(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114141237 B
(45)授权公告日 2025.05.06

(21)申请号 202111309392.6
(22)申请日 2021.11.06
(65)同一申请的已公布的文献号
申请公布号 CN 114141237 A
(43)申请公布日 2022.03.04
(73)专利权人 招联消费金融股份有限公司
地址 518000 广东省深圳市前海深港合作区前湾一路1号A栋201室(入驻深圳市前海商务秘书有限公司)
(72)发明人 詹维典 徐伟 林昊 张文锋 纳颖泉
(74)专利代理机构 华进联合专利商标代理有限公司 44224
专利代理师 伍健聪
(51)Int.Cl.
G10L 15/06(2013.01)
G10L 15/02(2006.01)
G10L 15/16(2006.01)
(56)对比文件
CN 116844529 A,2023.10.03
审查员 季英明

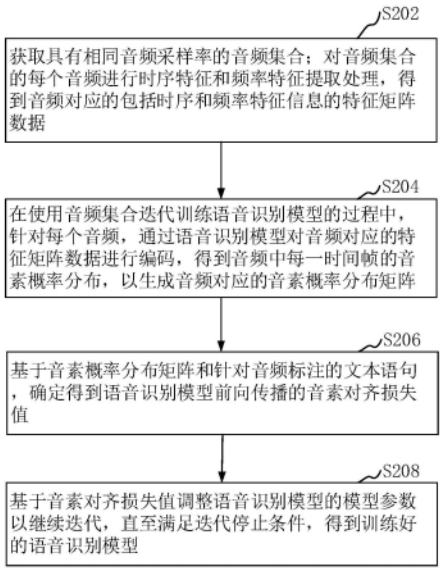
权利要求书2页 说明书12页 附图4页

(54)发明名称

语音识别方法、装置、计算机设备和存储介质

(57)摘要

本申请涉及一种语音识别方法、装置、计算机设备和存储介质。该方法包括:获取具有相同音频采样率的音频集合;对音频集合的每个音频进行时序特征和频率特征提取处理,得到音频对应的包括时序和频率特征信息的特征矩阵数据;在迭代训练语音识别模型的过程中,针对每个音频,通过语音识别模型对音频对应的特征矩阵数据进行编码,得到每一时间帧的音素概率分布,以生成音频对应的音素概率分布矩阵;基于音素概率分布矩阵和针对音频标注的文本语句,确定得到语音识别模型前向传播的音素对齐损失值;基于音素对齐损失值调整语音识别模型的模型参数以继续迭代,直至满足迭代停止条件,得到训练好的语音识别模型。采用本方法能够提高语音识别准确率。



CN 114141237 B

1.一种语音识别方法,其特征在于,所述方法包括:

获取具有相同音频采样率的音频集合;

对所述音频集合的每个音频进行时序特征和频率特征提取处理,得到所述音频对应的包括时序和频率特征信息的特征矩阵数据;

在使用所述音频集合迭代训练语音识别模型的过程中,针对每个音频,通过所述语音识别模型对所述音频对应的所述特征矩阵数据进行编码,得到所述音频中每一时间帧的音素概率分布,以生成所述音频对应的音素概率分布矩阵;

基于所述音素概率分布矩阵和针对所述音频标注的文本语句,确定得到所述语音识别模型前向传播的音素对齐损失值;

基于所述音素对齐损失值调整所述语音识别模型的模型参数以继续迭代,直至满足迭代停止条件,得到训练好的语音识别模型。

2.根据权利要求1所述的方法,其特征在于,在所述对所述音频集合的每个音频进行时序特征和频率特征提取处理之前,所述方法还包括:

从所述音频集合中随机抽取出部分音频;

针对随机抽取出的所述音频执行以下至少一种数据增强处理:

模拟不同发音人说话的声音大小之间的第一差异,针对随机抽取的所述音频,基于所述第一差异增强或减弱所述音频的音量;

模拟的不同发音人的说话语速之间的第二差异,针对随机抽取的所述音频,基于所述第二差异加快或减慢所述音频的语速;

模拟不同发音人在说话过程中语速韵律变化的第三差异,针对随机抽取的所述音频,基于所述第三差异在预设时间帧上扭曲音频波形数据;

模拟不同发音人的音色频率大小之间的第四差异,针对随机抽取的所述音频,基于所述第四差异在预设频率范围上扭曲音频频率。

3.根据权利要求1所述的方法,其特征在于,所述对所述音频集合的每个音频进行时序特征和频率特征提取处理,得到所述音频对应的包括时序和频率特征信息的特征矩阵数据包括:

基于多个具有三角滤波特征的带通滤波器,计算所述音频在各个时间帧上的不同频率的特征值,得到包括频率特征信息的梅尔频谱矩阵数据;

将所述梅尔频谱矩阵数据进行降维处理得到特征矩阵数据。

4.根据权利要求3所述的方法,其特征在于,待训练的所述语音识别模型包含多层卷积神经网络;所述将所述梅尔频谱矩阵数据进行降维处理得到特征矩阵数据包括:

将所述梅尔频谱矩阵数据输入至各层所述卷积神经网络,触发每层卷积神经网络做二维卷积计算,基于计算结果得到降维后的特征矩阵数据;

所述基于所述音素损失值调整所述语音识别模型的模型参数以继续迭代包括:

基于所述音素损失值调整所述语音识别模型的各层卷积神经网络的参数以继续迭代。

5.根据权利要求1所述的方法,其特征在于,所述针对每个音频,通过所述语音识别模型对所述音频对应的所述特征矩阵数据进行编码,得到所述音频中每一时间帧的音素概率分布,以生成所述音频对应的音素概率分布矩阵包括:

针对每个音频,将所述音频对应的所述特征矩阵数据输入所述语音识别模型的多层带

卷积网络的多头注意力模型,以基于所述多头注意力模型的每层依次针对该层所关注的信息进行特征提取;所述多头注意力模型中的每层关注的信息不同;

将所述多头注意力模型中各层所提取的特征进行拼接,得到所述音频中每一时间帧的音素概率分布;

根据所述音频中每一时间帧的音素概率分布,生成包括所有时间帧的音素概率分布矩阵。

6.根据权利要求1所述的方法,其特征在于,所述基于所述音素概率分布矩阵和针对所述音频标注的文本语句,确定得到所述语音识别模型前向传播的音素对齐损失值包括:

将所述音频标注的文本语句和所述音素概率分布矩阵进行音素对齐处理;

在执行音素对齐处理后,确定所述语音识别模型前向传播的音素对齐损失值。

7.一种语音识别装置,其特征在于,所述装置包括:

获取模板,用于获取具有相同音频采样率的音频集合;

特征计算模块,用于对所述音频集合的每个音频进行时序特征和频率特征提取处理,得到所述音频对应的包括时序和频率特征信息的特征矩阵数据;

损失值计算模块,用于在使用所述音频集合迭代训练语音识别模型的过程中,针对每个音频,通过所述语音识别模型对所述音频对应的所述特征矩阵数据进行编码,得到所述音频中每一时间帧的音素概率分布,以生成所述音频对应的音素概率分布矩阵;基于所述音素概率分布矩阵和针对所述音频标注的文本语句,确定得到所述语音识别模型前向传播的音素对齐损失值;

优化模块,用于基于所述音素对齐损失值调整所述语音识别模型的模型参数以继续迭代,直至满足迭代停止条件,得到训练好的语音识别模型。

8.一种计算机设备,包括存储器和处理器,所述存储器存储有计算机程序,其特征在于,所述处理器执行所述计算机程序时实现权利要求1至6中任一项所述的方法的步骤。

9.一种计算机可读存储介质,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权利要求1至6中任一项所述的方法的步骤。

