(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 113689846 B
(45)授权公告日 2022.02.08

(21)申请号 202111251829.5
(22)申请日 2021.10.27
(65)同一申请的已公布的文献号
申请公布号 CN 113689846 A
(43)申请公布日 2021.11.23
(73)专利权人 深圳市友杰智新科技有限公司
地址 518000 广东省深圳市南山区招商街道沿山社区沿山路22号火炬大厦501
(72)发明人 李杰 王广新 杨汉丹
(74)专利代理机构 深圳市明日今典知识产权代理事务所(普通合伙) 44343
代理人 王杰辉 曹勇
(51)Int.Cl.
G10L 15/02(2006.01)
G10L 15/06(2013.01)
G10L 15/16(2006.01)

(56)对比文件
CN 113327603 A,2021.08.31
CN 113129868 A,2021.07.16
CN 111564164 A,2020.08.21
US 2020402500 A1,2020.12.24
US 10777186 B1,2020.09.15
CN 111696526 A,2020.09.22

审查员 李国丽

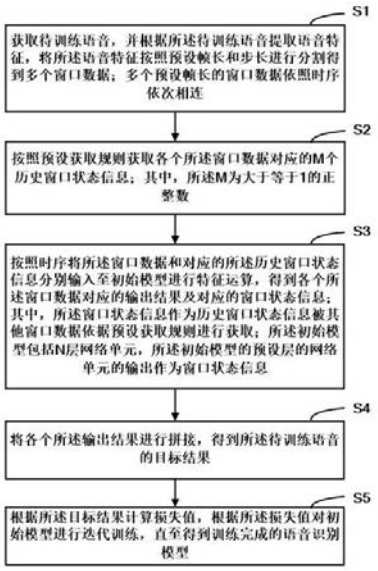
权利要求书2页 说明书10页 附图3页

(54)发明名称

语音识别模型训练方法、装置、计算机设备和存储介质

(57)摘要

本申请涉及人工智能领域，提供了一种语音识别模型训练方法、装置、计算机设备和存储介质，获取待训练语音，并根据待训练语音提取语音特征，将语音特征按照预设帧长和步长进行分割得到多个窗口数据；按照预设获取规则获取各个窗口数据对应的M个历史窗口状态信息；按照时序将窗口数据和对应的历史窗口状态信息分别输入初始模型进行特征运算得到各个窗口数据对应的输出结果及对应的窗口状态信息；将各个输出结果进行拼接得到待训练语音的目标结果；根据目标结果计算损失值，根据损失值对初始模型进行迭代训练，直至得到训练完成的语音识别模型。通过本申请提供的语音识别模型训练方法，训练得到的语音识别模型能够更加准确地进行语音识别。



CN 113689846 B

1.一种语音识别模型训练方法,其特征在于,包括以下步骤:

获取待训练语音,并根据所述待训练语音提取语音特征,将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据;多个预设帧长的窗口数据依照时序依次相连;

按照预设获取规则获取各个所述窗口数据对应的M个历史窗口状态信息;其中,所述M为大于等于1的正整数;

按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算,得到各个所述窗口数据对应的输出结果及对应的窗口状态信息;其中,所述窗口状态信息作为历史窗口状态信息被其他窗口数据依据预设获取规则进行获取;所述初始模型包括N层网络单元,所述初始模型的预设层的网络单元的输出作为窗口状态信息;

将各个所述输出结果进行拼接,得到所述待训练语音的目标结果;

根据所述目标结果计算损失值,根据所述损失值对初始模型进行迭代训练,直至得到训练完成的语音识别模型。

2.根据权利要求1所述的语音识别模型训练方法,其特征在于,所述按照预设获取规则获取各个所述窗口数据对应的M个历史窗口状态信息的步骤,包括:

根据所述窗口数据在所述待训练语音的位置信息确定对应的预设获取规则,并根据对应的所述预设获取规则获取M个历史窗口状态信息;

其中,所述预设获取规则包括:

获取各个所述窗口数据各自相邻的前M个历史窗口状态信息;

获取各个所述窗口数据各自相邻的后M个历史窗口状态信息;

获取各个所述窗口数据各自相邻的前E个的历史窗口状态信息和获取各个所述窗口数据各自相邻的后F个的历史窗口状态信息;其中,E+F=M。

3.根据权利要求1所述的语音识别模型训练方法,其特征在于,所述获取待训练语音,并根据所述待训练语音提取语音特征,将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据的步骤,包括:

将所述待训练语音进行数据增强处理;

根据数据增强处理后的待训练语音提取语音特征;

将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据。

4.根据权利要求1所述的语音识别模型训练方法,其特征在于,所述按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算,得到各个所述窗口数据对应的输出结果及对应的窗口状态信息的步骤,包括:

将所述窗口数据输入至所述初始模型依次进行计算,并将所述历史窗口状态信息和预设层的网络单元的输入一起输入至预设层的网络单元进行计算;其中,相邻的两层网络单元中,前一层的网络单元的输出作为后一层的网络单元的输入。

5.根据权利要求1所述的语音识别模型训练方法,其特征在于,所述按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算,得到各个所述窗口数据对应的输出结果及对应的窗口状态信息的步骤,包括:

将所述窗口数据和所述历史窗口状态信息输入至所述初始模型的网络单元进行处理,并在处理过程中按照预设规则跳连至对应层的网络单元进行计算,得到各个窗口数据对应的输出结果及对应的窗口状态信息。

6.根据权利要求5所述的语音识别模型训练方法,其特征在于,以第H层和第K层的网络单元的输出作为窗口状态信息;所述K大于H,且K和H小于N,窗口数据为第L个窗口数据;所述将所述窗口数据和所述历史窗口状态信息输入至所述初始模型的网络单元进行处理,并在处理过程中按照预设规则跳连至对应层的网络单元进行计算,得到各个窗口数据对应的输出结果及对应的窗口状态信息的步骤,包括:

将第L个窗口数据输入所述初始模型依次在各个网络单元进行处理;

当处理到第H层时,将第H-1层网络单元的输出加上在第H层获取到的历史窗口信息作为第H层网络单元的输入进行计算,得到第H层网络单元的第一目标输出;

