(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116913278 B
(45)授权公告日 2023.11.17

(21)申请号 202311171159.5
(22)申请日 2023.09.12
(65)同一申请的已公布的文献号
申请公布号 CN 116913278 A
(43)申请公布日 2023.10.20
(73)专利权人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层
(72)发明人 汤志远 黄申 商世东
(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 杜维
(51)Int.Cl.
G10L 15/22(2006.01)
G10L 15/02(2006.01)
G10L 15/26(2006.01)
G10L 15/18(2013.01)
G10L 15/06(2013.01)
G10L 15/16(2006.01)

(56)对比文件
US 2018033454 A1,2018.02.01
CN 110032742 A,2019.07.19
CN 115171731 A,2022.10.11
CN 114267324 A,2022.04.01
CN 114360551 A,2022.04.15

审查员 孟令鹏

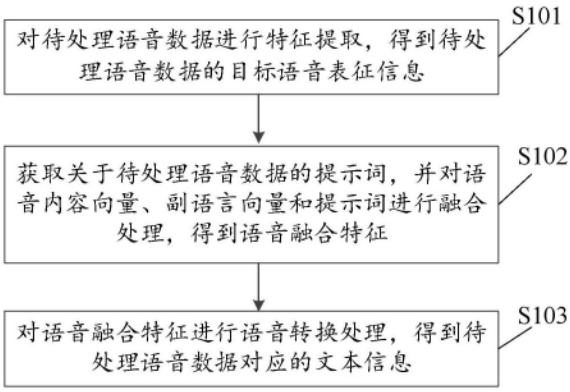
权利要求书3页 说明书20页 附图5页

(54)发明名称

语音处理方法、装置、设备和存储介质

(57)摘要

本申请实施例公开了一种语音处理方法、装置、设备和存储介质,涉及人工智能和云技术,该方法包括:对待处理语音数据进行特征提取,得到待处理语音数据的目标语音表征信息;该目标语音表征信息包括待处理语音数据对应的语音内容向量和副语言向量,该副语言向量用于辅助识别待处理语音数据对应的文本信息;获取关于待处理语音数据的提示词,并对该语音内容向量、该副语言向量和该提示词进行融合处理,得到语音融合特征;对该语音融合特征进行语音转换处理,得到待处理语音数据对应的文本信息。采用本申请实施例,可以提升语音识别的准确性。



CN 116913278 B

1.一种语音处理方法,其特征在于,所述方法包括:

对待处理语音数据进行特征提取,得到所述待处理语音数据的目标语音表征信息;所述目标语音表征信息包括所述待处理语音数据对应的语音内容向量和副语言向量,所述副语言向量用于辅助识别所述待处理语音数据对应的文本信息;

获取关于所述待处理语音数据的提示词,并对所述语音内容向量、所述副语言向量和所述提示词进行融合处理,得到语音融合特征;所述提示词是基于选择操作从显示界面输出的多个预设提示词中选择得到的,所述提示词用于反映对所述待处理语音数据的语音理解方式;

所述对所述语音内容向量、所述副语言向量和所述提示词进行融合处理,得到语音融合特征,包括:

采用特征转换参数对所述提示词进行特征转换,得到所述提示词对应的特征向量矩阵;

对所述语音内容向量、所述副语言向量和所述提示词对应的特征向量矩阵进行特征拼接,得到所述语音融合特征;

对所述语音融合特征进行语音转换处理,得到所述待处理语音数据对应的文本信息。

2.根据权利要求1所述的方法,其特征在于,所述对待处理语音数据进行特征提取,得到所述待处理语音数据的目标语音表征信息,包括:

获取用于对所述提示词进行特征转换的特征转换参数;

对所述待处理语音数据进行特征编码,得到所述待处理语音数据的语音向量矩阵;

采用所述特征转换参数对所述待处理语音数据的语音向量矩阵进行特征转换,得到所述待处理语音数据的目标语音表征信息;所述目标语音表征信息所表征的特征向量矩阵的维度与所述提示词对应的特征向量矩阵的维度相同。

3.根据权利要求2所述的方法,其特征在于,所述对所述待处理语音数据进行特征编码,得到所述待处理语音数据的语音向量矩阵,包括:

对所述待处理语音数据进行划分,得到多个音频帧;

对各个音频帧进行特征编码,得到所述各个音频帧的语音向量矩阵;

所述采用所述特征转换参数对所述待处理语音数据的语音向量矩阵进行特征转换,得到所述待处理语音数据的目标语音表征信息,包括:

采用所述特征转换参数对所述各个音频帧的语音向量矩阵进行特征转换,得到所述各个音频帧的候选语音表征信息;

遍历所述多个音频帧,基于当前遍历的音频帧的候选语音表征信息,预测所述当前遍历的音频帧映射为语音内容中的各个文字的概率;所述语音内容是指所述语音内容向量指示的内容;

若所述当前遍历的音频帧映射为语音内容中的各个文字的概率中的最大概率小于概率阈值,则从所述各个音频帧的候选语音表征信息中删除所述当前遍历的音频帧的候选语音表征信息;

在遍历结束后,基于剩余的候选语音表征信息,得到所述待处理语音数据的目标语音表征信息。

4.根据权利要求3所述的方法,其特征在于,所述方法还包括:

基于所述多个音频帧的划分顺序,确定各个音频帧的位置特征,所述位置特征用于指示相应音频帧在所述待处理语音数据中的位置;

所述对各个音频帧进行特征编码,得到所述各个音频帧的语音向量矩阵,包括:

对各个音频帧进行特征编码,得到所述各个音频帧的编码特征;

针对任一音频帧,对所述任一音频帧的位置特征和所述任一音频帧的编码特征进行特征拼接处理,得到所述任一音频帧的语音向量矩阵。

5.根据权利要求1所述的方法,其特征在于,所述待处理语音数据对应的文本信息是通过训练后的语音转换模型得到的,所述训练后的语音转换模型的训练方式包括:

获取样本语音数据对应的样本语音表征信息和样本提示词;所述样本语音表征信息包括所述样本语音数据对应的样本语音内容向量和样本副语言向量;

采用语音转换模型对所述样本语音内容向量、所述样本副语言向量和所述样本提示词进行融合处理,得到样本语音融合特征;

采用所述语音转换模型对所述样本语音融合特征进行语音转换处理,得到所述样本语音数据对应的文本信息;

获取所述样本语音数据对应的样本文本标签,基于所述样本文本标签和所述样本语音数据对应的文本信息训练所述语音转换模型,得到所述训练后的语音转换模型。

6.根据权利要求1所述的方法,其特征在于,所述待处理语音数据的目标语音表征信息是通过训练后的语音特征提取模型得到的,所述训练后的语音特征提取模型的训练方式包括:

获取样本语音数据,采用语音特征提取模型对所述样本语音数据进行特征提取,得到所述样本语音数据的样本语音表征信息;

获取所述样本语音数据的样本语音表征标签,基于所述样本语音表征标签和所述样本语音表征信息训练所述语音特征提取模型,得到所述训练后的语音特征提取模型。

7.根据权利要求6所述的方法,其特征在于,所述语音特征提取模型包括语音向量矩阵提取层和语音表征全连接层;

所述采用所述语音特征提取模型对所述样本语音数据进行特征提取,得到所述样本语音数据的样本语音表征信息,包括:

通过所述语音向量矩阵提取层对所述样本语音数据进行特征编码,得到所述样本语音数据的语音向量矩阵;

通过所述语音表征全连接层中的特征转换参数对所述样本语音数据的语音向量矩阵进行特征转换,得到所述样本语音数据的样本语音表征信息;

所述基于所述样本语音表征标签和所述样本语音表征信息训练所述语音特征提取模型,得到所述训练后的语音特征提取模型,包括:

基于所述样本语音表征标签和所述样本语音表征信息调整所述语音向量矩阵提取层的参数,得到所述训练后的语音特征提取模型。

8.一种语音处理装置,其特征在于,所述装置包括:

特征提取单元,用于对待处理语音数据进行特征提取,得到所述待处理语音数据的目标语音表征信息;所述目标语音表征信息包括所述待处理语音数据对应的语音内容向量和副语言向量,所述副语言向量用于辅助识别所述待处理语音数据对应的文本信息;

信息融合单元,用于获取关于所述待处理语音数据的提示词,并对所述语音内容向量、所述副语言向量和所述提示词进行融合处理,得到语音融合特征;所述提示词是基于选择操作从显示界面输出的多个预设提示词中选择得到的,所述提示词用于反映对所述待处理语音数据的语音理解方式;

所述信息融合单元,具体用于采用特征转换参数对所述提示词进行特征转换,得到所述提示词对应的特征向量矩阵;对所述语音内容向量、所述副语言向量和所述提示词对应的特征向量矩阵进行特征拼接,得到所述语音融合特征;

语音转换单元,用于对所述语音融合特征进行语音转换处理,得到所述待处理语音数据对应的文本信息。

9.一种计算机设备,其特征在于,包括处理器、存储器以及网络接口,其中,所述处理器与所述存储器、所述网络接口相连,其中,所述网络接口用于提供数据通信功能,所述存储器用于存储计算机程序,所述计算机程序包括程序指令,所述处理器被配置用于调用程序指令,以使得所述计算机设备执行权利要求1-7任一项所述的方法。

10.一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存储有计算机程序,所述计算机程序适于由处理器加载并执行,以使得具有所述处理器的计算机设备执行权利要求1-7任一项所述的方法。

语音处理方法、装置、设备和存储介质

技术领域

[0001] 本申请涉及人工智能技术领域,尤其涉及一种语音处理方法、装置、设备和存储介质。

背景技术

[0002] 语音识别技术应用在多种场景中,例如在智能对话场景中,通过对对话者的语音数据进行语音识别和理解,可以知道对话者想要表达的含义,从而选择对应的回复数据进行准确回复。然而对话者的语音数据中除了文字内容以外一般还包含有用于辅助识别对话者的语音数据的辅助信息,但是目前的语音识别和理解技术中只是将对话者的语音数据转换为文本信息,对于语音数据中的辅助信息无法体现在文本信息中,导致语音识别的准确性较低。

发明内容

[0003] 本申请实施例提供一种语音处理方法、装置、设备和存储介质,可以提升语音识别的准确性。

[0004] 第一方面,本申请提供一种语音处理方法,包括:

[0005] 对待处理语音数据进行特征提取,得到该待处理语音数据的目标语音表征信息;该目标语音表征信息包括该待处理语音数据对应的语音内容向量和副语言向量,该副语言向量用于辅助识别该待处理语音数据对应的文本信息;

[0006] 获取关于该待处理语音数据的提示词,并对该语音内容向量、该副语言向量和该提示词进行融合处理,得到语音融合特征;

[0007] 对该语音融合特征进行语音转换处理,得到该待处理语音数据对应的文本信息。

[0008] 第二方面,本申请提供一种语音处理装置,包括:

[0009] 特征提取单元,用于对待处理语音数据进行特征提取,得到该待处理语音数据的目标语音表征信息;该目标语音表征信息包括该待处理语音数据对应的语音内容向量和副语言向量,该副语言向量用于辅助识别该待处理语音数据对应的文本信息;

[0010] 信息融合单元,用于获取关于该待处理语音数据的提示词,并对该语音内容向量、该副语言向量和该提示词进行融合处理,得到语音融合特征;

