

语音识别模型的数据处理系统及方法、语音识别方法

5 本申请要求于 2022 年 10 月 27 日提交中国专利局、申请号为 202211329674.7、申请名称为“语音识别模型的数据处理系统及方法、语音识别方法”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本说明书实施例涉及计算机技术领域，特别涉及一种语音识别模型的数据处理系统及方法、语音识别方法。

10 背景技术

语音识别模型是作用是将输入的语音转换为文本的模型，其往往是基于语音-文本标注数据进行训练，为了提高训练得到的语音识别模型的精度，通常需要大量的语音-文本标注数据，但对数据进行标注通常是人工进行的，需要耗费大量人力物力，效率低且难实现。

因此，亟需一种高效的语音识别模型的数据处理方法。

15 发明内容

有鉴于此，本说明书实施例提供了一种语音识别模型的数据处理系统。本说明书一个或者多个实施例同时涉及一种语音识别模型的数据处理方法，一种语音识别方法，一种语音识别模型的数据处理装置，一种语音识别装置，一种计算设备，一种计算机可读存储介质以及一种计算机程序，以解决现有技术中存在的技术缺陷。

20 根据本说明书实施例的第一方面，提供了一种语音识别模型的数据处理系统，包括：

云侧设备，用于获取样本集，其中，样本集包括多个样本对，样本对包括样本语音数据和样本中文文本；利用编码器对样本语音数据进行编码，获得样本语音数据的语音特征，其中，编码器基于对预训练语音数据执行中文发音单元预测任务进行预训练；将语音特征输入解码器，获得预测中文文本，其中，解码器基于对预训练中文发音单元执行文本预测任务进行预训练；基于预测中文文本和样本中文文本，对包括编码器和解码器的模型进行预训练，在达到预训练停止条件的情况下，获取预训练得到的语音识别模型的模型参数；

云侧设备，还用于向端侧设备发送预训练得到的语音识别模型的模型参数；

端侧设备，用于利用语音识别模型对待识别语音数据进行语音识别，获得待识别语音数据对应的目标文本。

30 根据本说明书实施例的第二方面，提供了一种语音识别模型的数据处理方法，应用于云侧设备，云侧设备与多个端侧设备连接，方法包括：

获取样本集，其中，样本集包括多个样本对，样本对包括样本语音数据和样本中文文本；

利用编码器对样本语音数据进行编码，获得样本语音数据的语音特征，其中，编码器基于对预训练语音数据执行中文发音单元预测任务进行预训练；

35 将语音特征输入解码器，获得预测中文文本，其中，解码器基于对预训练中文发音单元执行文本预测任务进行预训练；

基于预测中文文本和样本中文文本，对包括编码器和解码器的模型进行预训练，在达到预训练停止条件的情况下，获取预训练得到的语音识别模型的模型参数；

向第一端侧设备发送预训练得到的语音识别模型的模型参数，其中，第一端侧设备是多个端侧设备中的任意一个。

40 根据本说明书实施例的第三方面，提供了一种语音识别方法，应用于端侧设备，端侧设备与云侧设备连接，方法包括：

获取待识别语音数据；

利用语音识别模型的编码器，对待识别语音数据进行编码，获得待识别语音数据的语音特征，其中，语音识别模型是云侧设备通过如第二方面提供的语音识别模型的数据处理方法预训练得到的；

45

将语音特征输入语音识别模型的解码器，获得待识别语音数据对应的目标文本。

根据本说明书实施例的第四方面，提供了一种语音识别模型的数据处理装置，应用于云侧设备，云侧设备与多个端侧设备连接，装置包括：

第一获取模块，被配置为获取样本集，其中，样本集包括多个样本对，样本对包括样本语音数据和样本中文文本；

第一编码模块，被配置为利用编码器对样本语音数据进行编码，获得样本语音数据的语音特征，其中，编码器基于对预训练语音数据执行中文发音单元预测任务进行预训练；

第一解码模块，被配置为将语音特征输入解码器，获得预测中文文本，其中，解码器基于对预训练中文发音单元执行文本预测任务进行预训练；

预训练模块，被配置为基于预测中文文本和样本中文文本，对包括编码器和解码器的模型进行预训练，在达到预训练停止条件的情况下，获取预训练得到的语音识别模型的模型参数；

第一发送模块，被配置为向第一端侧设备发送预训练得到的语音识别模型的模型参数，其中，第一端侧设备是多个端侧设备中的任意一个。

根据本说明书实施例的第五方面，提供了一种语音识别装置，应用于端侧设备，端侧设备与云侧设备连接，装置包括：

第二获取模块，被配置为获取待识别语音数据；

第二编码模块，被配置为利用语音识别模型的编码器，对待识别语音数据进行编码，获得待识别语音数据的语音特征，其中，语音识别模型是云侧设备通过如第二方面提供的语音识别模型的数据处理方法预训练得到的；

第二解码模块，被配置为将语音特征输入语音识别模型的解码器，获得待识别语音数据对应的目标文本。

根据本说明书实施例的第六方面，提供了一种计算设备，包括：

存储器和处理器；

存储器用于存储计算机可执行指令，处理器用于执行计算机可执行指令，该计算机可执行指令被处理器执行时实现上述第二方面的语音识别模型的数据处理方法的步骤，或者，实现上述第三方面的语音识别方法的步骤。

根据本说明书实施例的第六方面，提供了一种计算机可读存储介质，其存储有计算机可执行指令，该指令被处理器执行时实现上述第二方面的语音识别模型的数据处理方法的步骤，或者，实现上述第三方面的语音识别方法的步骤。

根据本说明书实施例的第七方面，提供了一种计算机程序，其中，当计算机程序在计算机中执行时，令计算机执行上述第二方面的语音识别模型的数据处理方法的步骤，或者，实现上述第三方面的语音识别方法的步骤。

本说明书实施例提供的语音识别模型的数据处理系统，包括端侧设备和云侧设备，云侧设备，用于获取样本集，其中，样本集包括多个样本对，样本对包括样本语音数据和样本中文文本；利用编码器对样本语音数据进行编码，获得样本语音数据的语音特征，其中，编码器基于对预训练语音数据执行中文发音单元预测任务进行预训练；将语音特征输入解码器，获得预测中文文本，其中，解码器基于对预训练中文发音单元执行文本预测任务进行预训练；基于预测中文文本和样本中文文本，对包括编码器和解码器的模型进行预训练，在达到预训练停止条件的情况下，获取预训练得到的语音识别模型的模型参数；云侧设备，还用于向端侧设备发送预训练得到的语音识别模型的模型参数；端侧设备，用于利用语音识别模型对待识别语音数据进行语音识别，获得待识别语音数据对应的目标文本。即本方案在对语音识别模型进行预训练的阶段，至少执行了根据语音数据预测中文文本的任务、根据语音数据预测中文发音单元的中文发音单元预测任务、以及根据中文发音单元预测中文文本的文本预测任务，则在实际应用过程中对语音识别模型进行训练时，需要的带标注的样本语音数据和样本中文文本的数量比较少，减轻了标注人员的负担，降低了获取标注数据的难度。另外，基于对预训练语音数据执行中文发音单元预测任务预训练编码器，使得编码器具备了通过语音数据预测中文发音单元的能力，由于中文是表意文字，中文文本和语音数据的差距较大，中文发音单元可以作为中文文本和语音数据的桥梁，缩小两者的差距，通过预训练编码器能够将语音数据转化成中文发音单元，可以得到一个对于语音识别任务来说较好的编码

器；并且，基于对预训练中文发音单元执行文本预测任务预训练解码器，使得解码器能够学习到根据中文发音单元构建中文文本的能力，提高了解码器的语言建模能力，即通过预训练使得编码器和解码器具备一定的语音识别能力，并且对编码器和解码器组成的模型执行语音-文本的预测任务进行预训练，将模型的参数朝着更加适应语音识别任务的方向调整，能够提高训练效率和训练精度。并且，在预训练过程中使用的模型的输入是预训练语音数据或预训练中文发音单元，二者与语音识别模型应用时输入的语音数据模态相近，使得语音识别模型更加适合下游的语音识别任务，在一定程度上可以提高预训练得到的语音识别模型的识别准确性。

附图说明

图 1 示出了根据本说明书一个实施例提供的一种语音识别模型的数据处理系统架构下语音识别模型的数据处理方法、语音识别方法的流程示意图；

图 2 示出了根据本说明书一个实施例提供的一种语音识别模型的数据处理系统示意图；

图 3 示出了本说明书一实施例提供的一种语音识别模型的数据处理方法的数据流向图；

图 4a 示出了本申请一实施例提供的一种确定中文发音单元的方法的数据流向图；

图 4b 示出了本申请一实施例提供的一种对编码器进行预训练的数据流向图；

图 5 示出了本申请一实施例提供的另一种对编码器进行预训练的数据流向图；

图 6 示出了本申请一实施例提供的一种对解码器进行预训练的数据流向图；

图 7 示出了本说明书一个实施例提供的另一种对解码器进行预训练的数据流向图；

图 8 示出了本说明书一实施例提供的一种对语音识别模型进行微调的方法的数据流向图；

图 9 示出了本说明书一个实施例提供的一种应用于云侧设备的语音识别模型的数据处理方法的流程图；

图 10 示出了本说明书一个实施例提供的一种应用于端侧设备的语音识别方法的流程图；

图 11 示出了本说明书一个实施例提供的一种语音识别模型执行语音识别任务的数据流向图；

图 12 示出了本说明书一个实施例提供的一种语音识别模型的数据处理方法的处理过程流程图；

图 13 示出了本申请实施例提供的一种对语音识别模型进行联合训练时的数据流向图；

图 14 示出了本说明书一个实施例提供的一种应用于云侧设备的语音识别模型的数据处理装置的结构示意图；

图 15 示出了本说明书一个实施例提供的一种语音识别装置的结构示意图；

图 16 示出了本说明书一个实施例提供的一种计算设备的结构框图。

具体实施方式

在下面的描述中阐述了很多具体细节以便于充分理解本说明书。但是本说明书能够以很多不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本说明书内涵的情况下做类似推广，因此本说明书不受下面公开的具体实施的限制。

在本说明书一个或多个实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本说明书一个或多个实施例。在本说明书一个或多个实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本说明书一个或多个实施例中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

应当理解，尽管在本说明书一个或多个实施例中可能采用术语第一、第二等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本说明书一个或多个实施例范围的情况下，第一也可以被称为第二，类似地，第二也可以被称为第一。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

需要说明的是，本说明书实施例所涉及的与用户相关的信息和与用户相关的数据，均为经用户授权或者经过各方充分授权的信息和数据。

首先，对本说明书一个或多个实施例涉及的名词术语进行解释。

语音识别模型：用于对输入的语音数据进行识别，获得文本的模型。

编码器：用于对输入的语音、文字、中文发音单元等进行编码，以特征向量的形式表征输入。

解码器：用于对输入的特征向量进行解码，得到语音、文字等。

特征编码层：用于对输入的特征进行编码，以捕捉特征之间的关联关系。

语音特征：是语音的一种向量化表示。

语音编码器：用于对语音进行编码，得到语音的向量化表示，即语音特征。

CTC(Connectionist temporal classification)：主要用于处理输入序列比输出序列长的问题，一种实现输入和输出序列标签的对齐的算法。

5 语音识别模型的作用是将语音数据转化为文本，其训练往往需要大量的标注数据。通过无标注数据进行预训练，能够以较低成本使模型更容易达到好的效果。端到端的语音识别模型常用的结构主要有两种，一种是时序连接关系预测（CTC）结构，另一种是编码器-解码器结构。其中编码器-解码器结构的语音识别模型同时考虑了语音数据和文本语法信息，而 CTC 结构只考虑了语音数据，所以编码器-解码器结构效果常常优于 CTC 结构。相比其他表音语言，比如英文，中文是一种表意语言，因此在进行语音识别过程中，文本语法信息的辅助更为重要。另一方面，端到端语音识别模型的训练往往需要大量的语音-文本对标标注数据，尤其对于中文这种表意语言，端到端识别难度更大。

10 目前，无监督预训练在各个领域已经为下游任务带来了显著的提升。关于语音识别模型的预训练，也提出了一系列利用无标注语音数据进行编码器预训练的方法，比如 HuBERT、Data2Vec。这些预训练方法往往只能应用在 CTC 结构的模型上进行调优，而应用在编码器-解码器结构上会存在问题。这是由于解码器没有参与预训练导致的。STPT 和 Speech2C 又提出可以利用无标注文本或者无标注语音数据对解码器进行预训练，SpeechT5 则提出同时利用无标注文本和无标注语音数据对编码器-解码器模型进行预训练。这些方法存在三点问题：一是它们没有验证不同预训练任务之间的互补性；二是它们都是基于英文表音语言设计的，忽略了中文表意语言的特性；三是它们对无标签文本数据的利用不充分，导致无标签文本数据带来的提升不明显。

20 对于语音识别任务，目前主流效果好的语音预训练方法分为如下两种。

第一种是语音表示单模态预训练，比如 Wav2vec 2.0、HuBERT、Data2vec、Speech2C。这种方法只利用了无标注语音数据，通过掩码预测的方式，得到更好的语音建模能力。缺陷在于这种方法缺乏对文本语义信息建模的预训练，应用在编码器-解码器结构的语音识别模型上往往效果较差。

25 第二种是语音-文本多模态预训练，比如 STPT、SpeechT5。除了使用无标注语音数据之外，还引入了无标注文本数据参与预训练，预训练了编码器-解码器结构模型对语音信息和文本语法信息建模的能力。

