

样本生成方法及装置

本申请要求于 2020 年 11 月 20 日提交中国专利局、申请号为 202011309190.7、发明名称为“样本生成方法及装置”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

本说明书涉及数据处理技术领域，特别涉及样本生成方法及装置。

背景技术

随着互联网技术的发展，语音合成被越来越多的场景所应用；语音合成（Text ToSpeech，TTS）又称为文语转换技术，是一种通过软、硬件结合的方式将文本转换为自然语音的技术，该技术可以通过波形拼接、基于参数的合成或使用神经网络的端到端合成方法实现；其中波形拼接方法需要较长时间的训练数据才能够完成语音合成；而基于参数的合成方法虽然可以完成语音合成，但是所参考的因素较少，导致最终的合成结果并不理想；现有技术中应用较为广泛的即为基于神经网络的端到端合成方法，该方法所需要的数据量小，无需人工调整大量参数即可实现语音合成；虽然端到端的语音合成方法在数据量的需求上要小于其他方法，但是结合基于神经网络的端到端合成方法的特性，对语音数据的质量要求要远高于其他方法，前期准备语音数据的成本将大大的增加，并且准备好的语音数据还可能存在不完善的问题，严重影响基于神经网络的端到端合成方法的实现，因此亟需一种有效的方案以解决上述问题。

发明内容

有鉴于此，本说明书实施例提供了一种样本生成方法。本说明书同时涉及一种样本生成装置，一种计算设备，以及一种计算机可读存储介质，以解决现有技术中存在的技术缺陷。

根据本说明书实施例的第一方面，提供了一种样本生成方法，包括：

获取多个文本音频对，其中每个文本音频对中包含文本片段和音频片段；
计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对和所述目标文本音频对对应的拼接文本音频对；

将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对，并对所述待检测文本音频对进行检测；

在所述待检测文本音频对满足预设检测条件的情况下，将所述待检测文本

音频对写入训练数据库。

可选地,所述获取多个文本音频对,包括:

获取目标文本以及所述目标文本对应的音频;

对所述音频进行预处理获得目标音频,并将所述目标文本转换为音素序列;

- 5 将所述音素序列与所述目标音频进行对齐处理,并根据对齐处理结果生成所述多个文本音频对。

可选地,所述根据对齐处理结果生成所述多个文本音频对,包括:

根据对齐处理结果得到音素音频文件,并确定所述音素音频文件的切分位置;

- 10 按照所述切分位置对所述音素音频文件进行切分,获得多个音素音频对,其中每个音素音频对中包含音素片段和音频片段;

基于所述目标文本确定所述多个音素音频对中的每个音素音频对的音素片段对应的文本片段;

- 15 根据每个音素音频对中音素片段对应的文本片段,以及每个音素音频对中的音频片段生成所述多个文本音频对。

可选地,所述计算所述多个文本音频对中每个文本音频对的音频片段的音频特征,包括:

提取所述多个文本音频对中每个文本音频对的音频片段,并对每个文本音频对的音频片段进行分帧处理,获得每个文本音频对的音频帧集合;

- 20 基于所述多个文本音频对中每个文本音频对的音频帧集合包含的音频帧,计算每个文本音频对的音频片段的基音频率特征和音频帧特征;

根据每个文本音频对的音频片段的所述基音频率特征和所述音频帧特征,确定每个文本音频对的音频片段的所述音频特征。

- 25 可选地,所述根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对和所述目标文本音频对对应的拼接文本音频对,包括:

将所述多个文本音频对中每个文本音频对的音频片段、文本片段和音频特征进行整合,获得每个文本音频对对应的文本音频包,并写入片段数据库;

在所述片段数据库中选择任意一个文本音频包作为目标文本音频包,并将所述目标文本音频包中的文本音频对确定为所述目标文本音频对;

- 30 基于所述片段数据库中除所述目标文本音频包外的文本音频包和所述音频特征确定拼接文本音频包,并将所述拼接文本音频包中的文本音频对作为所述拼接文本音频对。

可选地,所述基于所述片段数据库中除所述目标文本音频包外的文本音频包和所述音频特征确定拼接文本音频包,包括:

在所述片段数据库中选择除所述目标文本音频包外的文本音频包组成待筛选文本音频包集合;

- 5 将所述待筛选文本音频包集合中包含的各个待筛选文本音频包的文本音频对确定为待筛选文本音频对;

基于所述目标文本音频对的音频片段的音频特征和所述待筛选文本音频对的音频片段的音频特征,在所述待筛选文本音频包集合中筛选出所述拼接文本音频包。

- 10 可选地,所述基于所述目标文本音频对的音频片段的音频特征和所述待筛选文本音频对的音频片段的音频特征,在所述待筛选文本音频包集合中筛选出所述拼接文本音频包,包括:

根据所述目标文本音频包确定所述目标文本音频对的音频片段的第一音频特征,以及根据所述待筛选文本音频包确定所述待筛选文本音频对的音频片段的第二音频特征;

- 15 计算所述第一音频特征和所述第二音频特征之间的特征距离;

将所述特征距离小于预设距离阈值的待筛选文本音频对所属的待筛选文本音频包确定为所述拼接文本音频包。

- 20 可选地,所述将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对,并对所述待检测文本音频对进行检测步骤执行之前,还包括:

对所述目标文本音频对中的音频片段进行采样处理获得目标采样信息,以及确定所述目标文本音频对中的文本片段的目标文本信息;

判断所述目标采样信息和所述目标文本信息是否满足所述预设检测条件;

- 25 若否,则执行将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对,并对所述待检测文本音频对进行检测的步骤。

可选地,若所述判断所述采样信息和所述文本信息是否满足所述预设检测条件的判断结果为是,则将所述目标文本音频对写入所述训练数据库。

可选地,所述将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对,包括:

- 30 提取所述目标文本音频对中的目标文本片段和目标音频片段,以及所述拼接文本音频对中的拼接文本片段和拼接音频片段;

将所述目标文本片段和所述拼接文本片段拼接为待检测文本片段,以及将

所述目标音频片段和所述拼接音频片段拼接为待检测音频片段；

基于所述待检测文本片段和所述待检测音频片段组成所述待检测文本音频对。

可选地，所述对所述待检测文本音频对进行检测，包括：

- 5 对所述待检测音频片段进行采样处理获得待检测采样信息，以及确定所述待检测文本片段的待检测文本信息；

基于所述预设检测条件对所述待检测采样信息和所述待检测文本信息进行检测；

- 10 相应的，所述在所述待检测文本音频对满足预设检测条件的情况下，将所述待检测文本音频对写入训练数据库，包括：

在所述待检测采样信息和所述待检测文本信息均满足所述预设检测条件的情况下，将所述待检测文本音频对写入所述训练数据库。

可选地，所述将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对，并对所述待检测文本音频对进行检测步骤执行之后，还包括：

- 15 在所述待检测文本音频对未满足预设检测条件的情况下，根据所述音频特征在所述多个文本音频对中筛选出所述拼接文本音频对对应的多度拼接文本音频对；

将所述多度拼接文本音频对和所述待检测文本音频对拼接为多度待检测文本音频对，并判断所述多度待检测文本音频对是否满足所述预设检测条件；

- 20 若是，则将所述多度待检测文本音频对写入所述训练数据库；

若否，则将所述多度拼接文本音频对作为所述拼接文本音频对，以及将所述多度待检测文本音频对作为所述待检测文本音频对，并执行所述根据所述音频特征在所述多个文本音频对中筛选出所述拼接文本音频对对应的多度拼接文本音频对步骤。

- 25 可选地，所述将所述待检测文本音频对写入训练数据库步骤执行之后，还包括：

在所述训练数据库中提取样本文本音频对，所述样本文本音频对中包含样本文本片段和样本音频片段；

- 30 基于所述样本文本片段和所述样本音频片段对语音合成模型进行训练，获得目标语音合成模型。

根据本说明书实施例的第二方面，提供了一种样本生成装置，包括：

获取模块，被配置为获取多个文本音频对，其中每个文本音频对中包含文

本片段和音频片段；

计算模块，被配置为计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对和所述目标文本音频对对应的拼接文本音频对；

- 5 拼接模块，被配置为将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对，并对所述待检测文本音频对进行检测；

写入模块，被配置为在所述待检测文本音频对满足预设检测条件的情况下，将所述待检测文本音频对写入训练数据库。

根据本说明书实施例的第三方面，提供了一种计算设备，包括：

- 10 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：

获取多个文本音频对，其中每个文本音频对中包含文本片段和音频片段；

- 15 计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对和所述目标文本音频对对应的拼接文本音频对；

将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对，并对所述待检测文本音频对进行检测；

- 20 在所述待检测文本音频对满足预设检测条件的情况下，将所述待检测文本音频对写入训练数据库。

- 根据本说明书实施例的第四方面，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所述样本生成方法的步骤。本说明书提供一种样本生成方法，在获取到多个文本音频对后，计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对，以及所述目标文本音频对对应的拼接文本音频对，之后将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对，并对所述待检测文本音频对进行检测，在所述待检测文本音频对满足预设检测条件的情况下，将所述待检测文本音频对写入所述训练数据库，实现在样本数据准备阶段，可以通过拼接的方式得到质量高且满足下游业务使用需求的样本数据，节省数据准备阶段的资源消耗成本，并且拼接后写入所述训练数据库中的样本数据的数据量较多，有效的解决了下游业务样本数据量少、样本数据中音频长短分布不均匀导致语音合成效果不佳的问题，从而提高下游业务的业务处理效率。
- 25
- 30

附图说明

图 1 是本说明书一实施例提供的一种样本生成方法的流程图；

图 2 是本说明书一实施例提供的一种样本生成方法中对齐处理结果的示意图；

5 图 3 是本说明书一实施例提供的一种样本生成方法中切分处理结果的示意图；

图 4 是本说明书一实施例提供的一种样本生成方法中筛选拼接文本音频对的流程图；

10 图 5 是本说明书一实施例提供的一种应用于语音合成场景中的样本生成方法的流程图；

图 6 是本说明书一实施例提供的一种样本生成装置的结构示意图；

图 7 是本说明书一实施例提供的一种计算设备的结构框图。

具体实施方式

15 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。

20 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。

25 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”。

30 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。

F0（基音频率）：一般的声音都是由发音体发出的一系列频率、振幅各不相同的振动复合而成的；这些振动中有一个频率最低的振动，由它发出的音就是基音，其对应的频率即为基音频率。

强制对齐：一种得到给定音素的序列和语音在时间上的对应关系的技术，可以通过强制对齐工具实现，如通过 kaldil（一种开源语音识别工具(Toolkit)，它使用 WFST 来实现解码算法）或 HTK（HMM Toolkit，一款基于 hmm 模型的语音处理工具）等即可实现音素序列和音频的对齐。