[0011] 语音转换单元,用于对该语音融合特征进行语音转换处理,得到该待处理语音数据对应的文本信息。

[0012] 第三方面,本申请提供了一种计算机设备,包括处理器、存储器、网络接口,其中,上述处理器与存储器、网络接口相连,其中,网络接口用于提供数据通信功能,上述存储器用于存储计算机程序,上述计算机程序包括程序指令,上述处理器被配置用于调用程序指令,以使包含该处理器的计算机设备执行上述语音处理方法。

[0013] 第四方面,本申请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,该计算机程序适于由处理器加载并执行,以使得具有该处理器的计算机

设备执行上述语音处理方法。

[0014] 第五方面,本申请提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令被处理器执行时可实现上述语音处理方法。

[0015] 本申请实施例中,对待处理语音数据进行特征提取,得到待处理语音数据的目标语音表征信息。获取关于待处理语音数据的提示词,并对目标语音表征信息和提示词进行融合处理,得到语音融合特征;对语音融合特征进行语音转换处理,得到待处理语音数据对应的文本信息。由于目标语音表征信息包括待处理语音数据对应的语音内容向量和副语言向量,而副语言向量用于辅助识别待处理语音数据对应的文本信息。因此在对待处理语音数据进行语音识别时,既可以结合待处理语音数据的语音内容方面的信息,又可以结合待处理语音数据中的副语言方面的信息,还可以结合提示词对应的文本内容进行语音识别,可以实现对待处理语音数据进行深层次的语音识别和理解,从而提升语音识别的准确性。

附图说明

[0016] 为了更清楚地说明本申请实施例中的技术方案,下面将对实施例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申请的一些实施例,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,还可以根据这些附图获得其他的附图。

[0017] 图1是本申请实施例提供的一种语音处理系统的网络架构示意图;

[0018] 图2是本申请实施例提供的一种语音处理方法的应用场景示意图;

[0019] 图3是本申请实施例提供的一种语音处理方法的流程示意图;

[0020] 图4是本申请实施例提供的一种语音特征提取模型的架构示意图;

[0021] 图5是本申请实施例提供的一种语音特征提取模型训练的方法流程示意图;

[0022] 图6是本申请实施例提供的一种语音转换模型训练的方法流程示意图;

[0023] 图7是本申请实施例提供的一种语音转换模型中的参数调整示意图;

[0024] 图8是本申请实施例提供的一种语音处理装置的组成结构示意图;

[0025] 图9是本申请实施例提供的一种计算机设备的组成结构示意图。

具体实施方式

[0026] 下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅是本申请一部分实施例,而不是全部的实施例。基于本申请中的实施例,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,都属于本申请保护的范围。

[0027] 人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说,人工智能是计算机科学的一个综合技术,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。

[0028] 人工智能技术是一门综合学科,涉及领域广泛,既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、

大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。本申请实施例提供的方案属于人工智能领域下属的自然语言处理技术和机器学习技术。

[0029] 自然语言处理(Nature Language processing,NLP)是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此,这一领域的研究将涉及自然语言,即人们日常使用的语言,所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。例如,本申请中可以采用自然语言处理技术中的语义理解技术对语音融合特征进行语音转换处理,得到待处理语音数据对应的文本信息,等等。

[0030] 机器学习(Machine Learning,ML)是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。例如,本申请中可以采用机器学习技术中的人工神经网络对待处理语音数据进行特征提取,得到待处理语音数据的目标语音表征信息,以及对语音内容向量、副语言向量和提示词进行融合处理,得到语音融合特征,等等。

[0031] 云技术(Cloud technology)是指在广域网或局域网内将硬件、软件、网络等系列资源统一起来,实现数据的计算、储存、处理和共享的一种托管技术。云技术基于云计算商业模式应用的网络技术、信息技术、整合技术、管理平台技术、应用技术等的总称,可以组成资源池,按需所用,灵活便利。云计算技术将变成重要支撑。技术网络系统的后台服务需要大量的计算、存储资源,如视频网站、图片类网站和更多的门户网站。伴随着互联网行业的高度发展和应用,将来每个物品都有可能存在自己的识别标志,都需要传输到后台系统进行逻辑处理,不同程度级别的数据将会分开处理,各类行业数据皆需要强大的系统后盾支撑,只能通过云计算来实现。本申请实施例提供的方案属于云技术领域下属的云会议。

[0032] 云会议是基于云计算技术的一种高效、便捷、低成本的会议形式。使用者只需要通过互联网界面,进行简单易用的操作,便可快速高效地与全球各地团队及客户同步分享语音、数据文件及视频,而会议中数据的传输、处理等复杂技术由云会议服务商帮助使用者进行操作。目前国内云会议主要集中在以SaaS (Software as a Service,软件即服务)模式为主体的服务内容,包括电话、网络、视频等服务形式,基于云计算的视频会议就叫云会议。在云会议时代,数据的传输、处理、存储全部由视频会议厂家的计算机资源处理,用户完全无需再购置昂贵的硬件和安装繁琐的软件,只需打开浏览器,登录相应界面,就能进行高效的远程会议。云会议系统支持多服务器动态集群部署,并提供多台高性能服务器,大大提升了会议稳定性、安全性、可用性。近年来,视频会议因能大幅提高沟通效率,持续降低沟通成本,带来内部管理水平升级,而获得众多用户欢迎,已广泛应用在交通、运输、金融、运营商、教育、企业等各个领域。毫无疑问,视频会议运用云计算以后,在方便性、快捷性、易用性上具有更强的吸引力,必将激发视频会议应用新高潮的到来。例如,本申请中可以采用云会议技术获取会议过程中产生的待处理语音数据,从而可以对会议过程中产生的待处理语音数

据进行语音识别得到文本信息,等等。

[0033] 本申请技术方案可以适用于对语音数据进行语音识别转换为文本信息的场景中。例如可以应用于在线会议中的语音转录、社交应用程序中的语音输入、采访场景中的录音设备中的语音转文字、以及智能对话中的语音对话等场景中。例如针对于在线会议场景中,通过对会议中的语音进行录制得到语音数据,并对语音数据进行语音转录得到对应的会议纪要,可以提升会议重点内容获取的效率。又例如在社交应用程序中,通过获取使用者的语音数据进行语音识别得到文本信息,可以提升文本信息输入的效率。又例如在采访场景中,对录音设备录制的语音数据进行识别转换为文字,可以提升采访文本获取效率。再例如在智能对话场景中,通过对对话者的语音数据进行语音识别得到文本信息,可以实现准确文本对话。可选地,本申请技术方案还可应用于各种场景,包括但不限于云技术、人工智能、智慧交通、辅助驾驶等。

[0034] 需要特别说明的是,本申请实施例中涉及到对象信息相关的数据(例如待处理语音数据、样本语音数据、提示词,等等),当本申请实施例运用到具体产品或技术中时,需要获得对象许可或者同意,且相关数据的收集、使用和处理需要遵守相关地区的相关法律法规和标准。例如,对象可以是指终端设备或者计算机设备的使用者。

[0035] 请参见图1,图1是本申请实施例提供的一种语音处理系统的网络架构示意图,如图1所示,计算机设备可以与终端设备进行数据交互,终端设备的数量可以为一个或者至少两个。例如,当终端设备的数量为多个时,终端设备可以包括图1中的终端设备101a、终端设备101b及终端设备101c等。其中,以终端设备101a为例,计算机设备102可以对待处理语音数据进行特征提取,得到待处理语音数据的目标语音表征信息。进一步地,计算机设备102可以获取关于待处理语音数据的提示词,并对语音内容向量、副语言向量和提示词进行融合处理,得到语音融合特征。进一步地,计算机设备102还可以对语音融合特征进行语音转换处理,得到待处理语音数据对应的文本信息。可选地,计算机设备102可以向终端设备101a发送待处理语音数据对应的文本信息,以在终端设备101a上展示文本数据,或者计算机设备102还可以基于待处理语音数据对应的文本信息确定回复文本信息,并向终端设备101a发送回复文本信息,等等。

[0036] 可以理解的是,本申请实施例中所提及的计算机设备包括但不限于终端设备或服务器。换句话说,计算机设备可以是服务器或终端设备,也可以是服务器和终端设备组成的系统。其中,以上所提及的终端设备可以是一种电子设备,包括但不限于手机、平板电脑、台式电脑、笔记本电脑、掌上电脑、车载设备、智能语音交互设备、增强现实/虚拟现实(Augmented Reality/Virtual Reality,AR/VR)设备、头盔显示器、可穿戴设备、智能音箱、智能家电、飞行器、数码相机、摄像头及其他具备网络接入能力的移动互联网设备(mobile internet device,MID)等。其中,以上所提及的服务器可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、车路协同、内容分发网络(Content Delivery Network,CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0037] 请参见图2,图2是本申请实施例提供的一种语音处理方法的应用场景示意图;如图2所示,可以将待处理语音数据21输入语音特征提取模型22,采用语音特征提取模型22对

待处理语音数据21进行特征编码,得到待处理语音数据21的语音向量矩阵。例如待处理语音数据21为“嗯,对,嗯,我赞同,嗯”。通过语音特征提取模型22中语音表征全连接层对待处理语音数据21的语音向量矩阵进行特征转换,得到待处理语音数据21的目标语音表征信息23。其中,目标语音表征信息包括待处理语音数据21对应的语音内容向量和副语言向量。进一步地,可以通过获取关于待处理语音数据21的提示词24(如提示词为去口语化提示词),将目标语音表征信息23(即语音内容向量和副语言向量)和提示词24输入语音转换模型25,通过语音转换模型25对提示词24进行特征处理如特征编码后,得到提示词24对应的特征向量矩阵26,并通过语音转换模型25对提示词对应的特征向量矩阵26和待处理语音数据21的目标语音表征信息23进行融合处理,得到语音融合特征,则语音转换模型25可以将语音融合特征转换为待处理语音数据21对应的文本信息27,最终输出文本信息27,例如去口语化的文本信息27为“对,我赞同”。

[0038] 进一步地,请参见图3,图3是本申请实施例提供的一种语音处理方法的流程示意图;如图3所示,该语音处理方法可以应用于计算机设备,该语音处理方法包括但不限于以下步骤:

[0039] S101,对待处理语音数据进行特征提取,得到待处理语音数据的目标语音表征信息。

[0040] 在一些语音处理任务中,在进行语音识别和语音理解时,是对语音数据进行语音识别得到文本数据,对文本数据进行文本处理以实现语音理解。由于语音理解是在语音识别得到的文本数据的基础上进行的,而文本数据中只包括文字内容,因此在进行语音理解时只能根据文字内容进行理解,无法使用语音数据中包含的副语言信息,副语言信息例如可以是指反映情绪、性别、语调等隐藏在语音数据中的隐藏信息。因此在复杂场景下,语音识别效果较差时,将语音识别结果作为语音理解的前提条件,会造成语音理解的误差累积,无法进行错误纠正,从而导致语音识别的准确性较低。

[0041] 鉴于此,本申请实施例中提供的语音处理方法并非直接将语音数据识别为文本数据,对文本数据进行理解,而是提取语音数据中包含语音内容向量和副语言向量的语音表征信息,通过对语音内容向量、副语言向量和提示词进行进一步处理,得到最终的文本信息。由于语音数据中包含副语言信息,因此结合副语言信息对语音数据进行语音理解,可以提升语音理解的准确性。