STPT 的缺陷在于：1）无标注语音数据没有参与对解码器的参数更新；2）容易出现模型坍塌的问题。

30 SpeechT5 预训练主要是为了得到一个通用的语音模型，因此在预训练任务的设计上不完全只为语音识别任务考虑，因此对于语音识别模型预训练来说，它的缺陷在于：1）利用无标注语音数据设计的序列到序列任务是语音帧重建任务，这个任务对语音合成任务会更有利，而对语音识别任务会带来损害；2）利用无标注文本数据时，模型的输入是文本，对于中文这种象形文字来说，文本和语音两个模态差距较大，会给联合训练带来困难。

35 为此，本说明书提供了一种语音识别模型的数据处理系统，能够解决上述技术问题，其具体实现可以参见下述各个实施例的相关描述。

在本说明书中，提供了一种语音识别模型的数据处理系统，本说明书同时涉及一种语音识别模型的数据处理方法，一种语音识别方法，一种语音识别模型的数据处理装置，一种语音识别装置，一种计算设备，以及一种计算机可读存储介质，在下面的实施例中逐一进行详细说明。

40 参见图 1，图 1 示出了根据本说明书一个实施例提供的一种语音识别模型的数据处理系统架构下语音识别模型的数据处理方法、语音识别方法的流程示意图。

该系统可以包括云侧设备 101 和端侧设备 102，该云侧设备 101 用于对语音识别模型进行训练，该端侧设备 102 用于基于训练得到的语音识别模型执行语音识别任务。

45 云侧设备 101 可以是分布式云架构的中心云设备，端侧设备 102 可以是分布式云架构的边缘云设备，云侧设备 101 和端侧设备 102 可以是常规服务器、云服务器或服务器阵列等服务端设备，也可以是终端设备，本说明书实施例对此不做限定。并且，云侧设备 101 提供超强的计算和存储能力，距离用户较远，而端侧设备 102 部署范围大，距离用户较近。端侧设备 102 是云侧设备 101 的拓展，可将云侧设备 101 的计算能力向端侧设备 102 下沉，通过端云的一体化、协同管理来解决集中式云计算模式下无法满足的

业务需求。

在本说明书一个或多个实施例中，云侧设备 101 获取样本集，该样本集包括多个样本对，该样本对包括样本语音数据和样本中文文本；利用编码器对样本语音数据进行编码，获得样本语音数据的语音特征，该编码器基于对预训练语音数据执行中文发音单元预测任务进行预训练；将语音特征输入解码器，获得预测中文文本，该解码器基于对预训练中文发音单元执行文本预测任务进行预训练；基于预测中文文本和样本中文文本，对包括编码器和解码器的模型进行预训练，在达到预训练停止条件的情况下，获取预训练得到的语音识别模型的模型参数，并将语音识别模型的模型参数发送至端侧设备 102。

端侧设备 102 接收到语音识别模型后，利用语音识别模型对待识别语音数据进行语音识别，获得待识别语音数据对应的目标文本。具体的，端侧设备 102 获取待识别语音数据，将待识别语音数据输入语音识别模型的编码器，获得待识别语音数据的语音特征，将语音特征输入语音识别模型的解码器，得到待识别语音数据对应的目标文本。

本说明书实施例提供的语音识别模型的数据处理系统，在对语音识别模型进行预训练的阶段，云侧设备至少执行了根据语音数据预测中文文本的任务、根据语音数据预测中文发音单元的中文发音单元预测任务、以及根据中文发音单元预测中文文本的文本预测任务，则在实际应用过程中对语音识别模型进行训练时，需要的带标注的样本语音数据和样本中文文本的数量比较少，减轻了标注人员的负担，降低了获取标注数据的难度。另外，基于对预训练语音数据执行中文发音单元预测任务预训练编码器，使得编码器具备了通过语音数据预测中文发音单元的能力，由于中文是表意文字，中文文本和语音数据的差距较大，中文发音单元可以作为中文文本和语音数据的桥梁，缩小两者的差距，通过预训练编码器能够将语音数据转化成中文发音单元，可以得到一个对于语音识别任务来说较好的编码器；并且，基于对预训练中文发音单元执行文本预测任务预训练解码器，使得解码器能够学习到根据中文发音单元构建中文文本的能力，提高了解码器的语言建模能力，即通过预训练使得编码器和解码器具备一定的语音识别能力，并且对编码器和解码器组成的模型执行语音-文本的预测任务进行预训练，将模型的参数朝着更加适应语音识别任务的方向调整，能够提高训练效率和训练精度。并且，在预训练过程中使用的模型的输入是预训练语音数据或预训练中文发音单元，二者与语音识别模型应用时输入的语音数据模态相近，使得语音识别模型更加适合下游的语音识别任务，在一定程度上可以提高预训练得到的语音识别模型的识别准确性。

图 2 示出了根据本说明书一个实施例提供的一种语音识别模型的数据处理系统示意图。参见图 2，本说明书实施例提供的语音识别模型的数据处理系统，包括：端侧设备 201，以及与端侧设备 201 通信连接的云侧设备 202。

云侧设备 202，用于获取样本集，其中，样本集包括多个样本对，样本对包括样本语音数据和样本中文文本；利用编码器对样本语音数据进行编码，获得样本语音数据的语音特征，其中，编码器基于对预训练语音数据执行中文发音单元预测任务进行预训练；将语音特征输入解码器，获得预测中文文本，其中，解码器基于对预训练中文发音单元执行文本预测任务进行预训练；基于预测中文文本和样本中文文本，对包括编码器和解码器的模型进行预训练，在达到预训练停止条件的情况下，获取预训练得到的语音识别模型的模型参数；

云侧设备 202，还用于向端侧设备发送预训练得到的语音识别模型的模型参数；

端侧设备 201，用于利用语音识别模型对待识别语音数据进行语音识别，获得待识别语音数据对应的目标文本。

在本说明书一个或多个实施例中，由于语音识别模型需要执行的是根据语音数据预测中文文本的语音识别任务，则对语音识别模型的训练是有监督的训练，因此，需要获取由样本语音数据和样本文本组成的样本对，且由于本说明书提供的语音识别模型是针对中文语言的语音识别模型，因此，获取的样本文本是样本中文文本。并且，每个样本对中的样本语音数据和样本中文文本之间有对应关系。

作为一种示例，云侧设备 202 可以从开源的样本库中获取多个样本对，该多个样本对组成了样本集。示例性地，样本语音数据可以是任意语音数据，可以是一段话，也可以是一个词，如样本语音数据可以是语音聊天、演讲、音乐、会议录音等等。

在本说明书实施例中，云侧设备 202 获取包括多个样本对的样本集，且各样本对包括样本语音数据和样本中文文本，并根据样本集对包括编码器和解码器的模型进行预训练，得到语音识别模型，并向端

侧设备 201 下发预训练得到的语音识别模型的模型参数,以便端侧设备 201 可以基于该语音识别模型进行语音识别。

云侧设备 202,用于利用编码器对样本语音数据进行编码,获得样本语音数据的语音特征,该编码器基于对预训练语音数据执行中文发音单元预测任务进行预训练。

5 其中,中文发音单元预测任务是根据输入的预训练语音数据预测对应的中文发音单元。该中文发音单元是中文发音的组成单元,可以是拼音,也可以是音节,还可以是中文音素,如中文发音单元可以是 yin,也可以是 in、i、y 等。

其中,样本语音数据的语音特征是样本语音数据的向量化表示,是结合了样本语音数据中各个字之间的上下文关系后得到的。

10 需要说明的是,虽然该编码器基于对预训练语音数据执行中文发音单元预测任务进行预训练,但该预训练可以是在对解码器和编码器组成的模型进行预训练之前进行,或者,该预训练可以是在对解码器和编码器组成的模型进行预训练的过程中进行,本说明书实施例对此不做限定。

示例性地,该编码器可以是任何包括编码功能的模型中的编码器。例如,编码器可以是 transformer 模型的编码器,或者,编码器可以是 BERT、CNN (Convolutional Neural Network, 卷积神经网络),LSTM (Long Short Term Memory,长短期记忆神经网络),GRU (Gate Recurrent Unit, 门控循环单元结构) 15 等模型的编码器。并且,编码器可以包括 M+N 个 block (块),其中,M 和 N 均是大于 1 的正整数。

在本说明书第一种可能的实现方式中,可以直接将样本语音数据输入编码器,由编码器通过 M+N 个 block 对样本语音数据进行编码,得到样本语音数据对应的语音特征。

在本说明书第二种可能的实现方式中,编码器可以包括语音编码层和特征编码层,示例性地,语音 20 编码层包括 M 个 block,特征编码层包括 N 个 block。先将样本语音数据输入语音编码层,得到样本语音数据的初始语音特征,然后将初始语音特征输入特征编码层,得到样本语音数据的语音特征。其中,初始语音特征和语音特征的区别在于,初始语音特征是通过 M 个 block 后得到的语音上下文特征,语音特征是通过 M+N 个 block 处理后得到的语音上下文特征。

在本说明书第三种可能的实现方式中,编码器还可以包括特征提取层,且特征提取层与语音编码层 25 连接,将样本语音数据输入特征提取层中进行语音表示提取和下采样处理,提取样本语音数据中各个字的特征,得到样本语音数据的语音表示向量,再将该语音表示向量输入语音编码层,得到样本语音数据的初始语音特征,将初始语音特征输入特征编码层,得到样本语音数据的语音特征。

其中,特征提取层可以称为 feature extractor。作为一种示例,该特征提取层可以是编码器中的一个组件,则在对编码器进行预训练的过程中,该特征提取层也参与预训练。

30 在本说明书第四种可能的实现方式中,在利用编码器对样本语音数据进行编码之前,可以先提取样本语音数据的频谱特征,然后将该频谱特征输入编码器,得到样本语音数据的语音特征,该语音特征结合了样本语音数据中上下文的语义关系。

作为一种示例,可以使用线性预测倒谱系数算法或者梅尔频率倒谱系数或者预先训练得到的频谱特征提取模型,提取样本语音数据的频谱特征。并且线性预测倒谱系数算法或者梅尔频率倒谱系数算法均 35 是基于倒谱的,更符合人的听觉原理,是比较有效的频谱特征提取算法。

在本说明书实施例中,基于对预训练语音数据执行中文发音单元预测任务预训练编码器,则编码器能够学习到通过语音预测中文发音单元的能力,即对语音数据进行编码的能力,并且,在预训练编码器的同时对编码器和解码器组成的模型进行预训练,能够加快训练速度,提高训练效率。

40 云侧设备 202,用于将语音特征输入解码器,获得预测中文文本,该解码器基于对预训练中文发音单元执行文本预测任务进行预训练。

其中,文本预测任务是根据输入的预训练中文发音单元预测对应的中文文本。

示例性地,该解码器可以是任何具备解码功能的模型中的解码器,例如,解码器可以是 transformer 模型的解码器,或者,解码器可以是 BERT、CNN、LSTM、GRU 等模型的解码器。并且,解码器可以包括 X 个 block (块),其中,X 是大于或等于 1 的正整数。

45 在本说明书一个或多个实施例中,解码器可以包括解码层和文本嵌入层,先将编码器输出的语音特征输入解码层,得到预测文本特征,然后将预测文本特征输入文本嵌入层,将预测文本特征映射成概率分布向量,该概率分布向量表示了样本语音数据是某个中文文本的概率,将概率最大的中文文本确定为

样本语音数据对应的预测中文文本。

作为一种示例,解码器采用自回归的方式对语音特征进行解码,一个字一个字地进行预测,得到预测中文文本。在预训练阶段,解码器在预测当前中文文本时,其输入包括编码器输出的语音特征、以及之前的文本 ground-truth 特征,在测试阶段,解码器在预测当前中文文本时,其输入包括编码器输出的语音特征、以及文本嵌入层输出的上一个字的解码结果。

也就是说,解码器在做解码时考虑到了上下文之间的联系,得到的解码结果更加准确,进而得到的预测中文文本的准确率更高。

在本说明书一个或多个实施例中,基于对预训练中文发音单元执行文本预测任务预训练解码器,使得解码器能够学习到通过中文发音单元预测中文文本的能力,由于中文发音单元也是一种语音特征,则解码器具备了通过语音特征构建文本的能力,与此同时,对编码器和解码器组成的模型进行预训练时,便能够加快训练速度,提高训练效率。并且,由于中文语言是表意语言,根据中文文本很难确定发音,预训练中文发音单元相比中文文本更接近于语音数据的模态,因此,选择先将预训练文本数据转化成预训练中文发音单元作为模型的输入,来预测中文文本,使得预训练任务与语音识别模型的训练任务的输入比较相似,则预训练得到的能够更加适用于语音识别这一任务,提高了训练得到的语音识别模型的识别准确性。

云侧设备 202,用于基于预测中文文本和样本中文文本,对编码器和解码器组成的模型进行预训练,在达到预训练停止条件的情况下,获取预训练得到的语音识别模型的模型参数。

在本说明书一个或多个实施例中,可以根据预测中文文本和样本中文文本确定损失值,若损失值大于或等于损失阈值,则基于该损失值对编码器和解码器组成的模型的参数进行调整,即对编码器和解码器的参数进行调整,然后返回执行利用编码器对样本语音数据进行编码,直至达到预训练停止条件,停止预训练,得到语音识别模型的模型参数。

在本说明书一些实施例中,预训练停止条件可以是损失值小于损失阈值,或者,预训练停止条件可以是迭代预训练次数大于或等于次数阈值。

作为一种示例,若某次预训练后得到的损失值小于损失阈值,说明编码器和解码器组成的模型已经能够很好地进行语音识别,无需继续训练,因此停止预训练,获取预训练得到的语音识别模型的模型参数。

