(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117672176 A
(43)申请公布日 2024.03.08

(21)申请号 202311692184.8
(22)申请日 2023.12.08
(71)申请人 深圳市大数据研究院
地址 518000 广东省深圳市龙岗区龙城街道龙翔大道2001号道远楼225室
(72)发明人 薛浏蒙 潘诗锋 何磊 谢磊 武执政
(74)专利代理机构 深圳市恒和大知识产权代理有限公司 44479
专利代理师 邹航
(51)Int.Cl.
G10L 13/02(2013.01)
G10L 13/06(2013.01)
G10L 25/18(2013.01)
G10L 15/06(2013.01)
G10L 25/30(2013.01)

权利要求书2页 说明书11页 附图7页

(54)发明名称

基于语音自监督学习表征的重读可控语音合成方法及装置

(57)摘要

本发明公开了一种基于语音自监督学习表征的重读可控语音合成方法及装置,其方法实现,包括:获取原始语音信号,通过预设的语音自监督预训练模型从原始音频信号中提取帧级表征;通过预设的微调模块对帧级表征进行加权和处理以及池化处理,以得到字级表征,并对其进行重读分类预测,以得到每个字的字级重读表征;获取待处理文本数据,将文本数据转换为对应的离散音素序列,并编码为连续音素嵌入表征,并输入至预设的编码器中进行处理,以得到文本隐层表征;将文本隐层表征与字级重读表征输入至预设的自监督学习表征编码器中进行处理,以得到解码器输入特征;将解码器输入特征输入至预设的解码器中进行解码处理,以得到音频的梅尔频谱表征,并还原为语音波形。

CN 117672176 A

1.一种基于语音自监督学习表征的重读可控语音合成方法,其特征在于,所述方法,包括:

获取原始语音信号,通过预设的语音自监督预训练模型从所述原始音频信号中提取帧级表征;

通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征;

获取待处理文本数据,将所述待处理文本数据转换为对应的离散音素序列,将所述离散音素序列编码为连续音素嵌入表征,并输入至预设的编码器中进行处理,以得到文本隐层表征;

将所述文本隐层表征与所述字级表征输入至预设的自监督学习表征编码器中进行处理,以得到解码器输入特征;

将所述解码器输入特征输入至预设的解码器中进行解码处理,以得到音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形。

2.如权利要求1所述的基于语音自监督学习表征的重读可控语音合成方法,其特征在于,所述通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征之后,包括:

对所述字级表征进行重读分类预测,以得到每个字的字级重读表征;

基于所述字级重读表征,对所述预设的语音自监督预训练模型进行迭代训练。

3.如权利要求1所述的语音自监督学习表征的重读可控语音合成方法,其特征在于,所述通过预设的语音自监督预训练模型从所述原始音频信号中提取帧级表征,包括:

通过梅尔滤波器组对所述原始音频信号进行处理,以提取音频特征;

将所述音频特征依次输入至多个编码器层中进行特征提取,以得到多个所述帧级表征。

4.如权利要求3所述的语音自监督学习表征的重读可控语音合成方法,其特征在于,所述通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征,包括:

对多个所述帧级表征进行加权和操作,以得到最终的帧级表征;

对所述最终的帧级表征进行池化处理,以得到所述字级表征。

5.如权利要求4所述的语音自监督学习表征的重读可控语音合成方法,其特征在于,所述对所述最终的帧级表征进行池化处理,以得到所述字级表征,包括:

确定每个字对应的时间帧范围,所述时间帧范围包括起始时间帧以及结束时间帧;

基于所述起始时间帧以及结束时间帧,对每个字所属的帧级表征进行平均池化操作,以得到所述字级表征。

6.如权利要求1所述的语音自监督学习表征的重读可控语音合成方法,其特征在于,所述将所述离散音素序列编码为连续音素嵌入表征,包括:

获取预训练的音素查询表;

通过所述音素查询表查找所述离散音素序列中每个音素对应的音素嵌入向量,以得到音素嵌入向量序列,所述音素嵌入向量序列构成所述连续音素嵌入表征。

7.如权利要求1所述的语音自监督学习表征的重读可控语音合成方法,其特征在于,所

述输入至预设的编码器中进行处理,以得到文本隐层表征,包括:

获取说话人ID,并基于预设说话人查询表,得到说话人嵌入向量;

将所述连续音素嵌入表征与所述说话人嵌入向量进行相加的结果输入至所述预设的编码器中进行处理,以得到所述文本隐层表征。

8.如权利要求1所述的语音自监督学习表征的重读可控语音合成方法,其特征在于,所述将所述文本隐层表征与所述字级重读表征输入至预设的自监督学习表征编码器中进行处理,以得到解码器输入特征,包括:

根据每个字对应的音素数,将所述字级重读表征转换为音素级重读表征;

对所述音素级重读表征进行线性变换,得到线性变换后的音素级重读表征,所述线性变换后的音素级重读表征与所述文本隐层表征的维度一致;

对所述线性变换后的音素级重读表征进行卷积以及归一化处理,并与所述文本隐层表征进行相加,以得到所述解码器输入特征。

9.如权利要求1所述的语音自监督学习表征的重读可控语音合成方法,其特征在于,所述将所述解码器输入特征输入至预设的解码器中进行解码处理,以得到音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形,包括:

将所述解码器输入特征从音素级表征拓展为目标帧级表征;

将所述目标帧级表征输入至所述预设的解码器中进行处理,以得到所述音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形。

10.一种基于语音自监督学习表征的重读可控语音合成装置,其特征在于,所述装置,包括:

帧级表征提取单元,用于获取原始语音信号,通过预设的语音自监督预训练模型从所述原始音频信号中提取帧级表征;

字级表征获取单元,用于通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征;

文本隐层表征获取单元,用于获取待处理文本数据,将所述待处理文本数据转换为对应的离散音素序列,将所述离散音素序列编码为连续音素嵌入表征,并输入至预设的编码器中进行处理,以得到文本隐层表征;

预测特征获取单元,用于将所述文本隐层表征与所述字级表征输入至预设的自监督学习表征编码器中进行处理,以得到解码器输入特征;