[0042] 本申请实施例中,可以获取待处理语音数据,对待处理语音数据进行特征提取,得到待处理语音数据的目标语音表征信息。在获取待处理语音数据时,例如可以获取关联的录音装置录制得到的语音数据、或者获取关联的录像装置录制得到的视频数据并提取视频数据中的语音数据、或者获取本地存储的语音数据、或者获取终端设备上传的语音数据,等等。

[0043] 其中,目标语音表征信息包括待处理语音数据对应的语音内容向量和副语言向量,语音内容向量可以反映待处理数据中的语音内容。副语言向量用于辅助识别待处理语音数据对应的文本信息,副语言向量可以反映待处理数据中的副语言信息。例如,语音内容可以是指待处理语音数据中文字对应的语音,例如待处理语音数据为“你吃饭了吗”,则语音内容是指对“你”、“吃”、“饭”、“了”、“吗”这些文字进行发音得到的语音。副语言信息例如可以包括待处理语音数据中每个字发音的音量、音色、快慢等信息,或者还可以包括发音时

对应的情绪信息、以及说话者的性别、年龄等信息。这些副语言信息可以用于辅助识别待处理语音数据。

[0044] 本申请实施例中,待处理语音数据中的副语言向量和语音内容向量都是语音模态的信息对应的向量,而副语言信息和语音内容这两种语音模态的信息是难以分离开的,而将待处理语音数据直接转换为文本数据只能实现将语音内容转换为文本数据,而无法将副语言信息转换为文本数据。并且由于转换得到的文本数据中只包括文字,对于说话者对文字发音时所具备的情绪、发音者的性别、年龄等信息、以及发音者的音量等信息无法从文本数据中体现出来。而语音数据中的副语言信息是通过语音的方式体现出说话者的这些信息,因此通过语音内容和副语言信息可以完整表达出说话者的说话内容以及说话者的情绪、性别、年龄等信息。

[0045] 在一个实施例中,可以通过以下方式得到待处理语音数据的目标语音表征信息:获取用于对提示词进行特征转换的特征转换参数;对待处理语音数据进行特征编码,得到待处理语音数据的语音向量矩阵;采用特征转换参数对待处理语音数据的语音向量矩阵进行特征转换,得到待处理语音数据的目标语音表征信息;目标语音表征信息所表征的特征向量矩阵的维度与提示词对应的特征向量矩阵的维度相同。

[0046] 其中,特征转换参数可以用于对提示词进行特征转换,得到提示词对应的特征向量矩阵。通过对待处理语音数据进行特征编码,可以将待处理语音数据中的每个文字编码为语音向量。对待处理语音数据进行特征编码即是指将语音信号嵌入编码到固定维度的向量空间,得到语音向量。由于待处理语音数据中包括多个文字,因此可以编码成多个文字对应的语音向量,得到语音向量矩阵,从而得到待处理语音数据对应的语音向量矩阵。例如一个文字进行特征编码得到语音向量的维度为m维,包含n个文字的待处理语音数据进行特征编码得到m*n维的矩阵,即语音向量矩阵可以为m*n维的矩阵。采用特征转换参数对待处理语音数据的语音向量矩阵进行特征转换的目的是为了使得待处理语音数据的语音向量矩阵的维度与提示词对应的特征向量矩阵的维度相同,从而后续可以对待处理语音数据的语音向量矩阵和提示词对应的特征向量矩阵进行拼接。若两个向量矩阵的维度不相同,拼接后两个向量矩阵就难以起到融合效果。由于目标语音表征信息所表征的特征向量矩阵的维度与提示词对应的特征向量矩阵的维度相同,且目标语音表征信息包括语音内容向量和副语言向量,因此对目标语音表征信息与提示词对应的特征向量矩阵进行融合处理的实质上就是对语音内容向量、副语言向量和提示词进行融合处理。通过使用特征转换参数对待处理语音数据的语音向量矩阵进行特征转换,得到待处理语音数据的目标语音表征信息,便于后续更好地融合语音表征信息和提示词对应的文本信息,提升语音识别的准确性。

[0047] 在一个实施例中,对待处理语音数据进行特征编码,得到待处理语音数据的语音向量矩阵的方式可以包括:对待处理语音数据进行划分,得到多个音频帧;对各个音频帧进行特征编码,得到各个音频帧的语音向量矩阵。

[0048] 例如,在获取到待处理语音数据时,可以对待处理语音数据进行分帧处理,得到多个音频帧,分帧处理可以是指按照帧长对待处理语音数据进行划分,得到多个音频帧。例如帧长可以取任意数值,例如可以取10毫秒~30毫秒。一般来说,待处理语音信号的语音内容中一个文字可以对应多个音频帧,例如以语音内容中一个文字对应的语音信号为1秒,帧长为30毫秒进行举例,则该文字对应的音频帧个数约为33个。通过对待处理语音数据进行划

分得到多个音频帧,可以对各个音频帧进行特征编码,得到各个音频帧的语音向量矩阵。这里的特征编码是将音频帧由语音数据转换为特征向量。通过将语音数据转换为语音向量矩阵,便于后续进行计算。

[0049] 在一个实施例中,在对待处理语音数据的语音向量矩阵进行特征转换,得到待处理语音数据的目标语音表征信息时,还可以对待处理语音数据中的无意义的音频帧进行去除,从而减少计算量。具体地,可以采用特征转换参数对各个音频帧的语音向量矩阵进行特征转换,得到各个音频帧的候选语音表征信息;遍历多个音频帧,基于当前遍历的音频帧的候选语音表征信息,预测当前遍历的音频帧映射为语音内容中的各个文字的概率;语音内容是指语音内容向量指示的内容。若当前遍历的音频帧映射为语音内容中的各个文字的概率中的最大概率小于概率阈值,则从各个音频帧的候选语音表征信息中删除当前遍历的音频帧的候选语音表征信息。在遍历结束后,基于剩余的候选语音表征信息,得到待处理语音数据的目标语音表征信息。

[0050] 其中,采用特征转换参数对各个音频帧的语音向量矩阵进行特征转换的目的是:将各个音频帧的语音向量矩阵的维度转换为与提示词对应的特征向量矩阵的维度相同,因此各个音频帧的候选语音表征信息的维度与提示词对应的特征向量矩阵的维度相同。通过遍历待处理语音数据的多个音频帧,可以基于多个音频帧的候选语音表征信息预测多个音频帧映射为待处理语音数据的语音内容中的各个文字的概率。每个音频帧映射为待处理语音数据的语音内容中的各个文字的概率可以用于反映每个音频帧映射为待处理语音数据的语音内容中的各个文字的可能性。即概率越大表示该音频帧对应的文字为待处理语音数据的语音内容中的各个文字的可能性越大,概率越小表示该音频帧为待处理语音数据的语音内容中的各个文字的可能性越小。音频帧对应的文字可以是指对该文字进行发音得到的音频帧。若音频帧映射为待处理语音数据的语音内容中的各个文字的概率均小于概率阈值,则表示该音频帧为无意义的音频帧,即该音频帧中包含的语音信息较少,则为了加快计算效率可以省去该无意义的音频帧。

[0051] 本申请实施例中,若当前遍历的音频帧映射为语音内容中的某个文字的概率最大,并且大于概率阈值,则表示当前遍历的音频帧可以映射为语音内容中的该文字。若当前遍历的音频帧映射为语音内容中的各个文字的概率中的最大概率小于概率阈值,则表示当前遍历的音频帧无法映射为语音内容中的各个文字,即该音频帧中的语音信息较少,则可以从多个音频帧的候选语音表征信息中删除当前遍历的音频帧的候选语音表征信息。通过从多个音频帧中删除遍历的映射为待处理语音数据的语音内容中的各个文字的概率均小于概率阈值的音频帧的候选语音表征信息,在遍历结束后,就可以基于剩余的候选语音表征信息,得到待处理语音数据的目标语音表征信息。例如可以对剩余的候选语音表征信息进行拼接或者组合,得到待处理语音数据的目标语音表征信息。

[0052] 本申请实施例中,概率例如可以是指通过语音特征提取模型输出的后验概率,当音频帧对应的概率小于概率阈值,其对应的音频帧可以省去。通过将待处理语音数据划分为多个音频帧,每个文字对应的语音数据包含多个音频帧,因此一个文字对应的多个音频帧中存在语音信息较少的音频帧,而对这些音频帧中存在语音信息较少的音频帧进行删除对于整个语音识别结果产生的影响很小,因此通过删除待处理语音数据中包含信息较少的音频帧,可以减少计算量,提升计算效率。

[0053] 可选地实现方式中,可以结合各个音频帧的位置特征进行特征编码,得到音频帧的语音向量矩阵。例如,可以基于多个音频帧的划分顺序,确定各个音频帧的位置特征,位置特征用于指示相应音频帧在待处理语音数据中的位置;对各个音频帧进行特征编码,得到各个音频帧的编码特征;针对任一音频帧,对任一音频帧的位置特征和任一音频帧的编码特征进行特征拼接处理,得到任一音频帧的语音向量矩阵。

[0054] 其中,各个音频帧的位置特征可以用于指示各个音频帧在待处理语音数据中的位置。由于对待处理语音数据进行划分时,一般是按照待处理语音数据中发音的先后顺序进行划分,则对于待处理语音数据中发音靠前的语音数据对应的音频帧的划分顺序靠前,对于待处理语音数据中发音靠后的语音数据对应的音频帧的划分顺序靠后,因此可以基于多个音频帧的划分顺序,确定各个音频帧的位置特征。进而在对多个音频帧的编码特征进行拼接处理时,可以结合每个音频帧的位置特征对每个音频帧进行特征拼接,得到多个音频帧的语音向量矩阵。通过在特征编码时引入位置特征,后续在对音频帧的编码特征进行拼接处理时,可以结合音频帧的位置特征进行拼接,从而可以确保文本信息中文字顺序的准确性,使得语音识别结果更准确。

[0055] 可选地,可以采用训练后的语音特征提取模型对待处理语音数据进行特征提取,得到待处理语音数据的目标语音表征信息。通过采用训练后的语音特征提取模型进行特征提取,可以提升特征提取的效率。其中,语音特征提取模型例如可以包括但不限于自动语音识别模型(Automatic Speech Recognition,ASR)、基于自注意力机制的Transformer模型、卷积增强的Transformer模型、基于神经网络的时序类分类模型(Connectionist temporal classification,CTC),等等。

[0056] 可选地,语音特征提取模型中可以包括但不限于语音向量矩阵提取层和语音表征全连接层,语音向量矩阵提取层可以用于提取待处理语音数据的语音向量矩阵,语音表征全连接层可以用于将待处理语音数据的语音向量矩阵转换为待处理语音数据的目标语音表征信息。