作为另一种示例,可以记录迭代预训练次数,并在每次确定预测中文样本后将迭代预训练次数加1,若迭代预训练次数大于次数阈值,说明对编码器和解码器组成的模型的预训练已经足够多,若继续预训练可能无法达到更好地效果,因此,停止预训练,获取预训练得到的语音识别模型的模型参数。

在本申请实施例中,云侧设备 202 预训练得到语音识别模型后,将语音识别模型的模型参数发送至端侧设备 201,以便于端侧设备 201 可以基于该语音识别模型执行语音识别任务。

参见图 3,图 3 示出了本说明书一实施例提供的一种语音识别模型的数据处理方法的数据流向图。语音识别模型包括编码器和解码器,编码器包括特征提取层、语音编码层和特征编码层,解码器包括解码层和文本嵌入层,将样本语音数据输入特征提取层,获得样本语音数据的语音表示向量,将语音表示向量输入语音编码层进行编码,再将编码结果输入特征编码层,获得语音特征,将语音特征输入解码层,获得预测文本特征,将预测文本特征输入文本嵌入层,获得预测中文文本。在得到预测中文文本之后,根据预测中文文本和样本中文文本调整包括编码器和解码器的模型中各个组件的参数,直至达到预训练停止条件,得到语音识别模型。并且,由于解码器采用自回归的方式进行解码,即在预训练阶段,解码器的输入包括编码器输出的语音特征、以及之前的文本 ground-truth 特征,在测试阶段解码器的输入包括编码器输出的语音特征、以及文本嵌入层输出的解码结果。

需要说明的是,上述内容是基于语音识别任务对语音识别模型进行预训练的过程,但正如上述描述,编码器和解码器分别通过中文发音单元预测任务和文本预测任务进行预训练,接下来,对编码器和解码器的预训练过程进行说明,在本说明书实施例中,对编码器和解码器的预训练可以同时进行。

第一部分:对编码器的预训练过程进行说明。

在本说明书一个或多个实施例中,云侧设备 202,还用于:

获取第一预训练语音集,其中,该第一预训练语音集包括多个无监督的第一预训练语音数据;利用编码器,对第一预训练语音数据进行编码,获得第一预训练语音数据对应的第一语音特征,基于第一语

音特征确定第一发音单元；对第一预训练语音数据进行掩码处理；利用编码器，对掩码处理后的第一预训练语音数据进行编码，获得掩码处理后的第一预训练语音数据对应的第二语音特征，基于第二语音特征确定第二发音单元；基于第一预训练语音数据对应的第一发音单元和第二发音单元，对编码器进行预训练。

5 其中，第一预训练语音数据是不带标注的语音数据。

在本说明书实施例中，对编码器进行预训练时，采用的预训练任务是语音掩码预测任务，在该任务中编码器输出的是语音特征，即采用语音掩码预测任务可以根据语音数据确定语音特征，进一步确定对应的发音单元，以此对编码器的参数进行调整，使得编码器能够输出更准确的语音特征。

10 作为一种示例，云侧设备 202 可以从开源的数据库中获取多个无监督的第一预训练语音数据，并将第一预训练语音数据组成第一预训练语音集，然后基于该第一预训练语音数据对编码器进行预训练。获取的第一预训练语音数据是不带任何标注的，则减少了人工标注的成本。

在本说明书一些实施例中，可以直接将第一预训练语音数据输入编码器，获得第一预训练语音数据对应的第一语音特征。

15 在本说明书另一些实施例中，云侧设备 202，还用于：提取第一预训练语音数据的频谱特征；将第一预训练语音数据的频谱特征输入编码器，获得第一预训练语音数据对应的第一语音特征。

也就是说，在利用编码器对第一预训练语音数据进行编码之前，可以提取第一预训练语音数据的频谱特征，然后将频谱特征输入编码器进行编码，得到第一预训练语音数据对应的第一语音特征。并且，由于本说明书实施例中多种预训练任务同时对编码器和解码器进行预训练，且频谱特征相比波形特征（语音数据）引入了更少的声音细节，因此，使用频谱特征作为编码器的输入，使得编码器和解码器组成的模型难以区分不同预训练任务的语音数据，从而使得不同预训练任务之间不会相互独立，能够相互促进，提高训练效果。

作为一种示例，可以使用线性预测倒谱系数算法或者梅尔频率倒谱系数或者预先训练得到的频谱特征提取模型，提取第一预训练语音数据的频谱特征。

25 作为一种示例，编码器可以包括特征提取层、语音编码层和特征编码层，将频谱特征输入编码器进行编码可以包括：将频谱特征输入特征提取层进行语音表示提取和下采样处理，得到第一预训练语音数据对应的语音表示向量，再将该语音表示向量输入语音编码层，得到初始语音特征，然后将初始语音特征输入特征编码层，得到第一预训练语音数据对应的第一语音特征。该初始语音特征和第一语音特征的区别在于，初始语音特征是通过 M 个 block 后得到的语音上下文特征，第一语音特征是通过 $M+N$ 个 block 处理后得到的语音上下文特征。

30 在确定第一预训练语音数据对应的第一发音单元后，可以将第一发音单元作为该第一预训练语音数据的标签。然后通过掩码预测的方式预测第一预训练语音数据对应的第二发音单元，根据第二发音单元和第一发音单元对编码器进行预训练。

在本说明书一些实施例中，可以提取第一预训练语音数据的频谱特征，将频谱特征输入编码器的特征提取层确定语音表示向量，然后对语音表示向量进行随机掩码，将掩码处理后的语音表示向量输入语音编码层，得到初始语音特征，将初始语音特征输入特征编码层，得到第一预训练语音数据对应的第二语音特征，进一步确定对应的第二发音单元，然后根据第一发音单元和第二发音单元确定损失值，若损失值大于或等于损失阈值，说明第二发音单元与第一发音单元的相似度较低，预测的第二发音单元不准确，即编码器还没有很好地学习到根据语音数据预测中文发音单元的能力，需要继续训练，因此，继续对编码器进行预训练，直到达到预训练停止条件。

40 示例性地，预训练停止条件可以包括损失值小于损失阈值，或者，预训练停止条件可以包括预训练次数大于或等于次数阈值。

作为一种示例，若某次预训练得到的损失值小于损失阈值，说明第二发音单元与第一发音单元的相似度较高，预测的第二发音单元比较准确，即编码器已经能够很好地根据语音数据预测中文发音单元，不用再继续训练，因此，停止对编码器的预训练。

45 作为另一种示例，可以在预训练的过程中记录预训练次数，并在每次确定第二发音单元后将预训练次数加 1，若预训练次数大于次数阈值，说明对该编码器的预训练次数已经足够多，若继续预训练可能无法达到更好地效果，因此，停止对编码器的预训练。

在本说明书实施例中,先经编码器和发音单元嵌入层确定第一预训练语音数据的第一发音单元,然后提取第一预训练语音数据的语音表示向量,对语音表示向量进行掩码处理后经编码器和发音单元嵌入层确定第二发音单元,并基于第一发音单元和第二发音单元对编码器进行预训练,使得编码器在进行中文发音单元预测时能够多关注语音中字的发音信息以及字之间的前后关系,提高了编码器执行中文发音单元预测的准确性,进而提高了编码器对语音数据的编码能力。

参见图 4a,图 4a 示出了本申请一实施例提供的一种确定中文发音单元的方法的数据流向图。确定中文发音单元用到了编码器和发音单元嵌入层,该编码器包括特征提取层、语音编码层和特征编码层。获取第一预训练语音数据,将第一预训练语音数据输入特征提取层,获得第一预训练语音数据的语音表示向量,将语音表示向量输入语音编码层进行编码,再将编码结果输入至特征编码层,得到第一预训练语音数据对应的第一语音特征,将第一语音特征输入发音单元嵌入层,可以得到第一预训练语音数据对应的第一中文发音单元。

参见图 4b,图 4b 示出了本申请一实施例提供的一种对编码器进行预训练的数据流向图。对编码器进行预训练时,用到了语音识别模型的编码器和发音单元嵌入层,编码器包括特征提取层、语音编码层和特征编码层。获取第一预训练语音数据,将第一预训练语音数据输入特征提取层,获得第一预训练语音数据的语音表示向量,对语音表示向量进行掩码处理,将掩码后的语音表示向量输入语音编码层进行编码,再将编码结果输入至特征编码层,得到第二语音特征,将第二语音特征输入发音单元嵌入层,可以得到第二中文发音单元,根据第一中文发音单元和第二中文发音单元调整编码器的参数。

进一步地,上述根据第二预训练语音数据预测第二发音单元的语音掩码预测任务,使用第一发音单元作为第二预训练语音数据的标签,让编码器学习自己的输出,可能会导致模型坍塌,即编码器在不管输入什么语音数据的情况下都输出相同的特征,此时语音掩码预测任务变得没有意义。因此,为了提高编码器的预测准确性,可以在使用语音掩码预测任务的同时使用中文发音单元预测任务对编码器进行预训练,使得编码器能够更准确地基于语音数据预测中文发音单元。

在本说明书一个或多个实施例中,云侧设备 202,还用于:

获取多个第一预训练对,其中,第一预训练对包括第二预训练语音数据和第一预训练中文发音单元;利用编码器,对第二预训练语音数据进行中文发音单元预测,获得第二预训练语音数据对应的预测中文发音单元;

基于第一预训练中文发音单元和预测中文发音单元,对编码器进行预训练。

在一些实施例中,云侧设备 201 可以从开源的预训练数据集中获取第一预训练对,基于该第一预训练对对编码器进行预训练。并且,该第一预训练对包括的第二预训练语音数据和第一预训练中文发音单元的对应关系的准确性很高,因此,云侧设备 202 基于该预训练对对编码器进行预训练,能够提高编码器预测中文发音单元的准确性。即还可以基于中文发音单元预测这一有监督任务对编码器进行预训练,其中,第一预训练中文发音单元是第二预训练语音数据的标签。

在一些实施例中,云侧设备 202 利用编码器对第二预训练语音数据进行中文发音单元预测包括:将第二预训练语音数据输入编码器,得到该第二预训练语音数据的预测语音特征,将预测语音特征输入发音单元嵌入层,得到第二预训练语音数据对应的预测中文发音单元。

在另一些实施例中,云侧设备 202 在利用编码器对第二预训练语音数据进行中文发音单元预测之前,先提取第二预训练语音数据的频谱特征,再将频谱特征输入编码器,得到该第二预训练语音数据的预测语音特征,将预测语音特征输入发音单元嵌入层,得到第二预训练语音数据对应的预测中文发音单元。

在又一些实施例中,编码器包括特征提取层、语音编码层和特征编码层,云侧设备 202 在利用编码器对第二预训练语音数据进行中文发音单元预测之前,先提取第二预训练语音数据的频谱特征,将频谱特征输入特征提取层进行语音表示提取和下采样处理,得到第二预训练语音数据对应的语音表示向量,将语音表示向量输入语音编码层和特征编码层,得到第二预训练语音数据对应的预测语音特征,将预测语音特征输入发音单元嵌入层,得到第二预训练语音数据对应的预测中文发音单元。

需要说明的是,基于第一预训练中文发音单元和预测中文发音单元,对编码器进行预训练的实现过程与上述基于预测中文文本和样本中文文本,对包括编码器和解码器的模型进行预训练的实现过程类似,其具体实现可以参见上述实施例的相关描述,本实施例对此不再赘述。

在本说明书实施例中,可以采用语音掩码预测任务和中文发音单元预测任务对编码器进行预训练,

使得编码器能够学习到基于语音数据预测中文发音单元的能力,提高了训练得到的编码器的预测准确性。另外,由于频谱特征能够表示语音中各个字的音色、音高、以及各个字的中文发音单元,因此,对编码器预训练中采用中文发音单元做为预测目标,能够使得编码器更专注得捕捉语音的字发音信息。并且,由于预训练阶段中采用了多个任务同时进行预训练,原始语音数据包含更多噪声细节,会导致模型能够分辨不同的预训练任务使用的数据,从而使得任务之间的促进和约束作用减弱,导致训练不稳定,因此,采用频谱特征作为对编码器进行预训练时整个模型的输入,能够促进不同预训练任务之间的约束,提高模型训练的稳定性,避免模型坍塌的问题。

参见图5,图5示出了本申请一实施例提供的另一种对编码器进行预训练的数据流向图。对编码器进行预训练时,用到了编码器和发音单元嵌入层,编码器包括特征提取层、语音编码层和特征编码层。将第二预训练语音数据输入特征提取层,获得第二预训练语音数据的语音表示向量,将语音表示向量输入语音编码层进行编码,再将编码结果输入特征编码层,得到第二预训练语音数据对应的预测语音特征,将预测语音特征输入发音单元嵌入层,得到预测中文发音单元,根据预测中文发音单元与第一预训练中文发音单元调整编码器的参数。

第二部分:对解码器的预训练过程进行说明。

在本说明书一个或多个实施例中,编码器包括特征编码层;云侧设备202,还用于:

获取第一预训练文本集,其中,第一预训练文本集包括多个无监督的第一预训练中文文本;

将第一预训练中文文本转换为第二预训练中文发音单元,将第二预训练中文发音单元输入特征编码层,获得第二预训练中文发音单元的语音特征;

将第二预训练中文发音单元的语音特征输入解码器,获得第二预训练中文发音单元对应的预测中文文本;