梅尔频谱表征获取单元,用于将所述解码器输入特征输入至预设的解码器中进行解码处理,以得到音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形。

基于语音自监督学习表征的重读可控语音合成方法及装置

技术领域

[0001] 本发明涉及语音合成技术领域,尤其涉及一种基于语音自监督学习表征的重读可控语音合成方法及装置。

背景技术

[0002] 随着人工智能技术的发展,语音合成技术越来越受到人们的重视,语音合成技术被广泛应用在人机交互或者将文本转化成自然语言输出等领域。

[0003] 目前,在语音合成中建模重读的方法主要有以下两种:一种是使用人工标注的重读硬标签,例如使用数字0,1来分别代表重读和不重读;第二种是使用连续小波变化算法自动地从语音中提取韵律特征并分析得到重读标记。然而由于高质量语音合成数据录制的时间和金钱成本较高,语音合成数据量有限,导致语音重读标签不具有泛化性,因此这两种重读方法限制了重读语音合成的自然度和表现力。

发明内容

[0004] 基于此,有必要针对上述技术问题,提供一种基于语音自监督学习表征的重读可控语音合成方法、装置、计算机设备及存储介质,以解决上述现有技术中存在的至少一个问题。

[0005] 本申请实施例是这样实现的,提供了一种基于语音自监督学习表征的重读可控语音合成方法,包括如下步骤:

[0006] 获取原始语音信号,通过预设的语音自监督预训练模型从所述原始音频信号中提取帧级表征;

[0007] 通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征;

[0008] 获取待处理文本数据,将所述待处理文本数据转换为对应的离散音素序列,将所述离散音素序列编码为连续音素嵌入表征,并输入至预设的编码器中进行处理,以得到文本隐层表征;

[0009] 将所述文本隐层表征与所述字级表征输入至预设的自监督学习表征编码器中进行处理,以得到解码器输入特征;

[0010] 将所述解码器输入特征输入至预设的解码器中进行解码处理,以得到音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形。

[0011] 在一实施例中,所述通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征之后,包括:

[0012] 对所述字级表征进行重读分类预测,以得到每个字的字级重读表征;

[0013] 基于所述字级重读表征,对所述预设的语音自监督预训练模型进行迭代训练。

[0014] 在一实施例中,所述通过预设的语音自监督预训练模型从所述原始音频信号中提取帧级表征,包括:

[0015] 通过梅尔滤波器组对所述原始音频信号进行处理,以提取音频特征;

[0016] 将所述音频特征依次输入至多个编码器层中进行特征提取,以得到多个所述帧级表征。

[0017] 在一实施例中,所述通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征,包括:

[0018] 对多个所述帧级表征进行加权和操作,以得到最终的帧级表征;

[0019] 对所述最终的帧级表征进行池化处理,以得到所述字级表征。

[0020] 在一实施例中,所述对所述最终的帧级表征进行池化处理,以得到所述字级表征,包括:

[0021] 确定每个字对应的时间帧范围,所述时间帧范围包括起始时间帧以及结束时间帧;

[0022] 基于所述起始时间帧以及结束时间帧,对每个字所属的帧级表征进行平均池化操作,以得到所述字级表征。

[0023] 在一实施例中,所述将所述离散音素序列编码为连续音素嵌入表征,包括:

[0024] 获取预训练的音素查询表;

[0025] 通过所述音素查询表查找所述离散音素序列中每个音素对应的音素嵌入向量,以得到音素嵌入向量序列,所述音素嵌入向量序列构成所述连续音素嵌入表征。

[0026] 在一实施例中,所述输入至预设的编码器中进行处理,以得到文本隐层表征,包括:

[0027] 获取说话人ID,并基于预设说话人查询表,得到说话人嵌入向量;

[0028] 将所述连续音素嵌入表征与所述说话人嵌入向量进行相加的结果输入至所述预设的编码器中进行处理,以得到所述文本隐层表征。

[0029] 在一实施例中,所述将所述文本隐层表征与所述字级重读表征输入至预设的自监督学习表征编码器中进行处理,以得到解码器输入特征,包括:

[0030] 根据每个字对应的音素数,将所述字级重读表征转换为音素级重读表征;

[0031] 对所述音素级重读表征进行线性变换,得到线性变换后的音素级重读表征,所述线性变换后的音素级重读表征与所述文本隐层表征的维度一致;

[0032] 对所述线性变换后的音素级重读表征进行卷积以及归一化处理,并与所述文本隐层表征进行相加,以得到所述解码器输入特征。

[0033] 在一实施例中,所述将所述解码器输入特征输入至预设的解码器中进行解码处理,以得到音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形,包括:

[0034] 将所述解码器输入特征从音素级表征拓展为目标帧级表征;

[0035] 将所述目标帧级表征输入至所述预设的解码器中进行处理,以得到所述音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形。

[0036] 第二方面,提供了一种基于语音自监督学习表征的重读可控语音合成装置,包括:

[0037] 帧级表征提取单元,用于获取原始语音信号,通过预设的语音自监督预训练模型从所述原始音频信号中提取帧级表征;

[0038] 字级表征获取单元,用于通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征;

[0039] 文本隐层表征获取单元,用于获取待处理文本数据,将所述待处理文本数据转换为对应的离散音素序列,将所述离散音素序列编码为连续音素嵌入表征,并输入至预设的编码器中进行处理,以得到文本隐层表征;

[0040] 预测特征获取单元,用于将所述文本隐层表征与所述字级表征输入至预设的自监督学习表征编码器中进行处理,以得到解码器输入特征;

[0041] 梅尔频谱表征获取单元,用于将所述解码器输入特征输入至预设的解码器中进行解码处理,以得到音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形。

[0042] 第三方面,提供了一种计算机设备,包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令,所述处理器执行所述计算机可读指令时实现如上述所述基于语音自监督学习表征的重读可控语音合成方法的步骤;