10.一种计算机程序产品,包括计算机程序,其特征在于,该计算机程序被处理器执行时实现权利要求1至6任一项所述的方法的步骤。

语音识别方法、装置、计算机设备和存储介质

技术领域

[0001] 本申请涉及机器学习技术领域，特别是涉及一种语音识别方法、装置、计算机设备和存储介质。

背景技术

[0002] 随着机器学习技术的进步以及移动互联网的快速普及，计算机技术被广泛地运用到了社会的各个领域，随之而来的则是海量数据的产生。其中，语音识别技术即将进入工业、家电、通信、汽车电子、医疗、家庭服务、消费电子产品等各个领域，因此对语音识别技术有广泛的需求。

[0003] 目前的语音识别技术是通过基于概率统计的语音识别模型或循环神经网络模型进行训练得到的，不同的各个领域都需要采集大量的相关音频数据对语音识别模型进行训练。但是训练后的这两种模型识别的语义都不够流畅，识别准确率不够高。

发明内容

[0004] 基于此，有必要针对上述技术问题，提供一种能够提高语音识别准确率的语音识别方法、装置、计算机设备、存储介质和计算机程序产品，。

[0005] 第一方面，本申请提供了一种语音识别方法。所述方法包括：

[0006] 获取具有相同音频采样率的音频集合；

[0007] 对所述音频集合的每个音频进行时序特征和频率特征提取处理，得到所述音频对应的包括时序和频率特征信息的特征矩阵数据；

[0008] 在使用所述音频集合迭代训练语音识别模型的过程中，针对每个音频，通过所述语音识别模型对所述音频对应的所述特征矩阵数据进行编码，得到所述音频中每一时间帧的音素概率分布，以生成所述音频对应的音素概率分布矩阵；

[0009] 基于所述音素概率分布矩阵和针对所述音频标注的文本语句，确定得到所述语音识别模型前向传播的音素对齐损失值；

[0010] 基于所述音素对齐损失值调整所述语音识别模型的模型参数以继续迭代，直至满足迭代停止条件，得到训练好的语音识别模型。

[0011] 在其中一个实施例中，在所述对所述音频集合的每个音频进行时序特征和频率特征提取处理之前，所述方法还包括：

[0012] 从所述音频集合中随机抽取出部分音频；

[0013] 针对随机抽取出的所述音频执行以下至少一种数据增强处理：

[0014] 模拟不同发音人说话的声音大小之间的第一差异，针对随机抽取的所述音频，基于所述第一差异增强或减弱所述音频的音量；

[0015] 模拟的不同发音人的说话语速之间的第二差异，针对随机抽取的所述音频，基于所述第二差异加快或减慢所述音频的语速；

[0016] 模拟不同发音人在说话过程中语速韵律变化的第三差异，针对随机抽取的所述音

频,基于所述第三差异在预设时间帧上扭曲音频波形数据;

[0017] 模拟不同发音人的音色频率大小之间的第四差异,针对随机抽取的所述音频,基于所述第四差异在预设频率范围上扭曲音频频率。

[0018] 在其中一个实施例中,在所述对所述音频集合的每个音频进行时序特征和频率特征提取处理之前,所述方法还包括:

[0019] 统计在某业务场景下不同发音人的发音情况;

[0020] 对所述发音情况中不同发音人所产生的不同级别的声音大小、说话语速、语速韵律和音色频率提取各自对应的在所述业务场景下的出现概率;

[0021] 以达到相同的出现概率为目的、以全面覆盖为原则地选取音频集合中的音频执行音量扰动、速度扰动、时间扭曲和频率扭曲数据增强。

[0022] 在其中一个实施例中,所述对所述音频集合的每个音频进行时序特征和频率特征提取处理,得到所述音频对应的包括时序和频率特征信息的特征矩阵数据包括:

[0023] 基于多个具有三角滤波特征的带通滤波器,计算所述音频在各个时间帧上的不同频率的特征值,得到包括频率特征信息的梅尔频谱矩阵数据;

[0024] 将所述梅尔频谱矩阵数据进行降维处理得到特征矩阵数据。

[0025] 在其中一个实施例中,待训练的所述语音识别模型包含多层卷积神经网络;所述将所述梅尔频谱矩阵数据进行降维处理得到特征矩阵数据包括:

[0026] 将所述梅尔频谱矩阵数据输入至各层所述卷积神经网络,触发每层卷积神经网络做二维卷积计算,基于计算结果得到降维后的特征矩阵数据;

[0027] 所述基于所述音素损失值调整所述语音识别模型的模型参数以继续迭代包括:

[0028] 基于所述音素损失值调整所述语音识别模型的各层卷积神经网络的参数以继续迭代。

[0029] 在其中一个实施例中,所述针对每个音频,通过所述语音识别模型对所述音频对应的所述特征矩阵数据进行编码,得到所述音频中每一时间帧的音素概率分布,以生成所述音频对应的音素概率分布矩阵包括:

[0030] 针对每个音频,将所述音频对应的所述特征矩阵数据输入所述语音识别模型的多层带卷积网络的多头注意力模型,以基于所述多头注意力模型的每层依次针对该层所关注的信息进行特征提取;所述多头注意力模型中的每层关注的信息不同;

[0031] 将所述多头注意力模型中各层所提取的特征进行拼接,得到所述音频中每一时间帧的音素概率分布;

[0032] 根据所述音频中每一时间帧的音素概率分布,生成包括所有时间帧的音素概率分布矩阵。

[0033] 在其中一个实施例中,所述基于所述音素概率分布矩阵和针对所述音频标注的文本语句,确定得到所述语音识别模型前向传播的音素对齐损失值包括:

[0034] 将所述音频标注的文本语句和所述音素概率分布矩阵进行音素对齐;

[0035] 在执行音素对齐处理后,确定所述语音识别模型前向传播的音素对齐损失值。

[0036] 第二方面,本申请还提供了一种语音识别装置。所述装置包括:

[0037] 获取模板,用于获取具有相同音频采样率的音频集合;

[0038] 特征计算模块,用于对所述音频集合的每个音频进行时序特征和频率特征提取处

理,得到所述音频对应的包括时序和频率特征信息的特征矩阵数据;

[0039] 损失值计算模块,用于在使用所述音频集合迭代训练语音识别模型的过程中,针对每个音频,通过所述语音识别模型对所述音频对应的所述特征矩阵数据进行编码,得到所述音频中每一时间帧的音素概率分布,以生成所述音频对应的音素概率分布矩阵;基于所述音素概率分布矩阵和针对所述音频标注的文本语句,确定得到所述语音识别模型前向传播的音素对齐损失值;

[0040] 优化模块,用于基于所述音素对齐损失值调整所述语音识别模型的模型参数以继续迭代,直至满足迭代停止条件,得到训练好的语音识别模型。

[0041] 第三方面,本申请还提供了一种计算机设备。所述计算机设备包括存储器和处理器,所述存储器存储有计算机程序,所述处理器执行上述语音识别方法的步骤。

[0042] 第四方面,本申请还提供了一种计算机可读存储介质。所述计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行上述语音识别方法的步骤。