基于第二预训练中文发音单元对应的预测中文文本与第一预训练中文文本,对解码器进行预训练。

其中,特征编码层是使用中文发音单元作为输入对解码器进行预训练时使用的编码器中的编码层。由于第二预训练中文发音单元比语音数据更加抽象,因此,将第二预训练中文发音单元输入特征编码层做编码,而不需要经过编码器的其他编码层,即可得到第二预训练中文发音单元的语音特征。

在本说明书实施例中,对解码器进行预训练时,采用的预训练任务是文本预测任务,即根据输入的中文发音单元预测中文文本,以此对解码器的参数进行调整,提高解码器的文本构建能力,使得解码器能够预测出更准确的中文文本。

在一些实施例中,可以根据字典确定中文文本与中文发音单元之间的对应关系,并基于此将第一预训练中文文本转换为第二预训练中文发音单元;或者,可以预先训练得到能够实现中文文本-中文发音单元转换的模型,使用该模型将第一预训练中文文本转换为第二预训练中文发音单元。

示例性地,将第一预训练中文文本转换为第二预训练中文发音单元的过程中,不仅要转换得到中文文本的中文发音单元,还需要包括发音的声调(如阴平、阳平、上声、去声),且将不同的字的中文发音单元间隔开来,将声母与韵母间隔开来或者将各个音节间隔开来,以避免将不同的字混淆。

例如,假设第一预训练中文文本是“今天天气真不错,但下午可能下雨”,则转换得到的第二预训练中文发音单元可以是“j in1 t ian1 t ian1 qi4 zh en1 b u2 c uo4 d an4 x ia4 w u3 k e3 n eng2 x ia4 y u3”。其中,数字表示声调,1表示阴平,2表示阳平,3表示上声、4表示去声。

在一些实施例中,可以直接将第二预训练中文发音单元输入特征编码层进行语音表示提取和下采样处理,得到第二预训练中文发音单元的语音特征。

在另一些实施例中,将第二预训练中文发音单元输入特征编码层之前,可以先将第二预训练中文发音单元映射成特征矩阵,然后将该特征矩阵输入特征编码层,得到该第二预训练中文发音单元的语音特征。示例性地,可以通过中文发音单元嵌入层将第二预训练中文发音单元映射成特征矩阵。

在又一些实施例中,将第二预训练中文发音单元输入特征编码层之前,先对第二预训练中文发音单元进行掩码处理,将掩码处理后的第二预训练中文发音单元映射成特征矩阵,然后将该特征矩阵输入特征编码层,得到该第二预训练中文发音单元的语音特征。

需要说明的是,将第二预训练中文发音单元的语音特征输入解码器,获得第二预训练中文发音单元对应的预测中文文本、以及基于第二预训练中文发音单元对应的预测中文文本与第一预训练中文文本对解码器进行预训练,与上述实施例将语音特征输入解码器获得预测中文文本、以及对包括编码器和解码

器的模型进行预训练的的实现过程类似,其具体实现可以参见上述实施例的相关描述,本实施例在此不再赘述。

在本说明书实施例中,获取第一预训练文本集之后,先将第一预训练中文文本转换为第二预训练中文发音单元,再根据特征编码层确定第二预训练中文发音单元的语音特征,然后将第二预训练语音数据的语音特征输入解码器,解码器采用自回归的方式进行解码,使得解码器在进行文本预测时考虑了语音上下文之间的联系,则基于此对解码器进行预训练,提高了预训练得到的解码器做文本预测的准确性,且对解码器进行预训练时的输入是中文发音单元,与语音数据的模态更接近,更加符合下游任务,则基于此解码器做下游任务时能够提高任务的准确性。

参见图6,图6示出了本申请一实施例提供的一种对解码器进行预训练的数据流向图。解码器包括解码层和文本嵌入层,对解码器进行预训练时,还用到了发音单元嵌入层和编码器的特征编码层。对第二预训练中文发音单元进行掩码处理,将掩码后的第二预训练中文发音单元输入发音单元嵌入层,得到第二预训练中文发音单元对应的特征矩阵,然后将该特征矩阵输入编码器的特征编码层,得到该第二预训练中文发音单元的语音特征,将语音特征输入解码层,得到预测文本特征,将预测文本特征输入文本嵌入层,得到预测中文文本,根据预测中文文本和第一预训练中文文本调整解码器的参数。

进一步地,上述方法使用中文文本转换得到的中文发音单元和中文文本对解码器进行预训练,使得解码器学习了构建文本的语法规则,提高了解码器的语言建模能力。但语音识别任务与语音数据是紧密相关的,为了提高解码器对语音数据的定位和封装能力,还可以通过语音-伪标签预测任务对解码器进行预训练。

也即是,云侧设备202,还用于:

获取第二预训练语音集,其中,第二预训练语音集包括多个第三预训练语音数据,第三预训练语音数据携带目标伪标签;

利用编码器对第三预训练语音数据进行编码,获得第三预训练语音数据的语音特征;

将第三预训练语音数据的语音特征输入解码器,得到第三预训练语音数据对应的预测伪标签;

基于目标伪标签和预测伪标签,对解码器进行预训练。

在本说明书一种实施方式中,可以从开源的预训练数据库中获取到第二预训练语音集。

在本说明书另一种实施方式中,云侧设备202获取第二预训练语音集的具体实现可以包括:

获取多个无监督的第三预训练语音数据;

将多个第三预训练语音数据输入预训练的语音编码器,获得多个第三预训练语音数据的语音特征;

对多个第三预训练语音数据的语音特征进行聚类,获得各第三预训练语音数据的目标伪标签。

也就是说,可以使用预训练的语音编码器为第三预训练语音数据打上目标伪标签。其中,目标伪标签就是为语音特征之间的相似度比较高的多个预训练语音数据设置的标签,没有实际意义,若两个第三预训练语音数据的目标伪标签相同,说明该两个第三预训练语音数据的语音特征相似程度比较高。

作为一种示例,预训练的语音编码器可以是基于多个预训练语音数据预先训练好的语音编码器,可以用于提取语音数据的语音特征。

在一些实施例中,对多个第三预训练语音数据的语音特征进行聚类包括确定各第三预训练语音数据的语音特征与其他第三预训练语音数据的语音特征之间的相似度,然后将相似度大于相似度阈值的第三预训练语音数据聚类在一起,为该类别设置目标伪标签。

在本说明书实施例中,通过训练过的语音编码器确定第三预训练语音数据的语音特征,再将语音特征比较相似的第三预训练语音数据聚类在一起,并设置目标伪标签,则能够得到更加准确的目标伪标签。

在获取到第三预训练语音数据和各语音数据对应的目标伪标签之后,将第三预训练语音数据输入编码器提取语音特征。需要说明的是,利用编码器对第三预训练语音数据进行编码,获得第三预训练语音数据的语音特征的实现过程与上述利用编码器对样本语音数据进行编码,获得样本语音数据的语音特征的实现过程类似,其具体实现可以参见上述实施例的相关描述,本实施例在此不再赘述。

在本说明书一些实施例中,解码器包括解码层,将语音特征输入解码器,得到第三预训练语音数据对应的预测伪标签的具体实现可以包括:将语音特征输入解码器的解码层,得到第三预训练语音数据对应的预测文本特征,将多个第三预训练语音数据对应的预测文本特征输入伪码嵌入层,确定第三预训练语音数据与各伪标签存在对应关系的概率,将概率最大的伪标签确定为第三预训练语音数据对应的预测

伪标签。

需要说明的是,确定第三预训练语音数据对应的预测文本特征的实现过程与上述实施例确定预测文本特征的实现过程类似,其具体实现可以参见上述实施例的相关描述,本实施例在此不再赘述。

- 5 在确定预测伪标签之后,可以基于目标伪标签和预测伪标签,采用序列到序列的损失函数确定损失值,在损失值大于或等于损失阈值的情况下,基于该损失值调整解码器的参数,直至达到预训练停止条件,停止对解码器的预训练。

需要说明的是,预训练停止条件与上述对编码器进行预训练的预训练停止条件相同,实现过程也类似,因此,基于目标伪标签和预测伪标签对解码器进行预训练的具体实现可参见上述实施例的相关描述,本实施例在此不再赘述。

- 10 在本说明书实施例中,通过语音-伪标签预测任务对解码器进行预训练,且由于解码器采用自回归的方式进行解码,能够从编码器的输出中提取信息,该信息与解码器生成下一个预训练语音数据对应的预测文本相关,提高了解码器对语音数据的定位和封装能力。

- 15 参见图7,图7示出了本说明书一个实施例提供的另一种对解码器进行预训练的数据流向图。解码器包括解码层和伪码嵌入层,对解码器进行预训练时,还用到了编码器,编码器包括特征提取层、语音编码层和特征编码层。获取多个第三预训练语音数据,将第三预训练语音数据输入特征提取层,得到第三预训练语音数据的语音表示向量,将语音表示向量输入语音编码层进行编码,再将编码结果输入特征编码层,得到第三预训练语音数据的语音特征,将语音特征输入解码层,得到预测文本特征,将预测文本特征输入伪码嵌入层,确定第三预训练语音数据对应的预测伪标签,根据预测伪标签和目标伪标签调整解码器的参数。

- 20 值得注意的是,在实际应用中,可以将上述提及的语音识别任务、语音掩码预测任务、中文发音单元预测任务、文本预测任务和语音-伪标签预测任务联合起来对编码器和解码器进行预训练,将所有任务的损失值加权求和,根据求和后的损失值对编码器和解码器的参数进行调整。如此,在参数调整过程中,会考虑多个任务的效果,在尽可能保证多个任务的效果均比较好的前提下进行调整,使得预训练得到的语音识别模型能够适用于多种任务,提高了训练效率。并且,将语音识别任务加入预训练任务中,使得所有任务在优化的过程中会被语音识别任务的效果所影响,从而朝着语音识别效果更高的方向更新参数,提高了预训练得到的语音识别模型的语音识别效果。另外,将频谱特征作为编码器的输入而不是语音数据,是因为频谱特征忽略了一些语音细节,使得语音识别模型更难区分不同任务的数据,从而使
- 25 使得联合训练时多任务对语音识别模型的参数的调整能够相互约束,避免模型坍塌的问题。

- 需要说明的是,上述描述记载了将五种预训练任务联合起来对包括编码器和解码器的模型进行预训练(即对编码器和解码器进行预训练),以降低语音识别模型的预训练难度,提高训练精度。但在本申请实施例中,做联合预训练之前,可以先使用文本预测任务对编码器和解码器进行预训练直至收敛,得到包括编码器和解码器的模型。一方面,文本预测任务可以理解为是语音识别任务(即语音-文本的预测任务)的简化版本,使用文本预测任务作为预训练任务,模型更容易学习且为后续联合预训练奠定了好的基础,使得联合预训练更加稳定;另一方面,文本预测任务的学习初始化了发音单元嵌入层,使得
- 30 联合预训练阶段的语音掩码预测任务不容易出现模型坍塌的问题。

也即是,编码器包括特征编码层;云侧设备202,还用于:

获取多个第二预训练对,其中,第二预训练对包括第三预训练中文发音单元和第二预训练中文文本;

将第三预训练中文发音单元输入特征编码层,获得第三预训练中文发音单元的语音特征;

- 40 将第三预训练中文发音单元的语音特征输入解码器,获得第三预训练中文发音单元对应的预测中文文本;

基于第三预训练中文发音单元对应的预测中文文本与第二预训练中文文本,对特征编码层和解码器进行预训练,获得包括编码器和解码器的模型。

- 在本说明书实施例中,使用有监督的文本预测任务对模型进行预训练,因此,可以直接获取到包括第三预训练中文发音单元和第二预训练中文文本的多个第二预训练对,将第三预训练语音数据作为特征编码层的输入,将第二预训练中文文本作为标签,根据解码器输出的预测中文文本和该标签对特征编码层和解码器进行预训练,得到模型。
- 45

在本说明书一些实施例中,云侧设备202可以从开源的预训练数据集中获取多个第二预训练对。

在本说明书另一些实施例中，云侧设备 202 获取多个第二预训练对的具体实现可以包括：

获取多个第二预训练中文文本；

将多个第二预训练中文文本分别转换为第三预训练中文发音单元；

确定第二预训练中文文本与对应的第三预训练中文发音单元组成的第二预训练对。

5 也就是说，由于语音识别模型是根据语音数据预测中文文本，因此其训练数据通常是语音数据和中文文本，但由于中文文本与语音数据的模态差距较大，所以选择与语音数据的模态接近的中文发音单元做预训练，因此，可以先获取多个第二预训练中文文本，将多个第二预训练中文文本转换为第三预训练中文发音单元，将第二预训练中文文本和第三预训练语音数据组成第二预训练对，则该预训练对中第二预训练中文文本与第三预训练中文发音单元的对应关系的准确率是非常高的，基于该第二预训练对对模型进行预训练，能够使得模型学习到根据中文发音单元预测中文文本的能力，提高了模型的语音识别准确性。

10 在获取到第二预训练对之后，将第二预训练对中的第三预训练语音数据输入特征编码层进行处理。需要说明的是，将第三预训练中文发音单元输入特征编码层，获得第三预训练中文发音单元的语音特征、以及将第三预训练语音数据的语音特征输入解码器，获得第三预训练中文发音单元对应的预测中文文本，与上述实施例中使用文本预测任务对解码器进行预训练的实现过程类似，其具体实现可以参见上述实施例的相关描述，本实施例在此不再赘述。

15 在本说明书一些实施例中，确定第三预训练中文发音单元对应的预测中文文本之后，基于预测中文文本与第二预训练中文文本确定损失值，若损失值大于或等于损失阈值，说明模型的语音特征预测和文本预测效果均不太好，因此，继续对特征编码层和解码器进行预训练，直到达到预训练停止条件。