[0043] 第四方面,提供了一种可读存储介质,所述可读存储介质存储有计算机可读指令,所述计算机可读指令被处理器执行时实现如上述所述基于语音自监督学习表征的重读可控语音合成方法的步骤。

[0044] 上述基于语音自监督学习表征的重读可控语音合成方法、装置、计算机设备及存储介质,其方法实现,包括:获取原始语音信号,通过预设的语音自监督预训练模型从所述原始音频信号中提取帧级表征;通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征;获取待处理文本数据,将所述待处理文本数据转换为对应的离散音素序列,将所述离散音素序列编码为连续音素嵌入表征,并输入至预设的编码器中进行处理,以得到文本隐层表征;将所述文本隐层表征与所述字级表征输入至预设的自监督学习表征编码器中进行处理,以得到解码器输入特征;将所述解码器输入特征输入至预设的解码器中进行解码处理,以得到音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形。本申请实施例中,通过语音自监督预训练模型以及微调模块对每个字进行重读标记,然后基于该重读标记引入声学模型中进行重读建模以及重读语音合成处理,借助大规模数据训练的语音自监督模型知识帮助提升重度可控语音合成的自然度和表现力。

附图说明

[0045] 为了更清楚地说明本发明实施例的技术方案,下面将对本发明实施例的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性的前提下,还可以根据这些附图获得其他的附图。

[0046] 图1是本发明一实施例中基于语音自监督学习表征的重读可控语音合成方法的一应用环境示意图;

[0047] 图2是本发明一实施例中基于语音自监督学习表征的重读可控语音合成方法方法的一流程示意图;

[0048] 图3是本发明一实施例中微调模块的结构示意图;

[0049] 图4是本发明一实施例中语音合成模型的结构示意图;

[0050] 图5是本发明一实施例中Conformer编码器的结构示意图;

[0051] 图6是本发明一实施例中自监督学习表征编码器的结构示意图;

[0052] 图7是本发明一实施例中语音合成方法的应用场景图;

[0053] 图8是本发明一实施例中基于语音自监督学习表征的重读可控语音合成装置的一结构示意图;

[0054] 图9是本发明一实施例中计算机设备的一示意图。

具体实施方式

[0055] 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作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。

[0056] 本实施例提供的基于语音自监督学习表征的重读可控语音合成方法,可应用在如图1的应用环境中,具体地,其训练阶段可包括两个阶段,第一个阶段为微调语音自监督预训练模型,该语音自监督预训练模块可为WavLM模型,第二阶段为训练语音合成模型,通过该微调语音自监督预训练模型将语音预训练模型适配到重度标签分类任务上,从而从微调的语音自监督模型中提取重读表征。通过该语音合成模型将第一阶段提取的重音相关的自监督学习表征引入到语音合成模型中,以便训练语音合成模型,从而实现对重读建模和重读语音合成。

[0057] 在一实施例中,如图2所示,提供一种基于语音自监督学习表征的重读可控语音合成方法,包括如下步骤:

[0058] 在步骤S110中,获取原始语音信号,通过预设的语音自监督预训练模型从所述原始音频信号中提取帧级表征;

[0059] 其中,该原始语音信号可通过采集大量的用户语音信号得到,可形成语音信号样本集,用于对该语音自监督预训练模型进行训练。

[0060] 在本申请实施例中,该预设的语音自监督预训练模型可为WavLM模型,用于从原始语音信号中提取帧级表征,WavLM模型是一种基于Transformer的深度学习模型。具体可包括前端处理模块、基于Transformer的编码器,该编码器通常包含多个自注意力层,每层都包含多个头,此结构使得模型能够捕捉到输入特征之间的复杂关系和长距离依赖。此外,该WavLM模型的一个核心特点是它的自监督学习机制,通过掩蔽一部分输入然后让WavLM模型预测这些被掩蔽部分的内容来进行训练,这种训练方式使WavLM模型能够学习到音频数据的上下文化表示。

[0061] 在本申请一实施例中,所述通过预设的语音自监督预训练模型从所述原始音频信号中提取帧级表征,包括:

[0062] 通过梅尔滤波器组对所述原始音频信号进行处理,以提取音频特征;

[0063] 将所述音频特征依次输入至多个编码器层中进行特征提取,以得到多个所述帧级表征。

[0064] 具体地,该原始语音信号输入至WavLM模型中时,首先通过一个前端处理模块,如Log-Mel滤波器组,提取音频特征,提取的音频特征随后被送入至基于Transformer的编码器中,该基于Transformer的编码器包括多个编码器层,每一层均可产生一个帧级表征,因此可以提取出多个帧级表征,例如,输入的语音信号为x,第i层编码器输出的帧级表征可以表示为$E_i(x)$,其中,i=1,2.....,N。

[0065] 在步骤S120中,通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征,并对所述字级表征进行重读分类预测,以得到每个字的字级重读表征;

[0066] 参见图3,在本申请实施例中,该预设的微调模块可包括加权和层、字级平均池化层以及重读分类器,加权和层与微调任务联合训练得到一组可学习的权重参数,然后将学习到的权重与各层编码器输出的帧级表征进行加权和操作得到最终的帧级表征,然后进行池化处理,以得到该字级表征,并对该字级表征进行重读分类预测,以得到每个字的字级重读表征。

[0067] 在本申请一实施例中,所述通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征,包括:

[0068] 对多个所述帧级表征进行加权和操作,以得到最终的帧级表征;

[0069] 对所述最终的帧级表征进行池化处理,以得到所述字级表征。

[0070] 具体地,假设在微调阶段,通过联合训练得到了一组可学习的权重参数,表示为wi,其中i=1,2,.....,N。这些权重参数用于结合不同编码层输出的帧级表征。则通过该权重参数在加权和层对各编码层输出的帧级表征进行加权和操作后,可输出得到最终的帧级表征,该最终的帧级表征是各个编码层输出的加权和,具体可以表示为F(x),其可通过如下公式获取:

[0071] $$F(x)=\sum_{i=1}^{N} wi \cdot E_i(x)$$。

[0072] 进一步,为了将提取的帧级表征对齐到字级,则该字级平均池化层可以根据每个字的时间帧数对其所属的帧级表征进行平均池化操作以得到该字级表征。

[0073] 在本申请一实施例中,所述对所述最终的帧级表征进行池化处理,以得到所述字级表征,包括:

[0074] 确定每个字对应的时间帧范围,所述时间帧范围包括起始时间帧以及结束时间帧;

[0075] 基于所述起始时间帧以及结束时间帧,对每个字所属的帧级表征进行平均池化操作,以得到所述字级表征。

[0076] 具体地,对于每个字,都存在一定数量的时间帧,假设第j个字对应的时间帧范围是从t_{j1}到t_{j2},其中t_{j1}和t_{j2}分别是第j个字开始和结束的时间帧的索引。字级表征表示为W_j,其可以通过对其所属的帧级表征做平均池化操作得到,具体可通过如下公式计算得到:

[0077] $$W_j = \frac{1}{t_{j2}-t_{j1}} \sum_{t=t_{j1}}^{t_{j2}} F_t(x)$$

[0078] 其中,W_j表示第j个字的字级表征,$F_t(x)$表示时间帧t的帧级表征。

[0079] 然后,该重读分类器可将该字级表征作为输入,预测得到每个字的重读标记。具体地,该重读分类器的结构包括两层带有ReLU激活函数的线性层和一层线性输出层。字级表征作为输入:假设每个字的字级表征为W_j,其中,j表示字的索引,对于第一层线性层,可表示为$L_1(W_j)=ReLU(A_1 \cdot W_j+b_1)$,其中,$A_1$表示第一层线性输出层的权重矩阵,$b_1$表示偏置项,ReLU为激活函数。对于第二层线性层,同理,其可表示为$L_2(W_j)=ReLU(A_2 \cdot L_1(W_j)+b_2)$,其中,$A_2$表示第层线性输出层的权重矩阵,$b_2$表示偏置项,ReLU为激活函数。输出层可以表示为

$O(W_j)=A_3 \cdot L_2(W_j)+b_3$，其中，$A_3$为输出层的权重矩阵，$b_3$是偏置项。通过该输出层可以预测并输出每个字的重读标记，例如，个二分类问题(重读/非重读)，可以使用sigmoid函数来将输出转换为概率：$P_j=\text{sigmoid}(O(W_j))$，其中，$P_j$表示第j个字是重读的概率。

[0080] 通过该重读分类器可以预测出每个字的重读标记，然后可通过损失函数对该预测结果进行计算，例如交叉熵函数、均方差损失函数等，并基于该损失函数计算出的损失值，对该语音自监督预训练模型进行迭代训练，以对该字级表征进行学习，当该损失值达到预设阈值，例如90％，或者迭代次数达到预设次数时，则可得到训练后的语音自监督预训练模型，并该重读分类器预测得到每个字的重读标记，从而可以更好的对字级表征，即图3中所示的重音相关的自监督学习表征进行学习，以便可以更好的用于对第二阶段的语音合成模型进行训练。

[0081] 可以理解的，该第一阶段的微调模块以及语音自监督预训练模型进在训练过程中使用。

[0082] 在步骤S130中，获取待处理文本数据，将所述文本数据转换为对应的离散音素序列，将所述离散音素序列编码为连续音素嵌入表征，并输入至预设的编码器中进行处理，以得到文本隐层表征；

[0083] 参见图4，在本申请实施例中，第二阶段的语音合成模型具体可包括：文本转音素模块(Grapheme-to-Phoneme，G2P)、音素嵌入(Phoneme Embedding)、Conformer编码器(Conformer Encoder)、自监督学习表征编码器(SSL Representation Encoder)、变量适配器(Variance Adaptor)、Conformer解码器(Conformer Decoder)。

[0084] 其中，该预设的编码器为该Conformer编码器。

[0085] 具体地，该待处理文本数据可由字母或者字符组成，可称为图素(grapheme)，该文本转音素模块可用于将文本转换为对应的音素(phoneme)序列，假设，G为图素(grapheme)的集合，则对于给定的待处理文本数据为字符串T，可以表示为一个图素序列，$T=g_1, g_2, \ldots\ldots, g_n$，其中，每个$g_i \in G$，代表一个图素(字符或字母)。G2P转换的目标是找到一个函数f，将字符串T映射到相应的音素序列S：$S=f(T)=p_1, p_2 \cdots\cdots, p_n$，其中，每个$p_j \in P$，代表一个音素。

[0086] 在本申请一实施例中，将所述离散音素序列编码为连续音素嵌入表征，包括：

[0087] 获取预训练的音素查询表；

[0088] 通过所述音素查询表查找所述离散音素序列中每个音素对应的音素嵌入向量，以得到音素嵌入向量序列，所述音素嵌入向量序列构成所述连续音素嵌入表征。

[0089] 具体地，可通过一个和语音合成模型联合训练的音素查询表，查找离散音素序列中每个音素对应的音素嵌入向量，音素嵌入向量序列构成最终的音素嵌入表征。例如音素嵌入可定义为$P=(p_1, p_2, \ldots\ldots, p_m)$，其中，$p_m$可表示序列中的一个音素，m表示序列中音素的总数，该音素查询表可为一个函数E，可将离散的音素映射到连续嵌入向量，则对于序列中的每个音素p_m，其嵌入向量可表示为$E(p_m)$。该音素嵌入向量序列可以理解为将整个音素序列的嵌入表征表示为一个向量序列，例如$\{E(p_1), E(p_2), \ldots\ldots, E(p_m)\}$，其中，每个$E(p_m)$可为一个高维空间中的点，代表相应音素的嵌入。

[0090] 参见图5，进一步，在获取到连续音素嵌入表征之后，可将其输入至预设的编码器，例如，Conformer编码器中，该Conformer编码器可包括多个堆叠在一起的网络层组成，例如

4个,具体可包括括卷积前馈模块、深度卷积模块、自注意力模块和最后的第二个卷积前馈模块,其中每两个网络之间有一个相加和归一化操作。经过该Conformer内部各个网络层的处理,最后可输出该文本隐层表征。