[0057] 例如,请参见图4,图4是本申请实施例提供的一种语音特征提取模型的架构示意图,其中,语音特征提取模型可以包括语音向量矩阵提取层和语音表征全连接层,语音向量矩阵提取层可以用于输出待处理语音数据的语音向量矩阵,语音表征全连接层可以用于输出待处理语音数据的目标语音表征信息。进一步可选地,语音向量矩阵提取层中还可以包括编码层、多头注意力层、归一化层。具体地,待处理语音数据输入到语音特征提取模型中,由语音特征提取模型中的语音向量矩阵提取层对待处理语音数据进行处理,例如可以通过语音向量矩阵提取层中的编码层对待处理语音数据中的各个音频帧进行编码处理,得到编码特征。通过编码层获取待处理语音数据中各个音频帧的位置特征,对各个音频帧的编码特征和位置特征进行拼接,得到各个音频帧的拼接编码特征,拼接编码特征为带有位置信息的语音向量。通过多头注意力层计算各个音频帧的拼接编码特征之间的相似度,确定每两个音频帧之间的相似度分数,通过归一化层将相似度分数归一化到0~1范围内。对于每两个音频帧而言,两个音频帧之间的相似度分数越高,则两个音频帧之间的权重越大,两个音频帧之间的相似度分数越低,则两个音频帧之间的权重越小。通过结合音频帧之间的权重对每个音频帧与其他音频帧进行加权求和,得到的音频帧中包含了自身音频帧的语音信息,还包括其他音频帧的语音信息,即引入了待处理语音数据中上下文语音信息,进而使得

每个音频帧都包含整个待处理语音数据的信息。通过归一化层输出与拼接编码特征维度相同的矩阵,即待处理语音数据的语音向量矩阵。进一步地,可以预先固定语音表征全连接层中的特征转换参数,因此通过语音表征全连接层中的特征转换参数对待处理语音数据的语音向量矩阵进行转换,可以输出待处理语音数据的目标语音表征信息。

[0058] 可选地,语音特征提取模型还可以包括文本输出层,通过将待处理语音数据中每个音频帧的目标语音表征信息输入文本输出层,文本输出层可以预测待处理语音数据中每个音频帧的目标语音表征信息映射为待处理语音数据的语音内容中的各个文字的概率,即预测每个音频帧的目标语音表征信息为多个文字的概率,从而确定待处理语音数据对应的文本数据,并输出文本数据,例如你吃饭了吗。

[0059] 举例来说,例如语音特征提取模型为ASR模型时,可以将CTC全连接层的前一个全连接层确定为语音表征全连接层,例如可以获取特征转换参数,使用特征转换参数替换该全连接层中的原有参数,将替换参数后的全连接层称之为语音表征全连接层。语音表征全连接层的作用是为了与用于对提示词进行转换处理的特征转换参数进行参数共享,以便于对齐语音和文本的隐空间表征分布,从而实现后续语音与文本两种不同模态之间的信息融合。

[0060] S102,获取关于待处理语音数据的提示词,并对语音内容向量、副语言向量和提示词进行融合处理,得到语音融合特征。

[0061] 本申请实施例中,通过获取关于待处理语音数据的提示词,并对语音内容向量、副语言向量和提示词进行融合处理,得到语音融合特征,语音融合特征既包含待处理语音数据的提示词又包含语音内容和副语言信息,因此后续对语音融合特征进行处理得到的文本信息中既可以反映出待处理语音数据的语音内容,又可以反映出待处理语音数据中的副语言信息,还可以反映出提示词对应的文本内容,从而可以实现对语音数据进行深层次的语音理解,可以提升语音识别准确性。由于目标语音表征信息包括语音内容向量和副语言向量,因此对语音内容向量、副语言向量和提示词进行融合处理实质上就是对目标语音表征信息和提示词进行融合处理。

[0062] 在一个实施例中,可以通过以下方式获取关于待处理语音数据的提示词:基于显示界面输出多个预设提示词,响应于针对显示界面的选择操作,选择操作包括提示词;基于选择操作从显示界面输出的多个预设提示词中选择关于待处理语音数据的提示词。其中,提示词可以用于反映对待处理语音数据的语音理解方式。提示词可以包括但不限于重识别提示词、情绪识别提示词、去口语化提示词、性别判断提示词、加标点符号提示词、文本顺滑提示词、纠错提示词,等等。其中,情绪类别可以包括但不限于欢喜、悲哀、害怕、愤怒、意外和厌恶等类别。性别可以包括男、女。纠错提示词可以包括各种领域的专业术语。通过选择关于待处理语音数据的提示词,后续可以结合提示词输出对应的文本信息。

[0063] 举例来说,例如选择的提示词为重识别提示词,则可以对待处理语音数据进行重识别,输出对待处理语音数据重识别后的文本信息。或者选择的提示词为情绪识别提示词,输出的文本信息可以包括待处理语音数据对应的情绪类别,即说话者说话时的情绪类别,也可以同时输出待处理语音数据对应的文本信息。或者选择的提示词为去口语化提示词,则输出的文本信息可以为对待处理语音数据去口语化后的文本信息。或者选择的提示词为性别判断提示词,则输出的文本信息可以包括说话者的性别,也可以同时输出待处理语音

数据对应的文本信息。或者选择的提示词为加标点符号提示词,则输出的文本信息可以为对待处理语音数据对应的文本内容加标点符号后的文本信息。或者选择的提示词为文本顺滑提示词,则输出的文本信息可以为对待处理语音数据进行文本顺滑处理后的文本信息,即文本信息更通顺。或者选择的提示词为纠错提示词,则输出的文本信息可以为对待处理语音数据对应的文本内容进行文本纠错后的文本信息。

[0064] 可选地实现方式中,可以选择待处理语音数据对应的提示词,从而可以根据获取到的待处理语音数据的目标语音表征信息,输出与选择的提示词对应的文本信息。例如选择的提示词为重识别提示词,则可以根据获取到的待处理语音数据的目标语音表征信息,输出文本信息,输出的文本信息尽可能流畅、连贯。例如选择的提示词为情绪识别提示词,则可以根据获取到的待处理语音数据的目标语音表征信息,判断说话者的情绪类别,并根据说话者的情绪类别从以下多种情绪类别(如欢喜、悲哀、害怕、愤怒、意外和厌恶)中选择出待处理语音数据对应的情绪类别,输出选择出的情绪类别。例如选择的提示词为去口语化提示词,则可以根据获取到的待处理语音数据的目标语音表征信息,得到文本信息,去除文本信息中口语化的词,使得文本信息尽可能通顺、易读,从而输出去除文本信息中口语化的词的文本信息。例如选择的提示词为性别判断提示词,则可以根据获取到的待处理语音数据的目标语音表征信息,判断说话者的性别,选择“男”、“女”输出。例如选择的提示词为加标点提示词,则可以根据获取到的待处理语音数据的目标语音表征信息,得到文本信息,并在文本信息中添加标点后输出添加标点后的文本信息。

[0065] 本申请实施例中,通过结合待处理语音数据对应的需求选择提示词,可以在对待处理语音数据进行识别的同时进行语音理解,提升语音识别的准确性,从而得到更准确地文本信息。

[0066] 在一个实施例中,可以通过以下方式对提示词和目标语音表征信息进行融合处理:采用特征转换参数对提示词进行特征转换,得到提示词对应的特征向量矩阵;对语音内容向量、副语言向量和提示词对应的特征向量矩阵进行特征拼接,得到语音融合特征。

[0067] 其中,对语音内容向量、副语言向量和提示词对应的特征向量矩阵进行特征拼接的本质上就是对目标语音表征信息和提示词对应的特征向量矩阵进行特征拼接,两种拼接方法得到的语音融合特征一致。由于提示词对应的特征向量矩阵的维度与目标语音表征信息所表征的特征向量矩阵的维度相同,因此可以对目标语音表征信息和提示词对应的特征向量矩阵进行特征拼接,得到语音融合特征。语音融合特征中既可以反映出待处理语音数据的语音内容,又可以反映出待处理语音数据中的副语言信息,还可以反映出提示词对应的文本内容。因此后续对语音融合特征进行处理后,得到的文本信息既可以包含待处理语音数据的语音内容,又可以反映出待处理语音数据中的副语言信息,还可以反映出提示词对应的文本内容,从而可以提升语音识别准确性。

[0068] 可选地,特征转换参数可以是指词嵌入层的参数,词嵌入层可以对输入的文本数据如提示词进行特征转换,将文本数据转换成特征向量矩阵。通过词嵌入层中的参数对提示词进行特征转换的实质上是进行特征编码,即将文本维度的数据编码成特征向量。词嵌入是指将划分好的词语编码成稠密的向量,即将词语映射到数学空间的过程。例如可以预先设定词嵌入层的参数,即特征转换参数,通过将提示词输入到词嵌入层,词嵌入层的参数即特征转换参数可以将提示词转换为特征向量矩阵。通过使用词嵌入层对提示词进行特征

转换,便于后续使用特征向量矩阵进行特征融合和语音理解,提升语音识别的准确性。

[0069] 本申请实施例中,通过将待处理语音数据转换为维度与提示词对应的特征向量矩阵维度相等的语音表征信息,可以实现语音表征信息与文本语义对齐,即待处理语音数据的语音的隐空间表征与待处理语音数据对应的文本的隐空间表征是一致的,从而可以实现融合语音表征信息和文本语义信息,进而提升语音识别的准确性。

[0070] S103,对语音融合特征进行语音转换处理,得到待处理语音数据对应的文本信息。

[0071] 本申请实施例中,由于语音融合特征包含提示词对应的文本内容、待处理语音数据的语音内容和副语言信息,因此对语音融合特征进行语音转换处理,得到待处理语音数据对应的文本信息,待处理语音数据对应的文本信息既可以包含待处理语音数据的语音内容,又可以反映出待处理语音数据中的副语言信息,还可以反映出提示词对应的文本内容,即可以实现深层次的语音理解,从而提高语音识别的准确性。

[0072] 可选地,可以使用训练后的语音转换模型对语音融合特征进行语音转换处理,得到待处理语音数据对应的文本信息。其中,语音转换模型例如可以包括但不限于大规模语言模型(Large language model,LLM)、生成式对话模型(chat General Language Model,chat GLM)、开源对话语言模型(MOSS)、生成式的预训练模型(Generative Pre-Training,GPT),等等。

[0073] 例如,使用训练后的语音转换模型对语音融合特征进行语音转换处理,得到待处理语音数据对应的文本信息的过程可以如下:将语音融合特征划分为多个特征单元,基于语音转换模型的输入特征单元序列预测下一个特征单元,将输入的特征单元和预测出的特征单元加入特征单元序列,继续预测下一个特征单元,直到预测出语音融合特征对应的多个特征单元。每当预测出一个特征单元,则会将该特征单元和该特征单元之前的特征单元加入特征单元序列,预测该特征单元的下一个特征单元。其中,特征单元可以是指文字的基本单元,例如可以是指一个汉字,或者一个单词,等等。可选地,也可以使用BPE(Byte pair encoding)方法,将单词划分为更小的单元,例如,子字符串或字符作为基本单元。可选地实现方式中,可以根据文本语料训练出文本的基本组成单元,作为特征单元。

[0074] 具体实现中,由于目标语音表征信息所表征的特征向量矩阵为多维矩阵,且提示词对应的特征向量矩阵为多维矩阵,且两种矩阵的维度相同,因此融合得到的语音融合特征也为多维矩阵,且三种矩阵的维度均相同。在将语音融合特征对应的多维矩阵输入训练后的语音转换模型时,可以将语音融合特征对应的多维矩阵中的一列矩阵作为一个特征单元输入,一列矩阵可以包含待处理语音数据中的一个文字对应的语音,从而可以预测下一列特征单元,并在预测后一列特征单元时,将前面已经预测出的特征单元作为特征单元序列输入语音转换模型,从而实现预测语音融合特征对应的文本信息。