5 音素，是根据语音的自然属性划分出来的最小语音单位，依据音节里的发音动作来分析，一个动作构成一个音素。音素分为元音与辅音两大类。如汉语音节啊（ā）只有一个音素，爱（ài）有两个音素，代（dài）有三个音素等；在汉语中，音素即为拼音；在英语中，音素即为音标。

10 在本说明书中，提供了一种样本生成方法，本说明书同时涉及一种样本生成装置，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。

实际应用中，在基于神经网络的端到端语音合成方法中，由于该方法的特性，需要在模型训练前准备部分质量较高的样本数据，以实现训练出满足使用需求的语音合成模型；而这部分样本数据的往往需要在专业的录音棚中进行录
15 制后，再进行修剪和整理后才能够用于训练模型，不仅需要消耗较多的时间才能够完成数据的准备，而且成本消耗上也是一笔较大的开销；同时由于样本数据的要求较为严格，导致最终能够用于训练模型的数据少之更少，进而无法得到覆盖长度和韵律较为全面的样本数据，从而影响语音合成时的音色不像、韵律（音调起伏）不自然等问题。因此在样本数据的准备阶段如何生成质量高且
20 属性较为丰富的样本数据是亟需解决的问题。

本说明书提供一种样本生成方法，在获取到多个文本音频对后，计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对，以及所述目标文本音频对对应的拼接文本音频对，之后将所述目标文本音频对和所述拼接文本音频对拼
25 接为待检测文本音频对，并对所述待检测文本音频对进行检测，在所述待检测文本音频对满足预设检测条件的情况下，将所述待检测文本音频对写入所述训练数据库，实现在样本数据准备阶段，可以通过拼接的方式得到质量高且满足下游业务使用需求的样本数据，节省数据准备阶段的资源消耗成本，并且拼接后写入所述训练数据库中的样本数据的数据量较多，有效的解决了下游业务样
30 本数据量少、样本数据中音频长短分布不均匀导致语音合成效果不佳的问题，从而提高下游业务的业务处理效率。

图 1 示出了根据本说明书一实施例提供的一种样本生成方法的流程图，具体包括以下步骤：

步骤 S102，获取多个文本音频对，其中每个文本音频对中包含文本片段

和音频片段。

具体的,所述文本音频对具体是指具有对应关系的文本片段和音频片段组成的队列,所述文本片段包括但不限于字单元、词单元或句单元等,所述音频片段包括但不限于与字单元,词单元或句单元匹配的语音。

- 5 基于此,由于下游业务的处理过程为训练语音合成模型,即训练出能够将文本转换为音频的模型,因此需要在样本数据的准备阶段准备出大量满足下游业务使用需求的样本数据,并且为了能够训练出预测精准度较高的语音合成模型,将在数据准备阶段结合音频特征进行样本数据的准备,从而实现在较低的消耗成本下完成高质量、高数量的样本数据准备工作。
- 10 进一步的,由于下游业务对样本数据的需求量较大且质量要求较高,如果只是通过人工录制音频的方式构建样本数据将消耗较多的时间,并且属性覆盖范围较小,本申请为了解决该问题,采用拼接的方式实现样本数据的准备工作,实现可以拼接出大量的样本数据用于下游业务,并且为了保证样本数据的质量要求,将结合音频特征完成拼接处理,从而完成样本数据的准备工作;基于此,
- 15 在样本数据的准备阶段,可以准备长短不一的少量文本,并生成这部分文本对应的音频,之后基于少量的文本和其对应的少量的音频进行写入训练数据库中样本数据的构建。

- 本实施例将以少量文本中的任意一个文本作为目标文本为例,对所述样本生成方法进行描述,生成满足需求的样本数据过程均可参见本实施例相应的描述内容,在此不作过多赘述。
- 20

在生成所述样本数据的过程中,为了能够拼接出大量可以写入所述训练数据库的文本音频对,将对获取到的目标文本及其对应的音频进行切分后对齐,从而得到所述多个文本音频对,本实施例中,具体实现方式如下所述:

获取目标文本以及所述目标文本对应的音频;

- 25 对所述音频进行预处理获得目标音频,并将所述目标文本转换为音素序列;
- 将所述音素序列与所述目标音频进行对齐处理,并根据对齐处理结果生成所述多个文本音频对。

- 具体的,所述目标文本包括但不限于一篇文章或一段语句等,相应的,所述音频具体是指针对所述目标文本生成的语音,所述目标文本对应的音频可以采用录制的方式或语音合成的方式生成,本实施例在此不作任何限定,需要说明的是,所述音频与所述目标文本的匹配度较高,以保证后续拼接时可以得到较多能够写入所述训练数据库的样本数据;所述目标音频具体是指对所述音频进行标准化处理后得到的音频,所述音素序列具体是指构成所述目标文本的最
- 30

小单位组成的序列,所述对齐处理具体是指在音频中找到文本对应的时间区间。

基于此,为了保证文本音频对中的文本片段和音频片段相近程度较高,将在生成所述文本音频对时,从文本的最小单位出发完成对齐处理过程,即在获取到所述目标文本以及目标文本对应的音频之后,首先对所述音频进行预处理得到目标音频,实现去掉所述音频中对后续处理过程会造成干扰的部分,如音频起始和/或末尾处的空白音频片段(未发声的音频片段)或音频起始和/或末尾处的噪声较大的音频片段(无法分辨该音频片段的发音内容)等;其次将所述目标文本转换为音素序列,实现通过最小单位完成对齐文本和音频,提高对齐精准度;最后将所述音素序列和所述目标音频进行对齐处理,根据对齐处理结果即可得到所述多个文本音频对。

实际应用中,在将所述音素序列和所述目标音频进行对齐处理的过程中,可以使用 kald 对齐工具或者 HTK 对齐工具完成;此外还可以根据实际需求选择其他对齐工具完成所述音素序列与所述目标音频的对齐,本实施例在此不作任何限定。

进一步的,在根据对齐处理结果生成所述多个文本音频对的过程中,由于是从文本的最小单位完成对齐处理过程的,因此在完成音素与音频的对齐处理后,还需要将音素转换为文本,从而实现文本和音频精准的对齐,并经过切分以得到满足后续使用需求的多个文本音频对,本实施例中,具体实现方式如下所述:

根据对齐处理结果得到音素音频文件,并确定所述音素音频文件的切分位置;

按照所述切分位置对所述音素音频文件进行切分,获得多个音素音频对,其中每个音素音频对中包含音素片段和音频片段;

基于所述目标文本确定所述多个音素音频对中的每个音素音频对的音素片段对应的文本片段;

根据每个音素音频对中音素片段对应的文本片段,以及每个音素音频对中的音频片段生成所述多个文本音频对。

具体的,所述音素音频文件具体是指音素序列和目标音频经过对齐后得到的文件;所述切分位置可以是所述目标音频中的断句位置或发音中断时间超过设定时间阈值的位置;所述音素音频对具体是指由具有对应关系的音素片段和音频片段组成的队列。

基于此,在完成所述音素序列和所述目标音频的对齐之后,将得到所述音素音频文件,为了能够在后续拼接出大量的满足写入训练数据库的样本数据,

将根据所述音素音频文件中目标音频的切分位置对所述音素音频文件进行切分处理, 获得多个音素音频对, 其中每个音素音频对中包含音素片段和其对应的音频片段, 之后再基于所述目标文本将每个音素音频对的音素片段转换为文本片段, 从而实现根据每个音素音频对中音素片段对应的文本片段和音频片段组成所述文本音频对, 所述文本音频对中包含文本片段和其对应的音频片段。此时形成的多个文本音频对, 即可实现在后续处理时拼接出用于写入训练数据库的样本数据, 完成样本数据的准备工作。

具体实施时, 由于所述目标音频已经与所述音频序列完成精准的对齐, 因此在按照所述目标音频的切分位置对所述音素音频文件进行切分时, 可以实现被切分出的音素音频对中包含的音素片段和音频片段也是相对应的; 并且根据用户讲话的特性, 被切分后的音素音频中所包含的音素片段可以保证在目标文本中找到相应的文本片段, 不会出现音素片段被切分后不完整的问题。

例如, 目标文本为“我看了一场精彩的足球比赛”, 并针对该目标文本生成了12s的音频, 为了能够满足后续对齐处理, 将删除该音频起始处和末尾处的空白音频片段, 得到时间长度为10s的目标音频, 同时为了提高对齐精准度, 可以转换出目标文本“我看了一场精彩的足球比赛”对应的音素序列(wo kan le yi chang jing cai de zu qiu bi sai), 并通过kaldi对齐工具对音素序列和目标音频进行对齐处理, 得到如图2所示的对齐处理结果, 即由音素序列和目标音频组成的音素音频文件。

进一步的, 通过对音素音频文件中目标音频的检测, 确定目标音频中讲话的用户在录制这段语音时断句四次, 第一次断句在目标音频的3s处, 第二次断句在目标音频的4s处, 第三次断句在目标音频的6s处, 第四次断句在目标音频的8s处, 此时即可确定对音素音频文件的切分位置分别为 $T_1=3, T_2=4, T_3=6$ 和 $T_4=8$ 处, 按照切分位置对音素音频文件进行切分, 得到五个音素音频对, 第一音素音频对 P_1 由第一音素片段(wo kan le)和第一音频片段(0s~3s)组成; 第二音素音频对 P_2 由第二音素片段(yi chang)和第二音频片段(3s~4s)组成; 第三音素音频对 P_3 由第三音素片段(jing cai de)和第三音频片段(4s~6s)组成; 第四音素音频对 P_4 由第四音素片段(zu qiu)和第四音频片段(6s~8s)组成; 第五音素音频对 P_5 由第五音素片段(bi sai)和第五音频片段(8s~10s)组成。

更进一步的, 在得到音素音频对 $P_1 \sim P_5$ 之后, 还需要将各个音素音频对中的音素片段转换为文本片段, 从而得到能够用于后续拼接处理的文本音频对, 此时根据目标文本“我看了一场精彩的足球比赛”即可确定各个音素音频对中音素片段对应的文本片段, 即第一音素音频对 P_1 中包含的第一音素片段(wo

kan le) 对应的第一文本片段为 (我看了); 第二音素音频对 P_2 中包含的第二音素片段 (yi chang) 对应的第二文本片段为 (一场); 第三音素音频对 P_3 中包含的第三音素片段 (jing cai de) 对应的第三文本片段为 (精彩的); 第四音素音频对 P_4 中包含的第四音素片段 (zu qiu) 对应的第四文本片段为 (足球);

5 第五音素音频对 P_5 中包含的第五音素片段 (bi sai) 对应的第五文本片段为 (比赛)。

最后根据上述得到的文本片段和音频片段即可生成目标文本和目标音频对应的多个文本音频对, 如图 3 所示的切分结果, 其中, 第一文本音频对 TA_1 由第一文本片段 (我看了) 和第一音频片段 (0s~3s) 组成; 第二文本音频对 TA_2

