 200 可以包括: 音频识别单元 201, 被配置成将训练样本数据中的样本音频, 输入语气词识别模型中的语音识别子模型, 得到样本音频的预测识别结果, 其中, 样本音频为包含有语气词的音频数据; 意图预测单元 202, 被配置成将语音识别子模型中的目标网络层输出的中间层特征, 输入语气词识别模型中的语气词意图识别子模型, 得到预测意图, 其中, 目标网络层为除输入层和输出层以外的网络层; 模型调整单元 203, 被配置成基于预测识别结果、预测意图、训练样本数据中与样本音频对应的文本标签和意图标签, 对语气词识别模型进行调整训练, 直至训练完成。

[0071] 在一些实施例中, 语音识别子模型采用语音识别 ASR 预训练模型; 以及语气词意图识别子模型还用于, 对输入的中间层特征进行语气词识别, 输出得到预测语气词。

[0072] 在一些实施例中, 模型调整单元 203 可以进一步被配置成根据预测语气词、预测意图, 以及与样本音频对应的文本标签和意图标签, 对语气词意图识别子模型进行微调训练。

[0073] 在一些实施例中, 模型调整单元 203 可以进一步被配置成利用交叉熵损失函数, 确定损失值是否小于阈值; 响应于确定损失值不小于阈值, 利用自适应梯度优化算法, 调整语气词意图识别子模型中的相关参数, 以继续模型训练。

[0074] 在一些实施例中, 语气词意图识别子模型采用包括至少一个解码层的解码网络结

构。

[0075] 在一些实施例中,该生成装置200还可以包括训练样本数据生成单元(图中未示出),被配置成获取表示语气词的音频数据;根据音频数据所表示的语气词,进行文本标注以得到语气词文本标签;根据音频数据所表示的意图,进行意图类别标注以得到意图标签;基于音频数据、对应的语气词文本标签和意图标签,生成训练样本数据。

[0076] 在一些实施例中,训练样本数据生成单元可以进一步被配置成对音频数据进行处理,得到设定属性的音频频谱特征;将音频频谱特征、对应的语气词文本标签和意图标签,确定为标注数据;按照设定比例,将标注数据划分为训练样本数据和测试样本数据。

[0077] 可以理解的是,该生成装置200中记载的诸单元与参考图1描述的生成方法中的各个步骤相对应。由此,上文针对方法描述的操作、特征以及产生的有益效果同样适用于语气词识别模型的生成装置200及其中包含的单元,在此不再赘述。

[0078] 继续参考图3,其示出了根据本公开的音频数据分析方法的一些实施例的流程300。该方法包括以下步骤:

[0079] 步骤301,将待分析的音频数据,输入预先训练完成的语气词识别模型中的语音识别子模型,预测得到音频数据的识别结果。

[0080] 在一些实施例中,音频数据分析方法的执行主体(可以与图1实施例中的执行主体相同或不同),同样可以通过有线连接方式或无线连接方式,来接收待分析的音频数据。这里的音频数据可以是上述测试样本数据中的音频数据,也可以是基于实际语音对话场景而实时得到的音频数据。如执行主体可以将接收到的对话音频,转换为音频频谱特征以作为待分析的音频数据。

[0081] 在这里,执行主体可以将待分析的音频数据,输入预先训练完成的语气词识别模型中的语音识别子模型,从而预测得到音频数据的识别结果。其中,语气词识别模型可以采用上述图1实施例中,任一实现方式所描述的语气词识别模型的生成方法得到。

[0082] 步骤302,响应于确定识别结果表征为语气词,将语音识别子模型中目标网络层输出的中间层特征,输入语气词识别模型中的语气词意图识别子模型,预测得到语气词的意图。

[0083] 在一些实施例中,根据步骤301得到音频数据的识别结果,若识别结果表征为语气词,则可以将语音识别子模型中目标网络层输出的中间层特征,输入语气词识别模型中的语气词意图识别子模型,从而预测得到语气词的意图。