将所述第一目标输出跳连至第K层网络单元,将第一目标输出和在第K层获取到的历史窗口信息输出输入至第K层网络单元进行计算,得到第K层网络单元的第二目标输出,将第一目标输出和第二目标输出作为第L个窗口数据的窗口状态信息;

将所述第二目标输出输入第K层之后的网络单元依次进行计算得到第L个窗口数据的输出结果。

7.根据权利要求2所述的语音识别模型训练方法,其特征在于,所述获取各个所述窗口数据各自相邻的后M个历史窗口状态信息的步骤,包括:

将所述窗口数据相邻的后M个窗口数据分别输入至所述初始模型中经过各层网络单元处理,得到各层网络单元的输出,以所述初始模型的预设层的网络单元的输出作为窗口状态信息。

8.一种语音识别模型训练装置,其特征在于,包括:

第一获取单元,用于获取待训练语音,并根据所述待训练语音提取语音特征,将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据;多个预设帧长的窗口数据依照时序依次相连;

第二获取单元,用于按照预设获取规则获取各个所述窗口数据对应的M个历史窗口状态信息;其中,所述M为大于等于1的正整数;

特征运算单元,用于按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算,得到各个所述窗口数据对应的输出结果及对应的窗口状态信息;其中,所述窗口状态信息作为历史窗口状态信息被其他窗口数据依据预设获取规则进行获取;所述初始模型包括N层网络单元,所述初始模型的预设层的网络单元的输出作为窗口状态信息;

拼接单元,用于将各个所述输出结果进行拼接,得到所述待训练语音的目标结果;

训练单元,用于根据所述目标结果计算损失值,根据所述损失值对初始模型进行迭代训练,直至得到训练完成的语音识别模型。

9.一种计算机设备,包括存储器和处理器,所述存储器中存储有计算机程序,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7中任一项所述的语音识别模型训练方法的步骤。

10.一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7中任一项所述的语音识别模型训练方法的步骤。

语音识别模型训练方法、装置、计算机设备和存储介质

技术领域

[0001] 本申请涉及人工智能的技术领域，特别涉及一种语音识别模型训练方法、装置、计算机设备和存储介质。

背景技术

[0002] 语音识别算法在终端部署时，由于受到终端上的算力和内存大小的限制，一般处理流程是输入一小段语音(比如0.1s)，对其进行处理，获得对应的输出结果后，然后对每一小段语音的输出结果进行拼接，对拼接的结果进行最终的语音识别。而语音识别算法的模型，在训练时，是整个训练语句输入到模型中进行处理，跟部署时的处理流程有差异。这种训练和推理的不一致，会导致语音识别模型的性能有损失，导致语音识别模型的准确率较低。

发明内容

[0003] 本申请的主要目的为提供一种语音识别模型训练方法、装置、计算机设备和存储介质，旨在解决语音识别模型的准确率较低的技术问题。

[0004] 为实现上述目的，本申请提供了一种语音识别模型训练方法，包括以下步骤：

[0005] 获取待训练语音，并根据所述待训练语音提取语音特征，将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据；多个预设帧长的窗口数据依照时序依次相连；

[0006] 按照预设获取规则获取各个所述窗口数据对应的M个历史窗口状态信息；其中，所述M为大于等于1的正整数；

[0007] 按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算，得到各个所述窗口数据对应的输出结果及对应的窗口状态信息；其中，所述窗口状态信息作为历史窗口状态信息被其他窗口数据依据预设获取规则进行获取；所述初始模型包括N层网络单元，所述初始模型的预设层的网络单元的输出作为窗口状态信息；

[0008] 将各个所述输出结果进行拼接，得到所述待训练语音的目标结果；

[0009] 根据所述目标结果计算损失值，根据所述损失值对初始模型进行迭代训练，直至得到训练完成的语音识别模型。

[0010] 进一步地，所述按照预设获取规则获取各个所述窗口数据对应的M个历史窗口状态信息的步骤，包括：

[0011] 根据所述窗口数据在所述待训练语音的位置信息确定对应的预设获取规则，并根据对应的所述预设获取规则获取M个历史窗口状态信息；

[0012] 其中，所述预设获取规则包括：

[0013] 获取各个所述窗口数据各自相邻的前M个历史窗口状态信息；

[0014] 获取各个所述窗口数据各自相邻的后M个历史窗口状态信息；

[0015] 获取各个所述窗口数据各自相邻的前E个的历史窗口状态信息和获取各个所述窗口数据各自相邻的后F个的历史窗口状态信息；其中，E+F=M。

[0016] 进一步地，所述获取待训练语音，并根据所述待训练语音提取语音特征，将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据的步骤，包括：

[0017] 将所述待训练语音进行数据增强处理；

[0018] 根据数据增强处理后的待训练语音提取语音特征；

[0019] 将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据。

[0020] 进一步地，所述按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算，得到各个所述窗口数据对应的输出结果及对应的窗口状态信息的步骤，包括：

[0021] 将所述窗口数据输入至所述初始模型依次进行计算，并将所述历史窗口状态信息和预设层的网络单元的输入一起输入至预设层的网络单元进行计算；其中，相邻的两层网络单元中，前一层的网络单元的输出作为后一层的网络单元的输入。

[0022] 进一步地，所述按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算，得到各个所述窗口数据对应的输出结果及对应的窗口状态信息的步骤，包括：

[0023] 将所述窗口数据和所述历史窗口状态信息输入至所述初始模型的网络单元进行处理，并在处理过程中按照预设规则跳连至对应层的网络单元进行计算，得到各个窗口数据对应的输出结果及对应的窗口状态信息。

[0024] 进一步地，以第H层和第K层的网络单元的输出作为窗口状态信息；所述K大于H，且K和H小于N，窗口数据为第L个窗口数据；所述将所述窗口数据和所述历史窗口状态信息输入至所述初始模型的网络单元进行处理，并在处理过程中按照预设规则跳连至对应层的网络单元进行计算，得到各个窗口数据对应的输出结果及对应的窗口状态信息的步骤，包括：