10 由第二文本片段 (一场) 和第二音频片段 (3s~4s) 组成; 第三文本音频对 TA_3 由第三文本片段 (精彩的) 和第三音频片段 (4s~6s) 组成; 第四文本音频对 TA_4 由第四文本片段 (足球) 和第四音频片段 (6s~8s) 组成; 第五文本音频对 TA_5 由第五文本片段 (比赛) 和第五音频片段 (8s~10s) 组成; 以用于后续拼接出满足写入训练数据库的样本数据, 实现在训练语音合成模型时使用。

15 综上, 在构建所述多个文本音频对时, 通过从最小单位音素完成与目标音频的对齐, 不仅可以提高文本和音频对齐的精准度, 还能够在后续进行切分时, 保证被切分后的音素片段与音频片段的匹配程度是较高的, 从而实现切分出的多个文本音频对都可以用于后续的样本生成过程, 为后续生成样本数据提供充足的数量保障。

20 步骤 S104, 计算所述多个文本音频对中每个文本音频对的音频片段的音频特征, 并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对和所述目标文本音频对对应的拼接文本音频对。

具体的, 在上述获取到所述多个文本音频对的基础上, 进一步的, 由于写入所述训练数据库中的文本音频对是为了训练模型使用, 因此为了提高训练出的模型的预测精准度, 还需要保证训练模型所使用的样本数据的质量, 即在拼接出能够写入训练数据库的文本音频对时, 还需要考虑拼接前的文本音频对之间的音色和韵律等问题; 如果拼接前的两个文本音频对在音色和韵律等方面都不相同或相近, 或音调起伏不一致, 那么拼接出的文本音频对就存在着音频片段不匹配的问题, 以及文本片段前后文语义不一致的问题, 就无法用于训练模型。

25

30

基于此, 为了能够拼接出高质量的文本音频对 (能够写入训练数据库的样本数据), 以训练出预测精准度较高的模型, 本申请在拼接文本音频对前, 将计算各个文本音频对中的音频片段的音频特征, 之后基于该音频特征在多个文本音频对中选择出能够拼接的文本音频对, 实现将音调、韵律等属性相近的文

本音频对进行拼接，以得到音频片段连续、文本片段语义一致的文本音频对，从而得到高质量的文本音频对，以用于后续训练模型使用。

其中，所述音频特征包括但不限于表征音频片段的基音频率的特征，音频帧的特征和/或音频帧能量的特征等，通过文本音频对中音频片段的音频特征即可分析出各个需要拼接的文本音频对是否适合拼接，即通过基音频率特征、音频帧的特征和/或音频帧能量特征，确定需要拼接的文本音频对之间的音调、韵律等属性是否相近或相同，实现通过所述音频特征在所述多个文本音频对中筛选出所述拼接文本音频对；所述目标文本音频对具体是指基准文本音频对，所述拼接文本音频对即为满足与所述基准文本音频对进行拼接条件的文本音频对。

基于此，在得到所述多个文本音频对之后，为了能够得到能够相互拼接的文本音频对（即文本音频对之间音色、韵律相近或相同），以生成较多的样本数据，将计算各个文本音频对中音频片段的音频特征，并且在确定所述目标文本音频对之后，将基于目标文本音频对中音频片段的音频特征和多个文本音频对中各个文本音频对的音频片段的音频特征，在所述多个文本音频对中筛选出与所述目标文本音频对对应的拼接文本音频对，用于后续生成样本数据，实现在拼接出大量的样本数据时，不仅满足了样本数据的数量要求，还能够结合音频特征保证拼接前文本音频对之间的相近程度，进而提高拼接后的文本音频对的质量。

进一步的，在计算每个文本音频对的音频片段的音频特征的过程中，为了能够通过音频特征充分的反映出每个文本音频对中音频片段的属性特征，可以对每个文本音频对中音频片段进行分帧处理，通过音频帧分析出所述音频特征，本实施例中，具体实现方式如下所述：

提取所述多个文本音频对中每个文本音频对的音频片段，并对每个文本音频对的音频片段进行分帧处理，获得每个文本音频对的音频帧集合；

基于所述多个文本音频对中每个文本音频对的音频帧集合包含的音频帧，计算每个文本音频对的音频片段的基音频率特征和音频帧特征；

根据每个文本音频对的音频片段的所述基音频率特征和所述音频帧特征，确定每个文本音频对的音频片段的所述音频特征。

具体的，所述基音频率特征具体是指音频片段中发音体发出一系列频率、振幅各不相同的振动中，频率最低的振动所对应的频率数值；所述音频帧特征具体是指所述音频片段中音频帧经过傅里叶变换后在频谱上的点经过计算得到帧能量数值；相应的，所述基音频率特征可以用于分析文本音频对在进行拼

接时,彼此之间的发音振动幅度是否相近或相同;所述音频帧特征可以用于分析文本音频对在进行拼接时,彼此之间的能量分布是否相近或相同;以实现通过基音频率和帧能量选择拼接后效果较好的文本音频对进行拼接,以得到满足使用需求的样本数据。

5 基于此,首先提取所述每个文本音频对的音频片段,并对每个文本音频对的音频片段进行分帧处理,获得每个文本音频对的音频帧集合,其次基于音频帧集合中包含的音频帧计算每个文本音频对中音频片段的基音频率特征和音频帧特征,最后根据每个文本音频对的音频片段的基音频率特征和音频帧特征,即可确定每个文本音频对的音频片段的音频特征。

10 并且,由于任意两个文本音频对都存在拼接的可能,为了能够挖掘出更多可以写入训练数据库中的文本音频对,可以在计算所述音频特征时,计算每个文本音频对的音频片段的起始音频特征(起始基音频率特征和起始音频帧特征)和结束音频特征(结束基音频率特征和结束音频帧特征),之后在从所述多个文本音频对中筛选与所述目标文本音频对对应的拼接文本音频对时,可以将所
15 述目标文本音频对作为起始文本音频对,之后基于目标文本音频对中音频片段的结束音频特征和每个文本音频对中音频片段的起始音频特征,经过计算后筛选出所述拼接文本音频对,此后再对目标文本音频对和所述拼接文本音频对进行拼接时,则可以将所述目标文本音频对作为起始文本音频对,所述拼接文本音频对作为结束文本音频对,将二者按照先后顺序进行拼接,以得到后续需要进行检测的待检测文本音频对。
20

 还可以将所述目标文本音频对作为结束文本音频对,之后基于目标文本音频对中音频片段的起始音频特征和每个文本音频对中音频片段的结束音频特征,经过计算后筛选出所述拼接文本音频对,此后再对目标文本音频对和所述拼接文本音频对进行拼接时,则可以将所述目标文本音频对作为结束文本音频
25 对,所述拼接文本音频对作为起始文本音频对,将二者按照先后顺序进行拼接,以得到后续需要进行检测的待检测文本音频对;并且在此过程中,已经将目标文本音频对作为起始文本音频对和结束文本音频对与其他文本音频进行了可能性的拼接,当对其他文本音频对进行拼接时,就可以省略掉与目标文本音频对拼接的处理过程,以提高后续拼接处理过程中的处理效率。

30 具体实施时,计算所述基音频率特征可以通过时域估计法实现,即直接由音频波形来估计基音频率,如自相关法、并行处理法、平均幅度差法或数据减少法;还可以通过变换法实现,即将音频的语音信号变换到频域或者时域来估计基音频率,首先利用同态分析方法将声道的影响消除,得到属于激励部分的信息,然后再进行基音频率的计算,如倒谱法;再或者可以通过混合方法实现,

即先提取信号声道模型参数,然后利用它对信号进行滤波,得到音源序列,最后再利用自相关法或者平均幅度差法计算所述基音频率;计算所述音频片段的基音频率特征可以根据实际应用场景选择合适的方法实现,本实施例在此不作任何限定。

- 5 计算所述音频帧特征可以通过对每帧音频进行傅里叶变换,得到音频片段对应的频谱,之后统计频谱中各点的取值,并进行平方后相加即可得到各个音频帧的能量,最后再取平均值即可得到所述音频帧特征;或者将音频帧在复数域上的长度相加,得到每帧音频对应的帧能量,最后再取平均值即可得到所述音频帧特征;计算所述音频片段的音频帧特征可以根据实际应用场景选择合适的方法实现,本实施例在此不作任何限定。

此外,在对所述每个文本音频对中音频片段进行分帧时,可以按照固定的长度帧长进行分帧处理,如按照 32ms 或 64ms 进行分帧,具体帧长可以根据实际需求进行设定,本实施在此不作任何限定。

- 15 沿用上例,在得到文本音频对 $TA_1 \sim TA_5$ 的基础上,进一步的,为了后续能够拼接出大量满足质量要求的样本数据,可以提前计算各个文本音频对中音频片段的起始基音频率和起始帧能量,以及结束基音频率和结束帧能量;基于此,首先提起各个文本音频对中的音频片段,并对各个音频片段进行分帧处理,得到五个音频帧集合,分别对应文本音频对 $TA_1 \sim TA_5$;之后按照变换法和五个音频帧集合,计算出第一音频片段的起始基音频率 $F0_{s1}=N_{s1}$,结束基音频率 $F0_{e1}=N_{e1}$;第二音频片段的起始基音频率 $F0_{s2}=N_{s2}$,结束基音频率 $F0_{e2}=N_{e2}$;第三音频片段的起始基音频率 $F0_{s3}=N_{s3}$,结束基音频率 $F0_{e3}=N_{e3}$;第四音频片段的起始基音频率 $F0_{s4}=N_{s4}$,结束基音频率 $F0_{e4}=N_{e4}$;第五音频片段的起始基音频率 $F0_{s5}=N_{s5}$,结束基音频率 $F0_{e5}=N_{e5}$;同时按照傅里叶变换后统计频谱中点取值平方和的方法和五个音频帧集合,计算出第一音频片段的起始帧能量 $E_{s1}=M_{s1}$,结束帧能量 $E_{e1}=M_{e1}$;第二音频片段的起始帧能量 $E_{s2}=M_{s2}$,结束帧能量 $E_{e2}=M_{e2}$;第三音频片段的起始帧能量 $E_{s3}=M_{s3}$,结束帧能量 $E_{e3}=M_{e3}$;第四音频片段的起始帧能量 $E_{s4}=M_{s4}$,结束帧能量 $E_{e4}=M_{e4}$;第五音频片段的起始帧能量 $E_{s5}=M_{s5}$,结束帧能量 $E_{e5}=M_{e5}$ 。