[0043] 第五方面,本申请还提供了一种计算机程序产品。所述计算机程序产品,包括计算机程序,该计算机程序被处理器执行上述语音识别方法的步骤。

[0044] 上述语音识别方法、装置、计算机设备、存储介质和计算机程序产品,获取具有相同音频采样率的音频集合;对音频集合的每个音频进行时序特征和频率特征提取处理,得到音频对应的包括时序和频率特征信息的特征矩阵数据。在使用音频集合迭代训练语音识别模型的过程中,针对每个音频,通过语音识别模型对音频对应的特征矩阵数据进行编码,得到音频中每一时间帧的音素概率分布,以生成音频对应的音素概率分布矩阵,可以使得语音识别模型对音频的识别精准到音素级别。基于音素概率分布矩阵和针对音频标注的文本语句,确定得到语音识别模型前向传播的音素对齐损失值。基于损失值调整语音识别模型的模型参数以继续迭代,直至满足迭代停止条件,得到训练好的语音识别模型。因此,在使用语言识别模型时,语音识别模型从音素级别的精度上对音频进行识别,提高了语义流畅度,提高了识别准确率。

附图说明

[0045] 图1为一个实施例中语音识别方法的应用环境图;

[0046] 图2为一个实施例中语音识别方法的流程示意图;

[0047] 图3为一个实施例中语音识别方法的原理示意图;

[0048] 图4为一个实施例中语音识别装置的结构框图;

[0049] 图5为一个实施例中特征计算模块的结构框图;

[0050] 图6为一个实施例中计算机设备的内部结构图。

具体实施方式

[0051] 为了使本申请的目的、技术方案及优点更加清楚明白,以下结合附图及实施例,对本申请进行进一步详细说明。应当理解,此处描述的具体实施例仅仅用以解释本申请,并不用于限定本申请。

[0052] 本申请提供的语音识别方法,可以应用于如图1所示的应用环境中。其中,终端110通过网络与服务器120进行通信。其中,终端110可以但不限于是各种个人计算机、笔记本电

脑、智能手机、平板电脑和便携式可穿戴设备，服务器120可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

[0053] 终端110可以将对音频集合的音频进行采样率转换，得到具有相同采样率的音频集合，并将该音频集合发送给服务器120。服务器120获取具有相同音频采样率的音频集合；对音频集合的每个音频进行时序特征和频率特征提取处理，得到音频对应的包括时序和频率特征信息的特征矩阵数据。服务器120在使用音频集合迭代训练语音识别模型的过程中，针对每个音频，通过语音识别模型对音频对应的特征矩阵数据进行编码，得到音频中每一时间帧的音素概率分布，以生成音频对应的音素概率分布矩阵。服务器120基于音素概率分布矩阵和针对音频标注的文本语句，确定得到语音识别模型前向传播的音素对齐损失值，并基于音素对齐损失值调整语音识别模型的模型参数以继续迭代，直至满足迭代停止条件，最终得到训练好的语音识别模型。服务器120使用训练好的语音识别模型，对终端110发送的语音进行识别，并将对应的识别结果发送给终端110。

[0054] 在一个实施例中，如图2所示，提供了一种语音识别方法，本实施例以该方法应用于服务器进行举例说明，可以理解的是，该方法也可以应用于终端，还可以应用于包括终端和服务器的系统，并通过终端和服务器的交互实现。本实施例中，该方法包括以下步骤：

[0055] S202，获取具有相同音频采样率的音频集合；对音频集合的每个音频进行时序特征和频率特征提取处理，得到音频对应的包括时序和频率特征信息的特征矩阵数据。

[0056] 其中，音频采样率是指在生成音频过程中，将音频的模拟信号转为数字信号所使用的采用频率，比如44,100赫兹或48,000赫兹等。特征矩阵数据包括了音频在各个时间帧上的不同频率的特征值以及时序的特征值。

[0057] 在一个实施例中，在对音频集合的每个音频进行时序特征和频率特征提取处理之前，服务器可以对音频集合的音频进行包括音量扰动、速度扰动、时间扭曲、频率扭曲和随机噪声等中的至少一种数据增强方法。

[0058] 在一个实施例中，针对音频集合的音频，服务器随机抽取音频进行数据增强。

[0059] 在一个实施例中，服务器可以基于音频在各个时间帧上的不同频率的特征值得到包括频率特征信息的梅尔频谱矩阵数据，并对梅尔频谱矩阵数据进行降维处理得到特征矩阵数据。

[0060] 在一个实施例中，服务器可以是基于语音识别模型的三层卷积神经网络层对梅尔频谱矩阵进行降维处理得到特征矩阵数据。

[0061] 具体地，服务器获取具有相同音频采样率的音频集合；对音频集合的每个音频在各个时间帧上进行时序特征和频率特征提取处理，得到音频对应的包括各个时间帧的时序和频率特征信息的特征矩阵数据。

[0062] S204，在使用音频集合迭代训练语音识别模型的过程中，针对每个音频，通过语音识别模型对音频对应的特征矩阵数据进行编码，得到音频中每一时间帧的音素概率分布，以生成音频对应的音素概率分布矩阵。

[0063] 其中，编码用于将对特征矩阵数据进行处理，以得到音频每一时间帧的音素概率分布。编码是通过语音识别模型的编码器实现的。

[0064] 其中，音素是指根据语音的自然属性划分出来的最小语音单位。音素概率分布是指音素取值的概率分布，音素概率分布矩阵是指音频在所有时间帧上的音素概率分布。

[0065] 具体地,服务器在使用音频集合迭代训练语音识别模型的过程中,可以针对每个音频,通过语音识别模型的编码器对音频对应的特征矩阵数据进行编码,得到音频中每一时间帧的音素概率分布,基于所有时间帧的音素概率分布生成音频对应的音素概率分布矩阵。

[0066] 在一个实施例中,语音识别模型为Transformer模型(一种引入注意力机制的模型)。在其他实施例中,语音识别模型还可以是其它能够获取音频在每一个时间帧上的音素概率分布值的模型。

[0067] 在一个实施例中,服务器可以基于Transformer模型的七层带卷积的多头注意力模型得到音频对应的音素概率分布矩阵。

[0068] S206,基于音素概率分布矩阵和针对音频标注的文本语句,确定得到语音识别模型前向传播的音素对齐损失值。

[0069] 其中,文本语句是音频所表述的内容的文字形式,音素对齐损失值是指将真实的文本音素标签和音素概率分布矩阵中与标签对应的所有正确路径对齐处理后所计算出来的损失值。

[0070] 在一个实施例中,服务器是通过神经网络的时序类分类器将音素概率分布矩阵和文本语句的音素执行音素对齐处理后计算得到的音素对齐损失值。

[0071] 在一个实施例中,服务器是通过CTC(Connectionist Temporal Classification,一种可以避开输入与输出人工对齐的、用于解决时序类数据的分类问题的算法)算法计算得到的音素对齐损失值。

[0072] 具体地,服务器使用能够执行音素对齐的算法模型,将音素概率分布矩阵和针对音频标注的文本语句执行音素对齐处理,得到语音识别模型前向传播的音素对齐损失值。

[0073] S208,基于音素对齐损失值调整语音识别模型的模型参数以继续迭代,直至满足迭代停止条件,得到训练好的语音识别模型。

[0074] 具体地,服务器根据音素对齐损失值,在语音识别模型的各个参数上求导,得到语音识别模型的更新梯度,从而对语音识别模型进行更新和参数学习。服务器循环地使用音频集合来迭代训练语音识别模型,直到音素损失值大小表现稳定,则停止训练,得到训练好的语音识别模型。