[0025] 将第L个窗口数据输入所述初始模型依次在各个网络单元进行处理；

[0026] 当处理到第H层时，将第H-1层网络单元的输出加上在第H层获取到的历史窗口信息作为第H层网络单元的输入进行计算，得到第H层网络单元的第一目标输出；

[0027] 将所述第一目标输出跳连至第K层网络单元，将第一目标输出和在第K层获取到的历史窗口信息输出输入至第K层网络单元进行计算，得到第K层网络单元的第二目标输出，将第一目标输出和第二目标输出作为第L个窗口数据的窗口状态信息；

[0028] 将所述第二目标输出输入第K层之后的网络单元依次进行计算得到第L个窗口数据的输出结果。

[0029] 进一步地，所述获取各个所述窗口数据各自相邻的后M个历史窗口状态信息的步骤，包括：

[0030] 将所述窗口数据相邻的后M个窗口数据分别输入至所述初始模型中经过各层网络单元处理，得到各层网络单元的输出，以所述初始模型的预设层的网络单元的输出作为窗口状态信息。

[0031] 本申请还提供一种语音识别模型训练装置，包括：

[0032] 第一获取单元，用于获取待训练语音，并根据所述待训练语音提取语音特征，将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据；多个预设帧长的窗口数据依照时序依次相连；

[0033] 第二获取单元，用于按照预设获取规则获取各个所述窗口数据对应的M个历史窗

口状态信息;其中,所述M为大于等于1的正整数;

[0034] 特征运算单元,用于按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算,得到各个所述窗口数据对应的输出结果及对应的窗口状态信息;其中,所述窗口状态信息作为历史窗口状态信息被其他窗口数据依据预设获取规则进行获取;所述初始模型包括N层网络单元,所述初始模型的预设层的网络单元的输出作为窗口状态信息;

[0035] 拼接单元,用于将各个所述输出结果进行拼接,得到所述待训练语音的目标结果;

[0036] 训练单元,用于根据所述目标结果计算损失值,根据所述损失值对初始模型进行迭代训练,直至得到训练完成的语音识别模型。

[0037] 本申请还提供一种计算机设备,包括存储器和处理器,所述存储器中存储有计算机程序,所述处理器执行所述计算机程序时实现上述任一项所述的语音识别模型训练方法的步骤。

[0038] 本申请还提供一种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时实现上述任一项所述的语音识别模型训练方法的步骤。

[0039] 本申请提供的语音识别模型训练方法、装置、计算机设备和存储介质,通过将待训练语音分为多个窗口数据进行后续训练,在训练时即以较小的窗口数据进行处理,保证了训练和推理的一致,保证了部署时不会因为两者的差异导致的准确率降低,同时引入历史窗口状态信息,增加了模型的表达能力,可有效提升流式实时识别的准确率。进一步地,在将窗口数据当做二维图像进行特征运算,相比简单的拼接,特征更具有表现能力,使得最终的语音识别模型的识别准确率更优。

附图说明

[0040] 图1 是本申请一实施例中语音识别模型训练方法步骤示意图;

[0041] 图2是本申请一实施例中语音识别模型训练装置结构框图;

[0042] 图3为本申请一实施例的计算机设备的结构示意框图。

[0043] 本申请目的的实现、功能特点及优点将结合实施例,参照附图做进一步说明。

具体实施方式

[0044] 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。

[0045] 参照图1,本申请一实施例提供一种语音识别模型训练方法,包括以下步骤:

[0046] 步骤S1,获取待训练语音,并根据所述待训练语音提取语音特征,将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据;多个预设帧长的窗口数据依照时序依次相连;

[0047] 步骤S2,按照预设获取规则获取各个所述窗口数据对应的M个历史窗口状态信息;其中,所述M为大于等于1的正整数;

[0048] 步骤S3,按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算,得到各个所述窗口数据对应的输出结果及对应的窗口状态信息;其

中，所述窗口状态信息作为历史窗口状态信息被其他窗口数据依据预设获取规则进行获取；所述初始模型包括N层网络单元，所述初始模型的预设层的网络单元的输出作为窗口状态信息；

[0049] 步骤S4，将各个所述输出结果进行拼接，得到所述待训练语音的目标结果；

[0050] 步骤S5，根据所述目标结果计算损失值，根据所述损失值对初始模型进行迭代训练，直至得到训练完成的语音识别模型。

[0051] 本实施例中，如上述步骤S1所述，预先设置有训练集，训练集中包括有多条待训练语音，获取待训练语音，根据所述待训练语音提取语音特征，具体的，可采用fbank算法、mfcc算法等对待训练语音进行提取，得到语音特征，语音特征为一个多维的向量，如每帧为40维的向量。将语音特征按照预设帧长和步长进行分段，如预设帧长设置为11帧、15帧等，得到多个窗口数据。每一个窗口数据经过后续处理对应得到一个输出结果和一个窗口状态信息，输出结果为音素分类的概率分布向量，概率分布向量的和为1，步幅(stride)可设置为6或8，窗口状态信息是在进行特征运算时的一些高层次的特征，每个窗口数据得到的窗口状态信息可作为其他窗口数据所想要获取到的历史窗口状态信息。训练时各待训练语音的长度不一致，将待训练语音填充(padding)到最大长度，根据实际长度截取。比如训练语料，只选取长度小于6s的待训练语音填充到的最大长度为600即可(1帧为10ms)。如一个待训练语音长1s，100帧填充到600帧需要500帧补0，根据实际长度截取就是在使用它时，根据100帧进行计算，而不是600。在后续计算损失的时候，根据每个待训练语音的实际长度计算。将待训练语音填充至固定长度，使得每个待训练语音具有多少个窗口数据是确定值，方便流式窗口处理的循环处理。采用最大值填充的方式，使得窗口计算总次数确定且一致，方便进行batch训练。

[0052] 具体分段时，相邻的窗口数据之间具有一定的重叠比例，如重叠比例为50%，还可设置为其他重叠比例，通过将窗口数据重叠，保证语音的连续性。

[0053] 如上述步骤S2所述，按照预设获取规则为窗口数据获取M个历史窗口状态信息，M的数值可根据算力、内存、实时性、精确度等多方面平衡确定，如设置2、4、6等。