- 30 进一步的,根据各个音频片段的起始/结束基音频率和起始/结束帧能量即可确定各个音频片段对应的音频特征,得出第一音频片段的起始音频特征为 $(F0_{s1}=N_{s1}, E_{s1}=M_{s1})$,结束音频特征为 $(F0_{e1}=N_{e1}, E_{e1}=M_{e1})$;第二音频片段的起始音频特征为 $(F0_{s2}=N_{s2}, E_{s2}=M_{s2})$,结束音频特征为 $(F0_{e2}=N_{e2}, E_{e2}=M_{e2})$;第三音频片段的起始音频特征为 $(F0_{s3}=N_{s3}, E_{s3}=M_{s3})$,结束音频特征为 $(F0_{e3}=N_{e3}, E_{e3}=M_{e3})$;第四音频片段的起始音频特征为 $(F0_{s4}=N_{s4}, E_{s4}=M_{s4})$,

结束音频特征为 ($F0_{e4}=N_{e4}$, $E_{e4}=M_{e4}$) ; 第五音频片段的起始音频特征为 ($F0_{s5}=N_{s5}$, $E_{s5}=M_{s5}$) , 结束音频特征为 ($F0_{e5}=N_{e5}$, $E_{e5}=M_{e5}$) ; 以用于后续在筛选拼接文本音频对时, 可以根据音频特征筛选出拼接效果较高的文本音频对进行拼接处理。

- 5 除此之外, 在计算每个文本音频对中音频片段的音频特征时, 为了能够提高计算效率, 还可以按照前向连接和后向连接的方式实现高效计算, 即在确定需要计算任意一个音频片段的音频特征时, 可以选择该音频片段的前面连接的音频片段和后面连接的音频片段组成相邻音频片段, 对任意一个音频片段和相邻音频片段中的两个音频片段同时计算各自的音频特征, 以实现节省计算音频特征所消耗的时间, 提供训练数据库更新的效率。
- 10

- 综上, 为了能够在后续拼接出高质量的样本数据, 将提前计算出各个文本音频对中音频片段的音频特征, 实现在属性维度分析各文本音频对中音频片段的属性信息, 从而在筛选与目标文本音频对对应的拼接文本音频对时, 可以结合音频特征筛选出拼接后效果较好的文本音频对作为所述拼接文本音频对, 以提高样本数据的质量。
- 15

更进一步的, 在计算完成所述文本音频对中音频片段的音频特征之后, 将根据音频特征在所述文本音频对中筛选出目标文本音频对和所述拼接文本音频对, 以用于后续的拼接处理, 以得到满足写入需求的样本数据, 本实施例中, 具体实现方式如图 4 所示的步骤 S1042~步骤 S1052。

- 20 步骤 S1042, 将所述多个文本音频对中每个文本音频对的音频片段、文本片段和音频特征进行整合, 获得每个文本音频对对应的文本音频包, 并写入片段数据库;

步骤 S1044, 在所述片段数据库中选择任意一个文本音频包作为目标文本音频包, 并将所述目标文本音频包中的文本音频对确定为所述目标文本音频对;

- 25 步骤 S1046, 在所述片段数据库中选择除所述目标文本音频包外的文本音频包组成待筛选文本音频包集合;

步骤 S1048, 将所述待筛选文本音频包集合中包含的各个待筛选文本音频包的文本音频对确定为待筛选文本音频对;

- 30 步骤 S1050, 基于所述目标文本音频对的音频片段的音频特征和所述待筛选文本音频对的音频片段的音频特征, 在所述待筛选文本音频包集合中筛选出所述拼接文本音频包;

步骤 S1052, 将所述拼接文本音频包中的文本音频对作为所述拼接文本音频对。

具体的,所述文本音频包具体是指写入片段数据库的文本音频对和其对应的文本特征组成的集合,所述片段数据库具体是指临时存储文本音频对中文本片段、音频片段和其对应的音频特征的数据库,当获得所述多个文本音频对之后,由于后续针对所述目标文本音频对筛选关联的拼接文本音频对时还需要消耗一定时间,因此可以将所述文本音频包写入所述片段数据库,当需要进行拼接处理时,再从所述片段数据库中提取文本音频对进行后续的拼接处理即可。

进一步的,所述待筛选文本音频包集合中包含的待筛选文本音频包具体是指片段数据库中除所述目标文本音频包之外的其他文本音频包,相应的,所述待筛选文本音频对即为所述待筛选文本音频包中包含的文本音频对;所述拼接文本音频包具体是指可以与所述目标文本音频包进行拼接的文本音频对所属的文本音频包。

基于此,首先将所述多个文本音频对中每个文本音频的音频片段、文本片段和音频特征进行整合,得到多个文本音频包,并临时写入所述片段数据库;之后在需要进行文本音频对拼接时,从所述片段数据库中选择任意一个文本音频包作为目标文本音频包,并提取所述目标文本音频包中包含的文本音频对作为所述目标文本音频对,同时再从所述片段数据库中选择除所述目标文本音频包之外的其他文本音频包作为待筛选文本音频包,并组成待筛选文本音频包集合。

其次,提取所述待筛选文本音频包集合包含的各个待筛选文本音频包中的文本音频对作为所述待筛选文本音频对,并根据各个待筛选文本音频包中整合的音频特征确定各个待筛选文本音频对中音频片段的音频特征。

最后,基于目标文本音频对中音频片段的音频特征和各个待筛选文本音频对中音频片段的音频特征,即可计算出目标文本音频对与各个待筛选文本音频对的匹配程度,之后选择匹配程度较高的待筛选文本音频对所属的文本音频包作为所述拼接文本音频包即可,即将所述拼接文本音频包中的文本音频对作为与所述目标文本音频对对应的拼接文本音频对,以用于后续对二者进行拼接,得到满足写入训练数据库需求的样本数据。

更进一步的,在基于目标文本音频对中音频片段的音频特征和待筛选文本音频对中音频片段的音频特征,筛选所述拼接文本音频包的过程中,为了能够筛选出与所述目标文本音频对匹配度较高的文本音频对,本实施例中,可以采用如下的方式得到拼接文本音频包,从而将所述拼接文本音频包中的文本音频对作为所述拼接文本音频对,以用于后续与目标文本音频对进行拼接,得到满足写入训练数据库的样本数据,具体实现方式如下所述:

根据所述目标文本音频包确定所述目标文本音频对的音频片段的第一音频特征, 以及根据所述待筛选文本音频包确定所述待筛选文本音频对的音频片段的第二音频特征;

计算所述第一音频特征和所述第二音频特征之间的特征距离;

- 5 将所述特征距离小于预设距离阈值的待筛选文本音频对所属的待筛选文本音频包确定为所述拼接文本音频包。

具体的, 所述第一音频特征即为所述目标文本音频对中音频片段的音频特征, 所述第二音频特征即为所述待筛选文本音频对中音频片段的音频特征, 相应的, 所述特征距离具体是指评价文本音频对之间匹配程度的数值, 所述特征
10 距离越大, 表明所述文本音频对之间的匹配程度越低, 反之, 所述特征距离越小, 表明所述文本音频对之间的匹配程度越高。

基于此, 在确定所述目标文本音频对的音频片段的第一音频特征, 以及所述待筛选文本音频对的第二音频片段的音频特征的基础上, 可以根据所述
15 第一音频特征和所述第二音频特征, 计算出所述目标文本音频对与各个待筛选文本音频对之间的特征距离, 之后选择特征距离小于预设距离阈值的待筛选文本音频对作为所述拼接文本音频对即可, 以用于后续拼接处理。

在计算所述特征距离的过程中, 可以采用如下公式 (1) 实现:

$$L = (F0_e - F0_s)^2 + (E_e - E_s)^2 \quad (1)$$

其中, L 表示特征距离, $F0_e$ 表示目标文本音频对中音频片段的结束基音
20 频率特征, $F0_s$ 表示待筛选文本音频对中音频片段的起始基音频率特征, E_e 表示目标文本音频对中音频片段的结束音频帧特征, E_s 表示待筛选文本音频对中音频片段的起始音频帧特征。

沿用上例, 在得到文本音频对 $TA_1 \sim TA_5$, 以及各个文本音频对中音频片段的
25 音频特征之后, 可以将文本音频对及其对应的音频特征整合为文本音频包 ($TP_1 \sim TP_5$) 写入片段数据库 D 中, 以用于后续拼接时可以选择文本音频包进行拼接处理。

进一步的, 在片段数据库 D 选择文本音频包 TP_1 作为目标文本音频包, 此时
30 确定文本音频包 TP_1 中包含的文本音频对 TA_1 为目标文本音频对; 同时在片段数据库 D 中选择文本音频包 TP_2 、 TP_3 、 TP_4 、 TP_5 作为待筛选文本音频包, 则将各个待筛选文本音频包中的文本音频对 TA_2 、 TA_3 、 TA_4 、 TA_5 作为待筛选文本音频对, 并且根据文本音频包 TP_1 可以确定目标文本音频对 TA_1 的音频特征为 $[(F0_{s1}=N_{s1}, E_{s1}=M_{s1}), (F0_{e1}=N_{e1}, E_{e1}=M_{e1})]$; 根据文本音频包 TP_2 可以确定待筛选文本音频对 TA_2 的音频特征为 $[(F0_{s2}=N_{s2}, E_{s2}=M_{s2}), (F0_{e2}=N_{e2},$

$E_{e2}=M_{e2})$]; 根据文本音频包 TP_3 可以确定待筛选文本音频对 TA_3 的音频特征为 $[(F0_{s3}=N_{s3}, E_{s3}=M_{s3}), (F0_{e3}=N_{e3}, E_{e3}=M_{e3})]$; 根据文本音频包 TP_4 可以确定待筛选文本音频对 TA_4 的音频特征为 $[(F0_{s4}=N_{s4}, E_{s4}=M_{s4}), (F0_{e4}=N_{e4}, E_{e4}=M_{e4})]$; 根据文本音频包 TP_5 可以确定待筛选文本音频对 TA_5 的音频特征为 $[(F0_{s5}=N_{s5}, E_{s5}=M_{s5}), (F0_{e5}=N_{e5}, E_{e5}=M_{e5})]$ 。

更进一步的, 根据上述公式 (1) 计算目标文本音频对与各个待筛选文本音频对的特征距离, 确定目标文本音频对 TA_1 与待筛选文本音频对 TA_2 的特征距离 $L_1 = (F0_{e1}-F0_{s2})^2 + (E_{e1}-E_{s2})^2 = (N_{e1}-N_{s2})^2 + (M_{e1}-M_{s2})^2$; 确定目标文本音频对 TA_1 与待筛选文本音频对 TA_3 的特征距离 $L_2 = (F0_{e1}-F0_{s3})^2 + (E_{e1}-E_{s3})^2 = (N_{e1}-N_{s3})^2 + (M_{e1}-M_{s3})^2$; 确定目标文本音频对 TA_1 与待筛选文本音频对 TA_4 的特征距离 $L_3 = (F0_{e1}-F0_{s4})^2 + (E_{e1}-E_{s4})^2 = (N_{e1}-N_{s4})^2 + (M_{e1}-M_{s4})^2$; 确定目标文本音频对 TA_1 与待筛选文本音频对 TA_5 的特征距离 $L_4 = (F0_{e1}-F0_{s5})^2 + (E_{e1}-E_{s5})^2 = (N_{e1}-N_{s5})^2 + (M_{e1}-M_{s5})^2$ 。