[0075] 在一个实施例中,服务器可以在每次迭代训练前对音频集合重新执行数据增强处理。

[0076] 在一个实施例中,在每次迭代训练过程中,服务器可以逐渐减低语音识别模型的学习率。

[0077] 在一个实施例中,在每次迭代训练过程中,服务器可以逐渐减低语音识别模型中的多头注意力模型的学习率。

[0078] 在一个实施例中,在每次迭代训练过程中,服务器可以逐渐减低语音识别模型中的三层卷积神经网络层的学习率。

[0079] 在一个实施例中,语音识别模型的学习率随着训练步数的增加逐渐降低。

[0080] 在一个实施例中,语音识别模型的学习率基于当前训练步数的倒数来更新。

[0081] 上述语音识别方法、装置、计算机设备和存储介质,获取具有相同音频采样率的音频集合;对音频集合的每个音频进行时序特征和频率特征提取处理,得到音频对应的包括

时序和频率特征信息的特征矩阵数据。在使用音频集合迭代训练语音识别模型的过程中，针对每个音频，通过语音识别模型对音频对应的特征矩阵数据进行编码，得到音频中每一时间帧的音素概率分布，以生成音频对应的音素概率分布矩阵，可以使得语音识别模型对音频的识别精准到音素级别。基于音素概率分布矩阵和针对音频标注的文本语句，确定得到语音识别模型前向传播的音素对齐损失值。基于损失值调整语音识别模型的模型参数以继续迭代，直至满足迭代停止条件，得到训练好的语音识别模型。因此，在使用语言识别模型时，语音识别模型从音素级别的精度上对音频进行识别，提高了语义流畅度，提高了识别准确率。

[0082] 在一个实施例中，在对音频集合的每个音频进行时序特征和频率特征提取处理之前，方法还包括：从所述音频集合中随机抽取出部分音频；针对随机抽取出的所述音频执行以下至少一种数据增强处理：模拟不同发音人说话的声音大小之间的第一差异，针对随机抽取的所述音频，基于所述第一差异增强或减弱所述音频的音量；模拟的不同发音人的说话语速之间的第二差异，针对随机抽取的所述音频，基于所述第二差异加快或减慢所述音频的语速；模拟不同发音人在说话过程中语速韵律变化的第三差异，针对随机抽取的所述音频，基于所述第三差异在预设时间帧上扭曲音频波形数据；模拟不同发音人的音色频率大小之间的第四差异，针对随机抽取的所述音频，基于所述第四差异在预设频率范围上扭曲音频频率。

[0083] 具体地，服务器在对音频集合的每个音频进行时序特征和频率特征提取处理之前，随机抽取出音频集合的部分音频，针对部分音频执行包括音量扰动、速度扰动、时间扭曲、频率扭曲和随机噪声等中的至少一中数据增强处理。

[0084] 在一个实施例中，模拟第一差异、第二差异、第三差异和第四差异的过程包括对相应的指标对象进行分级，获取不同级别的指标对象之间的差异后，将差异特征模拟到音频的数据增强过程中，其中，指标对象分别对应为声音大小、语速、语速韵律和音色频率。比如，模拟不同发音人说话的声音大小之间的第一差异包括将不同发音人说话声音大小进行分级，获取不同级别的声音大小之间的第一差异，将第一差异特征模拟到音频的音量数据增强过程中。

[0085] 在一个实施例中，音量扰动的过程包括：服务器可以将不同发音人说话声音大小进行分级，获取不同级别的声音大小之间的第一差异，通过第一差异来增加或者减弱音频的音量。可以理解，通过对音量的增加和减弱可以增强语音识别模型对不同的音量音频的识别能力。

[0086] 在一个实施例中，速度扰动的过程包括：服务器可以将不同发音人说话语速大小进行分级，获取不同级别的语速之间的第二差异，通过第二差异来增加或者减弱音频的语速。可以理解，通过对语速的增加和减弱可以增强语音识别模型对不同语速的识别能力。

[0087] 在一个实施例中，时间扭曲的过程包括：服务器可以将在说话过程中语速韵律大小进行分级，获取不同级别的语速韵律之间的第三差异，基于第三差异在一个或多个随机片段的时间帧上对音频波形数据做非线性扭曲。可以理解，通过对音频波形的扭曲可以增强语音识别模型对于语速韵律变化情况的识别能力。

[0088] 在一个实施例中，频率扭曲的过程包括：服务器可以将不同发音人的音色频率大小进行分级，获取不同级别的音色频率之间的第四差异，基于第四差异在预设频率范围上

扭曲音频频率。可以理解,通过对频率的扭曲可以增强语音识别模型对于不同音色的识别能力。

[0089] 在一个实施例中,随机噪声的过程包括:服务器可以从既有的背景噪声库中随机取一段背景噪声叠加到音频数据上。可以理解,通过对音频叠加噪声可以增强语音识别模型在噪声下的抗干扰能力。

[0090] 在本实施例中,通过随机地对音频进行包括音量扰动、速度扰动、时间扭曲、频率扭曲和随机噪声等中的至少一种数据增强处理,以得到更丰富的音频,从而可以得到更为丰富的音频特征矩阵数据用于训练语音识别模型,以提高语音识别模型的泛化性能和识别准确率。并且,在本实施例中,由于对音频执行了多样化的数据增强,以使得基于小规模的数据集也能得到规模增大的多样化数据集,从而能提高模型泛化性能和识别准确性。

[0091] 在一个实施例中,在对音频集合的每个音频进行时序特征和频率特征提取处理之前,方法还包括:计算在目标业务场景下目标发音人的语音的第一发音范围特征;该第一发音范围特征包括不同级别的声音大小、说话语速、语速韵律和音色频率在目标发音人的语音中的所出现的目标概率;针对音频集合的每个音频,对音频执行音量扰动、速度扰动、时间扭曲和频率扭曲数据增强,以使得该音频和对应增强后的各个音频的第二发音范围特征和第一发音范围特征相同。

[0092] 其中,目标发音人可以是一个人或者多个人。目标发音人的语音包括多个音频,是目标业务场景下具有不同级别的声音大小、说话语速、语速韵律和音色频率的最小集合。

[0093] 在一个实施例中,不同级别的声音大小、说话语速、语速韵律和音色频率是基于对声音大小、说话语速、语速韵律和音色频率进行分级上得到的。比如在声音大小上,对1分贝的声音音量设置为1级,10分贝的声音音量设置为2级、18分贝的声音音量设置为3级等等。

[0094] 在本实施例中,通过计算在目标业务场景下目标发音人的语音的第一发音范围特征;针对音频集合的每个音频,对音频执行音量扰动、速度扰动、时间扭曲和频率扭曲数据增强,以使得该音频和对应增强后的各个音频的第二发音范围特征和第一发音范围特征相同,以得到更丰富和全面的音频,从而可以得到更为全面的音频特征矩阵数据用于训练语音识别模型,以提高语音识别模型的泛化性能和识别准确率。并且在语音采集困难或者语音采集成本较高的情况下,可以对小数据量的音频集合进行数据增强就可以获取丰富和全面的音频集合,降低人工采集成本并解决语音采集困难的问题。

[0095] 在一个实施例中,对音频集合的每个音频进行时序特征和频率特征提取处理,得到音频对应的包括时序和频率特征信息的特征矩阵数据包括:基于多个具有三角滤波特征的带通滤波器,计算音频在各个时间帧上的不同频率的特征值,得到包括频率特征信息的梅尔频谱矩阵数据;将梅尔频谱矩阵数据进行降维处理得到特征矩阵数据。

