(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105704538 A
(43)申请公布日 2016.06.22

(21)申请号 201610153291.7

(22)申请日 2016.03.17

(71)申请人 广东小天才科技有限公司
地址 523000 广东省东莞市长安镇乌沙步步高大道126号二楼

(72)发明人 王金龙 丁小响

(74)专利代理机构 深圳青年人专利商标代理有限公司 44350
代理人 傅俏梅

(51)Int.Cl.
H04N 21/43(2011.01)
H04N 21/439(2011.01)
H04N 21/81(2011.01)
H04N 21/845(2011.01)
G10L 15/26(2006.01)

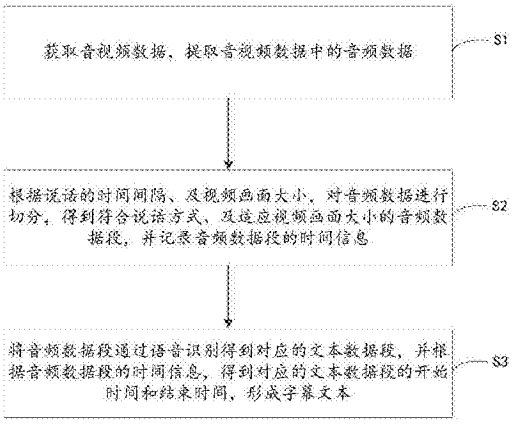
权利要求书1页 说明书4页 附图3页

(54)发明名称

一种音视频字幕生成方法及系统

(57)摘要

本发明适用计算机技术领域，提供了一种音视频字幕生成方法及系统，所述方法包括：获取音视频数据，提取音视频数据中的音频数据；根据说话的时间间隔、及视频画面大小，对音频数据进行切分，得到符合说话方式、及适应视频画面大小的音频数据段，并记录音频数据段的时间信息；将音频数据段通过语音识别得到对应的文本数据段，并根据音频数据段的时间信息，得到对应的文本数据段的开始时间和结束时间，形成字幕文本。本发明摆脱了人工录入字幕的繁杂工作量，实现通过识别音频数据得到文本数据，简单高效地生成完整的字幕。