最后, 由于特征距离越小表明目标文本音频对与待筛选文本音频对的匹配程度越高, 故在将特征距离 $L_1 \sim L_4$ 分别与预设距离阈值 L_T 进行比较时, 选择小于距离阈值 L_T 的待筛选文本音频对作为能够与目标文本音频对 TA_1 拼接的拼接文本音频对, 通过比较结果确定特征距离 L_1 、 L_3 和 L_4 小于距离阈值 L_T , 进一步表明在将目标文本音频对 TA_1 与待筛选文本音频对 TA_2 、 TA_4 和 TA_5 进行拼接时, 可以保证彼此之间的音色、韵律都是较为接近的, 以满足后续可以拼接出质量较高的样本数据, 即确定待筛选文本音频对 TA_2 、 TA_4 和 TA_5 可以与目标文本音频对 TA_1 进行拼接, 则将文本音频对 TA_2 、 TA_4 和 TA_5 确定为目标文本音频对 TA_1 的拼接文本音频对。

此外, 为了能够提高后续计算其他文本音频对作为目标文本音频对时特征距离的计算效率, 可以将所述目标文本音频对作为后向文本音频对, 待筛选文本音频对作为前向文本音频对, 再计算彼此之间的特征距离。如在计算目标文本音频对 TA_1 与待筛选文本音频对 TA_2 的特征距离 L_1 时, 还可以计算目标文本音频对 TA_1 与待筛选文本音频对 TA_2 的特征距离 L_{11} , 其中 $L_{11} = (F0_{e2}-F0_{s1})^2 + (E_{e2}-E_{s1})^2 = (N_{e2}-N_{s1})^2 + (M_{e2}-M_{s1})^2$, 特征距离 L_{11} 表征在将待筛选文本音频对 TA_2 作为拼接后的前向文本音频对, 以及将目标文本音频对 TA_1 作为拼接后的后向文本音频对时二者之间的匹配程度; 实现在计算目标文本音频对与各个待筛选文本音频对的特征距离时, 将目标文本音频对作为前向文本音频对和后向文本音频对的特征距离都计算完成, 节省在将待筛选文本音频对 TA_2 作为目标文本音频对时, 计算文本音频对 TA_1 与文本音频对 TA_2 特征距离的计算过程, 从而提高后续计算特征距离的效率。

5 综上,为了能够从所述多个文本音频对中筛选出与所述目标文本音频对对应的拼接文本音频对,将结合音频特征对所述拼接文本音频对进行筛选,从而实现筛选出的文本音频对与所述目标文本音频对在音色、韵律等属性上是彼此接近的,可以在后续拼接出满足使用需求的待检测文本音频对,以实现所述训练数据库进行扩充,供下游业务所使用。

步骤 S106,将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对,并对所述待检测文本音频对进行检测。

10 具体的,在上述基于所述音频特征得到与所述目标文本音频对对应的拼接文本音频对的基础上,进一步的,将对所述目标文本音频对和所述拼接文本音频对进行拼接,得到所述待检测文本音频对,并且由于所述待检测文本音频对是由两个文本音频对拼接而成,为了能够进一步保证写入所述训练数据库中的文本音频对是有质量保证的(拼接后的待检测文本音频对效果较好),在写入所述训练数据库之前,还可以对所述待检测文本音频对进行检测,检测所述待检测文本音频对中音频片段是否清晰,文本片段长度是否合适等等,从而得到
15 质量更佳的文本音频对写入所述训练数据库。

而在此之前,由于所述目标文本音频对可能满足写入所述训练数据库的要求,即无需与其他文本音频对进行拼接即可写入所述训练数据库,因此为了提高所述训练数据库的丰富程度,以及避免遗漏可以写入所述训练数据库的样本数据,可以在进行拼接处理前,判断目标文本音频对是否满足预设检测条件,
20 本实施例中,具体实现方式如下所述:

对所述目标文本音频对中的音频片段进行采样处理获得目标采样信息,以及确定所述目标文本音频对中的文本片段的目标文本信息;

判断所述目标采样信息和所述目标文本信息是否满足所述预设检测条件;

25 若否,则执行将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对,并对所述待检测文本音频对进行检测的步骤。

若是,则将所述目标文本音频对写入所述训练数据库。

30 具体的,所述目标采样信息具体是指对所述目标文本音频对中音频片段进行随机采样时的采样位数和采样频率,所述采样位数可以理解为处理音频片段时的解析度,采样位数越大,解析度越高,音频片段的真实性也就越高,反之采样位数越小,解析度越低,音频片段的真实性也就越低;所述采样频率是指在一秒内对音频片段的采样次数,采样频率越高音频片段的还原就越真实越自然,反之采样频率越低音频片段的还原就越不真实越不自然;所述目标文本信息具体是指所述目标文本音频对中文本片段的长度信息,字符数量信息等;相

应的,所述预设检测条件具体是指检测音频片段和文本片段是否符合写入训练数据库的条件,可以在文本音频对中的音频片段和文本片段都满足预设检测条件的情况下,将其写入所述训练数据库,也可以在文本音频对中的音频片段或文本片段满足预设检测条件的情况下,将其写入所述训练数据库。

- 5 基于此,在得到所述目标文本音频对和其对应的拼接文本音频对的基础上,进一步的,将对所述目标文本音频对中的音频片段在 $[0, 1]$ 之间进行随机采样处理,得到所述目标采样信息,同时确定所述目标文本音频对中文本片段的目标文本信息,之后判断所述目标采样信息和所述目标文本信息是否满足所述预设检测条件,若是,说明所述目标文本音频对已经满足了写入所述训练数据库的要求,则可以直接将所述目标文本音频对写入所述训练数据库,作为下游业务
- 10 的样本数据;若否,说明所述目标文本音频对并未满足写入所述训练数据库的要求,则可以执行步骤 S106,再对拼接后的待检测文本音频对进行检测,从而实现均衡写入训练数据库中文本音频对中音频片段和文本片段,保证训练数据库中文本音频对中文本片段和音频片段在形式上(音频长度,文本长度或音频能量等)都是相近或相同的,更加方便下游业务使用。
- 15

- 沿用上例,在从文本音频对 $TA_1 \sim TA_5$ 中选择文本音频对 TA_1 作为目标文本音频对的基础上,进一步的,此时将对文本音频对 TA_1 中的第一音频片段 ($0s \sim 3s$) 在 $[0, 1]$ 之间进行随机采样,得到第一音频片段的目标采样信息为 U ,同时确定文本音频对 TA_1 中的第一文本片段(我看了)的长度为 X 个字符,此时判断第一音频片段的目标采样信息 U 是否大于预设采样 U_t ,且第一文本片段的长度 X 是否小于预设文本长度 X_t ,若是,说明文本音频对 TA_1 中的第一音频片段和第一文本片段都满足了写入训练数据库 T 的要求,则此时可以将文本音频对 TA_1 写入训练数据库 T ,作为下游业务训练语音合成模型时的样本数据;若否,说明文本音频对 TA_1 中的第一音频片段或第一文本片段无法满足
- 20 写入训练数据库 T 的要求,则此时可以将拼接文本音频对 TA_2 、 TA_4 和 TA_5 与文本音频对 TA_1 进行拼接,得到多个待检测文本音频对,再通过对待检测文本音频对进行检测,以实现得到满足写入训练数据库 T 的要求的文本音频对。
- 25

- 综上,在对拼接文本音频对和目标文本音频对进行拼接之前,可以对所述目标文本音频对进行检测,以避免遗漏满足写入训练数据库的文本音频对,进而提高了训练数据库的丰富程度。
- 30

进一步的,在拼接所述目标文本音频对和所述拼接文本音频对的过程中,由于各文本音频对中包含文本片段和音频片段,因此需要拼接文本片段的同时拼接音频片段,以生成所述待检测文本音频段,本实施例中,具体实现方式如下所述:

提取所述目标文本音频对中的目标文本片段和目标音频片段,以及所述拼接文本音频对中的拼接文本片段和拼接音频片段;

将所述目标文本片段和所述拼接文本片段拼接为待检测文本片段,以及将所述目标音频片段和所述拼接音频片段拼接为待检测音频片段;

5 基于所述待检测文本片段和所述待检测音频片段组成所述待检测文本音频对。

具体的,首先提取所述目标文本音频对中的目标文本片段和目标音频片段,同时提取所述拼接文本音频对中的拼接文本片段和拼接音频片段,其次将所述目标文本片段和所述拼接文本片段拼接为待检测文本片段,以及将所述目标音频片段和所述拼接音频片段拼接为待检测音频片段;最后基于所述待检测文本片段和所述待检测音频片段即可组成所述待检测文本音频对。

更进一步的,在对所述待检测文本音频对进行检测的过程中,为了保证待检测文本音频对的质量,不仅可以对所述待检测文本音频对中的待检测文本片段进行检测,还可以同时对待检测文本音频对中的待检测音频片段进行检测,从而保证写入训练数据库的文本音频对中的文本片段和音频片段都是满足写入要求的,本实施例中,具体实现方式如下所述:

15 对所述待检测音频片段进行采样处理获得待检测采样信息,以及确定所述待检测文本片段的待检测文本信息;

20 基于所述预设检测条件对所述待检测采样信息和所述待检测文本信息进行检测;

在所述待检测采样信息和所述待检测文本信息均满足所述预设检测条件的情况下,将所述待检测文本音频对写入所述训练数据库。

具体的,在拼接出所述待检测文本音频对的基础上,进一步的,对所述待检测文本音频对中的待检测音频片段进行随机采样处理,得到所述待检测音频片段的待检测采样信息,同时确定所述待检测文本音频对所述待检测文本片段的待检测文本信息,之后基于所述预设的检测条件对所述待检测采样信息和所述待检测文本信息进行检测;若所述待检测采样信息和所述待检测文本信息均满足所述预设检测条件,说明所述待检测文本音频对可以写入所述训练数据库,则将所述待检测文本音频对作为样本数据写入所述训练数据库即可,若所述待检测采样信息或所述待检测文本信息不满足所述预设检测条件的情况下,说明所述待检测文本音频对不可以写入所述训练数据库,则舍弃所述待检测文本音频对即可。

沿用上例,在确定目标文本音频对 TA_1 和拼接文本音频对 TA_2 、 TA_4 和 TA_5

的基础上,进一步的,将对目标文本音频对和拼接文本音频对进行拼接处理,即提取目标文本音频对 TA_1 中的第一音频片段 (0s~3s) 和第一文本片段 (我看了),同时提取拼接文本音频对 TA_2 中的第二音频片段 (3s~4s) 和第二文本片段 (一场),提取拼接文本音频对 TA_4 中的第四音频片段 (6s~8s) 和第四文本片段 (足球),以及提取拼接文本音频对 TA_5 中第五音频片段 (8s~10s) 和第五文本片段 (比赛)。