[0096] 其中,带通滤波器是一个允许特定音频频段的波通过同时屏蔽其他音频频段的设备。梅尔频谱矩阵数据用于表征基于梅尔频谱对音频进行处理得到的矩阵数据。

[0097] 具体地,服务器在语音的频谱范围内设置多个带通滤波器,每个滤波器具有三角滤波特性,其中心频率均匀分布在人耳感知频率范围内。服务器基于每个滤波器计算音频在各个时间帧上的不同频率的特征值,得到包括频率特征信息的梅尔频谱矩阵数据,服务器对梅尔频谱矩阵数据进行降为处理得到特征矩阵数据。

[0098] 在一个实施例中,滤波器个数可以是为80。

[0099] 在本实施例中,通过计算音频在各个时间帧上的不同频率的特征值,得到包括频率特征信息的梅尔频谱矩阵数据,并将梅尔频谱矩阵数据进行降维处理得到特征矩阵数据。这样服务器就能得到包括音频频率特征信息的数据,并对数据进行降维,从而得到信息全面而且数据量较小的特征矩阵数据,减少语音识别模型的训练负荷,并提高准确率。

[0100] 在一个实施例中,待训练的语音识别模型包含多层卷积神经网络;将梅尔频谱矩阵数据进行降维处理得到特征矩阵数据包括:将梅尔频谱矩阵数据输入至各层卷积神经网络,触发每层卷积神经网络做二维卷积计算,基于计算结果得到降维后的特征矩阵数据;基于音素对齐损失值调整语音识别模型的模型参数以继续迭代包括:基于音素对齐损失值调整语音识别模型的各层卷积神经网络的参数以继续迭代。

[0101] 具体地,待训练地语音识别模型包括多层卷积神经网络,每层卷积神经网络可以做二维卷积计算。服务器将梅尔频谱矩阵数据输入语音识别模型的卷积神经网络,每层卷积神经网络对输入的梅尔频谱矩阵数据进行二维卷积计算,基于计算结果得到了降维后的特征矩阵数据。

[0102] 具体地,服务器在通过步骤S206得到音素对齐损失指后,还可以基于音素对齐损失指调整语音识别模型地各层卷积神经网络的参数。

[0103] 在一个实施例中,服务器可以对语音识别模型的卷积神经网络的卷积核设置(3,3),步进设置(2,2)。

[0104] 在本实施例中,通过对梅尔频谱矩阵数据输入到语音识别模型的各层卷积神经网络进行降维处理,并基于音素损失值来调整各层卷积神经网络的参数,以满足迭代训练的要求,从而提高语音识别模型的准确率。

[0105] 在一个实施例中,针对每个音频,通过语音识别模型对音频对应的特征矩阵数据进行编码,得到音频中每一时间帧的音素概率分布,以生成音频对应的音素概率分布矩阵包括:针对每个音频,将音频对应的特征矩阵数据输入语音识别模型的多层带卷积网络的多头注意力模型,以基于多头注意力模型的每层依次针对该层所关注的信息进行特征提取;多头注意力模型中的每层关注的信息不同;将多头注意力模型中各层所提取的特征进行拼接,得到音频中每一时间帧的音素概率分布;根据音频中每一时间帧的音素概率分布,生成包括所有时间帧的音素概率分布矩阵。

[0106] 其中,多头注意力模型是指对输入的梅尔频谱矩阵数据具有多种不同注意力处理、并相应的具有多层带卷积网络的模型。

[0107] 具体地,语音识别模型包括多头注意力模型,多头注意力模型中的每层关注的信息不同。服务器针对每个音频,将音频对应的特征矩阵数据输入语音识别模型的多层带卷积网络的多头注意力模型,以基于多头注意力模型的每层依次执行包括层归一化、前馈网络、卷积、多头注意力、前馈网络的计算,从而对该层所关注的信息进行特征提取,并对所提取的特征进行拼接,得到音频中每一时间帧的音素概率分布。服务器根据音频中每一时间帧的音素概率分布,生成包括所有时间帧的音素概率分布矩阵。

[0108] 在一个实施例中,语音识别模型包括了七层带卷积网络的多头注意力模型。

[0109] 在一个实施例中,多头注意力模型的每层所执行的层归一化的计算用于对输入本层的所有神经元的数据做归一化操作。

[0110] 在一个实施例中,多头注意力模型的每层所执行的前馈网络计算是用于实现层之

间的线性连接。

[0111] 在一个实施例中,多头注意力模型的每层所执行的卷积计算是基于卷积核设置为32、步进设置为1的一维卷积进行的。

[0112] 在一个实施例中,多头注意力模型的每层所执行的多头注意力计算是选取多个特征信息做平行地计算。

[0113] 在本实施例中,使用了每层关注的信息不同的多头注意力模型,来对所关注的信息进行特征提取,并对各层所提取的特征进行拼接,得到音频中每一时间帧的音素概率分布,从而生成包括所有时间帧的音素概率分布矩阵,进而可以生成准确而全面的音素概率分别矩阵,提高语音识别模型的准确率。并且,在本实施例中,是按照每一时间帧进行处理,从而实现顺序编码,实现对流式语音识别。

[0114] 在一个实施例中,基于音素概率分布矩阵和针对音频标注的文本语句,确定得到语音识别模型前向传播的音素对齐损失值包括:将音频标注的文本语句和音素概率分布矩阵进行音素对齐;在执行音素对齐处理后,确定语音识别模型前向传播的音素对齐损失值。

[0115] 具体地,服务器计算音频标注的文本语句和音素概率分布矩阵之间的损失值过程中,执行音素对齐处理,从而算出音素对齐损失值。

[0116] 在一个实施例中,服务器可以是基于CTC算法来处理音频标注的文本语句和音素概率分布矩阵,执行音素对齐处理,从而计算出音素对齐损失值。

[0117] 在本实施例中,通过对音频标注的文本语句和音素概率分布矩阵进行音素对齐;在执行音素对齐处理后,确定语音识别模型前向传播的音素对齐损失值。这样使得损失值的计算结果更为精准,以使得在基于音素对齐损失值调整语音识别模型参数时,能具备更准确的参考标准对象,从而提高语音识别模型的准确率。

[0118] 在一个实施例中,如图3所示,提供了语音识别方法的原理示意图。具体地,服务器对音频集合中的音频进行重采样,以获取具有相同采样率的音频。服务器还可以对音频集合中的音频做数据增强,数据增强的处理主要包括:音量扰动、速度扰动、时间扭曲、频率扭曲、随机噪声等中的至少一种,以提高音频数据的多样性和模型的泛化能力,并使得通过小数据量的音频集合就可以获取多样化全面化的数据集合。服务器对音频集合中的音频进行特征数据提取,并通过语音识别模型的三层卷积神经网络层进行下采样,得到音频对应的包括时序和频率特征信息的特征矩阵数据,以使得服务器在生成符合预设要求的训练数据的同时,能减少语音识别模型的系统负荷。服务器还将特征矩阵数据输入语音识别模型的七层带卷积网络的多头注意力模型(附图只画出一层,用7x表示7层),以使得每层依次执行包括层归一化、前馈网络、卷积、多头注意力、前馈网络的计算,每层提取不同的关注的信息,并进行拼接后最终生成音素概率分布矩阵。服务器通过使用时序类分类器对音素概率分布矩阵和该音频所标注的文本语句执行音素对齐处理,可以理解,时序类分类器可以是CTC算法。服务器基于时序类分类器得到音素对齐损失值后,基于该音素对齐损失值来调整语音模型的参数,迭代训练语音识别模型。

[0119] 应该理解的是,虽然本申请部分实施例中的流程图中的各个步骤按照箭头的指示依次显示,但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明,这些步骤的执行并没有严格的顺序限制,这些步骤可以以其它的顺序执行。而且,流程图中的至少一部分步骤可以包括多个步骤或者多个阶段,这些步骤或者阶段并不必然是