[0054] 如上述步骤S3所述，按照时序对窗口数据进行特征运算，即按照待训练语音的时序从左到右的顺序依次进行特征运算。将窗口数据作为二维图像，然后进行特征的提取，比如[11,40,1] 分别为[h,w,c](height,width,channels)。采用常规的Conv2D或者Depthwise卷积进行特征运算，最终的输出结果reshape为[1,embed_dim]，即一个1维的向量。窗口状态信息则是卷积过程中的高层次信息，如初始模型包括有6层卷积层，上一层的输出作为下一层的输入，可设置其中的第二层和第四层的输出作为对应的窗口数据的窗口状态信息，当然也可设置其他层的输出经过低秩分解后的结果作为窗口状态信息，经过低秩分解后能够减少窗口状态信息的参数量和运算量窗口，状态信息的个数也可根据实际需要进行确定。每个窗口数据在特征运算时权值复用。比如tensorflow 可用tf.get_variable_scope().reuse_variables()，通过权值复用可以大大减少网络的复杂度和空间，使得语音识别模型在资源受限的设备上更容易部署。

[0055] 一般的卷积，采用的是samepadding的方式，当由于本申请使用的小窗口，比如11，kernelsize=5 做samepadding的话，左右各padding 2 (一般padding的是0) ，相比于11占比较大，会引入噪声。本申请在特征运算时，卷积采用Valid的方式，能够更有效地利用窗

口数据,保证每个窗口数据的处理方式一致,且对于小窗口,不会引入噪声。在Valid卷积计算过程中,将卷积核与输入数据在通道方向分别卷积,之后将卷积后的数值相加,得到一个新的特征值,因此比完成单个样本的单层卷积操作的计算总量较大。将获取到的历史窗口状态信息与窗口数据输入到初始模型进行特征运算,通过结合历史窗口状态信息(可以结合当前窗口数据的前面一小段和/或后面一小段的窗口状态信息,一般在资源有限的嵌入式设备上,流式处理直接使用前面一小段的窗口状态信息),不是单独利用孤立的每个窗口数据,而是引入历史窗口状态信息,使得当前窗口数据的推理能够拥有更广的感受野,准确率更高。

[0056] 具体的,结合历史窗口状态信息和当前窗口数据时,可使用LSTM(Long Short-Term Memory,长短期记忆网络)的方式或者Attention的方式生成当前的窗口数据的输出结果和窗口状态信息,其中,Attention的方式包括FSMN(Feedforward Sequential Memory Networks,前馈序列记忆神经网络)及其变种:cFSMN、DFSMN、pyramidal-FSMN。在一实施例中,可通过[3,64] 表示利用当前窗口数据的前3段窗口数据的历史窗口状态信息,每个窗口状态信息的向量维度是64。

[0057] 如上述步骤S4-S5所述,单个窗口数据的标签很难获取,因此需要将整个待训练语音的窗口数据的输出结果按照顺序进行拼接,作为整个待训练语音的输出,然后基于整句,在训练时计算损失进行优化。损失函数可包括如ctc(Connectionist Temporal Classification)loss 、Rnn-t loss 等。通过损失函数计算损失值,将损失值与预设损失阈值进行比较,当损失值没有达到预设损失值阈值时,根据损失值进行迭代训练,当损失值小于预设损失值阈值时,表明训练后的语音识别模型能够准确地进行语音识别,结束训练。

[0058] 本实施例中,通过将待训练语音分为多个窗口数据进行后续训练,在训练时即以较小的窗口数据进行处理,保证了训练和推理的一致,保证了部署时不会因为两者的差异导致的准确率降低,同时引入历史窗口状态信息,增加了模型的表达能力,可有效提升流式实时识别的准确率。进一步地,在将窗口数据当做二维图像进行特征运算,相比简单的拼接,特征更具有表现能力,使得最终的语音识别模型的识别准确率更优。

[0059] 在一实施例中,所述按照预设获取规则获取各个所述窗口数据对应的M个历史窗口状态信息的步骤S2,包括:

[0060] 步骤S21,根据所述窗口数据在所述待训练语音的位置信息确定对应的预设获取规则,并根据对应的所述预设获取规则获取M个历史窗口状态信息;

[0061] 其中,所述预设获取规则包括:

[0062] 获取各个所述窗口数据各自相邻的前M个历史窗口状态信息;

[0063] 获取各个所述窗口数据各自相邻的后M个历史窗口状态信息;

[0064] 获取各个所述窗口数据各自相邻的前E个的历史窗口状态信息和获取各个所述窗口数据各自相邻的后F个的历史窗口状态信息;其中,E+F=M。

[0065] 本实施例中,待训练语音分为多个窗口数据后,如将待训练语音分为100个窗口数据,各个窗口数据按照位置顺序进行标号,即得到第一个窗口数据到第一百个窗口数据,预设获取规则具有三种,需要按照各个窗口数据在待训练语音中的位置信息确定对应的预设获取规则,如若M为5,那么第一个窗口数据到第五个窗口数据无法获得对应的相邻的前五个历史窗口数据,因此,可获得对应的相邻的后五个历史窗口数据。窗口数据处于第6个至

第95个时，可随意获取相邻的前和/或后的历史窗口数据，窗口数据处于第96个至第100个时，获取相邻的前五个历史窗口数据，各个窗口数据所获取到的历史窗口状态信息可根据实际情况进行设置。在另一实施例中，当M为5，第1个窗口数据至第5个窗口数据在获取前5个历史窗口状态信息时，第1个窗口数据无需借助历史窗口状态信息，直接将窗口数据输入初始模型得到对应的输出结果和窗口状态信息；第2个窗口数据借助第1个窗口数据对应得到的窗口状态信息作为历史窗口状态信息，将第2个窗口数据和1个历史窗口状态信息输入初始模型得到对应的输出结果和窗口状态信息；第3个窗口数据则可借助第1个窗口数据和第2个窗口数据对应的窗口状态信息；第4个窗口数据借助前面三个窗口数据对应的窗口状态信息；第5个窗口数据借助前面四个窗口数据的窗口状态信息。同理，第96个窗口数据至第100个窗口数据在获取后5个历史窗口装修信息时，第96个窗口数据可借助后4个窗口数据的窗口状态信息，第97个窗口数据可借助后3个窗口数据的窗口状态信息，第98个窗口数据可借助后2个窗口数据的窗口状态信息。

