

语音识别训练集的生成方法及装置

5 本专利申请要求于 2021 年 05 月 08 日提交的、申请号为 202110514350.X、发明名称为“语音识别训练集的生成方法及装置”的中国专利申请的优先权，该申请的全文以引用的方式并入本申请中。

技术领域

10 本申请实施例涉及计算机技术领域，具体涉及一种语音识别训练集的生成方法及装置。

背景技术

15 近年来，随着深度学习技术的高速发展，采用基于深度神经网络的自动语音识别（Automatic Speech Recognition，ASR）模型进行语音识别，已经成为当前语音识别技术领域的主流趋势。为了提高语音识别模型的泛化性能，需要广泛、大量地收集语音数据，并通过人工标注构建的训练集来优化语音识别模型。

发明内容

20 本公开实施例提出了一种语音识别训练集的生成方法及装置。
 在一种或多种实施例中，本公开提供了一种语音识别训练集的生成方法，包括：获取待处理音频和待处理视频，其中，待处理视频中包括对应于待处理音频的文本信息；识别待处理音频，得到音频文本；
25 识别待处理视频中的文本信息，得到视频文本；基于音频文本与视频文本的一致性，以待处理音频为语音样本，以视频文本为标签，得到语音识别训练集。

 在一种或多种实施例中，本公开提供了一种语音识别训练集的生成装置，包括：获取单元，被配置成获取待处理音频和待处理视频，
30 其中，待处理视频中包括对应于待处理音频的文本信息；第一识别单

元，被配置成识别待处理音频，得到音频文本；第二识别单元，被配置成识别待处理视频中的文本信息，得到视频文本；得到单元，被配置成基于音频文本与视频文本的一致性，以待处理音频为语音样本，以视频文本为标签，得到语音识别训练集。

5 在一种或多种实施例中，本公开提供了一种计算机可读介质，其上存储有计算机程序，其中，程序被处理器执行时实现如第一方面任一实现方式描述的方法。

 在一种或多种实施例中，本公开提供了一种电子设备，包括：一个或多个处理器；存储装置，其上存储有一个或多个程序，当一个或多个程序被一个或多个处理器执行，使得一个或多个处理器实现如第一方面任一实现方式描述的方法。

10 在一种或多种实施例中，本公开提供了一种包括计算机程序的计算机程序产品，该计算机程序在被处理器执行时能够实现如第一方面中任一实现方式描述的方法。

15

附图说明

 通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本公开的其它特征、目的和优点将会变得更明显：

 图 1 是本公开的一个实施例可以应用于其中的示例性系统架构图；

20 图 2 是根据本公开语音识别训练集的生成方法的一个实施例的流程图；

 图 3 是根据本实施例的文本拼接过程的示意图；

 图 4 是根据本实施例的语音识别训练集的生成方法的应用场景的示意图；

25 图 5 是根据本公开的语音识别训练集的生成方法的又一个实施例的流程图；

 图 6 是根据本公开的语音识别训练集的生成装置的一个实施例的结构图；

 图 7 是适于用来实现本公开实施例的计算机系统的结构示意图。

30

具体实施方式

下面结合附图和实施例对本公开作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关发明，而非对该发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。

需要说明的是，在不冲突的情况下，本公开中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本公开。

图 1 示出了可以应用本公开的语音识别训练集的生成方法及装置的示例性架构 100。

如图 1 所示，系统架构 100 可以包括终端设备 101、102、103，网络 104 和服务器 105。终端设备 101、102、103 之间通信连接构成拓扑网络，网络 104 用以在终端设备 101、102、103 和服务器 105 之间提供通信链路的介质。网络 104 可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

终端设备 101、102、103 可以是支持网络连接从而进行数据交互和数据处理硬件设备或软件。当终端设备 101、102、103 为硬件时，其可以是支持网络连接，信息获取、交互、显示、处理等功能的各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、膝上型便携计算机和台式计算机等等。当终端设备 101、102、103 为软件时，可以安装在上述所列举的电子设备中。其可以实现成例如用来提供分布式服务的多个软件或软件模块，也可以实现成单个软件或软件模块。在此不做具体限定。