在同一时刻执行完成,而是可以在不同的时刻执行,这些步骤或者阶段的执行顺序也不必然是依次进行,而是可以与其它步骤或者其它步骤中的步骤或者阶段的至少一部分轮流或者交替地执行。

[0120] 基于同样的发明构思,本申请实施例还提供了一种用于实现上述所涉及的语音识别方法的语音识别装置。该装置所提供的解决问题的实现方案与上述方法中所记载的实现方案相似,故下面所提供的一个或多个语音识别装置实施例中的具体限定可以参见上文中对于语音识别方法的限定,在此不再赘述。

[0121] 在一个实施例中,如图4所示,提供了一种语音识别装置400,包括:获取模块402、特征计算模块404、损失值计算模块406和优化模块408,其中:

[0122] 获取模块402,用于获取具有相同音频采样率的音频集合。

[0123] 特征计算模块404,用于对音频集合的每个音频进行时序特征和频率特征提取处理,得到音频对应的包括时序和频率特征信息的特征矩阵数据。

[0124] 损失值计算模块406,用于在使用音频集合迭代训练语音识别模型的过程中,针对每个音频,通过语音识别模型对音频对应的特征矩阵数据进行编码,得到音频中每一时间帧的音素概率分布,以生成音频对应的音素概率分布矩阵;基于音素概率分布矩阵和针对音频标注的文本语句,确定得到语音识别模型前向传播的音素对齐损失值。

[0125] 优化模块408,用于基于音素对齐损失值调整语音识别模型的模型参数以继续迭代,直至满足迭代停止条件,得到训练好的语音识别模型。

[0126] 在一个实施例中,在对音频集合的每个音频进行时序特征和频率特征提取处理之前,特征计算模块404还用于:从所述音频集合中随机抽取出部分音频;针对随机抽取出的所述音频执行以下至少一种数据增强处理:模拟不同发音人说话的声音大小之间的第一差异,针对随机抽取的所述音频,基于所述第一差异增强或减弱所述音频的音量;模拟的不同发音人的说话语速之间的第二差异,针对随机抽取的所述音频,基于所述第二差异加快或减慢所述音频的语速;模拟不同发音人在说话过程中语速韵律变化的第三差异,针对随机抽取的所述音频,基于所述第三差异在预设时间帧上扭曲音频波形数据;模拟不同发音人的音色频率大小之间的第四差异,针对随机抽取的所述音频,基于所述第四差异在预设频率范围上扭曲音频频率。

[0127] 在一个实施例中,如图5所示,特征计算模块404包括:特征提取模块404a、和降维模块404b,其中:

[0128] 特征提取模块404a用于基于多个具有三角滤波特征的带通滤波器,计算音频在各个时间帧上的不同频率的特征值,得到包括时序和频率特征信息的梅尔频谱矩阵数据。

[0129] 降维模块404b用于将梅尔频谱矩阵数据进行降维处理得到特征矩阵数据。

[0130] 在一个实施例中,待训练的语音识别模型包含多层卷积神经网络;损失值计算模块406还用于:将梅尔频谱矩阵数据输入至各层卷积神经网络,触发每层卷积神经网络做二维卷积计算,基于计算结果得到降维后的特征矩阵数据;优化模块408还用于基于音素损失值调整语音识别模型的各层卷积神经网络的参数以继续迭代。

[0131] 在一个实施例中,损失值计算模块406还用于针对每个音频,将音频对应的特征矩阵数据输入语音识别模型的多层带卷积网络的多头注意力模型,以基于多头注意力模型的每层依次针对该层所关注的信息进行特征提取;多头注意力模型中的每层关注的信息不

同;将多头注意力模型中各层所提取的特征进行拼接,得到音频中每一时间帧的音素概率分布;根据音频中每一时间帧的音素概率分布,生成包括所有时间帧的音素概率分布矩阵。

[0132] 在一个实施例中,损失值计算模块406还用于:将音频标注的文本语句和音素概率分布矩阵进行音素对齐;在执行音素对齐处理后,确定语音识别模型前向传播的音素对齐损失值。

[0133] 上述语音识别装置,获取具有相同音频采样率的音频集合;对音频集合的每个音频进行时序特征和频率特征提取处理,得到音频对应的包括时序和频率特征信息的特征矩阵数据。在使用音频集合迭代训练语音识别模型的过程中,针对每个音频,通过语音识别模型对音频对应的特征矩阵数据进行编码,得到音频中每一时间帧的音素概率分布,以生成音频对应的音素概率分布矩阵,可以使得语音识别模型对音频的识别精准到音素级别。基于音素概率分布矩阵和针对音频标注的文本语句,确定得到语音识别模型前向传播的音素对齐损失值。基于损失值调整语音识别模型的模型参数以继续迭代,直至满足迭代停止条件,得到训练好的语音识别模型。因此,在使用语言识别模型时,语音识别模型从音素级别的精度上对音频进行识别,提高了语义流畅度,提高了识别准确率。

[0134] 关于上述语音识别的具体限定可以参见上文中对于上述语音识别方法的限定,在此不再赘述。上述语音识别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。

[0135] 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服务器,其内部结构图可以如图6所示。该计算机设备包括通过系统总线连接的处理器、存储器和网络接口。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非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于存储音频集合数据。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。该计算机程序被处理器执行时以实现一种语音识别方法。

[0136] 本领域技术人员可以理解,图6中示出的结构,仅仅是与本申请方案相关的部分结构的框图,并不构成对本申请方案所应用于其上的计算机设备的限定,具体的计算机设备可以包括比图中所示更多或更少的部件,或者组合某些部件,或者具有不同的部件布置。

[0137] 在一个实施例中,还提供了一种计算机设备,包括存储器和处理器,存储器中存储有计算机程序,该处理器执行计算机程序时实现上述各方法实施例中的步骤。

[0138] 在一个实施例中,提供了一种计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现上述各方法实施例中的步骤。

[0139] 在一个实施例中,提供了一种计算机程序产品,包括计算机程序,该计算机程序被处理器执行时实现上述各方法实施例中的步骤。

[0140] 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,是可以通过计算机程序来指令相关的硬件来完成,的计算机程序可存储于一非易失性计算机可读取存储介质中,该计算机程序在执行时,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易失性存储器中的至少一种。非易失性存储器可包括只读存储器(Read-Only

Memory,ROM)、磁带、软盘、闪存或光存储器等。易失性存储器可包括随机存取存储器(Random Access Memory,RAM)或外部高速缓冲存储器。作为说明而非局限,RAM可以是多种形式,比如静态随机存取存储器(Static Random Access Memory,SRAM)或动态随机存取存储器(Dynamic Random Access Memory,DRAM)等。

[0141] 以上实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

[0142] 以上实施例仅表达了本申请的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对发明专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本申请构思的前提下,还可以做出若干变形和改进,这些都属于本申请的保护范围。因此,本申请专利的保护范围应以所附权利要求为准。