[0091] 进一步,在本申请一实施例中,所述输入至预设的编码器中进行处理,以得到文本隐层表征,包括:

[0092] 获取说话人ID,并基于预设说话人查询表,得到说话人嵌入向量;

[0093] 将所述连续音素嵌入表征与所述说话人嵌入向量进行相加的结果输入至所述预设的编码器中进行处理,以得到所述文本隐层表征。

[0094] 具体地,该Conformer编码器还可将连续的音素嵌入表征和说话人ID相加的结果作为输入,经过Conformer内部各个网络层的处理,最后输出文本隐层表征。通过配置该说话人ID,并敬爱那个该说话人ID该连续音素嵌入表征进行相加后在进行网络层处理,使得语音合成模型可以合成不同说话人的语音,得到多种语音效果。

[0095] 进一步地,该说话人ID即为说话人的编号,通过该编号,查询预设的说话人查询表,可得到说话人嵌入向量,该说话人嵌入向量的函数可以表示为$f_{lookup(\cdot)}$,则该过程可以通过如下公式进行表示:

[0096] 说话人嵌入向量$=f_{lookup(ID)}$

[0097] 可以理解的,在具体实现中,$f_{lookup(\cdot)}$可为一个训练好的嵌入表,其中,灭个说话人ID均可对应一个嵌入向量,该嵌入表可以通过训练一个神经网络或者使用预训练的嵌入模型得到,例如Word2Vec、GloVe等。

[0098] 在步骤S140中,将所述文本隐层表征与所述字级表征输入至预设的自监督学习表征编码器中进行处理,以得到解码器输入特征;

[0099] 参见图6,在本申请实施例中,该自监督学习表征编码器可包括两个分支网络,第一分支网络可包括采样层、瓶颈层以及卷积与归一化层,第二分支网络则可包括重音查询表、音素级自监督表征预测器以及卷积与归一化层。其中,该第二分支网络中还可包括重音标签预测器,用于对重音进行标记,或者也可用人工标记的方式实现。

[0100] 进一步,该第一分支网络用于在训练阶段使用,该第二分支网络则用于在测试阶段使用。

[0101] 在本申请一实施例中,所述将所述文本隐层表征与所述字级重读表征输入至预设的自监督学习表征编码器中进行处理,以得到解码器输入特征,包括:

[0102] 根据每个字对应的音素数,将所述字级重读表征转换为音素级重读表征;

[0103] 对所述音素级重读表征进行线性变换,得到线性变换后的音素级重读表征,所述线性变换后的音素级重读表征与所述文本隐层表征的维度一致;

[0104] 对所述线性变换后的音素级重读表征进行卷积以及归一化处理,并与所述文本隐层表征进行相加,以得到所述解码器输入特征。

[0105] 具体地,在训练阶段中,采样层可以根据每个字对应的音素数,例如N,将第一阶段输出的字级特征,即重音相关的自监督学习表征复制N份,得到音素级重读表征,此时,重读表征序列的长度和文本音素序列的长度相同,然后可通过瓶颈层对该音素级重读表征进行线性变换,变换后的音素级重读表征与文本隐层表征的维度一致,之后经过一个一维卷积和归一化操作后与Conformer编码器输出的文本隐层表征相加,以作为自监督学习表征编

码器的输出。

[0106] 在本申请一实施例中,在测试阶段,只包括第二阶段的语音合成模型,无需第一阶段的微调的语音自监督预训练模型,且预设的自监督学习表征编码器也仅包括第二分支网络,此时具体测试流程为:将Conformer编码器输出的文本隐层表征输入至该第二分支网络中,可通过重读标签预测器进行预测得到音素级重读标签,或者也可通过人工手动标记重读标签,然后可通过重读查询表(Emphasis Look-up Table,Emphasis LUT)查找得到每个音素离散重读标签对应的连续重读嵌入表征。之后,经过音素级自监督表征预测器预测得到重读自监督表征,经过一个一维卷积和归一化操作后,与Conformer编码器输出的文本隐层表征相加,以作为自监督学习表征编码器的输出。

[0107] 在步骤S150中,将所述解码器输入特征输入至预设的解码器中进行解码处理,以得到音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形。

[0108] 在本申请一实施例中,所述将所述解码器输入特征输入至预设的解码器中进行解码处理,以得到音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形,包括:

[0109] 将所述解码器输入特征从音素级表征拓展为目标帧级表征;

[0110] 将所述目标帧级表征输入至所述预设的解码器中进行处理,以得到所述音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形。

[0111] 具体地,可将自监督学习表征编码器输出的预测特征输入至变量适配器中进行预测,以得到预测音素时长,从而实现将所述解码器输入特征从音素级表征拓展为目标帧级表征。

[0112] 其中,该变量适配器可由包括两层带有ReLU激活函数的一维卷积网络组成,每个网络后面都有归一化层和dropout层,以及一个额外的线性层来输出标量,该标量即为预测的音素持续时间。

[0113] 其中,该音素持续时间代表每个音素持续的帧数,根据音素的帧数(例如第i个音素有N帧),对每个音素表征进行复制(例如将第i个音素表征复制N份),从而实现将自监督学习表征编码器结果从音素级到帧级表征的扩展。

[0114] 在本申请实施例中,该预设的解码器可为Conformer解码器,用于将变量适配器的帧级输出作为输入,经过Conformer内部各个网络层的结构如图6所示,从而可得到音频的梅尔频谱表征。

[0115] 进一步,在获取到音频的梅尔频谱表征后,可使用合成器,例如HFI-GAN合成器将梅尔频普还原回人耳可听的语音波形。

[0116] 参见图7,在一实施例场景中,通过训练完成的语音合成模型进行语音合成时,例如,当用户向模型输入文本,并给定文本中每个字对应的重读标记,其中0表示不重读,1表示重读,例如,输入的文本为“这次旅行者没有带书”其中,“这次”“没有”为重读,则该文本对应的重读标签则为110001100,通过可控重读语音合成模型可根据该文本和对应的重读标签合成指定字重读的语音,从而得到文本对应的语音波形。