[0066] 在一实施例中，所述获取待训练语音，并根据所述待训练语音提取语音特征，将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据的步骤S1，包括：

[0067] 步骤S11，将所述待训练语音进行数据增强处理；

[0068] 步骤S12，根据数据增强处理后的待训练语音提取语音特征；

[0069] 步骤S13，将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据。

[0070] 本实施例中，对于某个待训练语音，进行数据增强处理，扩充数据的多样性，使得初始模型具有更好的鲁棒性。具体的，可对整个待训练语音进行降采样，或对待训练语音中的部分进行降采样。对其进行分段处理，在某个待训练语音段进行采样。比如分段为3份待训练语音段，每份待训练语音段按一定概率选中进行抽样，选择预设个数的待训练语音段再进行分割，得到窗口数据进行后续的处理。在另一实施例中，训练集中包括有多个待训练语音，可使用随机的抽样选择待训练语音进行处理，比如隔一行取一个，增加初始模型的鲁棒性。

[0071] 在一实施例中，所述按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算，得到各个所述窗口数据对应的输出结果及对应的窗口状态信息的步骤，包括：

[0072] 将所述窗口数据输入至所述初始模型依次进行计算，并将所述历史窗口状态信息和预设层的网络单元的输入一起输入至预设层的网络单元进行计算；其中，相邻的两层网络单元中，前一层的网络单元的输出作为后一层的网络单元的输入。

[0073] 本实施例中，每一个窗口数据经过处理后得到的窗口状态信息可包括多个，如包括两个，初始模型可包括6层网络单元，如6层卷积层，前一层网络单元的输出作为后一层网络单元的输入，可设置第二层和第四层的网络单元的输出作为窗口状态信息，获取有M个历史窗口状态信息，M可设置为3，那么就包括有6个历史窗口状态信息，将窗口数据输入第一层网络单元并得到对应的输出后，将第一层的输出和3个同属于第二层的输出的历史窗口状态信息一起输入至第二层网络单元进行计算，得到第二层网络单元的输出，第二层网络单元的输出作为该窗口数据的窗口状态信息，将第二层网络单元的输出输入第三层网络单元依次进行计算，后续到达第四层网络单元时，将第三层的输出和3个同属于第四层的输出的历史窗口状态信息一起输入至第四层网络单元进行计算，得到输出，同样的，第四层网络

单元的输出也作为窗口状态信息。后续再依次进行计算，得到最终的输出结果。本实施例中，在横向上结合多个历史窗口状态信息，使得当前的窗口数据的输出结果能够具有更广的感受野，准确率更高。

[0074] 在一实施例中，所述按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算，得到各个所述窗口数据对应的输出结果及对应的窗口状态信息的步骤S3，包括：

[0075] 步骤S31，将所述窗口数据和所述历史窗口状态信息输入至所述初始模型的网络单元进行处理，并在处理过程中按照预设规则跳连至对应层的网络单元进行计算，得到各个窗口数据对应的输出结果及对应的窗口状态信息。

[0076] 本实施例中，在深度上，不只使用当前的窗口数据的信息，可利用不同层的网络单元之间的信息，比如通过跳连或多尺度等方式，将低层的信息跳连至高层去，从而训练更深的网络，使得最终训练得到的语音识别模型具有更强的表达能力。将窗口数据和对应的历史窗口状态信息输入至初始模型进行处理，预设规则中设置有跳连的初始层和目标层，如在一个具有五层网络单元的初始模型中，设置第二层为初始层，第五层为目标层，得到第二层网络单元的输出后，直接将输出跳连至第五层，将第二层的输出作为第二层的输入，不经过中间的第三层和第四层网络单元的处理。进一步的地，可按照预设规则进行多次跳连。

[0077] 在一实施例中，以第H层和第K层的网络单元的输出作为窗口状态信息；所述K大于H，且K和H小于N，窗口数据为第L个窗口数据；所述将所述窗口数据和所述历史窗口状态信息输入至所述初始模型的网络单元进行处理，并在处理过程中按照预设规则跳连至对应层的网络单元进行计算，得到各个窗口数据对应的输出结果及对应的窗口状态信息的步骤S31，包括：

[0078] 步骤S311，将第L个窗口数据输入所述初始模型依次在各个网络单元进行处理；

[0079] 步骤S312，当处理到第H层时，将第H-1层网络单元的输出加上在第H层获取到的历史窗口信息作为第H层网络单元的输入进行计算，得到第H层网络单元的第一目标输出；

[0080] 步骤S313，将所述第一目标输出跳连至第K层网络单元，将第一目标输出和在第K层获取到的历史窗口信息输出输入至第K层网络单元进行计算，得到第K层网络单元的第二目标输出，将第一目标输出和第二目标输出作为第L个窗口数据的窗口状态信息；

[0081] 步骤S314，将所述第二目标输出输入第K层之后的网络单元依次进行计算得到第L个窗口数据的输出结果。

[0082] 本实施例中，获取M个窗口数据的历史窗口状态信息，即M个窗口数据在第二层网络单元的输出和在第四层网络单元的输出。当前处理的窗口数据为第L个时，将第L个窗口数据输入第一层网络单元进行计算，得到第一层网络单元的输出，然后按照相邻的两层网络单元之间，上一层的输出作为下一层的输入依次进行计算，直到计算到第H层，将第H-1层的输出和M个窗口数据在第H层网络单元的输出一起输入至第H层网络单元进行处理，即这里输入第H层网络单元的有M+1个数据，然后得到第H层的第一目标输出，然后直接将第一目标输出跳连至第K层，不经过第H+1层到第K-1层这些网络单元的处理。跳连后，将第一目标输出和M个窗口数据在第K层网络单元的输出一起输入至第K层网络单元进行处理，得到第二目标输出，然后将第二目标输出输入第K+1层网络单元依次进行计算，得到第L个窗口数据的输出结果，而第一目标输出和第二目标输出则作为窗口状态信息。