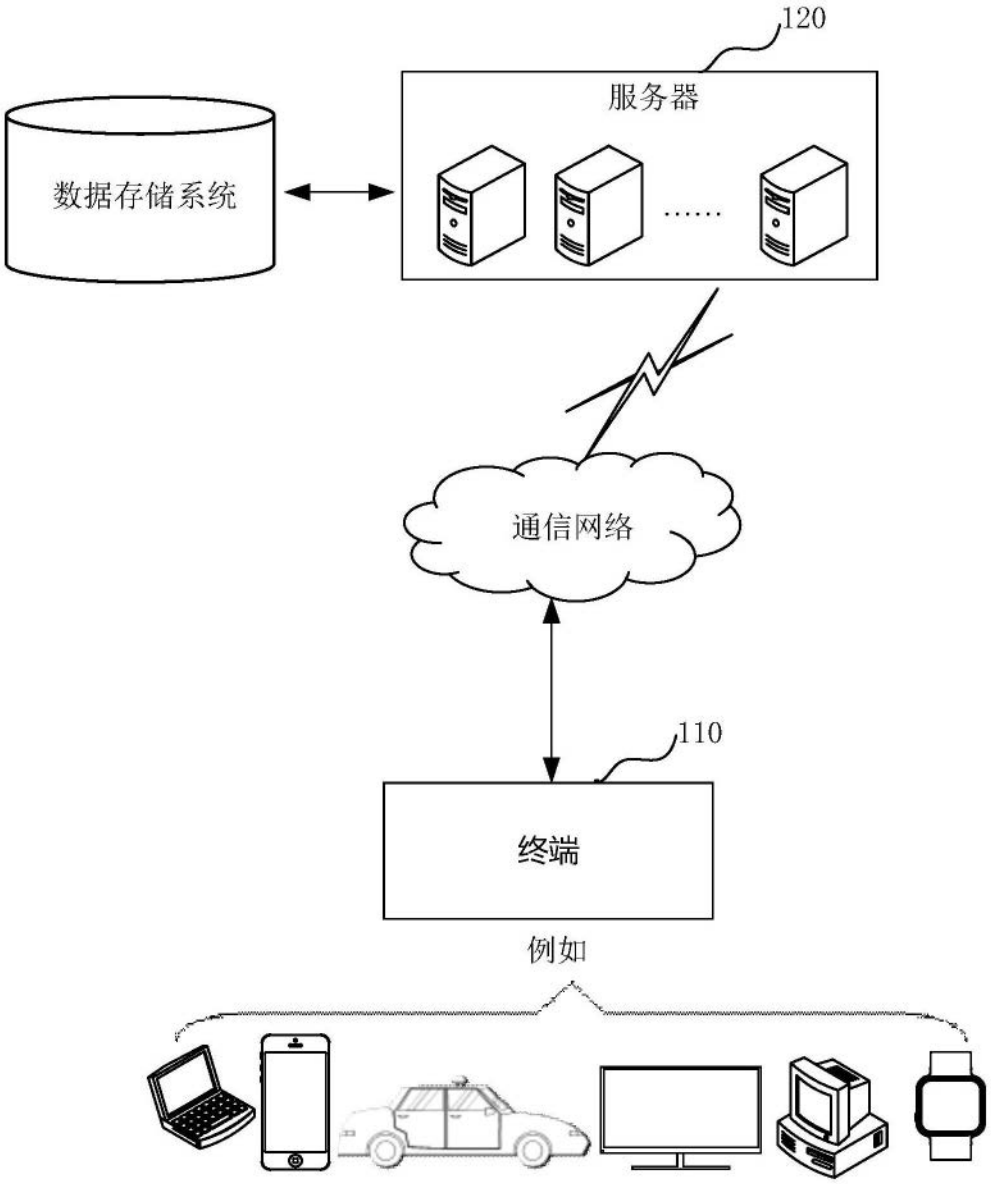


图1

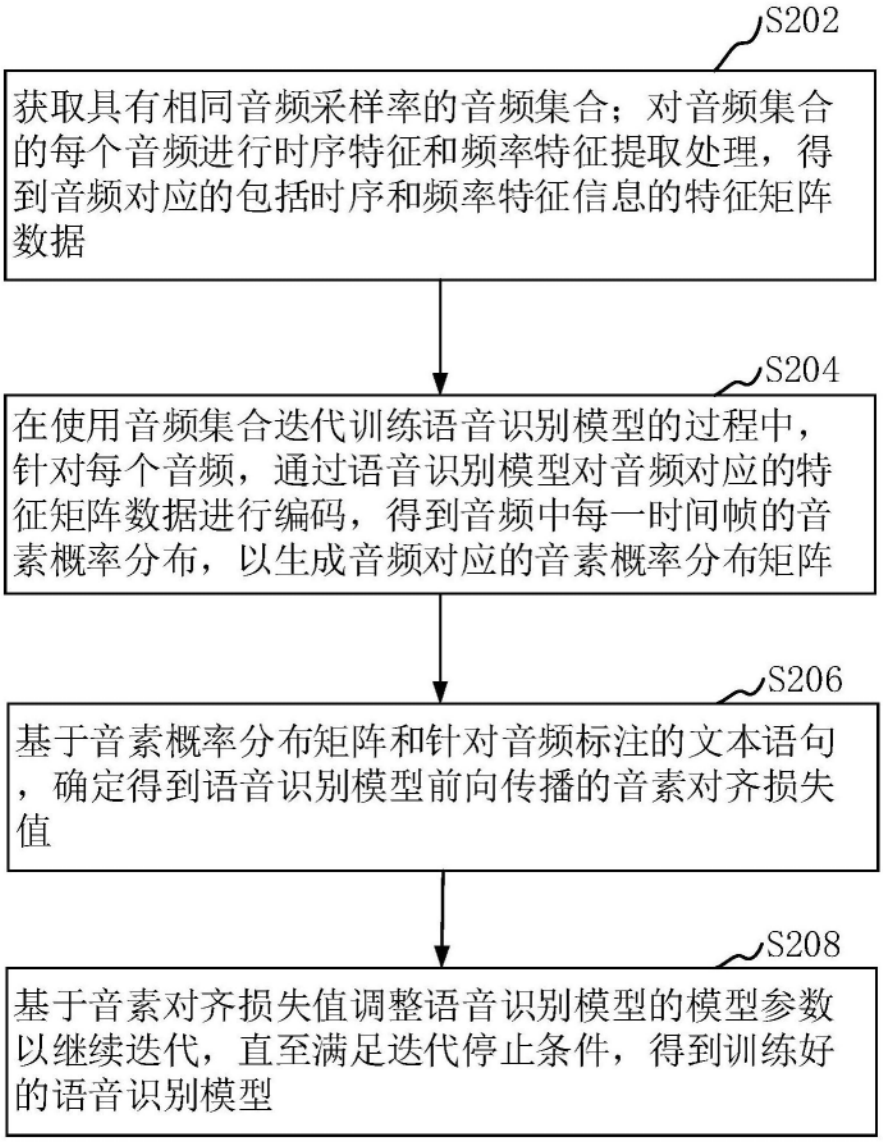

S202
获取具有相同音频采样率的音频集合；对音频集合的每个音频进行时序特征和频率特征提取处理，得到音频对应的包括时序和频率特征信息的特征矩阵数据
S204
在使用音频集合迭代训练语音识别模型的过程中，针对每个音频，通过语音识别模型对音频对应的特征矩阵数据进行编码，得到音频中每一时间帧的音素概率分布，以生成音频对应的音素概率分布矩阵
S206
基于音素概率分布矩阵和针对音频标注的文本语句，确定得到语音识别模型前向传播的音素对齐损失值
S208
基于音素对齐损失值调整语音识别模型的模型参数以继续迭代，直至满足迭代停止条件，得到训练好的语音识别模型