[0117] 上述基于语音自监督学习表征的重读可控语音合成方法,包括:获取原始语音信号,通过预设的语音自监督预训练模型从所述原始音频信号中提取帧级表征;通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征;获取待处理文本数据,将所述待处理文本数据转换为对应的离散音素序列,将所述离散音素序列编码为

连续音素嵌入表征,并输入至预设的编码器中进行处理,以得到文本隐层表征;将所述文本隐层表征与所述字级表征输入至预设的自监督学习表征编码器中进行处理,以得到解码器输入特征;将所述解码器输入特征输入至预设的解码器中进行解码处理,以得到音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形。本申请实施例中,通过语音自监督预训练模型以及微调模块对每个字进行重读标记,然后基于该重读标记引入声学模型中进行重读建模以及重读语音合成处理,借助大规模数据训练的语音自监督模型知识帮助提升重度可控语音合成的自然度和表现力。

[0118] 应理解,上述实施例中各步骤的序号的大小并不意味着执行顺序的先后,各过程的执行顺序应以其功能和内在逻辑确定,而不应对本发明实施例的实施过程构成任何限定。

[0119] 在一实施例中,提供一种基于语音自监督学习表征的重读可控语音合成装置,该基于语音自监督学习表征的重读可控语音合成装置与上述实施例中基于语音自监督学习表征的重读可控语音合成方法一一对应。如图8所示,该基于语音自监督学习表征的重读可控语音合成装置包括帧级表征提取单元10、字级表征获取单元20、文本隐层表征获取单元30、预测特征获取单元40和梅尔频谱表征获取单元50。各功能模块详细说明如下:

[0120] 帧级表征提取单元10,用于获取原始语音信号,通过预设的语音自监督预训练模型从所述原始音频信号中提取帧级表征;

[0121] 字级表征获取单元20,用于通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理,以得到字级表征;

[0122] 文本隐层表征获取单元30,用于获取待处理文本数据,将所述待处理文本数据转换为对应的离散音素序列,将所述离散音素序列编码为连续音素嵌入表征,并输入至预设的编码器中进行处理,以得到文本隐层表征;

[0123] 预测特征获取单元40,用于将所述文本隐层表征与所述字级表征输入至预设的自监督学习表征编码器中进行处理,以得到解码器输入特征;

[0124] 梅尔频谱表征获取单元50,用于将所述解码器输入特征输入至预设的解码器中进行解码处理,以得到音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形。

[0125] 在本申请实施例中,该装置还包括:重读标记分类预测单元,用于:

[0126] 对所述字级表征进行重读分类预测,以得到每个字的字级重读表征;

[0127] 基于所述字级重读表征,对所述预设的语音自监督预训练模型进行迭代训练。

[0128] 在本申请一实施例中,帧级表征提取单元10,还用于:

[0129] 通过梅尔滤波器组对所述原始音频信号进行处理,以提取音频特征;

[0130] 将所述音频特征依次输入至多个编码器层中进行特征提取,以得到多个所述帧级表征。

[0131] 在本申请实施例中,字级表征获取单元20,还用于:

[0132] 对多个所述帧级表征进行加权和操作,以得到最终的帧级表征;

[0133] 对所述最终的帧级表征进行池化处理,以得到所述字级表征。

[0134] 在本申请实施例中,字级表征获取单元20,还用于:

[0135] 确定每个字对应的时间帧范围,所述时间帧范围包括起始时间帧以及结束时间帧;

[0136] 基于所述起始时间帧以及结束时间帧,对每个字所属的所述最终的帧级表征进行平均池化操作,以得到所述字级表征。

[0137] 在本申请实施例中,文本隐层表征获取单元30,还用于:

[0138] 获取预训练的音素查询表;

[0139] 通过所述音素查询表查找所述离散音素序列中每个音素对应的音素嵌入向量,以得到音素嵌入向量序列,所述音素嵌入向量序列构成所述连续音素嵌入表征。

[0140] 在本申请实施例中,文本隐层表征获取单元30,还用于:

[0141] 获取说话人ID,并基于预设说话人查询表,得到说话人嵌入向量;

[0142] 将所述连续音素嵌入表征与所述说话人嵌入向量进行相加的结果输入至所述预设的编码器中进行处理,以得到所述文本隐层表征。

[0143] 在本申请一实施例中,预测特征获取单元40,还用于:

[0144] 根据每个字对应的音素数,将所述字级重读表征转换为音素级重读表征;

[0145] 对所述音素级重读表征进行线性变换,得到线性变换后的音素级重读表征,所述线性变换后的音素级重读表征与所述文本隐层表征的维度一致;

[0146] 对所述线性变换后的音素级重读表征进行卷积以及归一化处理,并与所述文本隐层表征进行相加,以得到所述解码器输入特征。

[0147] 在本申请一实施例中,梅尔频谱表征获取单元50,还用于:

[0148] 将所述解码器输入特征从音素级表征拓展为目标帧级表征;

[0149] 将所述目标帧级表征输入至所述预设的解码器中进行处理,以得到所述音频的梅尔频谱表征,并将所述梅尔频谱表征还原为语音波形。

[0150] 本申请实施例中,通过语音自监督预训练模型以及微调模块对每个字进行重读标记,然后基于该重读标记引入声学模型中进行重读建模以及重读语音合成处理,借助大规模数据训练的语音自监督模型知识帮助提升重度可控语音合成的自然度和表现力。

[0151] 关于基于语音自监督学习表征的重读可控语音合成装置的具体限定可以参见上文中对于基于语音自监督学习表征的重读可控语音合成方法的限定,在此不再赘述。上述基于语音自监督学习表征的重读可控语音合成装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中,也可以以软件形式存储于计算机设备中的存储器中,以便于处理器调用执行以上各个模块对应的操作。

[0152] 在一个实施例中,提供了一种计算机设备,该计算机设备可以是终端设备,其内部结构图可以如图9所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口。其中,该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括可读存储介质。该可读存储介质存储有计算机可读指令。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机可读指令被处理器执行时以实现一种基于语音自监督学习表征的重读可控语音合成方法。本实施例所提供的可读存储介质包括非易失性可读存储介质和易失性可读存储介质。