进一步的,将第一音频片段和第二音频片段进行拼接,得到第一待检测音频片段 (长度为 4s),将第一音频片段和第四音频片段进行拼接,得到第二待检测音频片段 (长度为 5s),以及将第一音频片段和第五音频片段进行拼接,得到第三待检测音频片段 (长度为 5s);同时将第一文本片段和第二文本片段进行拼接,得到第一待检测文本片段 (我看了一场),将第一文本片段和第四文本片段进行拼接,得到第二待检测文本片段 (我看了足球),将第一文本片段和第五文本片段进行拼接,得到第三待检测文本片段 (我看了比赛);此时将第一待检测音频片段和第一待检测文本片段组合为第一待检测文本音频对,将第二待检测音频片段和第二待检测文本片段组合为第二待检测文本音频对,将第三待检测音频片段和第三待检测文本片段组合为第三待检测文本音频对。

更进一步的,在得到第一待检测文本音频对、第二待检测文本音频对和第三待检测文本音频对的基础上,进一步的将对上述三个待检测文本音频对进行检测,选择能够写入训练数据库的文本音频对作为样本数据,基于此,分别对各个待检测文本音频对中的待检测音频片段在 $[0, 1]$ 之间进行随机采样,确定第一待检测文本音频对中第一待检测音频片段的采样结果为 U_1 ,确定第二待检测文本音频对中第二待检测音频片段的采样结果为 U_2 ,确定第三待检测文本音频对中第三待检测音频片段的采样结果为 U_3 ;同时确定第一待检测文本音频对中第一待检测文本片段的文本长度为 X_1 ,确定第二待检测文本音频对中第二待检测文本片段的文本长度为 X_2 ,确定第三待检测文本音频对中第三待检测文本片段的文本长度为 X_3 。

最后,分别判断采样结果 U_1 、 U_2 和 U_3 是否大于预设采样结果 U_t ,且文本长度 X_1 、 X_2 和 X_3 是否小于预设文本长度 X_t ,根据判断结果确定采样结果 U_2 大于预设采样结果 U_t ,且文本长度 X_2 小于预设文本长度 X_t ,以及采样结果 U_3 大于预设采样结果 U_t ,且文本长度 X_3 小于预设文本长度 X_t ,即第二待检测文本音频对和第三待检测文本音频对符合写入训练数据库 T,则将第二待检测文本音频对 (音频 5s,文本“我看了足球”),以及第三待检测文本音频对 (音频 5s,文本“我看了比赛”)作为样本数据写入训练数据库 T,以用于后续训练语音合成模型使用。

综上,通过音频维度和文本维度对所述待检测文本音频对进行检测,实现了写入所述训练数据库中的文本音频对都是满足写入要求的,有效的提高了所述训练数据库中样本数据的质量。

5 步骤 S108,在所述待检测文本音频对满足预设检测条件的情况下,将所述待检测文本音频对写入训练数据库。

具体的,在所述待检测文本音频对满足预设检测条件的情况下,说明所述待检测文本音频对满足写入所述训练数据库的要求,则将所述待检测文本音频对作为样本数据写入所述训练数据库即可,以用于后续在训练语音合成模型时,可以从训练数据库中提取满足训练要求的样本数据,提高训练出的语音合成模型
10 的预测精准度。

基于此,通过对多个文本中的各个文本进行上述处理,即可得到大量符合写入训练数据库的样本数据,并且写入所述训练数据库中的样本数据无论是数量和质量都满足下游训练模型的需求,从而实现在训练语音合成模型前,可以节省数据准备阶段的成本,提高样本数据的丰富程度。

15 此外,在向所述训练数据库写入样本数据的过程中,考虑到数据库的容量以及下游业务的需求,可以对写入训练数据库的数据条数进行限制,即在向所述训练数据库写入满足预设检测条件的待检测文本音频对时,检测所述训练数据库中文本音频对的条数是否小于等于预设数据量阈值,若小于,说明训练数据库还可以继续写入文本音频对,则将满足预设检测条件的文本音频对写入所
20 述训练数据库即可,若大于,说明训练数据库无法继续写入文本音频对,则停止后续拼接文本音频对的处理即可。

并且在可以向所述训练数据库中写入文本音频对时,为了避免重复存储占用过多的存储资源,还可以在写入训练数据库前,检测训练数据库中是否存在该文本音频对,若存在,则舍弃,继续进行其他文本音频对的拼接处理即可,
25 若不存在,则写入训练数据库后,再继续其他文本音频对的拼接处理即可。

进一步的,在完成所述训练数据库的扩增之后,此时所述训练数据库中的文本音频对即可作为样本数据(样本文本音频对)对下游业务中的语音合成模型进行训练,本实施例中,训练模型的过程如下所述:

30 在所述训练数据库中提取样本文本音频对,所述样本文本音频对中包含样本文本片段和样本音频片段;

基于所述样本文本片段和所述样本音频片段对语音合成模型进行训练,获得目标语音合成模型。

实际应用中,在对所述语音合成模型进行训练的过程中,可以从所述训练

数据库中提取大量的样本文本音频对,之后基于样本文本音频对中样本文本片段和样本音频片段对所述语音合成模型进行训练,直至得到满足训练停止条件的语音合成模型即可,将其作为目标语音合成模型进行存储,以用于在语音合成场景中可以将文本转换为音频。如文本为“我喜欢看足球比赛”,将该文本输入至语音合成模型进行处理,即可得到该文本对应的音频,实现将文本转换为语音的处理。

此外,在拼接后的待检测文本音频对未满足预设检测条件的情况下,说明所述待检测文本音频对中的待检测音频片段或待检测文本片段可能不满足所述预设检测条件,为了能够得到满足条件的样本数据,此时可以根据所述音频特征在所述多个文本音频对中筛选出与所述拼接文本音频对对应的多度拼接文本音频对,再进行拼接后检测处理,直至得到满足写入训练数据库的文本音频对,需要说明的是,如果一直进行拼接检测处理,所得到的待检测文本音频对可能一直无法满足写入训练数据库的要求,因此在持续进行拼接和检测的过程中,可以设置停止条件,当拼接次数达到一定条件的情况下,即可停止对该文本音频对进行处理,将其舍弃即可,本实施例中,具体实现方式如下所述:

在所述待检测文本音频对未满足预设检测条件的情况下,根据所述音频特征在所述多个文本音频对中筛选出所述拼接文本音频对对应的多度拼接文本音频对;

将所述多度拼接文本音频对和所述待检测文本音频对拼接为多度待检测文本音频对,并判断所述多度待检测文本音频对是否满足所述预设检测条件;

若是,则将所述多度待检测文本音频对写入所述训练数据库;

若否,则将所述多度拼接文本音频对作为所述拼接文本音频对,以及将所述多度待检测文本音频对作为所述待检测文本音频对,并执行所述根据所述音频特征在所述多个文本音频对中筛选出所述拼接文本音频对对应的多度拼接文本音频对步骤。

具体的,所述多度拼接文本音频对具体是指与所述拼接文本音频对可以进行拼接的文本音频对;基于此,在所述待检测文本音频对未满足预设检测条件的情况下,说明目标文本音频对和所述拼接文本音频对拼接后的待检测文本音频对不符合写入所述训练数据库的要求,为了能够得到满足写入要求的文本音频对,可以再根据音频特征在所述多个文本音频对中选择能够与拼接文本音频对进行拼接的多度拼接文本音频对,之后将待检测文本音频对和多度拼接文本音频对进行拼接,得到多度待检测文本音频对,再对所述多度待检测文本音频对进行检测,若所述多度待检测文本音频对满足预设检测条件,将其写入所述

训练数据库即可，若所述多度待检测文本音频对不满足预设检测条件，可以将多度待检测文本音频对作为待检测文本音频对，以及将多度拼接文本音频对作为拼接文本音频对，返回执行筛选多度拼接文本音频对的过程，直至得到满足写入训练数据库需求的文本音频对即可，或者直至达到停止拼接条件后舍弃该

5 文本音频对即可。

沿用上例，在得到第一待检测文本音频对不符合写入训练数据库 T 的要求情况下，由于第一待检测文本音频对中的第一待检测音频片段是由第一音频片段和第二音频片段组成，以及第一待检测文本片段是由第一文本片段和第二文本片段组成，因此可以选择与第二文本音频对 TA₂ 具有拼接可能的第三文本音频对 TA₃ 作为多度拼接文本音频对，之后将多度拼接文本音频对 TA₃ 与第一待检测文本音频对 (TA₁+TA₂) 进行拼接，得到多度待检测文本音频对 (TA₁+TA₂+TA₃)。

10

同时，确定多度待检测文本音频对中的多度待检测音频片段由第一音频片段、第二音频片段和第三音频片段组成，多度待检测文本音频对中的多度待检测文本片段由第一文本片段、第二文本片段和第三文本片段组成，即多度待检测文本音频对为 (音频片段 6s，文本片段“我看了一场精彩的”)，此时对多度待检测文本音频对进行检测，若该多度待检测文本音频对满足预设检测条件，则将多度待检测文本音频对写入训练数据库 T 中即可，若多度待检测文本音频对不满足预设检测条件，可以再选择与第三文本音频对 TA₃ 具有拼接可能的

15

20 文本音频对进行拼接和检测处理，或者舍弃该多度待检测文本音频对，选择其他文本音频对进行上述处理，以得到满足写入训练数据库 T 的样本数据即可。

综上，通过循环拼接的方式保证所述训练数据库中样本数据的均衡性，不仅可以方便下游业务在训练模型时使用，还能够提高所述训练数据库的丰富程度，进而有效的保证了下游业务的使用需求。

25 本说明书提供一种样本生成方法，在获取到多个文本音频对后，计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对，以及所述目标文本音频对对应的拼接文本音频对，之后将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对，并对所述待检测文本音频对进行检测，在所述待检测

30 文本音频对满足预设检测条件的情况下，将所述待检测文本音频对写入所述训练数据库，实现在样本数据准备阶段，可以通过拼接的方式得到质量高且满足下游业务使用需求的样本数据，节省数据准备阶段的资源消耗成本，并且拼接后写入所述训练数据库中的样本数据的数据量较多，有效的解决了下游业务样本数据量少、样本数据中音频长短分布不均匀导致语音合成效果不佳的问题，

从而提高下游业务的业务处理效率。

下述结合附图 5，以本说明书提供的样本生成方法在语音合成场景中的应用为例，对所述样本生成方法进行进一步说明。其中，图 5 示出了本说明书一实施例提供的一种应用于语音合成场景中的样本生成方法的处理流程图，具体包括以下步骤：