[0084] 作为示例,如图4A所示,执行主体可以将音频频谱特征输入ASR模型。经过ASR浅层网络可以计算获得中间层特征。接着,经过ASR深层网络可以计算获得ASR识别结果。此时,可以判断ASR模型的输出结果是否为语气词。当且仅当ASR识别结果为语气词时,触发计算语气词解码网络,并输出语气词识别结果和意图识别结果,作为模型最终的结果。

[0085] 从上面描述可以看出,本公开实施例中的方法,模型对于非语气词的音频识别不会产生影响。同时,模型在识别为语气词时,仅利用参数量较小的语气词解码网络进行计算,很好地限制了额外推理延时的产生。

[0086] 步骤303,根据语气词的意图,对音频数据进行处理。

[0087] 在一些实施例中,根据步骤302得到的语气词意图,执行主体可以对音频数据进行处理。例如,根据语气词表示的意图,对音频的识别文本进行填补或修正,从而输出音频数

据完整真实的文本数据。

[0088] 可选地,执行主体也可以根据语气词的意图,来生成与音频数据对应的回复信息。例如在实际对话场景中,可以根据用户说出的语气词表达的意图,来确定后面的对话内容,从而生成回复信息并发送给用户。

[0089] 需要说明的是,如下表所示的实验测试,即利用测试样本数据对不同方法的性能进行对比。可以表明本公开实施例中的语气词识别模型,能够显著提升对于语气词识别的准确率,尤其是最常用语气词嗯的识别精准率。并且能够直接给出意图识别结果。

[0090]	模型	嗯精准率	语气词准确率	意图准确率
	ASR 模型	48.1%	58.2%	NA
	本公开的语气词识别模型	76.2%	72.0%	84.4%

[0091] 可以理解的是,本公开实施例中的语气词识别模型的生产方法,通过利用预训练模型,在少量人工标注数据的情况下,提高了模型对语气词的识别与理解准确率。相比于现有的技术,本发明提出的方法主要有三个优势:

[0092] (1) 本发明在预训练ASR模型浅层网络的基础上进行优化,能够利用预训练模型的特征能力提高最终的识别性能,同时降低模型优化的人工数据标注的成本。

[0093] (2) 在ASR预训练模型的基础上,插入新的语气词分类与理解模块,能够独立ASR模型进行优化,避免了对线上非语气词识别性能的影响。

[0094] (3) 能够同时输出语气词和意图识别结果,辅助下游任务进行对话流程决策。

[0095] 进一步参考图4B,作为对上述图3所示音频数据分析方法的实现,本公开提供了一种音频数据分析装置的一些实施例。这些音频数据分析装置实施例与图3所示的那些方法实施例相对应。该音频数据分析装置具体可以应用于各种电子设备中。

[0096] 如图4B所示,一些实施例的音频数据分析装置400可以包括:识别单元401,被配置成将待分析的音频数据,输入预先训练完成的语气词识别模型中的语音识别子模型,预测得到音频数据的识别结果,其中,语气词识别模型采用上述第一方面中,任一实现方式所描述的语气词识别模型的生成方法得到;意图识别单元402,被配置成响应于确定识别结果表征为语气词,将语音识别子模型中目标网络层输出的中间层特征,输入语气词识别模型中的语气词意图识别子模型,预测得到语气词的意图;处理单元403,被配置成根据语气词的意图,对音频数据进行处理。

[0097] 在一些实施例中,处理单元403可以进一步被配置成根据语气词的意图,生成与音频数据对应的回复信息。

[0098] 可以理解的是,该音频数据分析装置400中记载的诸单元与参考图3描述的方法中的各个步骤相对应。由此,上文针对方法描述的操作、特征以及产生的有益效果同样适用于音频数据分析装置400及其中包含的单元,在此不再赘述。

[0099] 下面参考图5,其示出了适于用来实现本公开的一些实施例的电子设备500的结构示意图。图5示出的电子设备仅仅是一个示例,不应对本公开的实施例的功能和使用范围带来任何限制。

[0100] 如图5所示,电子设备500可以包括处理装置501(例如中央处理器、图形处理器等),其可以根据存储在只读存储器(ROM)502中的程序或者从存储装置508加载到随机访问存储器(RAM)503中的程序而执行各种适当的动作和处理。在RAM 503中,还存储有电子设备500操作所需的各种程序和数据。处理装置501、ROM 502以及RAM 503通过总线504彼此相连。输入/输出(I/O)接口505也连接至总线504。