[0153] 在本申请实施例中,提供了一种计算机设备,包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令,所述处理器执行所述计算机可读指令时实现如上述所述基于语音自监督学习表征的重读可控语音合成方法的步骤。

[0154] 在申请实施例中,提供了一种可读存储介质,所述可读存储介质存储有计算机可读指令,所述计算机可读指令被处理器执行时实现如上述所述基于语音自监督学习表征的重读可控语音合成方法的步骤。

[0155] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机可读指令来指令相关的硬件来完成,所述的计算机可读指令可存储于一非易失性可读取存储介质或易失性可读存储介质中,该计算机可读指令在执行时,可包括如上述各方法的实施例的流程。其中,本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限,RAM以多种形式可得,诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0156] 所属领域的技术人员可以清楚地了解到,为了描述的方便和简洁,仅以上述各功能单元、模块的划分进行举例说明,实际应用中,可以根据需要而将上述功能分配由不同的功能单元、模块完成,即将所述装置的内部结构划分成不同的功能单元或模块,以完成以上描述的全部或者部分功能。

[0157] 以上所述实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围,均应包含在本发明的保护范围之内。

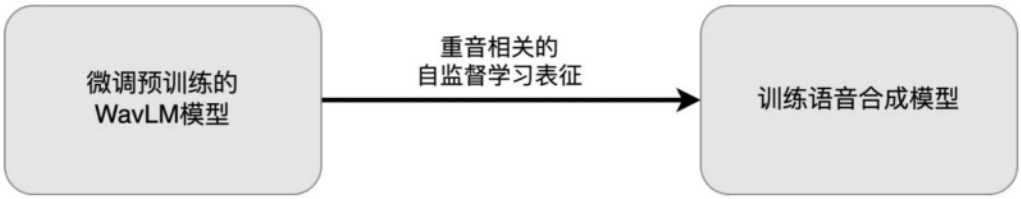
微调预训练的
WavLM模型
重音相关的
自监督学习表征
训练语音合成模型

图1


获取原始语音信号，通过预设的语音自监督预训练模型从所述原始音频信号中提取帧级表征
S110
通过预设的微调模块对所述帧级表征进行加权和处理以及池化处理，以得到字级表征
S120
获取待处理文本数据，将所述待处理文本数据转换为对应的离散音素序列，将所述离散音素序列编码为连续音素嵌入表征，并输入至预设编码器中进行处理，以得到文本隐层表征
S130
将所述文本隐层表征与所述字级表征输入至预设的自监督学习表征编码器中进行处理，以得到解码器输入特征
S140
将所述解码器输入特征输入至预设的解码器中进行解码处理，以得到音频的梅尔频谱表征，并将所述梅尔频谱表征还原为语音波形
S150

图2

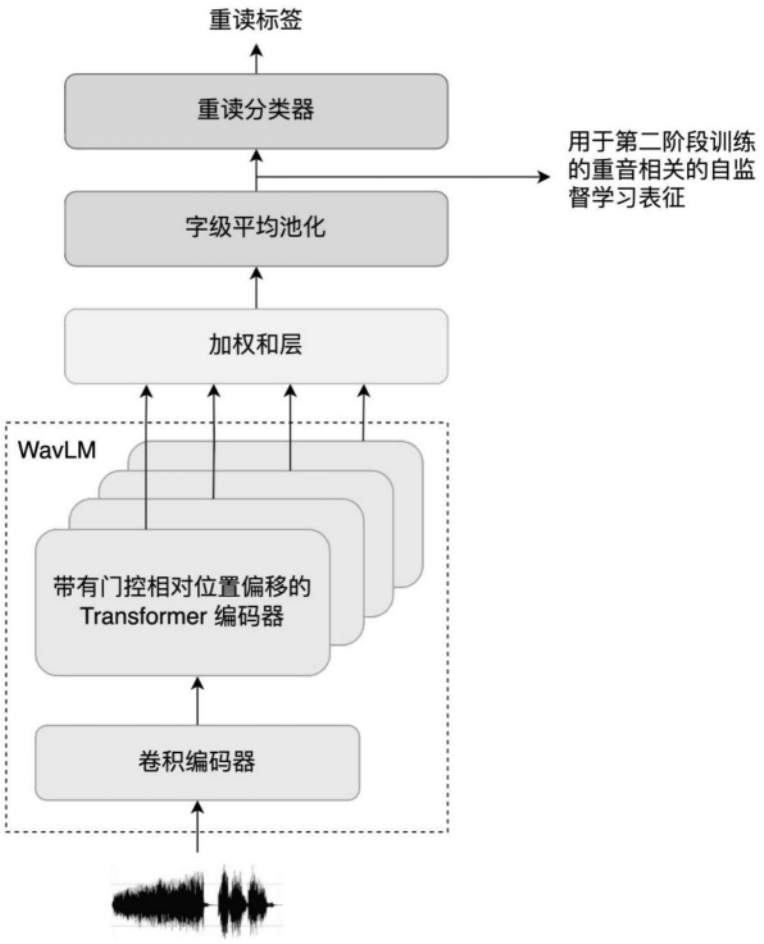
重读标签
重读分类器
用于第二阶段训练的重音相关的自监督学习表征
字级平均池化
加权和层
WavLM
带有门控相对位置偏移的Transformer 编码器
卷积编码器

图3

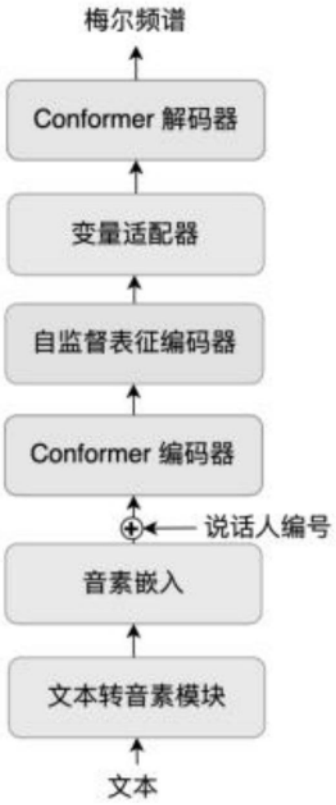
梅尔频谱
Conformer 解码器
变量适配器
自监督表征编码器
Conformer 编码器
说话人编号
音素嵌入
文本转音素模块
文本