服务器 105 可以是提供各种服务的服务器，例如获取用户通过终端设备 101、102、103 发送的相对应的待处理视频和待处理音频，进行信息处理，自动构建语音识别训练集的后台处理服务器。此外，服务器还可以基于语音识别训练集训练初始语音识别模型，或优化预训练的语音识别模型。作为示例，服务器 105 可以是云服务器。

需要说明的是，服务器可以是硬件，也可以是软件。当服务器为硬件时，可以实现成多个服务器组成的分布式服务器集群，也可以实

现成单个服务器。当服务器为软件时，可以实现成多个软件或软件模块（例如用来提供分布式服务的软件或软件模块），也可以实现成单个软件或软件模块。在此不做具体限定。

5 还需要说明的是，本公开的实施例所提供的语音识别训练集的生成方法可以由服务器执行，也可以由终端设备执行，还可以由服务器和终端设备彼此配合执行。相应地，语音识别训练集的生成装置包括的各个部分（例如各个单元）可以全部设置于服务器中，也可以全部设置于终端设备中，还可以分别设置于服务器和终端设备中。

10 应该理解，图 1 中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器的。当语音识别训练集的生成方法运行于其上的电子设备不需要与其他电子设备进行数据传输时，该系统架构可以仅包括语音识别训练集的生成方法运行于其上的电子设备（例如服务器或终端设备）。继续参考图 2，示出了语音识别训练集的生成方法的一个实施例的流程 200，包括
15 以下步骤：

步骤 201，获取待处理音频和待处理视频。

本实施例中，语音识别训练集的生成方法的执行主体（例如图 1 中的服务器）可以通过有线网络连接方式或无线网络连接方式从远程，或从本地获取待处理音频和待处理视频。其中，待处理视频中包
20 括对应于待处理音频的文本信息。

作为示例，包括相对应的待处理音频和待处理视频的数据可以是电影、电视剧、短视频等各种音视频数据。待处理视频中的文本信息是字幕信息，待处理音频是对应于字幕信息的语音信息。

25 本实施例中，待处理音频所表征的语音数据可以是各种类型的语音，包括但不限于外语音频、国语音频、方言音频。待处理音频和待处理视频可以是时长较长的数据，也可以是时长较短的数据。

步骤 202，识别待处理音频，得到音频文本。

本实施例中，上述执行主体可以识别待处理音频，得到音频文本。

30 作为示例，上述执行主体可以基于自动语音识别模型，处理待处理音频，得到音频文本。自动语音识别模型用于表征待处理音频与文

本之间的对应关系。

在本实施例的一些可选的实现方式中，上述执行主体可以通过如下方式执行上述步骤 202：

5 第一，根据静音检测算法，删除待处理音频中的静音部分，得到非静音的多个音频片段。

本实现方式中，上述执行主体可以待处理音频中的静音部分为分割点，将删除静音部分后的待处理音频进行分割，得到多个音频片段。

针对于所得到的音频片段较长的情况，上述执行主体可以设置时长阈值，将时长大于时长阈值的音频片段，以时长阈值所表征的时长为单
10 位进行进一步的切割，并记录每个音频片的起止时间。

作为示例，为了防止背景音乐等因素导致静音检测算法无法完全截断音频而导致获得的音频片段较长，设定时长阈值 T ，将时长大于时长阈值 T 的音频片段强行切割为多个时长为 T 的片段。其中，时长阈值可以根据实际情况具体设置，例如， $T = 10s$ 。

15 第二，识别多个音频片段，得到音频文本包括的多个音频片段文本。

本实现方式中，上述执行主体可以将多个音频片段中的每个音频片段输入自动语音识别模型，得到多个音频片段文本。其中，多个音频片段与多个音频片段文本一一对应，多个音频片段文本构成音频文
20 本。

步骤 203，识别待处理视频中的文本信息，得到视频文本。

本实施例中，上述执行主体可以识别待处理视频中的文本信息，得到视频文本。

作为示例，针对于组成待处理视频的每个视频帧，上述执行主体
25 可以利用 OCR（Optical Character Recognition，光学字符识别）技术，识别该视频帧中所包括的文本信息，并按照待处理视频中的视频帧的播放顺序，拼接每个视频帧对应的的文本信息，得到视频文本。其中，OCR 技术是目前较为成熟的技术，在此不做赘述。