20 示例性地，预训练停止条件可以包括损失值小于损失阈值，或者，预训练停止条件可以包括预训练次数大于或等于次数阈值。

作为一种示例，若某次预训练得到的损失值小于损失阈值，说明模型的语音特征预测和文本预测效果比较好，即模型已经能够很好地根据中文发音单元预测中文文本，不用再继续训练，因此，停止对模型的预训练，即停止对特征编码层和解码器的参数调整。

25 作为另一种示例，可以在预训练的过程中记录预训练次数，并在每次确定第二语音特征后将与预训练次数加 1，若预训练次数大于次数阈值，说明对模型的预训练次数已经足够多，若继续预训练可能无法达到更好地效果，因此，停止对模型的预训练，即停止对特征编码层和解码器的参数调整。

30 在本说明书实施例中，在使用五种预训练任务对编码器和解码器进行联合预训练之前，通过文本预测任务对特征编码层和解码器进行预训练，得到包括编码器和解码器的模型，使得编码器中的特征编码层具备预测语音特征的能力，解码器具备预测文本的能力，由于中文发音单元比语音数据少了说话人的情绪、噪音等干扰，训练效果更好，且使用文本预测任务提前预训练模型，提前初始化了发音单元嵌入层的处理规则，使得后续的预训练过程更加稳定。

35 需要说明的是，上述描述记载了使用五种预训练任务对包括编码器和解码器的模型进行预训练的过程，具体包括先使用文本预测任务对编码器的特征编码层和解码器进行预训练，得到包括编码器和解码器的模型，然后使用语音掩码预测任务、中文发音单元预测任务、语音识别任务、文本预测任务和语音-伪标签预测任务联合对包括编码器和解码器的模型进行预训练，得到语音识别模型，并将语音识别模型的模型参数发送至端侧设备 201，端侧设备 201 利用该语音识别模型对待识别语音数据进行语音识别，得到待识别语音数据对应的目标文本。

40 由于具体的下游任务不同，因此，端侧设备 201 在利用语音识别模型对待识别语音数据进行语音识别之前，可以对语音识别模型的参数进行微调。

也即是，端侧设备 201，还用于：

获取校验集，其中，校验集包括多个语音校验对和多个中文发音单元校验对，语音校验对包括校验语音数据和对应的校验中文文本，中文发音单元校验对包括校验语音数据和对应的校验中文发音单元；

45 利用语音识别模型的编码器，对校验语音数据进行中文发音单元预测，获得校验语音数据的语音特征和预测中文发音单元；

将校验语音数据的语音特征输入语音识别模型的解码器，得到校验语音数据对应的预测中文文本；

基于预测中文发音单元、校验中文发音单元、预测中文文本和校验中文文本，对语音识别模型进行

微调, 在达到微调停止条件的情况下, 得到微调完成的语音识别模型。

在本说明书一个或多个实施例中, 对语音识别模型的微调是基于有监督的微调任务实现的。由于语音识别模型的下游任务是语音识别任务, 因此, 在做微调时, 一个有监督的微调任务是语音识别任务, 且语音识别任务是根据语音数据确定中文文本, 因此需要获取包括校验语音数据和对应的校验中文文本的语音校验对。并且, 由于语音识别模型的编码器能够生成适用于语音识别的语音特征, 由于中文语音识别中, 中文发音单元为中文文本和语音数据建立了联系, 即中文文本和语音数据都可以唯一映射到同一个中文发音单元序列, 因此提高编码器的中文发音单元预测能力, 能够使编码器生成更适合语音识别的语音特征。因此, 可以设置另一个有监督的微调任务即中文发音单元预测任务, 且中文发音单元预测任务是根据语音数据确定中文发音单元, 因此需要获取包括校验语音数据和对应的校验中文发音单元的中文发音单元校验对。

作为一种示例, 可以从开源的校验数据库中获取语音校验对和中文发音单元校验对, 或者, 可以人工生成语音校验对和中文发音单元校验对。

具体实现中, 为了获得中文发音单元, 可以在语音识别模型中增加发音单元嵌入层, 该发音单元嵌入层与编码器连接, 用于将语音特征映射成中文发音单元, 则将编码器输出的语音特征输入发音单元嵌入层, 可以得到校验语音数据对应的预测中文发音单元。

在本说明书一些实施例中, 编码器可以包括特征提取层、语音编码层和特征编码层, 利用语音识别模型的编码器, 对校验语音数据进行中文发音单元预测包括: 先将校验语音数据输入特征提取层, 进行语音表示提取和下采样处理, 得到校验语音数据的语音表示向量, 然后将语音表示向量输入语音编码层和特征编码层, 得到校验语音数据的语音特征, 该语音特征是结合了语音数据上下文语音后得到的, 再将校验语音数据的语音特征输入发音单元嵌入层, 得到校验语音数据对应的预测中文发音单元。

需要说明的是, 将校验语音数据的语音特征输入语音识别模型的解码器, 得到校验语音数据对应的预测中文文本的实现与将语音特征输入解码器, 获得预测中文文本的实现过程类似, 具体实现可以参见上述实施例的相关描述, 本实施例在此不再赘述。

在本说明书一些实施例中, 获得预测中文发音单元和校验中文文本之后, 可以根据预测中文发音单元和样本中文发音单元确定第一损失值, 根据预测中文文本和校验中文文本确定第二损失值, 将第一损失值与第二损失值求和得到第三损失值, 若第三损失值大于或等于损失阈值, 则基于该损失值对语音识别模型的参数(包括解码器和编码器的参数)进行微调, 然后返回执行利用语音识别模型的编码器, 对校验语音数据进行中文发音单元预测, 获得校验语音数据的语音特征和预测中文发音单元的步骤, 直至达到微调停止条件, 停止对语音识别模型的参数的微调, 得到微调完成的语音识别模型。

在本说明书一些实施例中, 微调停止条件可以是损失值小于损失阈值, 或者, 微调停止条件可以是迭代微调次数大于或等于次数阈值。

作为一种示例, 若某次微调后得到的损失值小于损失阈值, 说明语音识别模型已经能够很好地进行语音识别, 无需再调整参数, 因此停止微调, 得到微调完成的语音识别模型。

作为另一种示例, 可以记录迭代微调次数, 并在每次确定预测中文样本后将迭代微调次数加1, 若迭代微调次数大于次数阈值, 说明对该语音识别模型的参数微调次数已经足够多, 若继续微调可能无法达到更好地效果, 因此, 停止微调, 得到微调完成的语音识别模型。

在本说明书实施例中, 使用语音识别任务和中文发音单元预测任务对语音识别模型进行微调, 即将两种有监督的任务联合起来对语音识别模型进行微调, 使得对语音识别模型的参数的微调受两个任务的影响, 不仅提高了语音识别模型的训练效率和语音识别模型的识别准确性, 而且使得语音识别模型能够适用于更多下游任务, 提高了训练得到的语音识别模型的适用性。

在本说明书另一个可选的实施方式中, 端侧设备 201 获取校验集的具体实现可以包括:

获取多个语音校验对, 其中, 语音校验对包括校验语音数据和校验中文文本;

对各校验中文文本分别进行中文发音单元转换, 得到各校验语音数据对应的校验中文发音单元;

确定校验语音数据与对应的校验中文发音单元组成的中文发音单元校验对;

确定多个语音校验对和多个中文发音单元校验对组成的校验集。

也就是说, 可以先获取多个语音校验对, 然后将语音校验对中的校验中文文本转换成校验中文发音单元, 由于校验中文文本与校验语音数据之间存在对应关系, 则校验中文发音单元与校验语音数据之间

也存在对应关系,则校验中文发音单元与对应的校验语音数据组成中文发音单元校验对,多个中文发音单元校验对和多个语音校验对便组成了校验集。

作为一种示例,可以根据字典确定中文文本与中文发音单元之间的对应关系,并基于此将校验中文文本转换为校验中文发音单元;或者,可以预先训练得到能够实现文本-中文发音单元转换的模型,使用该模型将校验中文文本转换成校验语音数据。

示例性地,中文发音单元可以是拼音、音节或中文音素,将校验中文文本转换为校验中文发音单元的过程中,不仅要转换得到中文文本的拼音,还需要包括发音的声调(如阴平、阳平、上声、去声),且将不同的字的拼音间隔开来,将声母与韵母间隔开来,以避免将不同的字混淆。

例如,假设校验中文文本是“今天天气真不错,但下午可能下雨”,则校验中文发音单元可以是“j in1 t ian1 t ian1 qi4 zh en1 b u2 c uo4 d an4 x ia4 w u3 k e3 n eng2 x ia4 y u3”。其中,数字表示声调,1表示阴平,2表示阳平,3表示上声、4表示去声。

在该种情况下,校验语音数据、校验中文发音单元和校验中文文本三者之间是存在对应关系的,则可以理解为语音识别任务和中文发音单元预测任务使用的校验对存在对应关系,如此能够提高联合微调的训练精度。

参见图8,图8示出了本说明书一实施例提供的一种对语音识别模型进行微调的方法的数据流向图。语音识别模型包括编码器、发音单元嵌入层和解码器,编码器包括特征提取层、语音编码层和特征编码层,解码器包括解码层和文本嵌入层。将校验语音数据输入特征提取层,获得校验语音数据的语音表示向量,将语音表示向量输入语音编码层进行编码,再将编码结果输入特征编码层,获得语音特征,将语音特征输入发音单元嵌入层,得到预测中文发音单元,将语音特征输入解码层,可以获得预测文本特征,将预测文本特征输入文本嵌入层,可以得到预测中文文本,在得到预测中文发音单元和预测中文文本之后,根据预测中文发音单元、校验中文发音单元、预测中文文本和样本中文文本,对语音识别模型中编码器和解码器的参数进行微调(即调整参数),直至达到微调停止条件,得到微调完成的语音识别模型。

需要说明的是,本方案对语音识别模型的预训练包括三个阶段。第一阶段:使用文本预测任务对模型整体进行训练,得到包括编码器和解码器的模型;第二阶段:使用中文发音单元预测任务、语音掩码预测任务对编码器进行预训练,使用文本预测任务和语音-伪标签预测任务对解码器进行预训练,使用有监督的语音识别任务对模型整体进行训练,且该五个任务可以分开执行也可以同时执行,得到语音识别模型;第三阶段:使用有监督的语音识别任务和有监督的中文发音单元预测任务对语音识别模型的参数进行微调,得到微调后的语音识别模型。并且,第一阶段和第二阶段由云侧设备202执行,第三阶段由端侧设备201执行,或者,三个阶段可以都由云侧设备202执行。

另外,第一阶段采用文本预测任务是由于中文发音单元比语音数据少了很多干扰,且文本预测任务的训练提前初始化了发音单元嵌入层的使用规则,使得第二阶段的预训练更加稳定;第二阶段使用5个任务联合训练,提高了训练效率;第三阶段提前将下游的语音识别任务加入到微调任务中,使得所有任务在优化模型参数的过程中会被语音识别任务影响,从而会朝着语音识别效果更好的方向更新参数,能够提前评测下游任务的效果,提高工作效率。

应用于本说明书实施例的方案,在对语音识别模型进行训练之前,已经对编码器和解码器进行预训练,则在训练得到语音识别模型时需要的样本语音数据和样本中文文本的数量比较少,减轻了标注人员的负担,降低了获取标注数据的难度。针对中文文本是表意语言这一特性,即语音数据和中文文本之间差距大,同一个发音可能对应上百个汉字,本方案将中文发音单元这个模态加入到模型的预训练过程中,这是因为中文发音单元是建立语音数据和中文文本之间关系的桥梁,即语音数据和中文文本都可以唯一地映射到一个中文发音单元序列。在预训练过程中,基于对预训练语音数据执行语音掩码预测任务和中文发音单元预测任务对编码器进行预训练,这两个任务都是将语音数据映射成中文发音单元序列,这使得编码器能够捕捉语音数据中的发音信息,有利于语音识别。并且,基于对预训练中文发音单元执行文本预测任务和语音-伪标签预测任务对编码器进行预训练,同时解码器具备了通过语音特征构建文本的能力,提高了解码器的语言建模能力。通过预训练使得编码器和解码器具备一定的语音识别能力,因此再对预训练之后的编码器和解码器进行训练,能够提高训练效率和训练精度。另外,在预训练过程中使用的模型的输入是预训练语音数据或预训练中文发音单元,二者与语音识别模型应用时输入的语音数据模态相近,因此,在使用语音识别模型对待识别语音数据进行识别时,可以提高识别准确性。并且,通

过大量低成本的无标注语音数据和无标注中文文本，只利用少量的语音-文本标注数据，就可以训练得到准确率高的针对中文语言的语音识别模型，减少了对标注数据的使用，降低了人力成本，提高了训练效率。

图 9 示出了本说明书一个实施例提供的一种应用于云侧设备的语音识别模型的数据处理方法的流程图，该云侧设备与多个端侧设备连接，该语音识别模型的数据处理方法具体包括如下步骤。