[0075] 本申请实施例中,由于语音识别属于感知任务,通过结合LLM模型,可以提升对语音数据的认知能力,从而结合语音、文本模态信息,提升对语音数据的理解能力,增强在更多语音、语义相关任务上的性能。由于LLM模型可以处理任意形式的文本任务,使用语音、语义相关的任务均可以在本申请技术方案上进行扩展,例如还可以在语音和文本的基础上进一步融合更多模态如视觉信息。例如可以将视觉信息转换为文本表征信息,进一步结合文本信息和语音表征信息输入LLM模型进行处理,从而丰富语音理解内容,提升语音识别的准确性。

[0076] 本申请实施例中,对待处理语音数据进行特征提取,得到待处理语音数据的目标语音表征信息。获取关于待处理语音数据的提示词,并对目标语音表征信息和提示词进行融合处理,得到语音融合特征;对语音融合特征进行语音转换处理,得到待处理语音数据对应的文本信息。由于目标语音表征信息包括待处理语音数据对应的语音内容向量和副语言向量,而副语言向量用于辅助识别待处理语音数据对应的文本信息。因此在对待处理语音数据进行语音识别时,既可以结合待处理语音数据的语音内容方面的信息,又可以结合待处理语音数据中的副语言方面的信息,还可以结合提示词对应的文本内容进行语音识别,可以实现对待处理语音数据进行深层次的语音识别和理解,从而提升语音识别的准确性。

[0077] 进一步地,请参见图5,图5是本申请实施例提供的一种语音特征提取模型训练的方法流程示意图。该方法可以应用于计算机设备;如图5所示,该方法包括但不限于以下步骤:

[0078] S201,获取样本语音数据,采用语音特征提取模型对样本语音数据进行特征提取,得到样本语音数据的样本语音表征信息。

[0079] 本申请实施例中,样本语音数据可以是预先获取的,例如可以是从语音数据存储网站下载得到、或者从终端设备上传得到、或者从本地存储的语音数据中获取到。为了增加训练数据的数量,还可以进一步对样本语音数据进行裁剪、旋转、调音、加噪等处理,从而实现扩充样本语音数据的数量。通过使用大量样本音频数据作为语音特征提取模型的训练数据进行训练,可以提升语音特征提取模型的准确性。

[0080] 本申请实施例中,例如可以采用语音特征提取模型对样本语音数据进行特征编码,得到待处理语音数据的语音向量矩阵,采用语音特征提取模型中的特征转换参数对待处理语音数据的语音向量矩阵进行特征转换,得到样本语音数据的样本语音表征信息。样本语音数据的样本语音表征信息可以包括样本语音数据对应的样本语音内容向量和样本副语言向量。

[0081] 在一个实施例中,语音特征提取模型可以包括语音向量矩阵提取层和语音表征全连接层,则可以结合语音向量矩阵提取层和语音表征全连接层确定样本语音数据的样本语音表征信息。例如,可以将样本语音数据输入语音向量矩阵提取层,通过语音向量矩阵提取层对样本语音数据进行特征编码,得到样本语音数据的语音向量矩阵,将样本语音数据的语音向量矩阵输入语音表征全连接层。通过语音表征全连接层中的特征转换参数对样本语音数据的语音向量矩阵进行特征转换,得到样本语音数据的样本语音表征信息。

[0082] 进一步可选地,语音向量矩阵提取层中还可以包括编码层、多头注意力层、归一化层。具体地,样本语音数据输入到语音特征提取模型中,由编码层对样本语音数据中的各个音频帧进行编码处理,得到编码特征,获取样本语音数据中各个音频帧的位置特征,对各个音频帧的编码特征和位置特征进行拼接,得到各个音频帧的拼接编码特征,拼接编码特征为带有位置信息的语音向量。通过多头注意力层计算样本语音数据中各个音频帧的拼接编码特征之间的相似度,确定每两个音频帧之间的相似度分数,通过归一化层将相似度分数归一化到0~1范围内。对于样本语音数据中每两个音频帧而言,两个音频帧之间的相似度分数越高,则两个音频帧之间的权重越大,两个音频帧之间的相似度分数越低,则两个音频帧之间的权重越小。通过结合样本语音数据中每个音频帧与其他音频帧之间的权重进行加权求和,得到的音频帧中包含了自身音频帧的语音信息,还包括样本语音数据中其他音频帧

的语音信息,从而引入样本语音数据中上下文语音信息,进而使得样本语音数据中每个音频帧都包含整个样本语音数据的信息。

[0083] S202,获取样本语音数据的样本语音表征标签。

[0084] 其中,样本语音表征标签可以是预先获取的用户与反映样本语音数据的真实值的语音表征标签。通过获取样本语音数据的样本语音表征标签,后续训练语音特征提取模型时,可以结合样本语音表征标签与语音特征提取模型输出的样本语音表征信息来调整语音特征提取模型。

[0085] S203,基于样本语音表征标签和样本语音表征信息训练语音特征提取模型,得到训练后的语音特征提取模型。

[0086] 这里,样本语音表征信息是指语音特征提取模型的模型输出值,样本语音表征标签是指样本真实值,训练语音特征提取模型的目的在于使得模型输出值与样本真实值尽可能一致。若模型输出值与样本真实值不一致,则继续调整语音特征提取模型中的模型参数,使得模型输出值与样本真实值一致。当模型输出值与样本真实值一致时,将此时的语音特征提取模型作为训练后的语音特征提取模型。

[0087] 其中,训练语音特征提取模型是指:比较样本语音表征标签和样本语音表征信息之间的差异,基于样本语音表征标签和样本语音表征信息之间的差异确定针对语音特征提取模型的损失函数。其中,样本语音表征标签和样本语音表征信息之间的差异可以基于相似度计算方法计算得到,即样本语音表征标签和样本语音表征信息之间的相似度越大,样本语音表征标签和样本语音表征信息之间的差异越小。样本语音表征标签和样本语音表征信息之间的相似度越小,样本语音表征标签和样本语音表征信息之间的差异越大。若样本语音表征标签和样本语音表征信息之间的差异大于差异阈值,则语音特征提取模型的损失函数大于第一损失阈值,则继续调整语音特征提取模型的模型参数,以降低语音特征提取模型的损失函数。当样本语音表征标签和样本语音表征信息之间的差异小于或等于差异阈值,则语音特征提取模型的损失函数小于或等于第一损失阈值,则可以保存此时的语音特征提取模型,作为训练后的语音特征提取模型。

[0088] 可选地,还可以在语音特征提取模型的迭代训练次数大于次数阈值、或者语音特征提取模型达到收敛条件时,停止调整语音特征提取模型中的模型参数,得到训练后的语音特征提取模型。

[0089] 可选地实现方式中,还可以通过以下方式训练语音特征提取模型:采用语音特征提取模型对样本语音数据进行特征提取,得到样本语音数据的样本语音表征信息;采用语音特征提取模型预测样本语音数据的样本语音表征信息对应的样本文本数据;获取样本语音数据的样本文本标签,基于样本文本标签和样本文本数据训练语音特征提取模型,得到训练后的语音特征提取模型。

[0090] 其中,通过预测样本语音数据的样本语音表征信息的文本数据,相当于将样本语音表征信息转换为文本模态的信息,从而根据两个文本之间的差异训练语音特征提取模型。样本文本标签可以是指样本语音数据的真实文本,样本语音表征信息的文本数据可以是指通过语音特征提取模型预测的文本,即模型输出的文本,通过比较样本文本标签与样本语音表征信息的文本数据之间的差异,从而基于样本文本标签与样本语音表征信息的文本数据之间的差异训练语音特征提取模型。样本文本标签与样本语音表征信息的文本数据

之间的差异可以基于文本相似度计算方法计算得到,本申请实施例对此不做限定。通过将样本语音表征信息转换为文本模态的信息进行比对,可以进行文本对比,确定文本之间的差异,从而调整语音特征提取模型。

[0091] 在一个实施例中,当语音特征提取模型包括语音向量矩阵提取层和语音表征全连接层时,则可以通过以下方式训练语音特征提取模型:

[0092] 基于样本语音表征标签和样本语音表征信息调整语音向量矩阵提取层的参数,得到训练后的语音特征提取模型。

[0093] 本申请实施例中,由于在训练语音特征提取模型时,使得语音表征全连接层中的参数固定,即语音表征全连接层中的参数固定为词嵌入层的特征转换参数。通过固定语音表征全连接层中的参数,可以使得语音特征提取模型能够进行语音识别的同时,其隐空间表征与语音转换模型一致,以便于将通过语音特征提取模型输出的待处理语音数据的目标语音表征信息输入语音转换模型。

[0094] 可选地实现方式中,在训练语音特征提取模型时,可以分别删除样本语音数据中的无意义帧,从而减少计算量,提升语音特征提取模型的训练效率。

[0095] 本申请实施例中,通过对待处理语音数据去除无意义帧后得到待处理语音数据的目标语音表征信息,可以减少计算量。通过对语音表征全连接层与语音转换模型中的词嵌入层的参数进行权重共享,可以使得语音表征全连接层输出的语音表征信息与词嵌入层输出的提示词的编码特征对齐,从而实现两种模态的信息之间的融合。语音表征全连接层的参数来自于语音转换模型中的词嵌入层。

[0096] 本申请实施例中,通过训练语音特征提取模型,可以使用训练后的语音特征提取模型对待处理语音数据进行特征提取,得到待处理语音数据的目标语音表征信息,提升语音数据处理效率。由于使用了大量的样本语音数据训练语音特征提取模型,可以提升语音特征提取模型的准确性。

[0097] 可选的,请参见图6,图6是本申请实施例提供的一种语音转换模型训练的方法流程示意图。该方法可以应用于计算机设备;如图6所示,该方法包括但不限于以下步骤:

[0098] S301,获取样本语音数据对应的样本语音表征信息和样本提示词。

[0099] 本申请实施例中,样本语音表征信息可以通过对样本语音数据进行特征提取得到,例如可以是对样本语音数据进行划分得到多个音频帧,对样本语音数据中多个音频帧进行特征编码得到多个音频帧的语音向量矩阵,并采用语音特征提取模型中的特征转换参数对多个音频帧的语音向量矩阵进行特征转换,得到多个音频帧的候选语音表征信息,通过从样本音频数据中的多个音频帧的候选语音表征信息中删除无意义的音频帧得到的样本语音表征信息。本申请实施例中可以包括多种样本提示词,每种样本提示词对应的场景可以不同。在准备训练数据时,可以结合实际需求选择对应的样本提示词。例如样本提示词可以包括文本顺滑、去口语化、纠错、重识别、情绪识别、性别判断、加标点符号等提示词。例如纠错提示词可以包括各个领域内的专业术语。去口语化提示词可以包括一些口语化的词语。文本顺滑提示词可以包括重复字组成的词语,等等。

[0100] 可以理解的是,由于提示词没有固定的格式,例如去口语化、性别判断、情绪识别等场景下的提示词为文字、而加标点符号场景下的提示词为标点符号、或者重识别场景下的提示词为指示用于进行重识别的提示信息,因此在实际使用场景中,只需要选择的提示