30 在本实施例的一些可选的实现方式中，上述执行主体可以通过如下方式执行上述步骤 203：

第一，从待处理视频中确定出与多个音频片段一一对应的多个视频帧序列。

本实现方式中，对于多个音频片段中的每个音频片段，上述执行主体从待处理视频中抽取该音频片段对应的多个视频帧，得到视频帧序列。

5 作为示例，该音频片段的起止时间分别为 t_{sk} 和 t_{ek} ，上述执行主体可以通过 $\lceil t_{sk} / f_p \rceil$ 、 $\lfloor t_{ek} / f_p \rfloor$ 依次确定该音频片段对应的视频帧序列的起止视频帧。其中， $\lceil \cdot \rceil$ 和 $\lfloor \cdot \rfloor$ 分别表征向上取整和向下取整， f_p 表征待处理视频的帧率。上述执行主体可以预先设置采样率，并基于采样率从起始帧 $\lceil t_{sk} / f_p \rceil$ 、终止 $\lfloor t_{ek} / f_p \rfloor$ 之间抽取视频帧，得到该音频片段对应的视频帧序列。

第二，识别多个视频帧序列中的每个视频帧中的文本信息，得到视频文本包括的视频帧文本。

本实现方式中，上述执行主体可以利用 OCR 技术，识别多个视频帧序列中的每个视频帧中的文本信息，得到视频文本包括的视频帧文本。

可以理解，对于每个视频帧，上述执行主体可能并未识别出文本信息，也即该视频帧中并不包括文本信息；也可能识别出多处文本信息，得到多个视频帧文本。例如，多个视频帧文本包括视频帧中的字幕信息，以及视频帧画面中的文本信息（例如是店铺招牌中的店铺名称信息、道路指示牌中的道路名称信息、广告语信息等）。

20 在另一些情形中，相邻帧中包括的文本信息中存在相同的文本信息。例如，相邻的视频帧中包括的字幕信息相同。

本实施例中，可以为视频帧添加预设标识，以表征视频帧中不包括文本信息，或包括与相邻视频帧相同的文本信息的情形。其中，预设标识可以是预先设置的任意标识，例如是“Blank”。

步骤 204，基于音频文本与视频文本的一致性，以待处理音频为语音样本，以视频文本为标签，得到语音识别训练集。

本实施例中，上述执行主体可以基于音频文本与视频文本的一致性，以待处理音频为语音样本，以视频文本为标签，得到语音识别训

练集。

作为示例，当音频文本与视频文本一致时，上述执行主体以待处理音频为语音样本，以视频文本为标签，得到语音识别训练集。

在本实施例的一些可选的实现方式中，上述执行主体可以通过如下方式执行上述步骤 204：

首先，对于多个视频帧序列中的每个视频帧序列，执行如下操作：

第一，以该视频帧序列中的视频帧中被识别出的至少一个视频帧文本中的一个视频帧文本为单位，对该视频帧序列中的每个视频帧所包括的文本信息进行拼接，得到该视频帧序列对应的多个视频帧序列文本。

作为示例，该视频帧序列包括 3 个视频帧，3 个视频帧对应的视频帧文本的数量依次为 3 个、4 个、3 个，则该视频帧序列对应的多个视频帧序列文本共有 36（ $3 \times 4 \times 3$ ）个。

在一些可选的实现方式中，每个视频帧对应的视频帧文本集合中，除了包括从该视频帧中识别到的视频帧文本，还包括表征该视频帧包括与相邻视频帧相同的文本信息的预设标识。预设标识还可以表征视频帧中不包括文本信息的情形。

继续参照上述示例，在每个视频帧对应的视频帧文本结合中添加预设标识后，3 个视频帧对应的视频帧文本的数量依次为 4 个、5 个、4 个，则该视频帧序列对应的多个视频帧序列文本共有 80（ $4 \times 5 \times 4$ ）个。