[0083] 在一实施例中，所述获取各个所述窗口数据各自相邻的后M个历史窗口状态信息的步骤，包括：

[0084] 将所述窗口数据相邻的后M个窗口数据分别输入至所述初始模型中经过各层网络单元处理，得到各层网络单元的输出，以所述初始模型的预设层的网络单元的输出作为窗口状态信息。

[0085] 本实施例中，当历史窗口状态信息是获取相邻的后面的窗口数据的窗口状态信息时，直接是将对应的窗口数据经过依次计算到预设层的网络单元处理后得到，无需借助历史窗口状态信息。如当前处理的是第J个窗口数据，将第J+1至第J+M的窗口数据分别输入至初始模型中，经过层层网络单元的处理，得到各层网络单元的输出，预先设置有哪些层的网络单元的输出可以作为窗口状态信息。当第J个窗口数据处理完毕得到窗口状态信息和输出结果后，第J+1个窗口状态信息结合历史窗口状态信息进行同样的处理。

[0086] 当第J个窗口数据获取的是相邻的前E个的历史窗口状态信息和相邻的后F个的历史窗口状态信息时，后F个历史窗口状态信息是直接输入至初始模型得到的，没有结合其他窗口状态信息，而前E个历史窗口状态信息则是有结合窗口状态信息。

[0087] 参见图2，本身一实施例提供一种语音识别模型训练装置，包括：

[0088] 第一获取单元10，用于获取待训练语音，并根据所述待训练语音提取语音特征，将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据；多个预设帧长的窗口数据依照时序依次相连；

[0089] 第二获取单元20，用于按照预设获取规则获取各个所述窗口数据对应的M个历史窗口状态信息；其中，所述M为大于等于1的正整数；

[0090] 特征运算单元30，用于按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算，得到各个所述窗口数据对应的输出结果及对应的窗口状态信息；其中，所述窗口状态信息作为历史窗口状态信息被其他窗口数据依据预设获取规则进行获取；所述初始模型包括N层网络单元，所述初始模型的预设层的网络单元的输出作为窗口状态信息；

[0091] 拼接单元40，用于将各个所述输出结果进行拼接，得到所述待训练语音的目标结果；

[0092] 训练单元50，用于根据所述目标结果计算损失值，根据所述损失值对初始模型进行迭代训练，直至得到训练完成的语音识别模型。

[0093] 在一实施例中，所述第二获取单元20，包括

[0094] 第一获取子单元，根据所述窗口数据在所述待训练语音的位置信息确定对应的预设获取规则，并根据对应的所述预设获取规则获取M个历史窗口状态信息；

[0095] 其中，所述预设获取规则包括：

[0096] 用于获取各个所述窗口数据各自相邻的前M个历史窗口状态信息；

[0097] 获取各个所述窗口数据各自相邻的后M个历史窗口状态信息；

[0098] 获取各个所述窗口数据各自相邻的前E个的历史窗口状态信息和获取各个所述窗口数据各自相邻的后F个的历史窗口状态信息；其中，E+F=M。

[0099] 在一实施例中，所述第一获取单元10，包括：

[0100] 数据增强子单元，用于将所述待训练语音进行数据增强处理；

[0101] 提取子单元，用于根据数据增强处理后的待训练语音提取语音特征；

[0102] 分割子单元，用于将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据。

[0103] 在一实施例中，所述特征运算单元30，包括：

[0104] 第一计算子单元，用于将所述窗口数据输入至所述初始模型依次进行计算，并将所述历史窗口状态信息和预设层的网络单元的输入一起输入至预设层的网络单元进行计算；其中，相邻的两层网络单元中，前一层的网络单元的输出作为后一层的网络单元的输入。

[0105] 在一实施例中，所述特征运算单元30，包括：

[0106] 第二计算子单元，用于将所述窗口数据和所述历史窗口状态信息输入至所述初始模型的网络单元进行处理，并在处理过程中按照预设规则跳连至对应层的网络单元进行计算，得到各个窗口数据对应的输出结果及对应的窗口状态信息。

[0107] 在一实施例中，所述第二计算子单元，包括：

[0108] 第一处理模块，用于将第L个窗口数据输入所述初始模型依次在各个网络单元进行处理；

[0109] 第一计算模块，用于当处理到第H层时，将第H-1层网络单元的输出加上在第H层获取到的历史窗口信息作为第H层网络单元的输入进行计算，得到第H层网络单元的第一目标输出；

[0110] 第二计算模块，用于将所述第一目标输出跳连至第K层网络单元，将第一目标输出和在第K层获取到的历史窗口信息输出输入至第K层网络单元进行计算，得到第K层网络单元的第二目标输出，将第一目标输出和第二目标输出作为第L个窗口数据的窗口状态信息；

[0111] 第三计算模块，用于将所述第二目标输出输入第K层之后的网络单元依次进行计算得到第L个窗口数据的输出结果。

[0112] 在一实施例中，所述第二获取子单元，包括：

[0113] 第二处理模块，用于将所述窗口数据相邻的后M个窗口数据分别输入至所述初始模型中经过各层网络单元处理，得到各层网络单元的输出，以所述初始模型的预设层的网络单元的输出作为窗口状态信息。

[0114] 在本实施例中，上述各个单元、子单元、模块的具体实现请参照上述方法实施例中所述，在此不再进行赘述。

[0115] 参照图3，本申请实施例中还提供一种计算机设备，该计算机设备可以是服务器，其内部结构可以如图3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、网络接口和数据库。其中，该计算机设计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数据等。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模型训练方法。

[0116] 本领域技术人员可以理解，图3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。

[0117] 本申请一实施例还提供一种计算机可读存储介质，其上存储有计算机程序，计算

机程序被处理器执行时实现一种语音识别模型训练方法。

[0118] 综上所述，为本申请实施例中提供的语音识别模型训练方法、装置、计算机设备和存储介质，获取待训练语音，并根据所述待训练语音提取语音特征，将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据；多个预设帧长的窗口数据依照时序依次相连；按照预设获取规则获取各个所述窗口数据对应的M个历史窗口状态信息；其中，所述M为大于等于1的正整数；按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算，得到各个所述窗口数据对应的输出结果及对应的窗口状态信息；其中，所述窗口状态信息作为历史窗口状态信息被其他窗口数据依据预设获取规则进行获取；所述初始模型包括N层网络单元，所述初始模型的预设层的网络单元的输出作为窗口状态信息；将各个所述输出结果进行拼接，得到所述待训练语音的目标结果；根据所述目标结果计算损失值，根据所述损失值对初始模型进行迭代训练，直至得到训练完成的语音识别模型。本申请通过将待训练语音分为多个窗口数据进行后续训练，在训练时即以较小的窗口数据进行处理，保证了训练和推理的一致，保证了部署时不会因为两者的差异导致的准确率降低，同时引入历史窗口状态信息，增加了模型的表达能力，可有效提升流式实时识别的准确率。进一步地，在将窗口数据当做二维图像进行特征运算，相比简单的拼接，特征更具有表现能力，使得最终的语音识别模型的识别准确率更优。