步骤 902: 获取样本集，其中，样本集包括多个样本对，样本对包括样本语音数据和样本中文文本。

步骤 904: 利用编码器对样本语音数据进行编码，获得样本语音数据的语音特征，其中，编码器基于对预训练语音数据执行中文发音单元预测任务进行预训练。

步骤 906: 将语音特征输入解码器，获得预测中文文本，其中，解码器基于对预训练中文发音单元执行文本预测任务进行预训练。

步骤 908: 基于预测中文文本和样本中文文本，对包括编码器和解码器的模型进行预训练，在达到预训练停止条件的情况下，获取预训练得到的语音识别模型的模型参数。

步骤 910: 向第一端侧设备发送预训练得到的语音识别模型的模型参数，其中，第一端侧设备是多个端侧设备中的任意一个。

在本说明书一个或多个实施例中，基于对预训练语音数据执行中文发音单元对编码器进行预训练的具体实现可以包括：

获取第一预训练语音集，其中，第一预训练语音集包括多个无监督的第一预训练语音数据；

利用编码器，对第一预训练语音数据进行编码，获得第一预训练语音数据对应的第一语音特征，基于第一语音特征确定第一发音单元；

对第一预训练语音数据进行掩码处理；

利用编码器，对掩码处理后的第一预训练语音数据进行编码，获得掩码处理后的第一预训练语音数据对应的第二语音特征，基于第二语音特征确定第二发音单元；

基于第一预训练语音数据对应的第一发音单元和第二发音单元，对编码器进行预训练。

在本说明书一个或多个实施例中，利用编码器对第一预训练语音数据进行编码，获得第一预训练语音数据对应的第一语音特征之前，还包括：

提取第一预训练语音数据的频谱特征；

利用编码器对第一预训练语音数据进行编码，获得第一预训练语音数据对应的第一语音特征，包括：

将第一预训练语音数据的频谱特征输入编码器，获得第一预训练语音数据对应的第一语音特征。

在本说明书一个或多个实施例中，基于对预训练语音数据执行中文发音单元对编码器进行预训练的具体实现可以包括：

获取多个第一预训练对，其中，第一预训练对包括第二预训练语音数据和第一预训练中文发音单元；

利用编码器，对第二预训练语音数据进行中文发音单元预测，获得第二预训练语音数据对应的预测中文发音单元；

基于第一预训练中文发音单元和预测中文发音单元，对编码器进行预训练。

在本说明书一个或多个实施例中，编码器包括特征编码层；基于对预训练中文发音单元执行文本预测任务对解码器进行预训练的具体实现可以包括：

获取第一预训练文本集，其中，第一预训练文本集包括多个无监督的第一预训练中文文本；

将第一预训练中文文本转换为第二预训练中文发音单元，将第二预训练中文发音单元输入特征编码层，获得第二预训练中文发音单元的语音特征；

将第二预训练中文发音单元的语音特征输入解码器，获得第二预训练中文发音单元对应的预测中文文本；

基于第二预训练中文发音单元对应的预测中文文本与第一预训练中文文本，对解码器进行预训练。

在本说明书一个或多个实施例中，基于对预训练中文发音单元执行文本预测任务对解码器进行预训练的具体实现可以包括：

获取第二预训练语音集，其中，第二预训练语音集包括多个第三预训练语音数据，第三预训练语音数据携带目标伪标签；

利用编码器对第三预训练语音数据进行编码，获得第三预训练语音数据的语音特征；

将第三预训练语音数据的语音特征输入解码器,得到第三预训练语音数据对应的预测伪标签;
基于目标伪标签和预测伪标签,对解码器进行预训练。

在本说明书一个或多个实施例中,获取第二预训练语音集的具体实现可以包括:

获取多个无监督的第三预训练语音数据;

5 将多个第三预训练语音数据输入预训练的语音编码器,获得多个第三预训练语音数据的语音特征;
对多个第三预训练语音数据的语音特征进行聚类,获得各第三预训练语音数据的目标伪标签。

在本说明书一个或多个实施例中,编码器包括特征编码层;还包括:

获取多个第二预训练对,其中,第二预训练对包括第三预训练中文发音单元和第二预训练中文文本;

将第三预训练中文发音单元输入特征编码层,获得第三预训练中文发音单元的语音特征;

10 将第三预训练中文发音单元的语音特征输入解码器,获得第三预训练中文发音单元对应的预测中文文本;

基于第三预训练中文发音单元对应的预测中文文本与第二预训练中文文本,对特征编码层和解码器进行预训练,获得包括编码器和解码器的模型。

需要说明的是,应用于云侧设备的语音识别模型的数据处理方法的具体实现与上述语音识别模型的数据处理系统中云侧设备执行的操作相同,具体实现可以参见上述实施例的相关描述,本实施例在此不再赘述。

应用于本说明书实施例的方案,在对进行训练得到语音识别模型之前,已经对的编码器和解码器进行预训练,则在训练得到语音识别模型时需要的样本语音数据和样本中文文本的数量比较少,减轻了标注人员的负担,降低了获取标注数据的难度。针对中文数据表意语言的特性,即语音和文本之间差距大,20 同一个发音可能对应上百个汉字,我们将发音单元这个模态加入到模型的预训练过程中。这是因为发音单元是建立语音和文本之间关系的桥梁,即语音和文本都可以唯一的映射到一个发音单元序列。在预训练过程中,通过对预训练语音数据执行语音掩码预测任务和发音单元预测任务预训练得到编码器,这两个任务都是将语音数据映射成发音单元序列,这使得编码器能够捕捉语音信号中的发音信息,有利于语音识别。并且,基于对预训练中文发音单元执行文本预测任务预训练得到解码器,解码器具备了通过语音特征构建文本的能力,提高了解码器的语言建模能力。通过预训练使得编码器和解码器具备一定的语音识别能力,因此再对预训练之后的编码器和解码器进行训练,能够提高训练效率和训练精度,并且,在预训练过程中使用的模型的输入是预训练语音数据或预训练中文发音单元,二者与语音识别模型应用时输入的语音数据模态相近,因此,在使用语音识别模型对待识别语音数据进行识别时,可以提高识别准确性。并且,通过大量低成本的无标注语音数据和无标注中文文本,只利用少量的语音-文本标注数据,30 就可以训练得到准确率高的针对中文语言的语音识别模型,减少了对标注数据的使用,降低了人力成本,提高了训练效率。

图 10 示出了本说明书一个实施例提供的一种应用于端侧设备的语音识别方法的流程图,该端侧设备与云侧设备连接,该语音识别方法具体包括如下步骤。

步骤 1002: 获取待识别语音数据。

35 步骤 1004: 利用语音识别模型的编码器,对待识别语音数据进行编码,获得待识别语音数据的语音特征,其中,语音识别模型是云侧设备通过上述语音识别模型的数据处理方法预训练得到的。

在本说明书一个或多个实施例中,在利用语音识别模型的编码器,对待识别语音数据进行编码之前,还包括:

40 获取校验集,其中,校验集包括多个语音校验对和多个中文发音单元校验对,语音校验对包括校验语音数据和对应的校验中文文本,中文发音单元校验对包括校验语音数据和对应的校验中文发音单元;利用语音识别模型的编码器,对校验语音数据进行中文发音单元预测,获得校验语音数据的语音特征和预测中文发音单元;将校验语音数据的语音特征输入语音识别模型的解码器,得到校验语音数据对应的预测中文文本;基于预测中文发音单元、校验中文发音单元、预测中文文本和校验中文文本,对语音识别模型进行微调,在达到微调停止条件的情况下,得到微调完成的语音识别模型。

45 在微调完成后,使用微调完成的语音识别模型对待识别语音数据进行识别,可以得到待识别语音数据对应的目标文本。

步骤 1006: 将语音特征输入语音识别模型的解码器,获得待识别语音数据对应的目标文本。

在本说明书一个或多个实施例中，参见图 11，图 11 示出了本说明书一个实施例提供的一种语音识别模型执行语音识别任务的数据流向图。语音识别模型包括编码器和解码器，将待识别语音数据输入语音识别模型的编码器，得到待识别语音数据的语音特征，解码器包括解码层和文本嵌入层，将语音特征输入解码层，得到预测文本特征，将预测文本特征输入文本嵌入层，输出待识别语音数据对应的目标文本。

在本说明书一个或多个实施例中，获取待识别语音数据的具体实现可以包括：

接收语音识别请求，其中，语音识别请求携带待识别语音数据；

从语音识别请求中获取待识别语音数据；

相应的，步骤 1006 之后，还包括：

将目标文本发送至前端显示；

接收用户在前端输入的目标文本对应的修订文本；

根据修订文本和待识别语音数据，对语音识别模型进行更新，获得更新后的语音识别模型。

作为一种示例，将目标文本反馈给前端显示后，用户可以对该目标文本进行校正，接收用户在前端输入的目标文本对应的修订文本，然后可以根据修订文本和待识别语音数据对语音识别模型进行更新，进而提高语音识别模型的语音识别准确性。

需要说明的是，应用于端侧设备的语音识别方法的具体实现与上述语音识别模型的数据处理系统中端侧设备执行的操作相同，具体实现可以参见上述实施例的相关描述，本实施例在此不再赘述。

应用于本说明书实施例的方案，在对进行训练得到语音识别模型之前，已经对编码器和解码器进行预训练，则在训练得到语音识别模型时需要的样本语音数据和样本中文文本的数量比较少，减轻了标注人员的负担，降低了获取标注数据的难度。针对中文数据表意语言的特性，即语音和文本之间差距大，同一个发音可能对应上百个汉字，我们将发音单元这个模态加入到模型的预训练过程中。这是因为发音单元是建立语音和文本之间关系的桥梁，即语音和文本都可以唯一的映射到一个发音单元序列。在预训练过程中，通过对预训练语音数据执行语音掩码预测任务和发音单元预测任务预训练得到编码器，这两个任务都是将语音数据映射成发音单元序列，这使得编码器能够捕捉语音信号中的发音信息，有利于语音识别。并且，基于对预训练中文发音单元执行文本预测任务预训练得到解码器，解码器具备了通过语音特征构建文本的能力，提高了解码器的语言建模能力。通过预训练使得编码器和解码器具备一定的语音识别能力，因此再对预训练之后的编码器和解码器进行训练，能够提高训练效率和训练精度，并且，在预训练过程中使用的模型的输入是预训练语音数据或预训练中文发音单元，二者与语音识别模型应用时输入的语音数据模态相近，因此，在使用语音识别模型对待识别语音数据进行识别时，可以提高识别准确性。并且，通过大量低成本的无标注语音数据和无标注中文文本，只利用少量的语音-文本标注数据，就可以训练得到准确率高的针对中文语言的语音识别模型，减少了对标注数据的使用，降低了人力成本，提高了训练效率。

下述结合附图 12，以本说明书提供的语音识别模型的数据处理方法在针对中文语言的语音识别模型的应用为例，对所述语音识别模型的数据处理方法进行进一步说明。其中，图 12 示出了本说明书一个实施例提供的一种语音识别模型的数据处理方法的处理过程流程图，具体包括以下步骤。

步骤 1202：获取多个预训练对，该预训练对包括预训练中文发音单元和预训练中文文本。

步骤 1204：对预训练中文发音单元进行掩码处理，得到掩码后的预训练中文发音单元。

步骤 1206：将掩码后的预训练中文发音单元输入发音单元嵌入层，得到掩码后的预训练中文发音单元对应的特征矩阵；将特征矩阵输入特征编码层，得到该掩码后的预训练中文发音单元的语音特征。

步骤 1208：将语音特征输入解码层，得到预测文本特征；将预测文本特征输入文本嵌入层，得到预测中文文本。

步骤 1210：根据预训练中文文本和预测中文文本对模型进行预训练，得到包括编码器和解码器的模型。

示例性地，步骤 1202-步骤 1210 是第一阶段使用文本预测任务对包括编码器和解码器的模型进行预训练的过程。

步骤 1212：获取预训练语音数据和对应的第一中文发音单元。

步骤 1214：将预训练语音数据输入特征提取层，得到语音表示向量，对语音表示向量进行掩码处

理。

步骤 1216: 将掩码处理后的语音表示向量输入语音编码层和特征编码层, 得到第二语音特征, 将第二语音特征输入发音单元嵌入层, 得到第二中文发音单元。

5 步骤 1218: 根据第一中文发音单元和第二中文发音单元确定损失值, 基于该损失值对编码器进行预训练。

示例性地, 步骤 1212-步骤 1218 是第二阶段中, 使用语音掩码预测任务对编码器进行预训练的过程。

步骤 1220: 获取预训练语音数据和对应的预训练中文发音单元, 将预训练语音数据输入特征提取层, 获得预训练语音数据的语音表示向量。

10 步骤 1222: 将语音表示向量输入语音编码层和特征编码层, 得到预训练语音数据对应的预测中文发音单元。

步骤 1224: 根据预测中文发音单元和预训练中文发音单元对编码器进行预训练。

示例性地, 步骤 1220-步骤 1224 是第二阶段中, 使用中文发音单元预测任务对编码器进行预训练的过程。

15 步骤 1226: 获取预训练中文发音单元和对应的预训练中文文本, 对预训练中文发音单元进行掩码处理。

步骤 1228: 将掩码处理后的预训练中文发音单元输入发音单元嵌入层, 得到预训练中文发音单元对应的特征矩阵, 将该特征矩阵输入特征编码层, 得到该预训练中文发音单元的语音特征。