词与模型训练时保持一致即可。

[0101] S302,采用语音转换模型对样本语音内容向量、样本副语言向量和样本提示词进行融合处理,得到样本语音融合特征。

[0102] 这里,由于样本语音数据的样本语音表征信息包括样本语音数据对应的样本语音内容向量和样本副语言向量,因此对样本语音内容向量、样本副语言向量和样本提示词进行融合处理的本质上就是对样本语音表征信息和样本提示词进行融合处理。由于语音表征信息是通过特征向量矩阵表示的,而提示词是通过文字来表示的,因此在对两者进行融合处理之前,可以将提示词转换为特征向量矩阵,便于进行特征融合,例如特征拼接。

[0103] S303,采用语音转换模型对样本语音融合特征进行语音转换处理,得到样本语音数据对应的文本信息。

[0104] 例如,使用语音转换模型对样本语音融合特征进行语音转换处理,得到样本语音数据对应的文本信息的过程可以如下:将样本语音融合特征划分为多个特征单元,基于输入语音转换模型的特征单元序列预测下一个特征单元,将输入的特征单元和预测出的特征单元加入特征单元序列,继续预测下一个特征单元,直到预测出样本语音融合特征对应的多个特征单元。

[0105] 可选地实现方式中,由于样本语音表征信息所表征的特征向量矩阵为多维矩阵,且样本提示词对应的特征向量矩阵为多维矩阵,且两种矩阵的维度相同,因此融合得到的样本语音融合特征也为多维矩阵,且三种矩阵的维度均相同。在将样本语音融合特征对应的多维矩阵输入训练后的语音转换模型时,可以将样本语音融合特征对应的多维矩阵中的一列矩阵作为一个特征单元输入,一列矩阵可以表示样本语音数据中一个文字对应的语音数据,从而可以预测下一列特征单元,并在预测后一列特征单元时,将前面已经预测出的特征单元作为特征单元序列输入语音转换模型,从而实现预测样本语音融合特征对应的文本信息。

[0106] 可选地,语音转换模型例如可以使用当下开源的大规模语言模型,如使用较为广泛的基于Transformer结构的模型,采用自回归形式,即基于输入的token(特征单元)序列,预测下一个token,然后基于输入及已预测出的token,预测下一个token,以此类推实现预测出待处理语音数据对应的文本信息。

[0107] S304,获取样本语音数据对应的样本文本标签,基于样本文本标签和样本语音数据对应的文本信息训练语音转换模型,得到训练后的语音转换模型。

[0108] 本申请实施例中,样本文本标签可以是指样本语音数据的真实文本标签,样本语音数据对应的文本信息可以是指基于语音转换模型输出的模型输出值,训练语音转换模型的目的在于使得样本文本标签和样本语音数据对应的文本信息尽可能一致,当样本文本标签和样本语音数据对应的文本信息一致时,可以将此时的语音转换模型确定为训练后的语音转换模型。样本文本标签和样本语音数据对应的文本信息可以通过文本相似度计算方法计算得到。

[0109] 其中,基于样本文本标签和样本语音数据对应的文本信息训练语音转换模型是指:基于样本文本标签和样本语音数据对应的文本信息之间的差异确定针对语音转换模型的损失函数。在语音转换模型的损失函数大于第二损失阈值时,继续调整语音转换模型的模型参数,以降低语音转换模型的损失函数。当语音转换模型的损失函数小于或等于第二

损失阈值时,将此时的语音转换模型确定为训练后的语音转换模型。

[0110] 可选地,训练语音转换模型的过程实际上是一个调整语音转换模型中的参数的过程,由于语音转换模型中包括大量参数,训练时对语音转换模型中的所有参数进行调整需要耗费大量的时间,降低训练效率,因此可以针对语音转换模型中的部分参数进行调整,实现提升语音转换模型效率的目的。

[0111] 请参见图7,图7是本申请实施例提供的一种语音转换模型中的参数调整示意图,图7中左边部分为语音转换模型(如LLM模型)中预训练的模型参数W(即预训练权重),在预训练的模型结构旁加上一个分支,这个分支包含A、B两个结构,A、B这两个参数分别初始化为高斯分布和0。在训练刚开始,附加的参数就是0,A的输入维度和B的输出维度分别与原始模型的输入输出维度相同,而A的输出维度和B的输入维度是一个远小于原始模型输入输出维度的值,这样就可以极大的减少LLM模型中待训练的参数。在训练LLM模型时只更新A、B的参数,预训练好的模型参数W固定不变,将AB与原始模型参数矩阵W合并,这样就不会在推断中引入额外的计算,对于不同的下游任务只需要在预训练模型的基础上重新训练A、B就可以。当训练好新的参数后,将新的参数与老的参数合并,利用重参的方式,既能在新的任务上达到微调效果,又不会在模型推理中增加耗时,可以提升模型训练效率。由于LLM模型本身参数量较大,微调时仅增加小部分参数进行训练,可以提升训练效率。

[0112] 在训练语音转换模型时,通过语音转换模型的输入数据预测语音转换模型的输出数据,关键在于训练集的准备。针对给定任务如(文本顺滑、去口语化、纠错等),可以准备如下训练集,样本语音数据通过上述语音特征转换模型去除无意义帧后得到样本语音数据的语音表征信息,加上样本语音数据对应的提示词,一并输入语音转换模型,可以输出对应任务上的文本信息。

[0113] 举例1为文本顺滑场景:

[0114] 输入:你你吃过了吗(输入为语音模态的语音表征信息),加上对应的提示词如“你你、我我”等重复词;

[0115] 输出:你吃过了吗。

[0116] 举例2为去口语化场景:

[0117] 输入:嗯,对,嗯,我赞同,嗯(输入为语音模态的语音表征信息),加上对应的提示词如“嗯嗯、是是是”等口语化词语;

[0118] 输出:对,我赞同。

[0119] 举例3为纠错场景:

[0120] 输入:这套系统的原厂效果不太好(输入为语音模态的语音表征信息),加上对应的提示词如“远场”等对应场景下的专业术语;

[0121] 输出:这套系统的远场效果不太好。

[0122] 举例4为加标点符号场景:

[0123] 输入:这是一句很长的话需要加标点吗我觉得需要(输入为语音模态的语音表征信息),加上对应的提示词如“逗号、问号、句号”等标点符号;

[0124] 输出:这是一句很长的话,需要加标点吗?我觉得需要。

[0125] 在本申请实施例中,语音特征提取模型如ASR模型通过复用LLM模型的词嵌入(word embedding)机制,将语音表征信息对齐到LLM模型的文本语义空间,使得语音表征信

息可以直接作为LLM模型的输入,由于语音表征信息包含文本内容和副语言信息,因此LLM模型可以充分利用语音模态的信息,进一步提升语音识别和理解的能力。通过LLM模型可以充分利用语音数据中除文本内容以外的其他信息,从而可以增强语音识别与语音理解能力。语音识别时通过设置不同的提示词,在语音识别环节直接提升语音识别的去口语化、热词替换、情绪识别、加标点符号等能力,形成端到端模型,从而输出对应的文本信息。而不是在语音识别环节先识别成文本,再将文本拿给LLM模型进行处理。本申请技术方案还可以扩展至其他模态的信息,如视觉信息,共同作为LLM模型的输入,进一步增强模型的语音识别与语音理解能力。

[0126] 本申请实施例中,通过训练语音转换模型,可以使用训练后的语音转换模型对语音融合特征进行语音转换处理,得到待处理语音数据对应的文本信息,提升语音数据处理效率。由于使用了大量的样本语音数据训练语音转换模型,可以提升语音转换模型的准确性。

[0127] 上面介绍了本申请实施例的方法,下面介绍本申请实施例的装置。

[0128] 参见图8,图8是本申请实施例提供的一种语音处理装置的组成结构示意图,上述语音处理装置可以部署于计算机设备上;该语音处理装置可以用于执行本申请实施例提供的语音处理方法中的相应步骤。该语音处理装置80包括:

[0129] 特征提取单元801,用于对待处理语音数据进行特征提取,得到该待处理语音数据的目标语音表征信息;该目标语音表征信息包括该待处理语音数据对应的语音内容向量和副语言向量,该副语言向量用于辅助识别该待处理语音数据对应的文本信息;

[0130] 信息融合单元802,用于获取关于该待处理语音数据的提示词,并对该语音内容向量、该副语言向量和该提示词进行融合处理,得到语音融合特征;

[0131] 语音转换单元803,用于对该语音融合特征进行语音转换处理,得到该待处理语音数据对应的文本信息。

[0132] 可选地,该信息融合单元802,具体用于:

[0133] 采用特征转换参数对该提示词进行特征转换,得到该提示词对应的特征向量矩阵;

[0134] 对该语音内容向量、该副语言向量和该提示词对应的特征向量矩阵进行特征拼接,得到该语音融合特征。

[0135] 可选地,该特征提取单元801,具体用于:

[0136] 获取用于对该提示词进行特征转换的特征转换参数;

[0137] 对该待处理语音数据进行特征编码,得到该待处理语音数据的语音向量矩阵;

[0138] 采用该特征转换参数对该待处理语音数据的语音向量矩阵进行特征转换,得到该待处理语音数据的目标语音表征信息;该目标语音表征信息所表征的特征向量矩阵的维度与该提示词对应的特征向量矩阵的维度相同。

[0139] 可选地,该特征提取单元801,具体用于:

[0140] 对该待处理语音数据进行划分,得到多个音频帧;

[0141] 对各个音频帧进行特征编码,得到该各个音频帧的语音向量矩阵;

[0142] 采用该特征转换参数对该各个音频帧的语音向量矩阵进行特征转换,得到该各个音频帧的候选语音表征信息;

[0143] 遍历该多个音频帧,基于当前遍历的音频帧的候选语音表征信息,预测该当前遍历的音频帧映射为语音内容中的各个文字的概率;语音内容是指语音内容向量指示的内容;

[0144] 若该当前遍历的音频帧映射为语音内容中的各个文字的概率中的最大概率小于概率阈值,则从该各个音频帧的候选语音表征信息中删除该当前遍历的音频帧的候选语音表征信息;

[0145] 在遍历结束后,基于剩余的候选语音表征信息,得到该待处理语音数据的目标语音表征信息。

[0146] 可选地,该特征提取单元801,具体还用于:

[0147] 基于该多个音频帧的划分顺序,确定各个音频帧的位置特征,该位置特征用于指示相应音频帧在该待处理语音数据中的位置;

[0148] 该特征提取单元801,具体用于:

[0149] 对各个音频帧进行特征编码,得到该各个音频帧的编码特征;

[0150] 针对任一音频帧,对该任一音频帧的位置特征和该任一音频帧的编码特征进行特征拼接处理,得到该任一音频帧的语音向量矩阵。

[0151] 可选地,该待处理语音数据对应的文本信息是通过训练后的语音转换模型得到的,该语音处理装置80还包括:第一训练单元804,该第一训练单元804,用于:

[0152] 获取样本语音数据对应的样本语音表征信息和样本提示词;该样本语音表征信息包括样本语音数据对应的样本语音内容向量和样本副语言向量;

[0153] 采用语音转换模型对该样本语音内容向量、该样本副语言向量和该样本提示词进行融合处理,得到样本语音融合特征;

[0154] 采用该语音转换模型对该样本语音融合特征进行语音转换处理,得到该样本语音数据对应的文本信息;