如图 3 所示，预设标识为“Blank”。对应于视频帧 301、302、303 的识别结果依次为 304、305、306。对于每一个视频帧中的每一个视频帧文本均可以与其他视频帧中视频帧文本组合，得到多个视频帧序列文本。例如，可以将视频帧 301、302、303 的识别结果组合为“今天的天气怎么样”。

第二，根据多个视频帧序列文本中的每个视频帧序列文本与目标音频片段文本之间的编辑距离，确定目标视频帧序列文本。

其中，目标音频片段文本为对应于该视频帧序列的音频片段所对应的音频片段文本。编辑距离是指两个字符串之间，由一个转成另一

个所需要的最少编辑操作次数。

作为示例，上述执行主体可以将多个视频帧序列文本中与目标音频片段文本之间的编辑距离最小的视频帧序列文本，确定为目标视频帧序列文本。

5 然后，以多个音频片段中的每个音频片段为语音样本，以该音频片段对应的目标视频帧序列文本为标签，得到语音识别训练集。

在本实施例的一些可选的实现方式中，上述执行主体可以通过如下方式执行上述第一步骤：

对于该视频帧序列中包括文本信息的每个视频帧，执行如下操作：

10 首先，确定该视频帧对应的多个待拼接文本，并将多个待拼接文本与该视频帧中的至少一个视频帧文本进行拼接，得到多个拼接后文本。

15 然后，基于多个拼接后文本与目标音频片段文本之间的编辑距离，从多个拼接后文本选取出预设数量个拼接后文本，作为该视频帧的下一视频帧所对应的多个待拼接文本。

作为示例，上述执行主体可以将编辑距离由小到大进行排序，选取前预设数量个拼接后文本，作为该视频帧的下一视频帧所对应的多个待拼接文本。其中，预设数量可以根据实际情况进行具体设置，例如，可以是 10。

20 在所得到的拼接后文本的数量较小（例如，拼接后文本的数量小于预设数量）的情况下，上述执行主体可以设置预设距离阈值，将编辑距离小于预设距离阈值的拼接后文本删除。

25 可以理解，上述执行主体针对于拼接后文本的数量较大的情况，也可以结合选取预设数量个文本、删除编辑距离小于预设距离阈值的文本的方式，确定该视频帧的下一视频帧所对应的多个待拼接文本。

作为又一示例，上述执行主体可以通过如下公式确定所保留的多个拼接后文本与音频文本之间的匹配度：

$$Q_i = -|d(p_{ki}, S_k) - ||S_k|| - ||p_{ki}|||$$

30 其中， $d(\bullet, \bullet)$ 表示两个文本的编辑距离计算函数， $||\bullet||$ 表示文本的长度， p_{ki} 表示拼接后文本， S_k 表示音频文本， Q_i 表示两个文本之间匹

配度。为了进一步减少每次拼接后得到的拼接后文本的数量，设计匹配度阈值 T_h ，当 $Q_i < T_h$ 时，则删除该拼接后文本。例如 $T_h = -3$ 。

在本实施例的一些可选的实现方式中，上述执行主体还可以对于多个视频帧序列中的每个视频帧序列，响应于确定该视频帧序列对应的目标视频帧序列文本与目标音频片段文本之间的编辑距离大于预设距离阈值，删除语音识别训练集中对应于该视频帧序列的训练样本，从而过滤掉低质量的训练样本。

继续参见图 4，图 4 是根据本实施例的语音识别训练集的生成方法的应用场景的一个示意图 400。在图 4 的应用场景中，首先，服务器 401 获取待处理音频 402 和待处理视频 403。其中，待处理视频 403 中包括对应于待处理音频 402 的文本信息。然后，服务器 401 识别待处理音频 402，得到音频文本 404；识别待处理视频 403 中的文本信息，得到视频文本 405。最后，服务器 401 基确定音频文本 404 与视频文本 405 的一致性，以待处理音频 402 为语音样本，以视频文本 405 为标签，得到语音识别训练集 406。