[0101] 通常,以下装置可以连接至I/O接口505:包括例如触摸屏、触摸板、键盘、鼠标、摄像头、麦克风、加速度计、陀螺仪等的输入装置506;包括例如扬声器、振动器等输出装置507;包括例如磁盘、硬盘等的存储装置508;以及通信装置509。通信装置509可以允许电子设备500与其他设备进行无线或有线通信以交换数据。虽然图5示出了具有各种装置的电子设备500,但是应理解的是,并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。图5中示出的每个方框可以代表一个装置,也可以根据需要代表多个装置。

[0102] 特别地,根据本公开的一些实施例,上文参考流程图描述的过程可以被实现为计算机软件程序。例如,本公开的一些实施例包括一种计算机程序产品,其包括承载在计算机可读介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的一些实施例中,该计算机程序可以通过通信装置509从网络上被下载和安装,或者从存储装置508被安装,或者从ROM 502被安装。在该计算机程序被处理装置501执行时,执行本公开的一些实施例的方法中限定的上述功能。

[0103] 需要说明的是,本公开的一些实施例中记载的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是一——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开的一些实施例中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开的一些实施例中,计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:电线、光缆、RF(射频)等等,或者上述的任意合适的组合。

[0104] 在一些实施方式中,客户端、服务器可以利用诸如HTTP(Hyper Text Transfer Protocol,超文本传输协议)之类的任何当前已知或未来研发的网络协议进行通信,并且可以与任意形式或介质的数字数据通信(例如,通信网络)互连。通信网络的示例包括局域网(“LAN”),广域网(“WAN”),网际网(例如,互联网)以及端对端网络(例如,ad hoc端对端网络),以及任何当前已知或未来研发的网络。

[0105] 上述计算机可读介质可以是上述电子设备中所包含的;也可以是单独存在,而未

装配入该电子设备中。上述计算机可读介质承载有一个或者多个程序,当上述一个或者多个程序被该电子设备执行时,使得该电子设备:将训练样本数据中的样本音频,输入语气词识别模型中的语音识别子模型,得到样本音频的预测识别结果,其中,样本音频为包含有语气词的音频数据;将语音识别子模型中的目标网络层输出的中间层特征,输入语气词识别模型中的语气词意图识别子模型,得到预测意图,其中,目标网络层为除输入层和输出层以外的网络层;基于预测识别结果、预测意图、训练样本数据中与样本音频对应的文本标签和意图标签,对语气词识别模型进行调整训练,直至训练完成。

[0106] 或者,使得该电子设备:将待分析的音频数据,输入预先训练完成的语气词识别模型中的语音识别子模型,预测得到音频数据的识别结果,其中,语气词识别模型采用上述第一方面中,任一实现方式所描述的语气词识别模型的生成方法得到;响应于确定识别结果表征为语气词,将语音识别子模型中目标网络层输出的中间层特征,输入语气词识别模型中的语气词意图识别子模型,预测得到语气词的意图;根据语气词的意图,对音频数据进行处理。

[0107] 此外,可以以一种或多种程序设计语言或其组合来编写用于执行本公开的一些实施例的操作的计算机程序代码,程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++,还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)——连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0108] 附图中的流程图和框图,图示了按照本公开各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0109] 描述于本公开的一些实施例中的单元可以通过软件的方式实现,也可以通过硬件的方式来实现。所描述的单元也可以设置在处理器中,例如,可以描述为:一种处理器包括音频识别单元、意图预测单元和模型调整单元;或识别单元、意图识别单元和处理单元。其中,这些单元的名称在某种情况下并不构成对该单元本身的限定,例如,音频识别单元还可以被描述为“得到样本音频的预测识别结果的单元”。

[0110] 本文中以上描述的功能可以至少部分地由一个或多个硬件逻辑部件来执行。例如,非限制性地,可以使用的示范类型的硬件逻辑部件包括:现场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、片上系统(SOC)、复杂可编程逻辑设备(CPLD)等等。

[0111] 本公开的一些实施例还提供一种计算机程序产品,包括计算机程序,计算机程序在被处理器执行时,实现上述的任一种语气词识别模型的生成方法,或音频数据分析方法。