图2

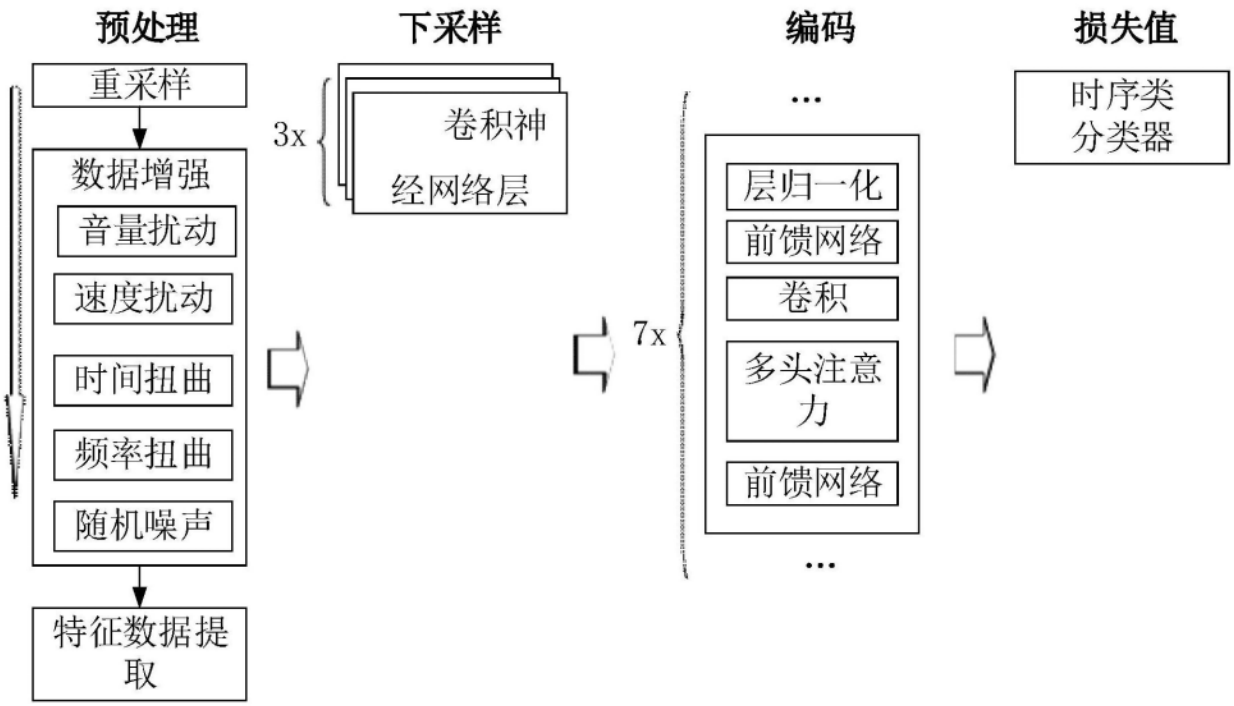

预处理
下采样
编码
损失值
重采样
数据增强
音量扰动
速度扰动
时间扭曲
频率扭曲
随机噪声
特征数据提取
3x
卷积神经网络层
...
层归一化
前馈网络
卷积
多头注意力
前馈网络
7x
...
时序类分类器

图3

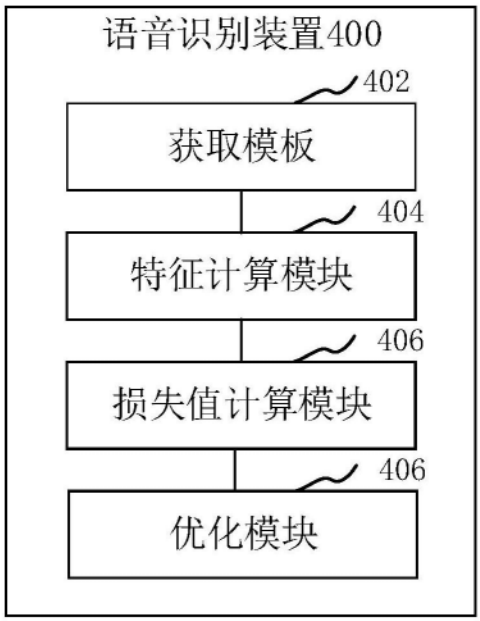

语音识别装置400
402
获取模板
404
特征计算模块
406
损失值计算模块
406
优化模块

图4

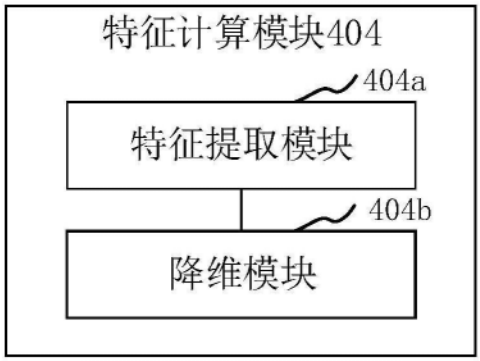

特征计算模块404
404a
特征提取模块
404b
降维模块

图5

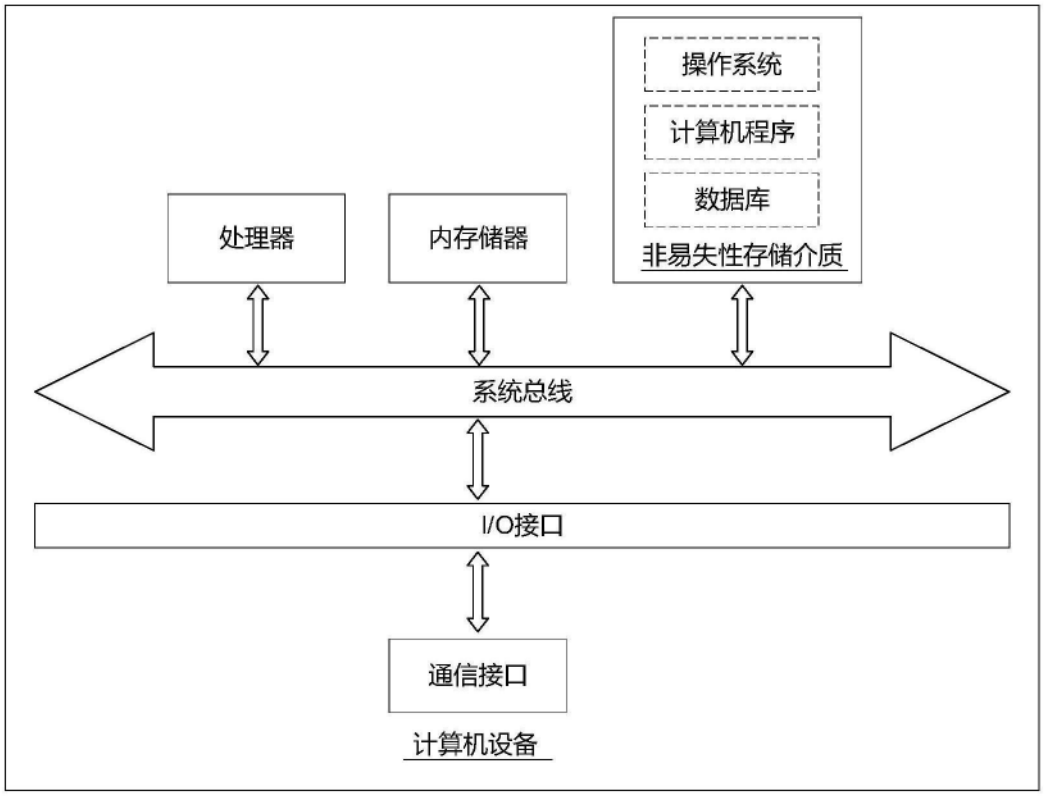

操作系统
计算机程序
数据库
非易失性存储介质
处理器
内存储器
系统总线
I/O接口
通信接口
计算机设备

图6