本公开的上述实施例提供的方法，通过获取待处理音频和待处理视频，其中，待处理视频中包括对应于待处理音频的文本信息；识别待处理音频，得到音频文本；识别待处理视频中的文本信息，得到视频文本；基于音频文本与视频文本的一致性，以待处理音频为语音样本，以视频文本为标签，得到语音识别训练集，从而提供了一种语音识别训练集的自动获取方法，提高了构建语音识别训练集的灵活性和效率。

在本实施例的一些可选的实现方式中，上述执行主体可以基于语音识别训练集，训练未经训练的初始语音识别模型，或优化预训练的语音识别模型。

具体的，上述执行主体采用机器学习算法，以训练样本中的待处理音频为输入，以所输入的待处理音频为期望输出，训练未经训练的初始语音识别模型，或优化预训练的语音识别模型，得到最终的语音识别模型。

继续参考图 5，示出了根据本公开的语音识别训练集的生成方法的一个实施例的示意性流程 500，包括以下步骤：

步骤 501，获取待处理音频和待处理视频。

5 其中，待处理视频中包括对应于待处理音频的文本信息；

步骤 502，根据静音检测算法，删除待处理音频中的静音部分，得到非静音的多个音频片段。

步骤 503，识别多个音频片段，得到音频文本包括的多个音频片段文本。

10 步骤 504，从待处理视频中确定出与多个音频片段一一对应的多个视频帧序列。

步骤 505，识别多个视频帧序列中的每个视频帧中的文本信息，得到视频文本包括的视频帧文本。

15 步骤 506，对于多个视频帧序列中的每个视频帧序列，执行如下操作：

步骤 5061，以该视频帧序列中的视频帧中被识别出的至少一个视频帧文本中的一个视频帧文本为单位，对该视频帧序列中的每个视频帧所包括的文本信息进行拼接，得到该视频帧序列对应的多个视频帧序列文本。

20 步骤 5062，根据多个视频帧序列文本中的每个视频帧序列文本与目标音频片段文本之间的编辑距离，确定目标视频帧序列文本，其中，目标音频片段文本为对应于该视频帧序列的音频片段所对应的音频片段文本。

25 步骤 507。以多个音频片段中的每个音频片段为语音样本，以该音频片段对应的目标视频帧序列文本为标签，得到语音识别训练集。

从本实施例中可以看出，与图 2 对应的实施例相比，本实施例中的语音识别训练集的生成方法的流程 400 具体说明了待处理音频和待处理视频的分割过程，以及视频帧文本的拼接过程，提高了语音识别训练集中的训练样本的准确性。

继续参考图 6，作为对上述各图所示方法的实现，本公开提供了一种语音识别训练集的生成装置的一个实施例，该装置实施例与图 2 所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

如图 6 所示，语音识别训练集的生成装置包括：包括：获取单元 601，被配置成获取待处理音频和待处理视频，其中，待处理视频中包括对应于待处理音频的文本信息；第一识别单元 602，被配置成识别待处理音频，得到音频文本；第二识别单元 603，被配置成识别待处理视频中的文本信息，得到视频文本；得到单元 604，被配置成基于音频文本与视频文本的一致性，以待处理音频为语音样本，以视频文本为标签，得到语音识别训练集。

在本实施例的一些可选的实现方式中，第一识别单元 602，进一步被配置成：根据静音检测算法，删除待处理音频中的静音部分，得到非静音的多个音频片段；识别多个音频片段，得到音频文本包括的多个音频片段文本。

在本实施例的一些可选的实现方式中，第二识别单元 603，进一步被配置成：从待处理视频中确定出与多个音频片段一一对应的多个视频帧序列；识别多个视频帧序列中的每个视频帧中的文本信息，得到视频文本包括的视频帧文本。

在本实施例的一些可选的实现方式中，得到单元 604，进一步被配置成：对于多个视频帧序列中的每个视频帧序列，执行如下操作：以该视频帧序列中的视频帧中被识别出的至少一个视频帧文本中的一个视频帧文本为单位，对该视频帧序列中的每个视频帧所包括的文本信息进行拼接，得到该视频帧序列对应的多个视频帧序列文本；根据多个视频帧序列文本中的每个视频帧序列文本与目标音频片段文本之间的编辑距离，确定目标视频帧序列文本，其中，目标音频片段文本为对应于该视频帧序列的音频片段所对应的音频片段文本；以多个音频片段中的每个音频片段为语音样本，以该音频片段对应的目标视频帧序列文本为标签，得到语音识别训练集。