[0155] 获取该样本语音数据对应的样本文本标签,基于该样本文本标签和该样本语音数据对应的文本信息训练该语音转换模型,得到该训练后的语音转换模型。

[0156] 可选地,该待处理语音数据的目标语音表征信息是通过训练后的语音特征提取模型得到的,该语音处理装置80还包括:第二训练单元805,该第二训练单元805,用于:

[0157] 获取样本语音数据,采用语音特征提取模型对该样本语音数据进行特征提取,得到该样本语音数据的样本语音表征信息;

[0158] 获取该样本语音数据的样本语音表征标签,基于该样本语音表征标签和该样本语音表征信息训练该语音特征提取模型,得到该训练后的语音特征提取模型。

[0159] 可选地,该语音特征提取模型包括语音向量矩阵提取层和语音表征全连接层;该第二训练单元805,具体用于:

[0160] 通过该语音向量矩阵提取层对该样本语音数据进行特征编码,得到该样本语音数据的语音向量矩阵;

[0161] 通过该语音表征全连接层中的特征转换参数对该样本语音数据的语音向量矩阵进行特征转换,得到该样本语音数据的样本语音表征信息;

[0162] 基于该样本语音表征标签和该样本语音表征信息调整该语音向量矩阵提取层的参数,得到该训练后的语音特征提取模型。

[0163] 需要说明的是,图8对应的实施例中未提及的内容可参见方法实施例的描述,这里不再赘述。

[0164] 本申请实施例中,对待处理语音数据进行特征提取,得到待处理语音数据的目标语音表征信息。获取关于待处理语音数据的提示词,并对目标语音表征信息和提示词进行融合处理,得到语音融合特征;对语音融合特征进行语音转换处理,得到待处理语音数据对应的文本信息。由于目标语音表征信息包括待处理语音数据对应的语音内容向量和副语言向量,而副语言向量用于辅助识别待处理语音数据对应的文本信息。因此在对待处理语音数据进行语音识别时,既可以结合待处理语音数据的语音内容方面的信息,又可以结合待处理语音数据中的副语言方面的信息,还可以结合提示词对应的文本内容进行语音识别,可以实现对待处理语音数据进行深层次的语音识别和理解,从而提升语音识别的准确性。

[0165] 参见图9,图9是本申请实施例提供的一种计算机设备的组成结构示意图。如图9所示,上述计算机设备90可以包括:处理器901和存储器902以及网络接口903。处理器901连接到存储器902和网络接口903,例如处理器901可以通过总线连接到存储器902和网络接口903。其中,计算机设备可以是终端设备,也可以是服务器。

[0166] 处理器901被配置为支持语音处理装置执行上述的语音处理方法中相应的功能。该处理器901可以是中央处理器(Central Processing Unit,CPU),网络处理器(Network Processor,NP),硬件芯片或者其任意组合。上述硬件芯片可以是专用集成电路(Application-Specific Integrated Circuit,ASIC),可编程逻辑器件(Programmable Logic Device,PLD)或其组合。上述PLD可以是复杂可编程逻辑器件(Complex Programmable Logic Device,CPLD),现场可编程逻辑门阵列(Field-Programmable Gate Array,FPGA),通用阵列逻辑(Generic Array Logic,GAL)或其任意组合。

[0167] 存储器902存储器用于存储程序指令和数据等。存储器902可以包括易失性存储器(Volatile Memory,VM),例如随机存取存储器(Random Access Memory,RAM);存储器902也可以包括非易失性存储器(Non-Volatile Memory,NVM),例如只读存储器(Read-Only Memory,ROM),快闪存储器(flash memory),硬盘(Hard Disk Drive,HDD)或固态硬盘(Solid-State Drive,SSD);存储器902还可以包括上述种类的存储器的组合。

[0168] 网络接口903用于提供网络通讯功能。

[0169] 处理器901可以调用该程序代码以执行以下操作:

[0170] 对待处理语音数据进行特征提取,得到该待处理语音数据的目标语音表征信息;该目标语音表征信息包括该待处理语音数据对应的语音内容向量和副语言向量,该副语言向量用于辅助识别该待处理语音数据对应的文本信息;

[0171] 获取关于该待处理语音数据的提示词,并对该语音内容向量、该副语言向量和该提示词进行融合处理,得到语音融合特征;

[0172] 对该语音融合特征进行语音转换处理,得到该待处理语音数据对应的文本信息。

[0173] 应当理解,本申请实施例中所描述的计算机设备90可执行前文图3、图5和图6所对应实施例中对上述方法的描述,也可执行前文图8所对应实施例中对上述语音处理装置的描述,在此不再赘述。另外,对采用相同方法的有益效果描述,也不再进行赘述。

[0174] 本申请实施例中,对待处理语音数据进行特征提取,得到待处理语音数据的目标语音表征信息。获取关于待处理语音数据的提示词,并对目标语音表征信息和提示词进行

融合处理,得到语音融合特征;对语音融合特征进行语音转换处理,得到待处理语音数据对应的文本信息。由于目标语音表征信息包括待处理语音数据对应的语音内容向量和副语言向量,而副语言向量用于辅助识别待处理语音数据对应的文本信息。因此在对待处理语音数据进行语音识别时,既可以结合待处理语音数据的语音内容方面的信息,又可以结合待处理语音数据中的副语言方面的信息,还可以结合提示词对应的文本内容进行语音识别,可以实现对待处理语音数据进行深层次的语音识别和理解,从而提升语音识别的准确性。

[0175] 可选的,该程序指令被处理器执行时还可实现上述实施例中方法的其他步骤,这里不再赘述。

[0176] 本申请实施例还提供一种计算机可读存储介质,该计算机可读存储介质存储有计算机程序,该计算机程序包括程序指令,该程序指令当被计算机执行时使该计算机执行如前述实施例的方法,该计算机可以为上述提到的计算机设备的一部分。作为示例,程序指令可被部署在一个计算机设备上执行,或者被部署位于一个地点的多个计算机设备上执行,又或者,在分布在多个地点且通过通信网络互连的多个计算机设备上执行,分布在多个地点且通过通信网络互连的多个计算机设备可以组成区块链网络。

[0177] 本申请实施例还提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令被处理器执行时可实现上述方法中的部分或全部步骤。例如,该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得该计算机设备执行上述各方法的实施例中所执行的步骤。

[0178] 本领域普通技术人员可以理解实现上述实施例的方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,该的程序可存储于计算机可读取存储介质中,该程序在执行时,可包括如上述各方法的实施例的流程。其中,该的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory,ROM)或随机存储记忆体(Random Access Memory,RAM)等。

[0179] 以上所揭露的仅为本申请较佳实施例而已,当然不能以此来限定本申请之权利范围,因此依本申请权利要求所作的等同变化,仍属本申请所涵盖的范围。

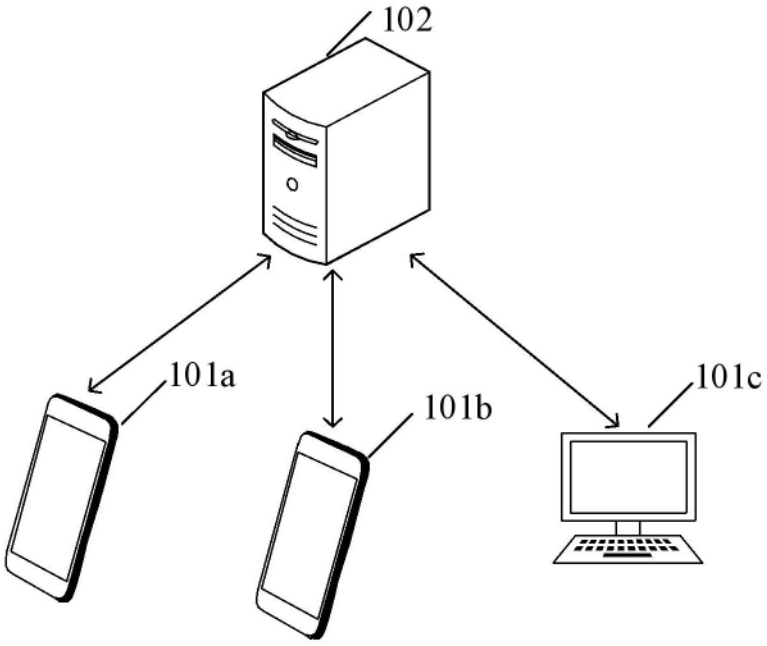

102
101a
101b
101c

图1

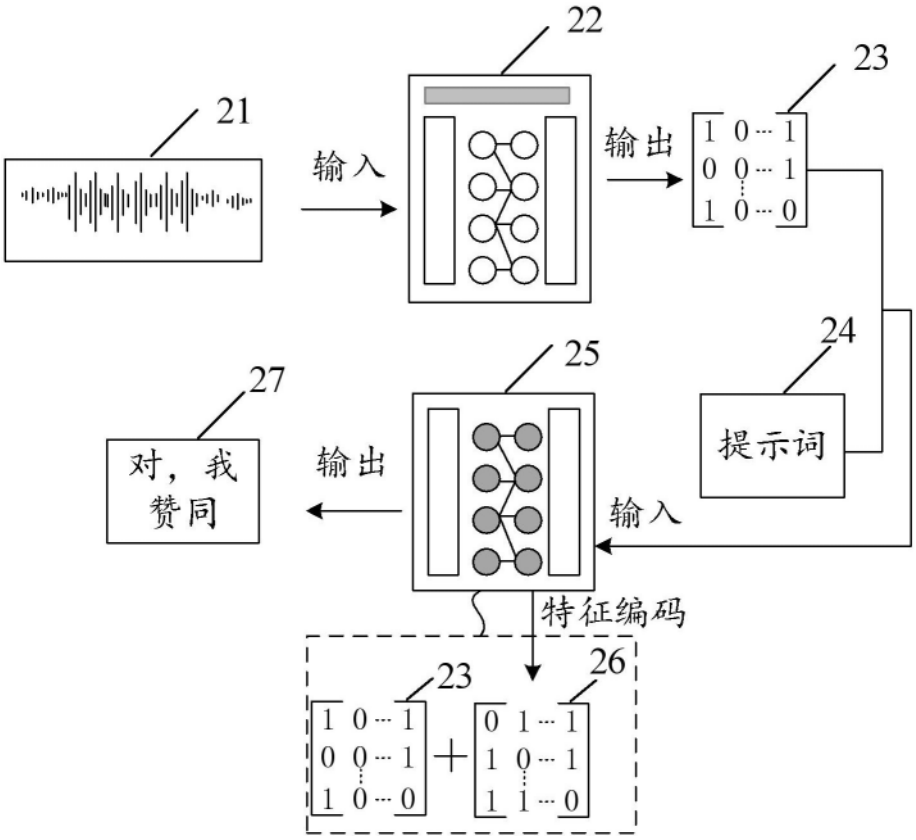

22
21
输入
输出
23
24
25
27
对，我
赞同
输出
提示词
输入
特征编码
23
26

图2

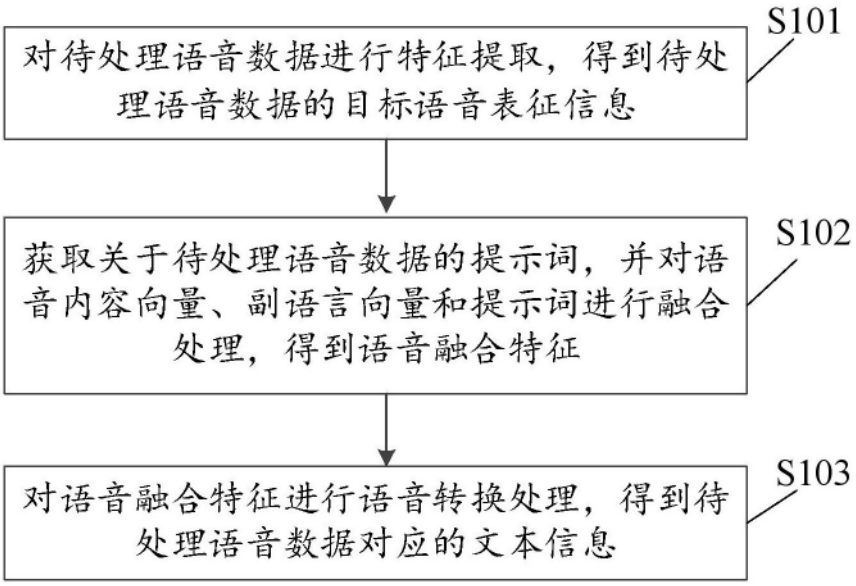

对待处理语音数据进行特征提取，得到待处理语音数据的目标语音表征信息
S101
获取关于待处理语音数据的提示词，并对语音内容向量、副语言向量和提示词进行融合处理，得到语音融合特征
S102
对语音融合特征进行语音转换处理，得到待处理语音数据对应的文本信息
S103