CN 105704538 A

1.一种音视频字幕生成方法，其特征在于，所述方法包括如下步骤：

获取音视频数据，提取所述音视频数据中的音频数据；

根据说话的时间间隔、及视频画面大小，对所述音频数据进行切分，得到符合说话方式、及适应视频画面大小的音频数据段，并记录所述音频数据段的时间信息；

将所述音频数据段通过语音识别得到对应的文本数据段，并根据所述音频数据段的时间信息，得到对应的文本数据段的开始时间和结束时间，形成字幕文本。

2.如权利要求1所述的方法，其特征在于，所述将所述音频数据段通过语音识别得到对应的文本数据段，包括：

将所述音频数据段与词库进行匹配，得到对应所述音频数据段的分类词库；

根据所匹配的分类词库进行语音识别。

3.如权利要求2所述的方法，其特征在于，所述分类词库包括：两种以上的语种分类词库、及两种以上的专业学科分类词库。

4.如权利要求1所述的方法，其特征在于，所述将所述音频数据段通过语音识别得到对应的文本数据段，还包括：

根据说话的时间间隔的长短，在对应的文本数据段中添加间隔标识符。

5.如权利要求1所述的方法，其特征在于，所述方法还包括：

根据音频数据段的时间信息，将各音频数据段与其对应的文本数据段进行同步，所述时间信息包括各文本数据段的开始时间和结束时间。

6.一种音视频字幕生成系统，其特征在于，所述系统包括：

音频数据提取单元，用于获取音视频数据，提取所述音视频数据中的音频数据；

切分单元，用于根据说话的时间间隔、及视频画面大小，对所述音频数据进行切分，得到符合说话方式、及适应视频画面大小的音频数据段，并记录所述音频数据段的时间信息；

字幕文本形成单元，用于将所述音频数据段通过语音识别得到对应的文本数据段，并根据所述音频数据段的时间信息，得到对应的文本数据段的开始时间和结束时间，形成字幕文本。

7.如权利要求6所述的系统，其特征在于，所述字幕文本形成单元包括：

词库匹配单元，用于将所述音频数据段与词库进行匹配，得到对应所述音频数据段的分类词库；根据所匹配的分类词库进行语音识别。

8.如权利要求7所述的系统，其特征在于，所述分类词库包括：两种以上的语种分类词库、及两种以上的专业学科分类词库。

9.如权利要求6所述的系统，其特征在于，所述字幕文本形成单元还包括：

间隔标识符添加单元，用于根据说话的时间间隔的长短，在对应的文本数据段中添加间隔标识符。

10.如权利要求6所述的系统，其特征在于，所述系统还包括：

同步单元，用于根据音频数据段的时间信息，将各音频数据段与其对应的文本数据段进行同步，所述时间信息包括各文本数据段的开始时间和结束时间。

一种音视频字幕生成方法及系统

技术领域

[0001] 本发明属于计算机技术领域，尤其涉及一种音视频字幕生成方法及系统。

背景技术

[0002] 随着互联网技术的不断发展，音视频以其便捷的访问体验、多样化的影片来源以及实时的更新速度吸引了广大的用户，使得音视频成为了用户生活中不可或缺的重要组成部分。字幕的出现，使音视频以一种更加直观可靠的方式帮助人们了解音视频的内容，越来越多的用户习惯了添加字幕的音视频文件，然而由于字幕的制作较复杂，尤其是使用大段的语音数据与大段的文本数据来生成字幕，对于没有字幕的音视频，用户只能依靠听到的内容进行理解，用户体验效果较差。

[0003] 在无文本稿条件下，现有的音视频字幕的生成方法是通过人工一边看视频听音频，一边录入字幕文本，然后拍好时间轴，最后字幕文本与音视频合成，生成有字幕的音视频，而对于长度较长的音视频文件，需要花费大量的时间成本进行制作，造成音视频字幕生成效率较低，同时依赖人工进行录入，无疑人力成本较大。

发明内容

[0004] 本发明的目的在于提供一种音视频字幕生成方法及系统，旨在解决现有技术中依赖人工进行录入字幕文本，所造成的人力成本较大、字幕生成效率较低的问题。

[0005] 一方面，本发明提供了一种音视频字幕生成方法，所述方法包括下述步骤：

[0006] 获取音视频数据，提取所述音视频数据中的音频数据；

[0007] 根据说话的时间间隔、及视频画面大小，对所述音频数据进行切分，得到符合说话方式、及适应视频画面大小的音频数据段，并记录所述音频数据段的时间信息；

[0008] 将所述音频数据段通过语音识别得到对应的文本数据段，并根据所述音频数据段的时间信息，得到对应的文本数据段的开始时间和结束时间，形成字幕文本。

[0009] 另一方面，本发明提供了一种音视频字幕生成系统，所述系统包括：

[0010] 音频数据提取单元，用于获取音视频数据，提取所述音视频数据中的音频数据；

[0011] 切分单元，用于根据说话的时间间隔、及视频画面大小，对所述音频数据进行切分，得到符合说话方式、及适应视频画面大小的音频数据段，并记录所述音频数据段的时间信息；

[0012] 字幕文本形成单元，用于将所述音频数据段通过语音识别得到对应的文本数据段，并根据所述音频数据段的时间信息，得到对应的文本数据段的开始时间和结束时间，形成字幕文本。

[0013] 在本发明实施例中，根据说话的时间间隔、及视频画面大小对音频数据进行切分，符合人们沟通讲话的语言断句方式，且适应视频画面大小，保证了音视频画面中呈现出的字幕阅读量能够使得观看者感到舒适、方便消化理解字幕内容，同时提高了字幕生成效率，减少大量的人力成本。