[0112] 以上描述仅为本公开的一些较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解,本公开的实施例中所涉及的发明范围,并不限于上述技术特征的特定组合而成的技术方案,同时也应涵盖在不脱离上述发明构思的情况下,由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本公开的实施例中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

100

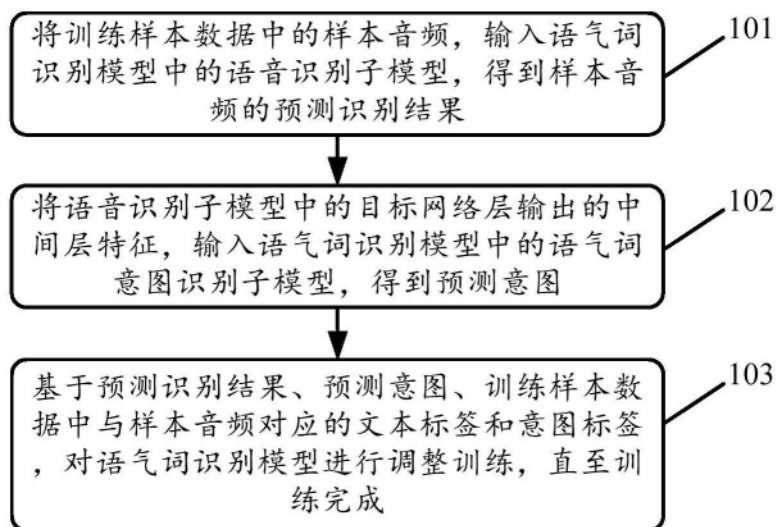


图1

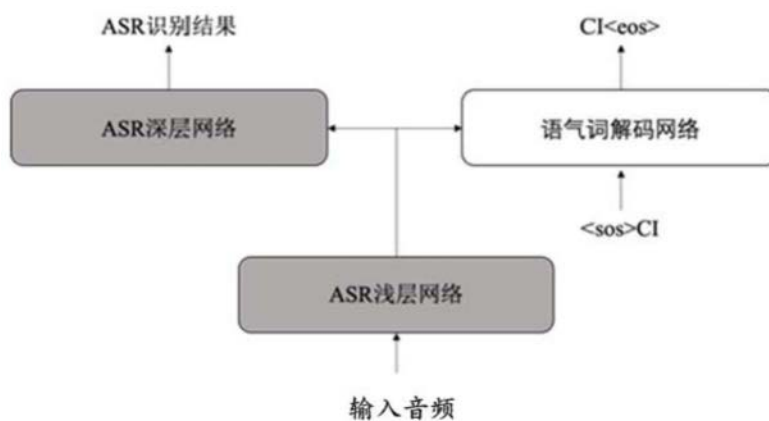


图2A

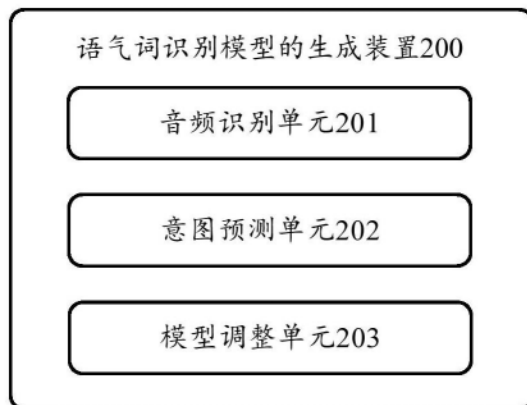


图2B

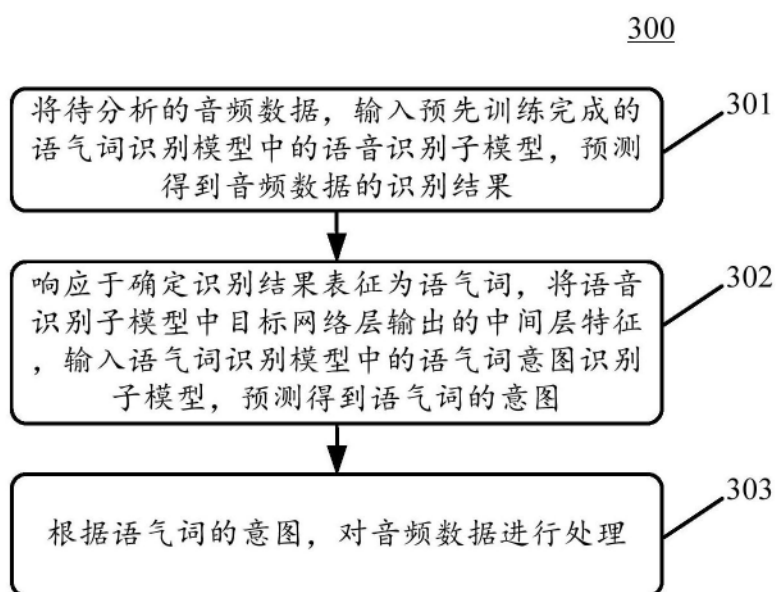


图3

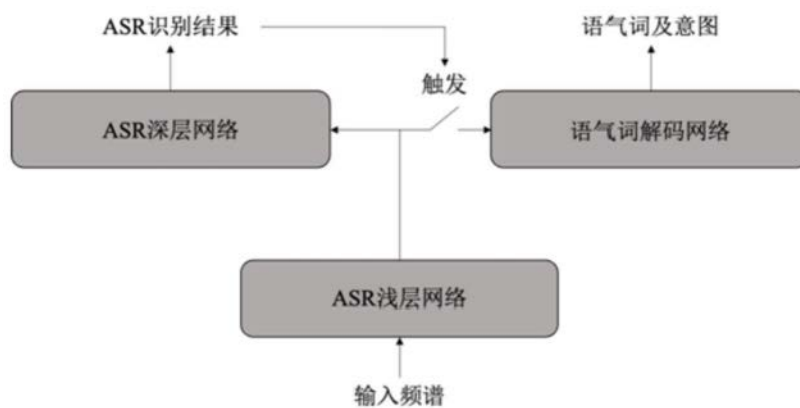


图4A



图4B

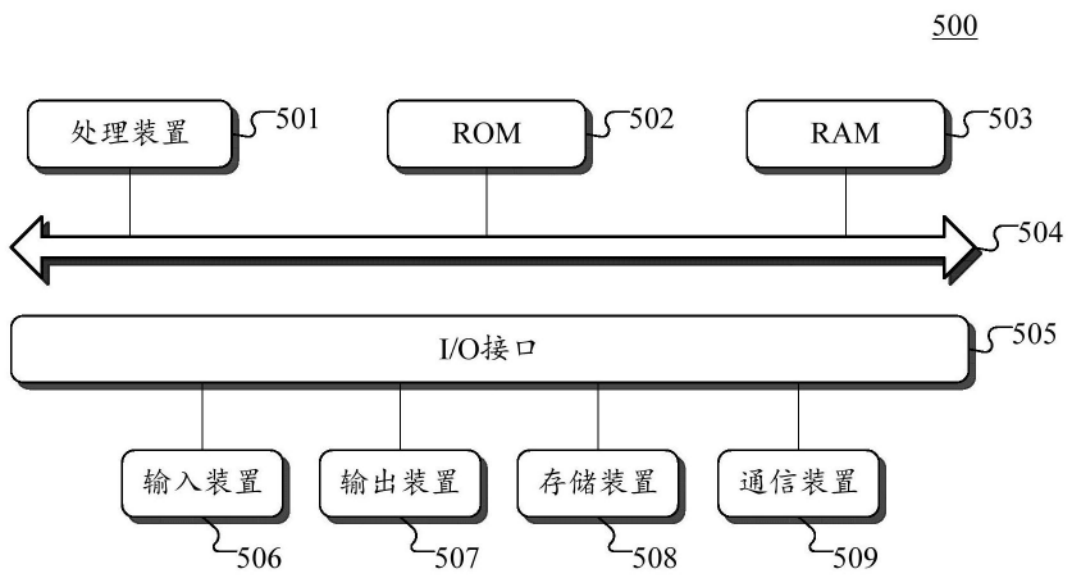


图5