图4

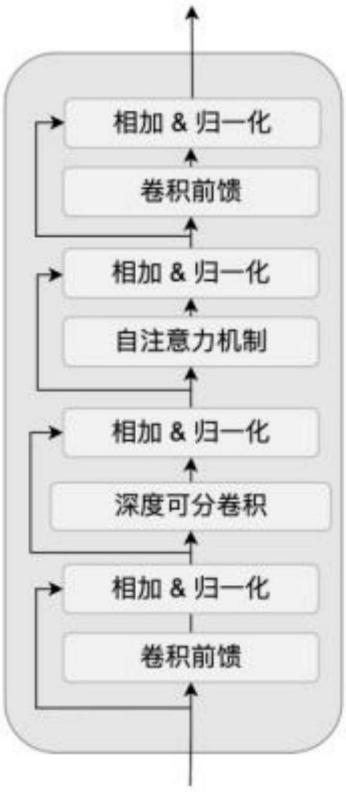

图5

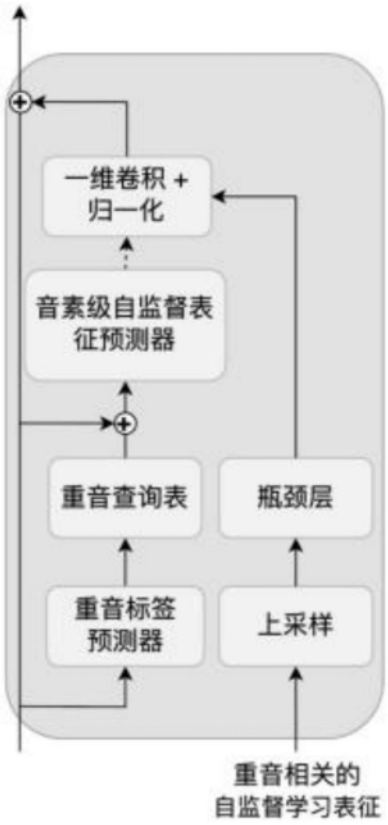
一维卷积 +
归一化
音素级自监督表
征预测器
重音查询表
瓶颈层
重音标签
预测器
上采样
重音相关的
自监督学习表征

图6

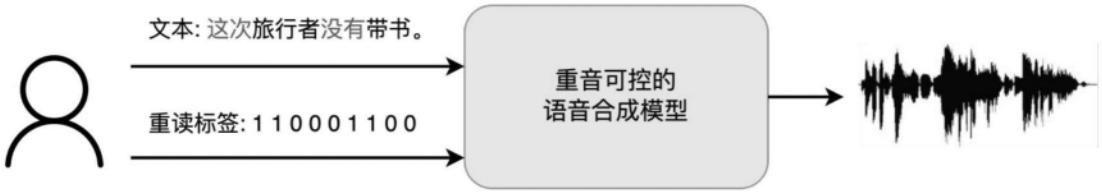
文本: 这次旅行者没有带书。
重读标签: 1 1 0 0 0 1 1 0 0
重音可控的
语音合成模型

图7


帧级表征提取单元
10
字级表征获取单元
20
文本隐层表征获取单元
30
预测特征获取单元
40
梅尔频谱表征获取单元
50

图8

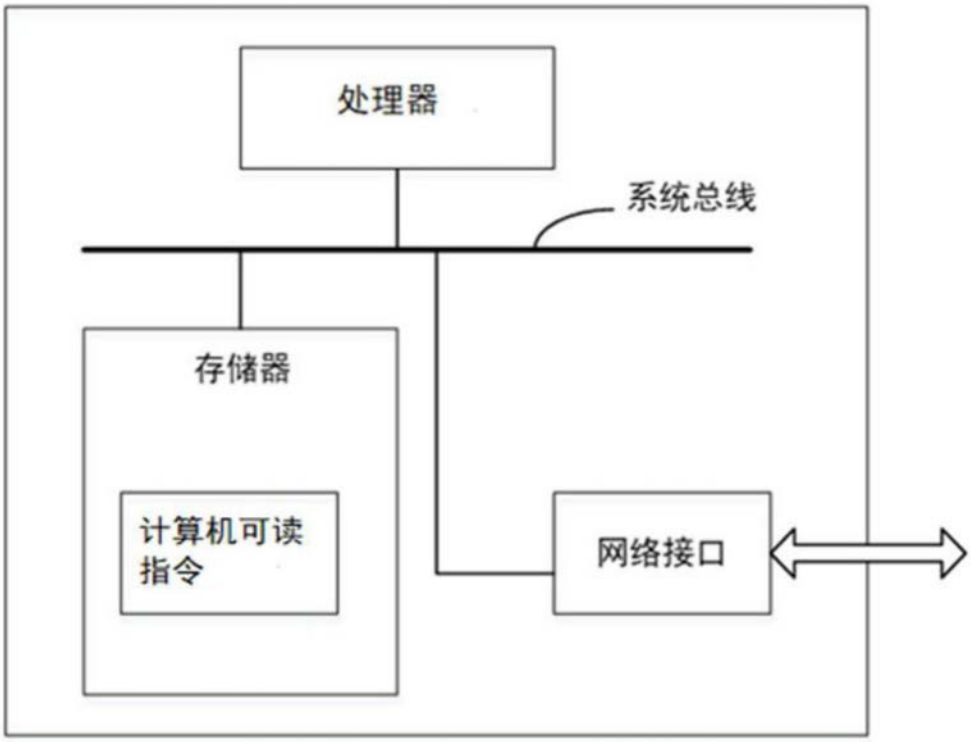

图9