图3

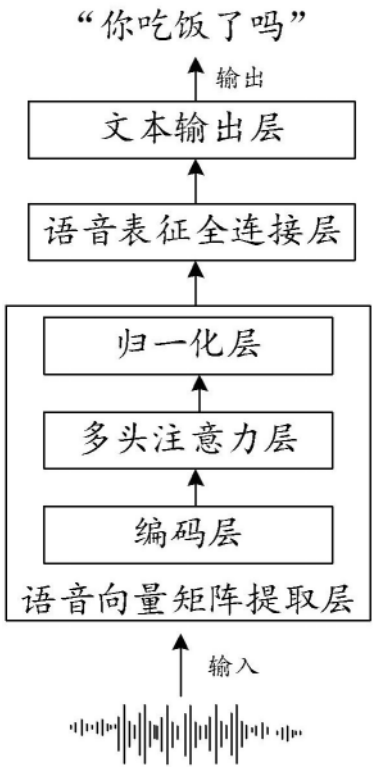

“你吃饭了吗”
输出
文本输出层
语音表征全连接层
归一化层
多头注意力层
编码层
语音向量矩阵提取层
输入

图4

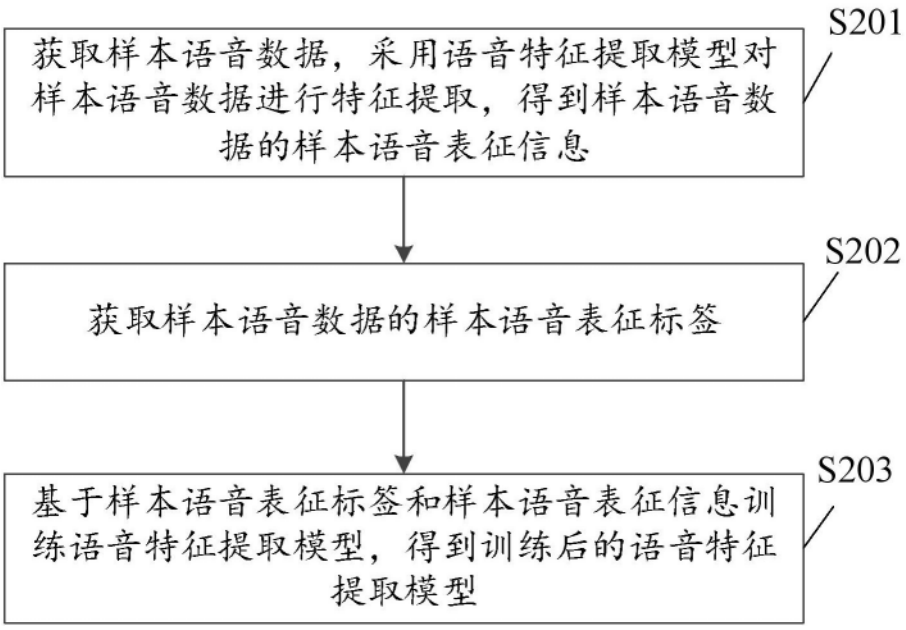

S201
获取样本语音数据，采用语音特征提取模型对样本语音数据进行特征提取，得到样本语音数据的样本语音表征信息
S202
获取样本语音数据的样本语音表征标签
S203
基于样本语音表征标签和样本语音表征信息训练语音特征提取模型，得到训练后的语音特征提取模型

图5

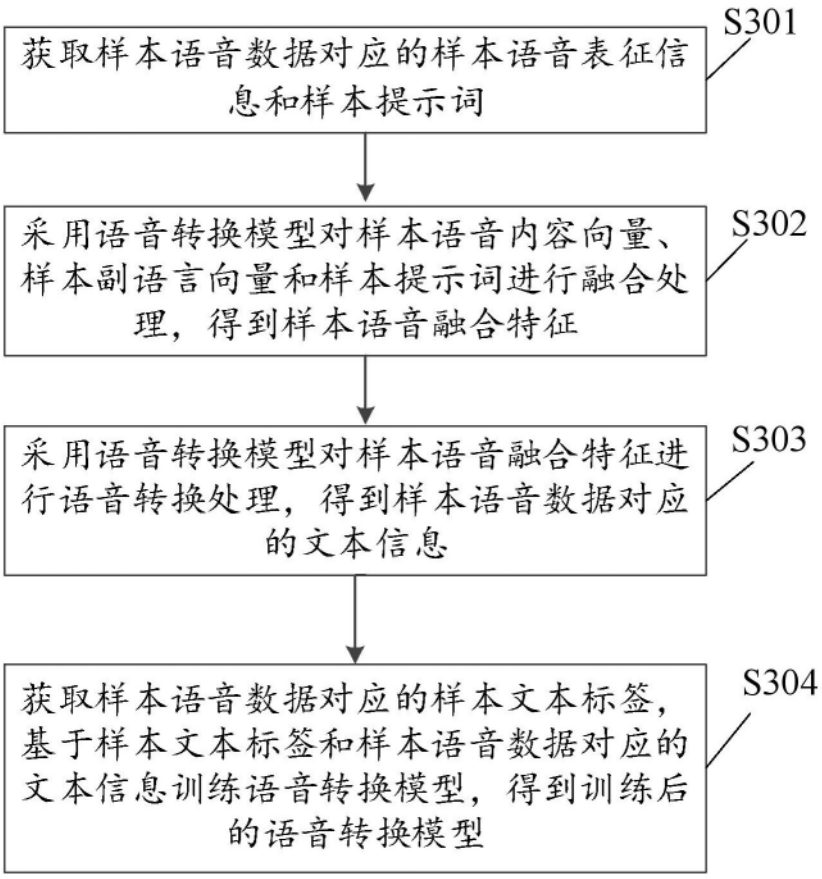

S301
获取样本语音数据对应的样本语音表征信息和样本提示词
S302
采用语音转换模型对样本语音内容向量、样本副语言向量和样本提示词进行融合处理，得到样本语音融合特征
S303
采用语音转换模型对样本语音融合特征进行语音转换处理，得到样本语音数据对应的文本信息
S304
获取样本语音数据对应的样本文本标签，基于样本文本标签和样本语音数据对应的文本信息训练语音转换模型，得到训练后的语音转换模型

图6

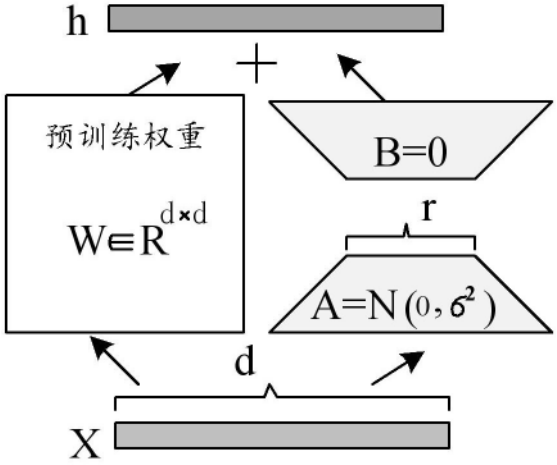

h
预训练权重
W∈R^{d×d}
B=0
r
A=N(0,σ²)
d
X

图7

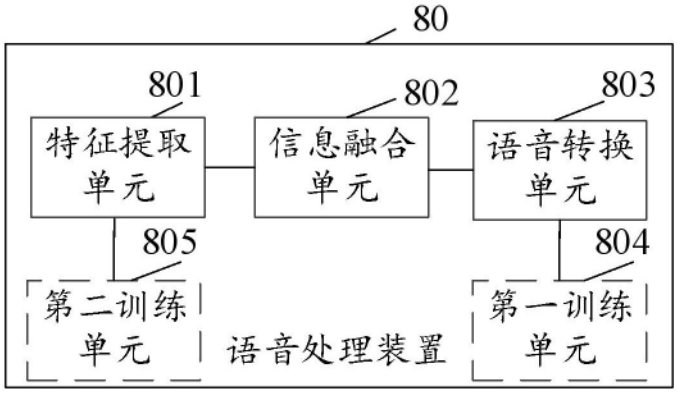

80
801
特征提取单元
802
信息融合单元
803
语音转换单元
805
第二训练单元
804
第一训练单元
语音处理装置

图8

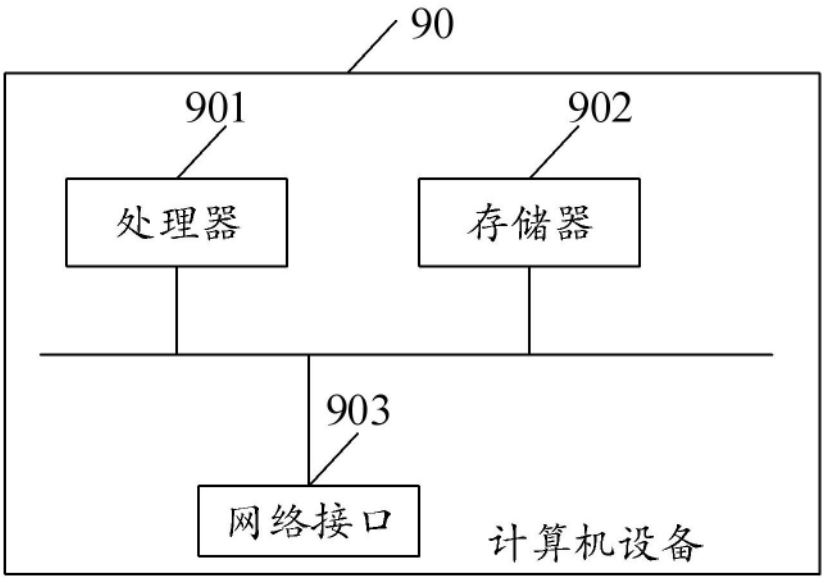


图9