20 步骤 1230: 将语音特征输入解码层, 得到预测文本特征, 将预测文本特征输入文本嵌入层, 得到预测中文文本。

步骤 1232: 根据预测中文文本和预训练中文文本对解码器进行预训练。

示例性地, 步骤 1226-步骤 1232 是第二阶段中, 使用文本预测任务对解码器进行预训练的过程。

步骤 1234: 获取预训练语音数据, 将预训练语音数据输入特征提取层, 得到预训练语音数据的语音表示向量。

25 步骤 1236: 将语音表示向量输入语音编码层和特征编码层, 得到预训练语音数据的语音特征。

步骤 1238: 将语音特征输入解码层, 得到预测文本特征, 将预测文本特征输入伪码嵌入层, 确定预训练语音数据对应的预测伪标签。

步骤 1240: 根据预测伪标签和目标伪标签对解码器进行预训练。

30 示例性地, 步骤 1234-步骤 1240 是第二阶段中, 使用语音-伪标签预测任务对解码器进行预训练的过程。

步骤 1242: 获取样本语音数据和样本中文文本, 将样本语音数据输入特征提取层, 获得样本语音数据的语音表示向量。

步骤 1244: 将语音表示向量输入语音编码层和特征编码层, 获得语音特征。

35 步骤 1246: 将语音特征输入解码层, 获得预测文本特征, 将预测文本特征输入文本嵌入层, 获得预测中文文本。

步骤 1248: 根据预测中文文本和样本中文文本对进行训练。

示例性地, 步骤 1242-步骤 1248 是第二阶段中, 使用语音识别任务对包括编码器和解码器的模型进行预训练的过程。

40 步骤 1250: 获取语音校验对和中文发音单元校验对, 该语音校验对包括校验语音数据和对应的校验中文文本, 该中文发音单元校验对包括校验语音数据和对应的校验中文发音单元。

步骤 1252: 将校验语音数据输入特征提取层, 获得校验语音数据的语音表示向量。

步骤 1254: 将语音表示向量输入语音编码层和特征编码层, 得到语音特征, 将语音特征输入中文发音单元嵌入层, 得到预测中文发音单元。

45 步骤 1256: 将语音特征输入解码层, 获得预测文本特征, 将预测文本特征输入文本嵌入层, 可以得到预测中文文本。

步骤 1258: 根据预测中文发音单元、校验中文发音单元、预测中文文本和样本中文文本对语音识别模型的参数进行微调。

示例性地,步骤 1250-步骤 1258 是第三阶段使用语音识别任务和中文发音单元预测任务对语音识别模型的参数进行微调的过程。

参见图 13,图 13 示出了本申请实施例提供的一种对语音识别模型进行联合训练时的数据流向图。图中线 1 表示执行语音掩码预测任务时的数据流向,线 2 表示执行中文发音单元预测任务时的数据流向,线 3 表示执行语音-伪标签预测任务时的数据流向,线 4 表示执行语音识别任务时的数据流向,线 5 表示执行文本预测任务时的数据流向,各个任务中数据流向的具体过程可以参见上述图 3、图 4a、图 4b、图 5、图 6 和图 7 的相关描述,本实施例在此不再赘述。

应用于本说明书实施例的方案,在对进行训练得到语音识别模型之前,已经对编码器和解码器进行预训练,则在训练得到语音识别模型时需要的样本语音数据和样本中文文本的数量比较少,减轻了标注人员的负担,降低了获取标注数据的难度。针对中文数据表意语言的特性,即语音和文本之间差距大,同一个发音可能对应上百个汉字,我们将发音单元这个模态加入到模型的预训练过程中。这是因为发音单元是建立语音和文本之间关系的桥梁,即语音和文本都可以唯一的映射到一个发音单元序列。在预训练过程中,通过对预训练语音数据执行语音掩码预测任务和发音单元预测任务预训练得到编码器,这两个任务都是将语音数据映射成发音单元序列,这使得编码器能够捕捉语音信号中的发音信息,有利于语音识别。并且,基于对预训练中文发音单元执行文本预测任务预训练得到解码器,解码器具备了通过语音特征构建文本的能力,提高了解码器的语言建模能力。通过预训练使得编码器和解码器具备一定的语音识别能力,因此再对预训练之后的编码器和解码器进行训练,能够提高训练效率和训练精度,并且,在预训练过程中使用的模型的输入是预训练语音数据或预训练中文发音单元,二者与语音识别模型应用时输入的语音数据模态相近,因此,在使用语音识别模型对待识别语音数据进行识别时,可以提高识别准确性。并且,通过大量低成本的无标注语音数据和无标注中文文本,只利用少量的语音-文本标注数据,就可以训练得到准确率高的针对中文语言的语音识别模型,减少了对标注数据的使用,降低了人力成本,提高了训练效率。

与上述应用于云侧设备的语音识别模型的数据处理方法实施例相对应,本说明书还提供了应用于云侧设备的语音识别模型的数据处理装置实施例,图 14 示出了本说明书一个实施例提供的一种应用于云侧设备的语音识别模型的数据处理装置的结构示意图。如图 14 所示,该装置包括:

第一获取模块 1402,被配置为获取样本集,其中,样本集包括多个样本对,样本对包括样本语音数据和样本中文文本;

第一编码模块 1404,被配置为利用编码器对样本语音数据进行编码,获得样本语音数据的语音特征,其中,编码器基于对预训练语音数据执行中文发音单元预测任务进行预训练;

第一解码模块 1406,被配置为将语音特征输入解码器,获得预测中文文本,其中,解码器基于对预训练中文发音单元执行文本预测任务进行预训练;

预训练模块 1408,被配置为基于预测中文文本和样本中文文本,对包括编码器和解码器的模型进行预训练,在达到预训练停止条件的情况下,获取预训练得到的语音识别模型的模型参数;

第一发送模块 1410,被配置为向第一端侧设备发送预训练得到的语音识别模型的模型参数,其中,第一端侧设备是多个端侧设备中的任意一个。

在本说明书一个或多个实施例中,该装置还包括编码器预训练模块,该编码器预训练模块,被配置为:

获取第一预训练语音集,其中,第一预训练语音集包括多个无监督的第一预训练语音数据;

利用编码器,对第一预训练语音数据进行编码,获得第一预训练语音数据对应的第一语音特征,基于第一语音特征确定第一发音单元;

对第一预训练语音数据进行掩码处理;

利用编码器,对掩码处理后的第一预训练语音数据进行编码,获得掩码处理后的第一预训练语音数据对应的第二语音特征,基于第二语音特征确定第二发音单元;

基于第一预训练语音数据对应的第一发音单元和第二发音单元,对编码器进行预训练。

在本说明书一个或多个实施例中,该编码器预训练模块,进一步被配置为:

提取第一预训练语音数据的频谱特征;

将第一预训练语音数据的频谱特征输入编码器,获得第一预训练语音数据对应的第一语音特征。

- 在本说明书一个或多个实施例中，该编码器预训练模块，被配置为：
 获取多个第一预训练对，其中，第一预训练对包括第二预训练语音数据和第一预训练中文发音单元；
 利用编码器，对第二预训练语音数据进行中文发音单元预测，获得第二预训练语音数据对应的预测中文发音单元；
- 5 基于第一预训练中文发音单元和预测中文发音单元，对编码器进行预训练。
- 在本说明书一个或多个实施例中，编码器包括特征编码层；
 该装置还包括解码器预训练模块，该解码器预训练模块，被配置为：
 获取第一预训练文本集，其中，第一预训练文本集包括多个无监督的第一预训练中文文本；
 将第一预训练中文文本转换为第二预训练中文发音单元，将第二预训练中文发音单元输入特征编码层，获得第二预训练中文发音单元的语音特征；
- 10 将第二预训练中文发音单元的语音特征输入解码器，获得第二预训练中文发音单元对应的预测中文文本；
 基于第二预训练中文发音单元对应的预测中文文本与第一预训练中文文本，对解码器进行预训练。
- 在本说明书一个或多个实施例中，该解码器预训练模块，被配置为：
- 15 获取第二预训练语音集，其中，第二预训练语音集包括多个第三预训练语音数据，第三预训练语音数据携带目标伪标签；
 利用编码器对第三预训练语音数据进行编码，获得第三预训练语音数据的语音特征；
 将第三预训练语音数据的语音特征输入解码器，得到第三预训练语音数据对应的预测伪标签；
 基于目标伪标签和预测伪标签，对解码器进行预训练。
- 20 在本说明书一个或多个实施例中，该解码器预训练模块，进一步被配置为：
 获取多个无监督的第三预训练语音数据；
 将多个第三预训练语音数据输入预训练的语音编码器，获得多个第三预训练语音数据的语音特征；
 对多个第三预训练语音数据的语音特征进行聚类，获得各第三预训练语音数据的目标伪标签。
- 在本说明书一个或多个实施例中，编码器包括特征编码层；
- 25 该第一获取模块，进一步被配置为：
 获取多个第二预训练对，其中，第二预训练对包括第三预训练中文发音单元和第二预训练中文文本；
 将第三预训练中文发音单元输入特征编码层，获得第三预训练中文发音单元的语音特征；
 将第三预训练中文发音单元的语音特征输入解码器，获得第三预训练中文发音单元对应的预测中文文本；
- 30 基于第三预训练中文发音单元对应的预测中文文本与第二预训练中文文本，对特征编码层和解码器进行预训练，获得包括编码器和解码器的模型。
- 应用于本说明书实施例的方案，在对进行训练得到语音识别模型之前，已经对的编码器和解码器进行预训练，则在训练得到语音识别模型时需要的样本语音数据和样本中文文本的数量比较少，减轻了标注人员的负担，降低了获取标注数据的难度。针对中文数据表意语言的特性，即语音和文本之间差距大，
- 35 同一个发音可能对应上百个汉字，我们将发音单元这个模态加入到模型的预训练过程中。这是因为发音单元是建立语音和文本之间关系的桥梁，即语音和文本都可以唯一的映射到一个发音单元序列。在预训练过程中，通过对预训练语音数据执行语音掩码预测任务和发音单元预测任务预训练得到编码器，这两个任务都是将语音数据映射成发音单元序列，这使得编码器能够捕捉语音信号中的发音信息，有利于语音识别。并且，基于对预训练中文发音单元执行文本预测任务预训练得到解码器，解码器具备了通过语音特征构建文本的能力，提高了解码器的语言建模能力。通过预训练使得编码器和解码器具备一定的语音识别能力，因此再对预训练之后的编码器和解码器进行训练，能够提高训练效率和训练精度，并且，在预训练过程中使用的模型的输入是预训练语音数据或预训练中文发音单元，二者与语音识别模型应用时输入的语音数据模态相近，因此，在使用语音识别模型对待识别语音数据进行识别时，可以提高识别准确性。并且，通过大量低成本的无标注语音数据和无标注中文文本，只利用少量的语音-文本标注数据，就可以训练得到准确率高的针对中文语言的语音识别模型，减少了对标注数据的使用，降低了人力成本，提高了训练效率。
- 45 上述为本实施例的一种应用于云侧设备的语音识别模型的数据处理装置的示意性方案。需要说明的

是,该应用于云侧设备的语音识别模型的数据处理装置的技术方案与上述应用于云侧设备的语音识别模型的数据处理方法的技术方案属于同一构思,应用于云侧设备的语音识别模型的数据处理装置的技术方案未详细描述的细节内容,均可以参见上述应用于云侧设备的语音识别模型的数据处理方法的技术方案的描述。

5 与上述语音识别方法实施例相对应,本说明书还提供了语音识别装置实施例,图15示出了本说明书一个实施例提供的一种语音识别装置的结构示意图。如图15所示,该装置包括:

第二获取模块1502,被配置为获取待识别语音数据;

第二编码模块1504,被配置为利用语音识别模型的编码器,对待识别语音数据进行编码,获得待识别语音数据的语音特征,其中,语音识别模型是云侧设备通过上述语音识别模型的数据处理方法预训练得到的;

10 第二解码模块1506,被配置为将语音特征输入语音识别模型的解码器,获得待识别语音数据对应的目标文本。

在本说明书一个或多个实施例中,该装置还包括微调模块,被配置为:

15 获取校验集,其中,校验集包括多个语音校验对和多个中文发音单元校验对,语音校验对包括校验语音数据和对应的校验中文文本,中文发音单元校验对包括校验语音数据和对应的校验中文发音单元;

利用语音识别模型的编码器,对校验语音数据进行中文发音单元预测,获得校验语音数据的语音特征和预测中文发音单元;

将校验语音数据的语音特征输入语音识别模型的解码器,得到校验语音数据对应的预测中文文本;

20 基于预测中文发音单元、校验中文发音单元、预测中文文本和校验中文文本,对语音识别模型进行微调,在达到微调停止条件的情况下,得到微调完成的语音识别模型。

在本说明书一个或多个实施例中,该装置还包括显示模块、输入模块和更新模块;

显示模块,被配置为将目标文本发送至前端显示;

输入模块,被配置为接收用户在前端输入的目标文本对应的修订文本;

25 更新模块,被配置为根据修订文本和待识别语音数据,对语音识别模型进行更新,获得更新后的语音识别模型。

应用于本说明书实施例的方案,在对进行训练得到语音识别模型之前,已经对的编码器和解码器进行预训练,则在训练得到语音识别模型时需要的样本语音数据和样本中文文本的数量比较少,减轻了标注人员的负担,降低了获取标注数据的难度。针对中文数据表意语言的特性,即语音和文本之间差距大,同一个发音可能对应上百个汉字,我们将发音单元这个模态加入到模型的预训练过程中。这是因为发音单元是建立语音和文本之间关系的桥梁,即语音和文本都可以唯一的映射到一个发音单元序列。在预训练过程中,通过对预训练语音数据执行语音掩码预测任务和发音单元预测任务预训练得到编码器,这两个任务都是将语音数据映射成发音单元序列,这使得编码器能够捕捉语音信号中的发音信息,有利于语音识别。并且,基于对预训练中文发音单元执行文本预测任务预训练得到解码器,解码器具备了通过语音特征构建文本的能力,提高了解码器的语言建模能力。通过预训练使得编码器和解码器具备一定的语音识别能力,因此再对预训练之后的编码器和解码器进行训练,能够提高训练效率和训练精度,并且,在预训练过程中使用的模型的输入是预训练语音数据或预训练中文发音单元,二者与语音识别模型应用时输入的语音数据模态相近,因此,在使用语音识别模型对待识别语音数据进行识别时,可以提高识别准确性。并且,通过大量低成本的无标注语音数据和无标注中文文本,只利用少量的语音-文本标注数据,就可以训练得到准确率高的针对中文语言的语音识别模型,减少了对标注数据的使用,降低了人力成本,提高了训练效率。