步骤 S502，获取目标文本以及目标文本对应的音频。

实际应用中，在基于神经网络的端到端语音合成方法中，由于该方法的特性，需要在模型训练前准备部分质量较高的样本数据，以实现训练出满足使用需求的语音合成模型；而这部分样本数据的往往需要在专业的录音棚中进行录制后，再进行修剪和整理后才能够用于训练模型，不仅需要消耗较多的时间才能够完成数据的准备，而且成本消耗上也是一笔较大的开销；同时由于样本数据的要求较为严格，导致最终能够用于训练模型的数据少之更少，进而无法得到覆盖长度和韵律较为全面的样本数据，从而影响语音合成时的音色不像、韵律（音调起伏）不自然等问题。因此在样本数据的准备阶段如何生成质量高且属性较为丰富的样本数据是亟需解决的问题。

本实施例提供一种应用于语音合成场景中的样本生成方法，以解决上述问题。

步骤 S504，对音频进行预处理获得目标音频，并将目标文本转换为音素序列。

步骤 S506，将音素序列与目标音频进行对齐处理，根据对齐处理结果得到音素音频文件，并确定音素音频文件的切分位置。

步骤 S508，按照切分位置对音素音频文件进行切分获得多个音素音频对，并基于目标文本确定多个音素音频对中的每个音素音频对的音素片段对应的文本片段。

步骤 S510，根据每个音素音频对中音素片段对应的文本片段，以及每个音素音频对中的音频片段生成多个文本音频对。

步骤 S512，提取多个文本音频对中每个文本音频对的音频片段，并对每个文本音频对的音频片段进行分帧处理，获得每个文本音频对的音频帧集合。

步骤 S514，基于多个文本音频对中每个文本音频对的音频帧集合包含的音频帧，计算每个文本音频对的音频片段的基音频率特征和音频帧特征。

步骤 S516，将多个文本音频对中每个文本音频对的音频片段、文本片段、基音频率特征和音频帧特征进行整合，获得每个文本音频对对应的文本音频包，并写入片段数据库。

步骤 S518, 在片段数据库中选择任意一个文本音频包作为目标文本音频包, 并将目标文本音频包中的文本音频对确定为目标文本音频对。

步骤 S520, 在片段数据库中选择除目标文本音频包外的文本音频包组成待筛选文本音频包集合。

- 5 步骤 S522, 将待筛选文本音频包集合中包含的各个待筛选文本音频包的文本音频对确定为待筛选文本音频对。

步骤 S524, 根据目标文本音频包确定目标文本音频对的音频片段的基音频率特征和音频帧特征, 以及根据待筛选文本音频包确定待筛选文本音频对的音频片段的基音频率特征和音频帧特征。

- 10 步骤 S526, 基于目标文本音频对的音频片段的基音频率特征和音频帧特征, 以及待筛选文本音频对的音频片段的基音频率特征和音频帧特征计算特征距离。

步骤 S528, 将特征距离小于预设距离阈值的待筛选文本音频对所属的待筛选文本音频包确定为拼接文本音频包。

- 15 步骤 S530, 将拼接文本音频包中的文本音频对作为拼接文本音频对。

步骤 S532, 提取目标文本音频对中的目标文本片段和目标音频片段, 以及拼接文本音频对中的拼接文本片段和拼接音频片段。

步骤 S534, 将目标文本片段和拼接文本片段拼接为待检测文本片段, 以及将目标音频片段和拼接音频片段拼接为待检测音频片段。

- 20 步骤 S536, 基于待检测文本片段和待检测音频片段组成待检测文本音频对。

步骤 S538, 对待检测文本音频对中的待检测音频片段进行采样处理获得待检测采样信息, 以及确定待检测文本音频对中的待检测文本片段的待检测文本信息。

- 25 步骤 S540, 在待检测采样信息和待检测文本信息均满足预设检测条件的情况下, 将待检测文本音频对写入训练数据库。

- 30 本说明书提供一种样本生成方法, 实现在样本数据准备阶段, 可以通过拼接的方式得到质量高且满足下游业务使用需求的样本数据, 节省数据准备阶段的资源消耗成本, 并且拼接后写入所述训练数据库中的样本数据的数据量较多, 有效的解决了下游业务样本数据量少、样本数据中音频长短分布不均匀导致语音合成效果不佳的问题, 从而提高下游业务的业务处理效率。

与上述方法实施例相对应, 本说明书还提供了样本生成装置实施例, 图 5

示出了本说明书一实施例提供的一种样本生成装置的结构示意图。如图 5 所示，该装置包括：

获取模块 602，被配置为获取多个文本音频对，其中每个文本音频对中包括文本片段和音频片段；

- 5 计算模块 604，被配置为计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对和所述目标文本音频对对应的拼接文本音频对；

拼接模块 606，被配置为将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对，并对所述待检测文本音频对进行检测；

- 10 写入模块 608，被配置为在所述待检测文本音频对满足预设检测条件的情况下，将所述待检测文本音频对写入训练数据库。

一个可选的实施例中，所述获取模块 602 进一步被配置为：

- 获取目标文本以及所述目标文本对应的音频；对所述音频进行预处理获得目标音频，并将所述目标文本转换为音素序列；将所述音素序列与所述目标音频进行对齐处理，并根据对齐处理结果生成所述多个文本音频对。
- 15

一个可选的实施例中，所述获取模块 602 进一步被配置为：

- 根据对齐处理结果得到音素音频文件，并确定所述音素音频文件的切分位置；按照所述切分位置对所述音素音频文件进行切分，获得多个音素音频对，其中每个音素音频对中包含音素片段和音频片段；基于所述目标文本确定所述多个音素音频对中的每个音素音频对的音素片段对应的文本片段；根据每个音素音频对中音素片段对应的文本片段，以及每个音素音频对中的音频片段生成所述多个文本音频对。
- 20

一个可选的实施例中，所述计算模块 604 进一步被配置为：

- 提取所述多个文本音频对中每个文本音频对的音频片段，并对每个文本音频对的音频片段进行分帧处理，获得每个文本音频对的音频帧集合；基于所述多个文本音频对中每个文本音频对的音频帧集合包含的音频帧，计算每个文本音频对的音频片段的基音频率特征和音频帧特征；根据每个文本音频对的音频片段的所述基音频率特征和所述音频帧特征，确定每个文本音频对的音频片段的所述音频特征。
- 25

- 30 一个可选的实施例中，所述计算模块 604 进一步被配置为：

将所述多个文本音频对中每个文本音频对的音频片段、文本片段和音频特征进行整合，获得每个文本音频对对应的文本音频包，并写入片段数据库；在所述片段数据库中选择任意一个文本音频包作为目标文本音频包，并将所述目

标文本音频包中的文本音频对确定为所述目标文本音频对;基于所述片段数据库中除所述目标文本音频包外的文本音频包和所述音频特征确定拼接文本音频包,并将所述拼接文本音频包中的文本音频对作为所述拼接文本音频对。

一个可选的实施例中,所述计算模块 604 进一步被配置为:

- 5 在所述片段数据库中选择除所述目标文本音频包外的文本音频包组成待筛选文本音频包集合;将所述待筛选文本音频包集合中包含的各个待筛选文本音频包的文本音频对确定为待筛选文本音频对;基于所述目标文本音频对的音频片段的音频特征和所述待筛选文本音频对的音频片段的音频特征,在所述待筛选文本音频包集合中筛选出所述拼接文本音频包。

- 10 一个可选的实施例中,所述计算模块 604 进一步被配置为:

根据所述目标文本音频包确定所述目标文本音频对的音频片段的第一音频特征,以及根据所述待筛选文本音频包确定所述待筛选文本音频对的音频片段的第二音频特征;计算所述第一音频特征和所述第二音频特征之间的特征距离;将所述特征距离小于预设距离阈值的待筛选文本音频对所属的待筛选文本音频包确定为所述拼接文本音频包。

- 15

一个可选的实施例中,所述样本生成装置,还包括:

采样模块,被配置为对所述目标文本音频对中的音频片段进行采样处理获得目标采样信息,以及确定所述目标文本音频对中的文本片段的目标文本信息;判断所述目标采样信息和所述目标文本信息是否满足所述预设检测条件;

- 20 若否,则运行所述拼接模块 606。

一个可选的实施例中,若所述采样模块的判断结果为是,则将所述目标文本音频对写入所述训练数据库。

一个可选的实施例中,所述拼接模块 606 进一步被配置为:

- 25 提取所述目标文本音频对中的目标文本片段和目标音频片段,以及所述拼接文本音频对中的拼接文本片段和拼接音频片段;将所述目标文本片段和所述拼接文本片段拼接为待检测文本片段,以及将所述目标音频片段和所述拼接音频片段拼接为待检测音频片段;基于所述待检测文本片段和所述待检测音频片段组成所述待检测文本音频对。

一个可选的实施例中,所述拼接模块 606 进一步被配置为:

- 30 对所述待检测音频片段进行采样处理获得待检测采样信息,以及确定所述待检测文本片段的待检测文本信息;基于所述预设检测条件对所述待检测采样信息和所述待检测文本信息进行检测;

相应的，所述写入模块 608 进一步被配置为：

在所述待检测采样信息和所述待检测文本信息均满足所述预设检测条件的情况下，将所述待检测文本音频对写入所述训练数据库。

一个可选的实施例中，所述样本生成装置，还包括：

- 5 筛选模块，被配置为在所述待检测文本音频对未满足预设检测条件的情况下，根据所述音频特征在所述多个文本音频对中筛选出所述拼接文本音频对对应的多度拼接文本音频对；将所述多度拼接文本音频对和所述待检测文本音频对拼接为多度待检测文本音频对，并判断所述多度待检测文本音频对是否满足所述预设检测条件；

- 10 若是，则将所述多度待检测文本音频对写入所述训练数据库；

若否，则将所述多度拼接文本音频对作为所述拼接文本音频对，以及将所述多度待检测文本音频对作为所述待检测文本音频对，并运行所述筛选模块。

一个可选的实施例中，所述样本生成装置，还包括：

- 15 训练模块，被配置为在所述训练数据库中提取样本文本音频对，所述样本文本音频对中包含样本文本片段和样本音频片段；基于所述样本文本片段和所述样本音频片段对语音合成模型进行训练，获得目标语音合成模型。

- 20 本实施例提供的样本生成装置，在获取到多个文本音频对后，计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对，以及所述目标文本音频对对应的拼接文本音频对，之后将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对，并对所述待检测文本音频对进行检测，在所述待检测文本音频对满足预设检测条件的情况下，将所述待检测文本音频对写入所述训练数据库，实现在样本数据准备阶段，可以通过拼接的方式得到质量高且满足下游业务使用需求的样本数据，节省数据准备阶段的资源消耗成本，并且拼接后
- 25 写入所述训练数据库中的样本数据的数据量较多，有效的解决了下游业务样本数据量少、样本数据中音频长短分布不均匀导致语音合成效果不佳的问题，从而提高下游业务的业务处理效率。