在本实施例的一些可选的实现方式中，得到单元 604，进一步被配置成：对于该视频帧序列中包括文本信息的每个视频帧，执行如下

操作：确定该视频帧对应的多个待拼接文本，并将多个待拼接文本与该视频帧中的至少一个视频帧文本进行拼接，得到多个拼接后文本；基于多个拼接后文本与目标音频片段文本之间的编辑距离，从多个拼接后文本选取出预设数量个拼接后文本，作为该视频帧的下一视频帧所对应的多个待拼接文本。

在本实施例的一些可选的实现方式中，上述装置还包括：删除单元(图中未示出)，被配置成对于多个视频帧序列中的每个视频帧序列，响应于确定该视频帧序列对应的目标视频帧序列文本与目标音频片段文本之间的编辑距离大于预设距离阈值，删除语音识别训练集中对应于该视频帧序列的训练样本。

本实施例中，语音识别训练集的生成装置中的获取单元获取待处理音频和待处理视频，其中，待处理视频中包括对应于待处理音频的文本信息；第一识别单元识别待处理音频，得到音频文本；第二识别单元识别待处理视频中的文本信息，得到视频文本；得到单元基于音频文本与视频文本的一致性，以待处理音频为语音样本，以视频文本为标签，得到语音识别训练集，从而提供了一种语音识别训练集的自动获取装置，提高了构建语音识别训练集的灵活性和效率。

下面参考图 7，其示出了适于用来实现本公开实施例的设备（例如如图 1 所示的设备 101、102、103、105）的计算机系统 700 的结构示意图。图 7 示出的设备仅仅是一个示例，不应对本公开实施例的功能和使用范围带来任何限制。

如图 7 所示，计算机系统 700 包括处理器（例如 CPU，中央处理器）701，其可以根据存储在只读存储器（ROM）702 中的程序或者从存储部分 708 加载到随机访问存储器（RAM）703 中的程序而执行各种适当的动作和处理。在 RAM703 中，还存储有系统 700 操作所需的各种程序和数据。处理器 701、ROM702 以及 RAM703 通过总线 704 彼此相连。输入/输出（I/O）接口 705 也连接至总线 704。

以下部件连接至 I/O 接口 705：包括键盘、鼠标等的输入部分 706；包括诸如阴极射线管（CRT）、液晶显示器（LCD）等以及扬声器等的

输出部分 707; 包括硬盘等的存储部分 708; 以及包括诸如 LAN 卡、调制解调器等网络接口卡的通信部分 709。通信部分 709 经由诸如因特网的网络执行通信处理。驱动器 710 也根据需要连接至 I/O 接口 705。可拆卸介质 711, 诸如磁盘、光盘、磁光盘、半导体存储器等等, 5 根据需要安装在驱动器 710 上, 以便于从其上读出的计算机程序根据需要被安装入存储部分 708。

特别地, 根据本公开的实施例, 上文参考流程图描述的过程可以被实现为计算机软件程序。例如, 本公开的实施例包括一种计算机程序产品, 其包括承载在计算机可读介质上的计算机程序, 该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中, 10 该计算机程序可以通过通信部分 709 从网络上被下载和安装, 和/或从可拆卸介质 711 被安装。在该计算机程序被处理器 701 执行时, 执行本公开的方法中限定的上述功能。

需要说明的是, 本公开的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是一一但不限于一一电、磁、光、电磁、红外线、或半导体的系统、装置或器件, 或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于: 具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器 (RAM)、只读存储器 (ROM)、可擦式可编程只读存储器 (EPROM 或闪存)、光纤、便携式紧凑磁盘只读存储器 (CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开中, 计算机可读存储介质可以是任何包含或存储程序的有形介质, 该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开中, 计算机可读的信号 25 介质可以包括在基带中或者作为载波一部分传播的数据信号, 其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式, 包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质, 该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含 30