[0119] 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，所述的计算机程序可存储与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，本申请所提供的和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。非易失性存储器可以包括只读存储器(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通过多种形式可得，诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双速据率SDRAM(SSRSDRAM)、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。

[0120] 需要说明的是，在本文中，术语“包括”、“包含”或者其任何其它变体意在涵盖非排他性地包含，从而使得包括一系列要素的过程、装置、物品或者方法不仅包括那些要素，而且还包括没有明确列出的其它要素，或者是还包括为这种过程、装置、物品或者方法所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

[0121] 以上所述仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其它相关的技术领域，均同理包括在本申请的专利保护范围内。

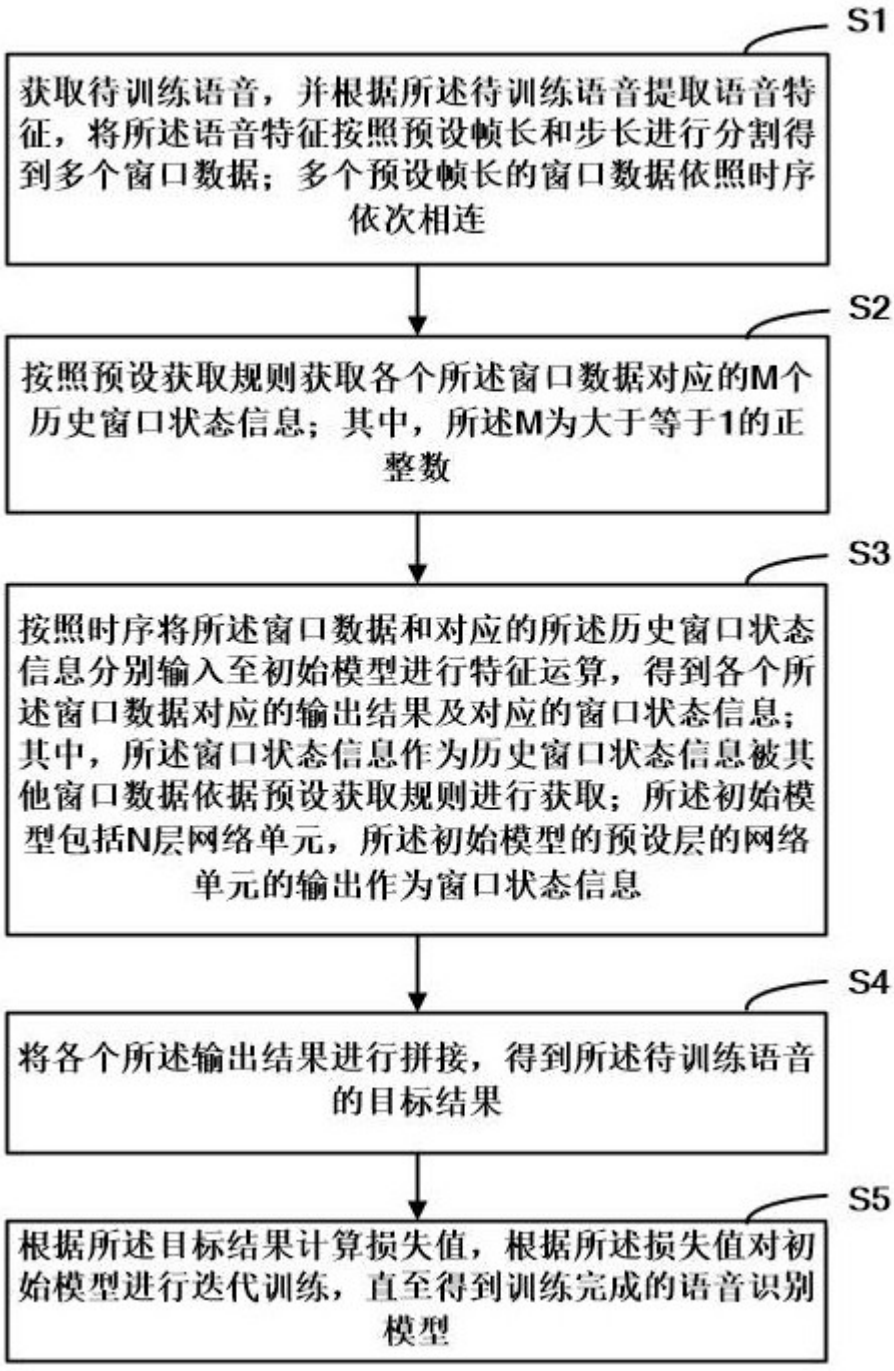

S1
获取待训练语音，并根据所述待训练语音提取语音特征，将所述语音特征按照预设帧长和步长进行分割得到多个窗口数据；多个预设帧长的窗口数据依照时序依次相连
S2
按照预设获取规则获取各个所述窗口数据对应的M个历史窗口状态信息；其中，所述M为大于等于1的正整数
S3
按照时序将所述窗口数据和对应的所述历史窗口状态信息分别输入至初始模型进行特征运算，得到各个所述窗口数据对应的输出结果及对应的窗口状态信息；其中，所述窗口状态信息作为历史窗口状态信息被其他窗口数据依据预设获取规则进行获取；所述初始模型包括N层网络单元，所述初始模型的预设层的网络单元的输出作为窗口状态信息
S4
将各个所述输出结果进行拼接，得到所述待训练语音的目标结果
S5
根据所述目标结果计算损失值，根据所述损失值对初始模型进行迭代训练，直至得到训练完成的语音识别模型