上述为本实施例的一种语音识别装置的示意性方案。需要说明的是,该语音识别装置的技术方案与上述的语音识别方法的技术方案属于同一构思,语音识别装置的技术方案未详细描述的细节内容,均可以参见上述语音识别方法的技术方案的描述。

45 图16示出了本说明书一个实施例提供的一种计算设备1600的结构框图。该计算设备1600的部件包括但不限于存储器1610和处理器1620。处理器1620与存储器1610通过总线1630相连接,数据库1650用于保存数据。

计算设备1600还包括接入设备1640,接入设备1640使得计算设备1600能够经由一个或多个网络

1660 通信。这些网络的示例包括公用交换电话网 (PSTN, Public Switched Telephone Network)、局域网 (LAN, Local Area Network)、广域网 (WAN, Wide Area Network)、个域网 (PAN, Personal Area Network) 或诸如因特网的通信网络的组合。接入设备 1640 可以包括有线或无线的任何类型的网络接口 (例如, 网络接口卡 (NIC, network interface controller)) 中的一个或多个, 诸如 IEEE802.11 无线局域网 (WLAN, Wireless Local Area Network) 无线接口、全球微波互联接入 (Wi-MAX, Worldwide Interoperability for Microwave Access) 接口、以太网接口、通用串行总线 (USB, Universal Serial Bus) 接口、蜂窝网络接口、蓝牙接口、近场通信 (NFC, Near Field Communication) 接口, 等等。

在本说明书的一个实施例中, 计算设备 1600 的上述部件以及图 16 中未示出的其他部件也可以彼此相连接, 例如通过总线。应当理解, 图 16 所示的计算设备结构框图仅仅是出于示例的目的, 而不是对本说明书范围的限制。本领域技术人员可以根据需要, 增添或替换其他部件。

计算设备 1600 可以是任何类型的静止或移动计算设备, 包括移动计算机或移动计算设备 (例如, 平板计算机、个人数字助理、膝上型计算机、笔记本计算机、上网本等)、移动电话 (例如, 智能手机)、可佩戴的计算设备 (例如, 智能手表、智能眼镜等) 或其他类型的移动设备, 或者诸如台式计算机或个人计算机 (PC, Personal Computer) 的静止计算设备。计算设备 1600 还可以是移动式或静止式的服务器。

其中, 处理器 1620 用于执行如下计算机可执行指令, 该计算机可执行指令被处理器执行时实现上述语音识别模型的数据处理方法的步骤, 或者, 实现上述语音识别方法的步骤。

上述为本实施例的一种计算设备的示意性方案。需要说明的是, 该计算设备的技术方案与上述的语音识别模型的数据处理方法或语音识别方法的技术方案属于同一构思, 计算设备的技术方案未详细描述的细节内容, 均可以参见上述语音识别模型的数据处理方法或语音识别方法的技术方案的描述。

本说明书一实施例还提供一种计算机可读存储介质, 其存储有计算机可执行指令, 该计算机可执行指令被处理器执行时实现上述语音识别模型的数据处理方法的步骤, 或者, 实现上述语音识别方法的步骤。

上述为本实施例的一种计算机可读存储介质的示意性方案。需要说明的是, 该存储介质的技术方案与上述的语音识别模型的数据处理方法或语音识别方法的技术方案属于同一构思, 存储介质的技术方案未详细描述的细节内容, 均可以参见上述语音识别模型的数据处理方法或语音识别方法的技术方案的描述。

本说明书一实施例还提供一种计算机程序, 其中, 当所述计算机程序在计算机中执行时, 令计算机执行上述语音识别模型的数据处理方法的步骤, 或者, 实现上述语音识别方法的步骤。

上述为本实施例的一种计算机程序的示意性方案。需要说明的是, 该计算机程序的技术方案与上述的语音识别模型的数据处理方法或语音识别方法的技术方案属于同一构思, 计算机程序的技术方案未详细描述的细节内容, 均可以参见上述语音识别模型的数据处理方法或语音识别方法的技术方案的描述。

上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下, 在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外, 在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中, 多任务处理和并行处理也是可以的或者可能是有利的。

所述计算机指令包括计算机程序代码, 所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括: 能够携带所述计算机程序代码的任何实体或装置、记录介质、U 盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器 (ROM, Read-Only Memory)、随机存取存储器 (RAM, Random Access Memory)、电载波信号、电信信号以及软件分发介质等。

需要说明的是, 对于前述的各方法实施例, 为了简便描述, 故将其都表述为一系列的动作组合, 但是本领域技术人员应该知悉, 本说明书实施例并不受所描述的动作顺序的限制, 因为依据本说明书实施例, 某些步骤可以采用其它顺序或者同时进行。其次, 本领域技术人员也应该知悉, 说明书中所描述的实施例均属于优选实施例, 所涉及的动作和模块并不都是本说明书实施例所必须的。

在上述实施例中, 对各个实施例的描述都各有侧重, 某个实施例中没有详述的部分, 可以参见其它实施例的相关描述。

以上公开的本说明书优选实施例只是用于帮助阐述本说明书。可选实施例并没有详尽叙述所有的细

节，也不限制该发明仅为所述的具体实施方式。显然，根据本说明书实施例的内容，可作很多的修改和变化。本说明书选取并具体描述这些实施例，是为了更好地解释本说明书实施例的原理和实际应用，从而使所属技术领域技术人员能很好地理解和利用本说明书。本说明书仅受权利要求书及其全部范围和等效物的限制。

权利要求书

- 1、一种语音识别模型的数据处理系统，包括：
云侧设备，用于获取样本集，其中，所述样本集包括多个样本对，所述样本对包括样本语音数据和样本中文文本；利用编码器对所述样本语音数据进行编码，获得所述样本语音数据的语音特征，其中，
5 所述编码器基于对预训练语音数据执行中文发音单元预测任务进行预训练；将所述语音特征输入解码器，获得预测中文文本，其中，所述解码器基于对预训练中文发音单元执行文本预测任务进行预训练；基于所述预测中文文本和所述样本中文文本，对包括所述编码器和所述解码器的模型进行预训练，在达到预训练停止条件的情况下，获取预训练得到的语音识别模型的模型参数；
所述云侧设备，还用于向端侧设备发送所述预训练得到的语音识别模型的模型参数；
10 所述端侧设备，用于利用所述语音识别模型对待识别语音数据进行语音识别，获得所述待识别语音数据对应的目标文本。
- 2、根据权利要求1所述的数据处理系统，所述云侧设备，还用于：
获取第一预训练语音集，其中，所述第一预训练语音集包括多个无监督的第一预训练语音数据；
利用编码器，对所述第一预训练语音数据进行编码，获得所述第一预训练语音数据对应的第一语音
15 特征，基于所述第一语音特征确定第一发音单元；
对所述第一预训练语音数据进行掩码处理；
利用所述编码器，对掩码处理后的第一预训练语音数据进行编码，获得掩码处理后的第一预训练语音数据对应的第二语音特征，基于所述第二语音特征确定第二发音单元；
基于所述第一预训练语音数据对应的第一发音单元和第二发音单元，对所述编码器进行预训练。
20 3、根据权利要求2所述的数据处理系统，所述云侧设备，具体用于：
提取所述第一预训练语音数据的频谱特征；
将所述第一预训练语音数据的频谱特征输入编码器，获得所述第一预训练语音数据对应的第一语音特征。
- 4、根据权利要求2所述的数据处理系统，所述云侧设备，还用于：
25 获取多个第一预训练对，其中，所述第一预训练对包括第二预训练语音数据和第一预训练中文发音单元；
利用所述编码器，对所述第二预训练语音数据进行中文发音单元预测，获得所述第二预训练语音数据对应的预测中文发音单元；
基于所述第一预训练中文发音单元和预测中文发音单元，对所述编码器进行预训练。
30 5、根据权利要求1所述的数据处理系统，所述编码器包括特征编码层；所述云侧设备，还用于：
获取第一预训练文本集，其中，所述第一预训练文本集包括多个无监督的第一预训练中文文本；
将所述第一预训练中文文本转换为第二预训练中文发音单元，将所述第二预训练中文发音单元输入所述特征编码层，获得所述第二预训练中文发音单元的语音特征；
将所述第二预训练中文发音单元的语音特征输入解码器，获得所述第二预训练中文发音单元对应的
35 预测中文文本；
基于所述第二预训练中文发音单元对应的预测中文文本与所述第一预训练中文文本，对所述解码器进行预训练。
- 6、根据权利要求1所述的数据处理系统，所述云侧设备，还用于：
40 获取第二预训练语音集，其中，所述第二预训练语音集包括多个第三预训练语音数据，所述第三预训练语音数据携带目标伪标签；
利用所述编码器对所述第三预训练语音数据进行编码，获得所述第三预训练语音数据的语音特征；
将所述第三预训练语音数据的语音特征输入所述解码器，得到所述第三预训练语音数据对应的预测伪标签；
基于所述目标伪标签和预测伪标签，对所述解码器进行预训练。
- 7、根据权利要求6所述的数据处理系统，所述云侧设备，具体用于：
45 获取多个无监督的第三预训练语音数据；

- 将所述多个第三预训练语音数据输入预训练的语音编码器,获得所述多个第三预训练语音数据的语音特征;
- 对所述多个第三预训练语音数据的语音特征进行聚类,获得各第三预训练语音数据的目标伪标签。
- 8、根据权利要求 1-7 中任一项所述的数据处理系统,所述编码器包括特征编码层;所述云侧设备,
- 5 还用于:
- 获取多个第二预训练对,其中,所述第二预训练对包括第三预训练中文发音单元和第二预训练中文文本;
- 将所述第三预训练中文发音单元输入所述特征编码层,获得所述第三预训练中文发音单元的语音特征;
- 10 将所述第三预训练中文发音单元的语音特征输入所述解码器,获得所述第三预训练中文发音单元对应的预测中文文本;
- 基于所述第三预训练中文发音单元对应的预测中文文本与所述第二预训练中文文本,对所述特征编码层和所述解码器进行预训练,获得包括编码器和解码器的模型。
- 9、一种语音识别模型的数据处理方法,应用于云侧设备,所述云侧设备与多个端侧设备连接,所述方法包括:
- 15 获取样本集,其中,所述样本集包括多个样本对,所述样本对包括样本语音数据和样本中文文本;
- 利用编码器对所述样本语音数据进行编码,获得所述样本语音数据的语音特征,其中,所述编码器基于对预训练语音数据执行中文发音单元预测任务进行预训练;
- 将所述语音特征输入解码器,获得预测中文文本,其中,所述解码器基于对预训练中文发音单元执行文本预测任务进行预训练;
- 20 基于所述预测中文文本和所述样本中文文本,对包括所述编码器和所述解码器的模型进行预训练,在达到预训练停止条件的情况下,获取预训练得到的语音识别模型的模型参数;
- 向第一端侧设备发送所述预训练得到的语音识别模型的模型参数,其中,所述第一端侧设备是所述多个端侧设备中的任意一个。
- 25 10、一种语音识别方法,应用于端侧设备,所述端侧设备与云侧设备连接,所述方法包括:
- 获取待识别语音数据;
- 利用语音识别模型的编码器,对所述待识别语音数据进行编码,获得所述待识别语音数据的语音特征,其中,所述语音识别模型是所述云侧设备通过如权利要求 9 所述的语音识别模型的数据处理方法预训练得到的;
- 30 将所述语音特征输入所述语音识别模型的解码器,获得所述待识别语音数据对应的目标文本。
- 11、根据权利要求 10 所述的语音识别方法,所述方法还包括:
- 获取校验集,其中,所述校验集包括多个语音校验对和多个中文发音单元校验对,所述语音校验对包括校验语音数据和对应的校验中文文本,所述中文发音单元校验对包括校验语音数据和对应的校验中文发音单元;
- 35 利用所述语音识别模型的编码器,对所述校验语音数据进行中文发音单元预测,获得所述校验语音数据的语音特征和预测中文发音单元;
- 将所述校验语音数据的语音特征输入所述语音识别模型的解码器,得到所述校验语音数据对应的预测中文文本;
- 基于所述预测中文发音单元、所述校验中文发音单元、所述预测中文文本和所述校验中文文本,对所述语音识别模型进行微调,在达到微调停止条件的情况下,得到微调完成的语音识别模型。
- 40 12、根据权利要求 10 所述的语音识别方法,在所述将所述语音特征输入所述语音识别模型的解码器,获得所述待识别语音数据对应的目标文本之后,还包括:
- 将所述目标文本发送至前端显示;
- 接收用户在所述前端输入的所述目标文本对应的修订文本;
- 45 根据所述修订文本和所述待识别语音数据,对所述语音识别模型进行更新,获得更新后的语音识别模型。
- 13、一种计算设备,包括:

存储器和处理器；

所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指令，该计算机可执行指令被处理器执行时实现权利要求 9 所述的语音识别模型的数据处理方法的步骤或实现权利要求 10-12 任一项所述的语音识别方法的步骤。

- 5 14、一种计算机可读存储介质，其存储有计算机可执行指令，该计算机可执行指令被处理器执行时实现权利要求 9 所述的语音识别模型的数据处理方法的步骤或实现权利要求 10-12 任一项所述的语音识别方法的步骤。