的程序代码可以用任何适当的介质传输，包括但不限于：无线、电线、光缆、RF 等等，或者上述的任意合适的组合。

可以以一种或多种程序设计语言或其组合来编写用于执行本公开的操作的计算机程序代码，程序设计语言包括面向目标的设计语言—诸如 Java、Smalltalk、C++，还包括常规的过程式程序设计语言—诸如“C”语言或类似的设计语言。程序代码可以完全地在客户计算机上执行、部分地在客户计算机上执行、作为一个独立的软件包执行、部分在客户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)——连接到客户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。

附图中的流程图和框图，图示了按照本公开各种实施例的装置、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

描述于本公开实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元也可以设置在处理器中，例如，可以描述为：一种处理器，包括获取单元、第一识别单元、第二识别单元和得到单元。其中，这些单元的名称在某种情况下并不构成对该单元本身的限定，例如，得到单元还可以被描述为“基于音频文本与视频文本的一致性，以待处理音频为语音样本，以视频文本为标签，得到语音识别训练集的单元”。

5 作为另一方面，本公开还提供了一种计算机可读介质，该计算机可读介质可以是上述实施例中描述的设备中所包含的；也可以是单独存在，而未装配入该设备中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被该装置执行时，使得该计算机设备：获取待处理音频和待处理视频，其中，待处理视频中包括对应于待处理音频的文本信息；识别待处理音频，得到音频文本；识别待处理视频中的文本信息，得到视频文本；基于音频文本与视频文本的一致性，以待处理音频为语音样本，以视频文本为标签，得到语音识别训练集。

10 以上描述仅为本公开的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本公开中所涉及的发明范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述发明构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本公开中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

权 利 要 求 书

1、一种语音识别训练集的生成方法，包括：

5 获取待处理音频和待处理视频，其中，所述待处理视频中包括对
应于所述待处理音频的文本信息；

识别所述待处理音频，得到音频文本；

识别所述待处理视频中的文本信息，得到视频文本；以及

基于所述音频文本与所述视频文本的一致性，以所述待处理音频
为语音样本，以所述视频文本为标签，得到所述语音识别训练集。

10

2、根据权利要求 1 所述的方法，其中，所述识别所述待处理音频，
得到音频文本，包括：

根据静音检测算法，删除所述待处理音频中的静音部分，得到非
静音的多个音频片段；以及

15 识别所述多个音频片段，得到所述音频文本包括的多个音频片段
文本。

3、根据权利要求 2 所述的方法，其中，所述识别所述待处理视频
中的文本信息，得到视频文本，包括：

20 从所述待处理视频中确定出与所述多个音频片段一一对应的多个
视频帧序列；以及

识别所述多个视频帧序列中的每个视频帧中的文本信息，得到所
述视频文本包括的视频帧文本。

25 4、根据权利要求 3 所述的方法，其中，所述基于所述音频文本与
所述视频文本的一致性，以所述待处理音频为语音样本，以所述视频
文本为标签，得到所述语音识别训练集，包括：

对于所述多个视频帧序列中的每个视频帧序列，执行如下操作：
以该视频帧序列中的视频帧中被识别出的至少一个视频帧文本中的一
30 个视频帧文本为单位，对该视频帧序列中的每个视频帧所包括的文本

信息进行拼接，得到该视频帧序列对应的多个视频帧序列文本；根据所述多个视频帧序列文本中的每个视频帧序列文本与目标音频片段文本之间的编辑距离，确定目标视频帧序列文本，其中，所述目标音频片段文本为对应于该视频帧序列的音频片段所对应的音频片段文本；

5 以及

以所述多个音频片段中的每个音频片段为语音样本，以该音频片段对应的目标视频帧序列文本为标签，得到所述语音识别训练集。

5、根据权利要求4所述的方法，其中，所述以该视频帧序列中的视频帧中被识别出的至少一个视频帧文本中的一个视频帧文本为单位，
10 对该视频帧序列中的每个视频帧所包括的文本信息进行拼接，得到该视频帧序列对应的多个视频帧序列文本，包括：