图1

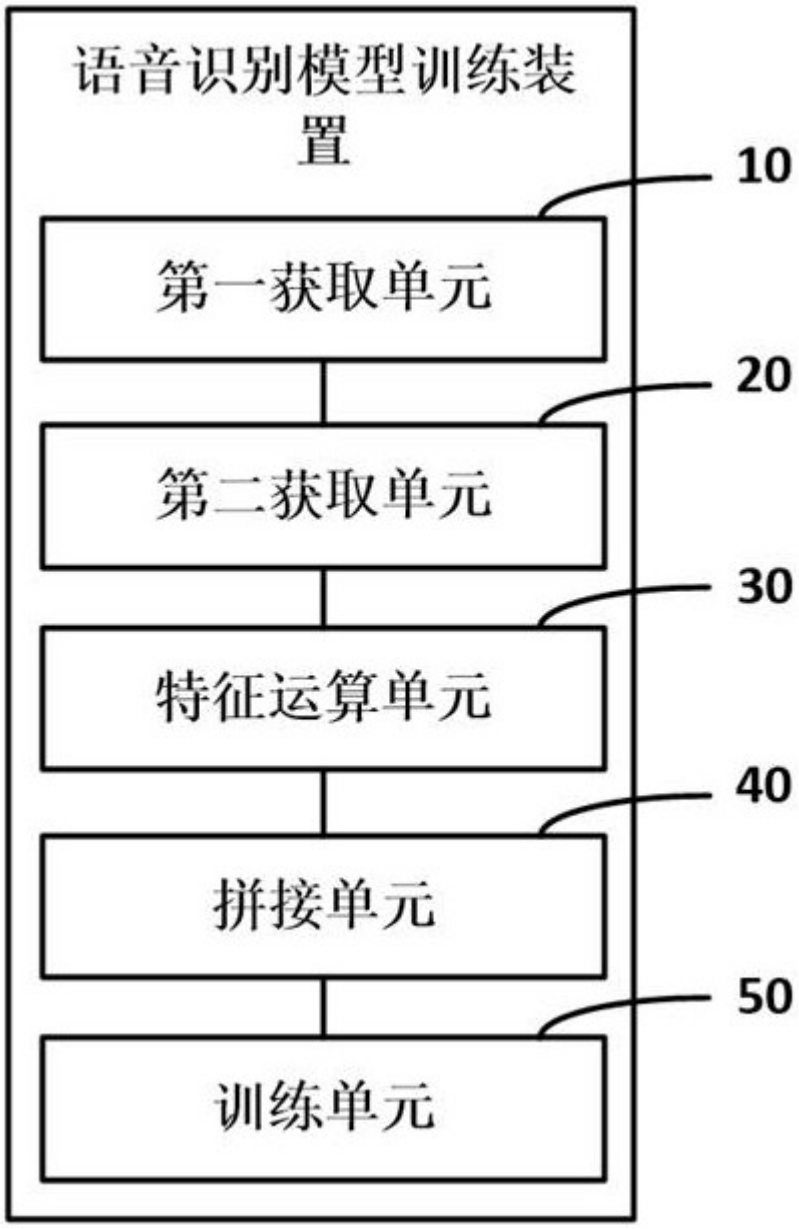


图2

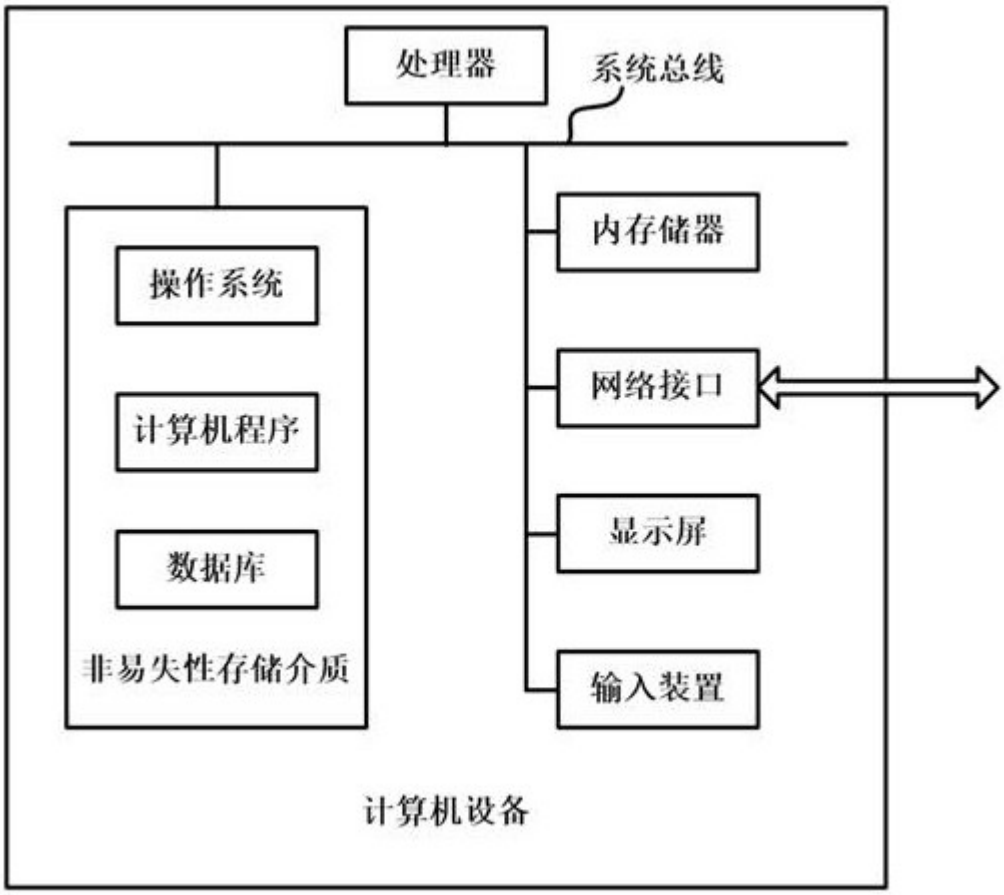


图3