- 30 上述为本实施例的一种样本生成装置的示意性方案。需要说明的是，该样本生成装置的技术方案与上述的样本生成方法的技术方案属于同一构思，样本生成装置的技术方案未详细描述的细节内容，均可以参见上述样本生成方法的技术方案的描述。

图 7 示出了根据本说明书一实施例提供的一种计算设备 700 的结构框图。该计算设备 700 的部件包括但不限于存储器 710 和处理器 720。处理器 720 与

存储器 710 通过总线 730 相连接，数据库 750 用于保存数据。

5 计算设备 700 还包括接入设备 740，接入设备 740 使得计算设备 700 能够经由一个或多个网络 760 通信。这些网络的示例包括公用交换电话网(PSTN)、局域网(LAN)、广域网(WAN)、个域网(PAN)或诸如因特网的通信网络的组合。接入设备 740 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))中的一个或多个，诸如 IEEE502.11 无线局域网(WLAN)无线接口、全球微波互联接入(Wi-MAX)接口、以太网接口、通用串行总线(USB)接口、蜂窝网络接口、蓝牙接口、近场通信(NFC)接口，等等。

10 在本说明书的一个实施例中，计算设备 700 的上述部件以及图 7 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 7 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。

15 计算设备 700 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、笔记本电脑、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 PC 的静止计算设备。计算设备 700 还可以是移动式或静止式的服务器。

其中，处理器 720 用于执行如下计算机可执行指令：

20 获取多个文本音频对，其中每个文本音频对中包含文本片段和音频片段；
计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对和所述目标文本音频对对应的拼接文本音频对；

将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对，并对所述待检测文本音频对进行检测；

25 在所述待检测文本音频对满足预设检测条件的情况下，将所述待检测文本音频对写入训练数据库。

30 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样本生成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，均可以参见上述样本生成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，其存储有计算机指令，该指令被处理器执行时以用于：

获取多个文本音频对，其中每个文本音频对中包含文本片段和音频片段；

计算所述多个文本音频对中每个文本音频对的音频片段的音频特征,并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对和所述目标文本音频对对应的拼接文本音频对;

5 将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对,并对所述待检测文本音频对进行检测;

在所述待检测文本音频对满足预设检测条件的情况下,将所述待检测文本音频对写入训练数据库。

10 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样本生成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,均可以参见上述样本生成方法的技术方案的描述。

15 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。

20 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,所述计算机可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减,例如在某些司法管辖区,根据立法和专利实践,计算机可读介质不包括电载波信号和电信信号。

25 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,本说明书并不受所描述的动作顺序的限制,因为依据本说明书,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一定都是本说明书所必须的。

30 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

1、一种样本生成方法，其特征在于，包括：

获取多个文本音频对，其中每个文本音频对中包含文本片段和音频片段；

5 计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对和所述目标文本音频对对应的拼接文本音频对；

将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对，并对所述待检测文本音频对进行检测；

10 在所述待检测文本音频对满足预设检测条件的情况下，将所述待检测文本音频对写入训练数据库。

2、根据权利要求1所述的样本生成方法，其特征在于，所述获取多个文本音频对，包括：

获取目标文本以及所述目标文本对应的音频；

对所述音频进行预处理获得目标音频，并将所述目标文本转换为音素序列；

15 将所述音素序列与所述目标音频进行对齐处理，并根据对齐处理结果生成所述多个文本音频对。

3、根据权利要求2所述的样本生成方法，其特征在于，所述根据对齐处理结果生成所述多个文本音频对，包括：

20 根据对齐处理结果得到音素音频文件，并确定所述音素音频文件的切分位置；

按照所述切分位置对所述音素音频文件进行切分，获得多个音素音频对，其中每个音素音频对中包含音素片段和音频片段；

基于所述目标文本确定所述多个音素音频对中的每个音素音频对的音素片段对应的文本片段；

25 根据每个音素音频对中音素片段对应的文本片段，以及每个音素音频对中的音频片段生成所述多个文本音频对。

4、根据权利要求1所述的样本生成方法，其特征在于，所述计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，包括：

30 提取所述多个文本音频对中每个文本音频对的音频片段，并对每个文本音频对的音频片段进行分帧处理，获得每个文本音频对的音频帧集合；

基于所述多个文本音频对中每个文本音频对的音频帧集合包含的音频帧，计算每个文本音频对的音频片段的基音频率特征和音频帧特征；

根据每个文本音频对的音频片段的所述基音频率特征和所述音频帧特征，确定每个文本音频对的音频片段的所述音频特征。

5、根据权利要求1所述的样本生成方法，其特征在于，所述根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对和所述目标文本音频对对应的拼接文本音频对，包括：

将所述多个文本音频对中每个文本音频对的音频片段、文本片段和音频特征进行整合，获得每个文本音频对对应的文本音频包，并写入片段数据库；

在所述片段数据库中选择任意一个文本音频包作为目标文本音频包，并将所述目标文本音频包中的文本音频对确定为所述目标文本音频对；

10 基于所述片段数据库中除所述目标文本音频包外的文本音频包和所述音频特征确定拼接文本音频包，并将所述拼接文本音频包中的文本音频对作为所述拼接文本音频对。

6、根据权利要求5所述的样本生成方法，其特征在于，所述基于所述片段数据库中除所述目标文本音频包外的文本音频包和所述音频特征确定拼接文本音频包，包括：

在所述片段数据库中选择除所述目标文本音频包外的文本音频包组成待筛选文本音频包集合；

将所述待筛选文本音频包集合中包含的各个待筛选文本音频包的文本音频对确定为待筛选文本音频对；

20 基于所述目标文本音频对的音频片段的音频特征和所述待筛选文本音频对的音频片段的音频特征，在所述待筛选文本音频包集合中筛选出所述拼接文本音频包。

7、根据权利要求6所述的样本生成方法，其特征在于，所述基于所述目标文本音频对的音频片段的音频特征和所述待筛选文本音频对的音频片段的音频特征，在所述待筛选文本音频包集合中筛选出所述拼接文本音频包，包括：

根据所述目标文本音频包确定所述目标文本音频对的音频片段的第一音频特征，以及根据所述待筛选文本音频包确定所述待筛选文本音频对的音频片段的第二音频特征；

计算所述第一音频特征和所述第二音频特征之间的特征距离；

30 将所述特征距离小于预设距离阈值的待筛选文本音频对所属的待筛选文本音频包确定为所述拼接文本音频包。

8、根据权利要求1所述的样本生成方法，其特征在于，所述将所述目标

文本音频对和所述拼接文本音频对拼接为待检测文本音频对,并对所述待检测文本音频对进行检测步骤执行之前,还包括:

对所述目标文本音频对中的音频片段进行采样处理获得目标采样信息,以及确定所述目标文本音频对中的文本片段的目标文本信息;

5 判断所述目标采样信息和所述目标文本信息是否满足所述预设检测条件;

若否,则执行将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对,并对所述待检测文本音频对进行检测的步骤。

9、根据权利要求 8 所述的样本生成方法,其特征在于,若所述判断所述采样信息和所述文本信息是否满足所述预设检测条件的判断结果为是,则将所述目标文本音频对写入所述训练数据库。

10、根据权利要求 1 所述的样本生成方法,其特征在于,所述将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对,包括:

提取所述目标文本音频对中的目标文本片段和目标音频片段,以及所述拼接文本音频对中的拼接文本片段和拼接音频片段;

15 将所述目标文本片段和所述拼接文本片段拼接为待检测文本片段,以及将所述目标音频片段和所述拼接音频片段拼接为待检测音频片段;

基于所述待检测文本片段和所述待检测音频片段组成所述待检测文本音频对。

11、根据权利要求 10 所述的样本生成方法,其特征在于,所述对所述待检测文本音频对进行检测,包括:

对所述待检测音频片段进行采样处理获得待检测采样信息,以及确定所述待检测文本片段的待检测文本信息;

基于所述预设检测条件对所述待检测采样信息和所述待检测文本信息进行检测;

25 相应的,所述在所述待检测文本音频对满足预设检测条件的情况下,将所述待检测文本音频对写入训练数据库,包括:

在所述待检测采样信息和所述待检测文本信息均满足所述预设检测条件的情况下,将所述待检测文本音频对写入所述训练数据库。

12、根据权利要求 1 所述的样本生成方法,其特征在于,所述将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对,并对所述待检测文本音频对进行检测步骤执行之后,还包括:

在所述待检测文本音频对未满足预设检测条件的情况下,根据所述音频特

征在所述多个文本音频对中筛选出所述拼接文本音频对对应的多度拼接文本音频对；

将所述多度拼接文本音频对和所述待检测文本音频对拼接为多度待检测文本音频对，并判断所述多度待检测文本音频对是否满足所述预设检测条件；

5 若是，则将所述多度待检测文本音频对写入所述训练数据库；

若否，则将所述多度拼接文本音频对作为所述拼接文本音频对，以及将所述多度待检测文本音频对作为所述待检测文本音频对，并执行所述根据所述音频特征在所述多个文本音频对中筛选出所述拼接文本音频对对应的多度拼接文本音频对步骤。

10 13、根据权利要求1所述的样本生成方法，其特征在于，所述将所述待检测文本音频对写入训练数据库步骤执行之后，还包括：

在所述训练数据库中提取样本文本音频对，所述样本文本音频对中包含样本文本片段和样本音频片段；

15 基于所述样本文本片段和所述样本音频片段对语音合成模型进行训练，获得目标语音合成模型。

14、一种样本生成装置，其特征在于，包括：

获取模块，被配置为获取多个文本音频对，其中每个文本音频对中包含文本片段和音频片段；

20 计算模块，被配置为计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对和所述目标文本音频对对应的拼接文本音频对；

拼接模块，被配置为将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对，并对所述待检测文本音频对进行检测；

25 写入模块，被配置为在所述待检测文本音频对满足预设检测条件的情况下，将所述待检测文本音频对写入训练数据库。

15、一种计算设备，其特征在于，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，以实现下述方法：

30 获取多个文本音频对，其中每个文本音频对中包含文本片段和音频片段；

计算所述多个文本音频对中每个文本音频对的音频片段的音频特征，并根据所述音频特征在所述多个文本音频对中筛选出目标文本音频对和所述目标

文本音频对对应的拼接文本音频对；

将所述目标文本音频对和所述拼接文本音频对拼接为待检测文本音频对，并对所述待检测文本音频对进行检测；

5 在所述待检测文本音频对满足预设检测条件的情况下，将所述待检测文本音频对写入训练数据库。

16、一种计算机可读存储介质，其存储有计算机指令，其特征在于，该指令被处理器执行时实现权利要求 1 至 13 任意一项所述样本生成方法的步骤。