对于该视频帧序列中包括文本信息的每个视频帧，执行如下操作：

确定该视频帧对应的多个待拼接文本，并将所述多个待拼接文本
15 与该视频帧中的至少一个视频帧文本进行拼接，得到多个拼接后文本；
以及

基于所述多个拼接后文本与所述目标音频片段文本之间的编辑距离，从所述多个拼接后文本选取出预设数量个拼接后文本，作为该视频帧的下一视频帧所对应的多个待拼接文本。

20

6、根据权利要求4所述的方法，其中，还包括：

对于所述多个视频帧序列中的每个视频帧序列，响应于确定该视频帧序列对应的目标视频帧序列文本与目标音频片段文本之间的编辑距离大于预设距离阈值，删除所述语音识别训练集中对应于该视频帧
25 序列的训练样本。

7、一种语音识别训练集的生成装置，包括：

获取单元，被配置成获取待处理音频和待处理视频，其中，所述待处理视频中包括对应于所述待处理音频的文本信息；

30 第一识别单元，被配置成识别所述待处理音频，得到音频文本；

第二识别单元，被配置成识别所述待处理视频中的文本信息，得到视频文本；

5 得到单元，被配置成基于所述音频文本与所述视频文本的一致性，以所述待处理音频为语音样本，以所述视频文本为标签，得到所述语音识别训练集。

8、根据权利要求 7 所述的装置，其中，所述第一识别单元，进一步被配置成：

10 根据静音检测算法，删除所述待处理音频中的静音部分，得到非静音的多个音频片段；识别所述多个音频片段，得到所述音频文本包括的多个音频片段文本。

9、根据权利要求 8 所述的装置，其中，所述第二识别单元，进一步被配置成：

15 从所述待处理视频中确定出与所述多个音频片段一一对应的多个视频帧序列；识别所述多个视频帧序列中的每个视频帧中的文本信息，得到所述视频文本包括的视频帧文本。

20 10、根据权利要求 9 所述的装置，其中，所述得到单元，进一步被配置成：

对于所述多个视频帧序列中的每个视频帧序列，执行如下操作：
以该视频帧序列中的视频帧中被识别出的至少一个视频帧文本中的一个视频帧文本为单位，对该视频帧序列中的每个视频帧所包括的文本信息进行拼接，得到该视频帧序列对应的多个视频帧序列文本；根据
25 所述多个视频帧序列文本中的每个视频帧序列文本与目标音频片段文本之间的编辑距离，确定目标视频帧序列文本，其中，所述目标音频片段文本为对应于该视频帧序列的音频片段所对应的音频片段文本；

以所述多个音频片段中的每个音频片段为语音样本，以该音频片段对应的目标视频帧序列文本为标签，得到所述语音识别训练集。

30

11、根据权利要求 10 所述的装置，其中，所述得到单元，进一步被配置成：

5 对于该视频帧序列中包括文本信息的每个视频帧，执行如下操作：确定该视频帧对应的多个待拼接文本，并将所述多个待拼接文本与该视频帧中的至少一个视频帧文本进行拼接，得到多个拼接后文本；基于所述多个拼接后文本与所述目标音频片段文本之间的编辑距离，从所述多个拼接后文本选取出预设数量个拼接后文本，作为该视频帧的下一视频帧所对应的多个待拼接文本。

10 12、根据权利要求 10 所述的装置，其中，还包括：

删除单元，被配置成对于所述多个视频帧序列中的每个视频帧序列，响应于确定该视频帧序列对应的目标视频帧序列文本与目标音频片段文本之间的编辑距离大于预设距离阈值，删除所述语音识别训练集中对应于该视频帧序列的训练样本。

15

13、一种计算机可读介质，其上存储有计算机程序，其中，所述程序被处理器执行时实现如权利要求 1-6 中任一所述的方法。

14、一种电子设备，包括：

20

一个或多个处理器；

存储装置，其上存储有一个或多个程序，

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求 1-6 中任一所述的方法。

25

15、一种计算机程序产品，包括计算机程序，所述计算机程序在被处理器执行时实现根据权利要求 1-6 中任一项所述的方法。

30