附图说明

[0014] 图1是本发明实施例一提供的音视频字幕生成方法的流程图；

[0015] 图2是本发明实施例二提供的音视频字幕生成方法的流程图；

[0016] 图3是本发明实施例三提供的音视频字幕生成系统的结构示意图；以及

[0017] 图4是本发明实施例四提供的音视频字幕生成系统的结构示意图。

具体实施方式

[0018] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0019] 以下结合具体实施例对本发明的具体实现进行详细描述：

[0020] 实施例一：

[0021] 图1示出了本发明实施例一提供的音视频字幕生成方法的流程图，为了便于描述，仅示出了与本发明实施例相关的部分，本发明实施例提供的音视频字幕生成方法，该方法包括如下步骤：

[0022] 步骤S1，获取音视频数据，提取音视频数据中的音频数据。

[0023] 在本实施例中，获取待处理的音视频数据可以是视频文件或者视频流，该视频文件或者视频流的来源包括但不限于：检测到的下载文件、对存储设备进行搜索所发现的视频文件、检测到的视频流(例如：直播视频流、http视频流)。提取音视频数据中的音频数据可以是不经切分处理的音频数据，也可以是经过切分处理后的音频。

[0024] 步骤S2，根据说话的时间间隔、及视频画面大小，对音频数据进行切分，得到符合说话方式、及适应视频画面大小的音频数据段，并记录音频数据段的时间信息。

[0025] 在本实施例中，根据说话的时间间隔对音频数据进行切分是根据音频数据中音频的波形图通过语音识别来判断应该断句位置。为了达到断句的精确性，可以设置停顿时间间隔、每段语音的时间间隔，使得在音频的波形图比较紧密的情况下能够精确断句。由于人声的语速快慢不同，有一般语速、较快语速以及较慢语速，为了进一步的实现断句的精确性，可以根据音频数据中人声的语速分别设置停顿时间间隔、每段语音的时间间隔。其中，对音频数据进行切分以得到适应视频画面大小的音频数据段保证了音视频画面中呈现出的字幕阅读量能够使得观看者感到舒适、方便消化理解字幕内容。

[0026] 步骤S3，将音频数据段通过语音识别得到对应的文本数据段，并根据音频数据段的时间信息，得到对应的文本数据段的开始时间和结束时间，形成字幕文本。

[0027] 在本实施例中，音频数据段通过语音识别得到文本数据段，可以是根据每段文本数据段的开始时间和结束时间将文本数据进行分割和换行，形成音频数据的字幕文本。具体地，将文本数据进行分割和换行的标准主要依据音视频中字幕与音频的配合。需要说明的是，生成音视频数据的字幕文本后，可以根据实际情况选择字幕文本的输出方式，字幕文本的输出方式包括但不限于：生成特定格式、符合字幕格式标准的字幕文本；在播放视频时，将字幕文本整合到音视频输出流中，让播放器去做字幕显示工作。

[0028] 在本实施例中，将音频数据段通过语音识别得到对应的文本数据段，包括：将所述

音频数据段与词库进行匹配，得到对应音频数据段的分类词库；根据所匹配的分类词库进行语音识别。该分类词库包括：两种以上的语种分类词库、及两种以上的专业学科分类词库。通过将音频数据段与词库进行匹配可以得到与音频数据中原声语种对应语种分类词库，并可以利用该语种分类词库中的词汇进一步加快语音识别得到对应的文本数据、还可以通过将音频数据段与词库进行匹配得到与音频数据中的专业学科对应专业学科分类词库，例如历史题材的音频数据可以匹配到历史专业学科分类词库，可利用该专业学科分类词库中的词汇进一步加快语音识别得到对应的文本数据。

[0029] 具体地，将音频数据段通过语音识别得到对应的文本数据段可以是将音频数据段中的音频内容直接识别成原声对应语言的文本数据，当然，也可将音频数据段中的音频内容识别成其它语言的文字。将音频数据段中的音频内容识别成其它语言的文字的具体过程为：获取用户选择的语言类别，将音频数据段识别成原声对应语言的文本数据，然后将识别出的原声对应语言的文本数据翻译成用户所选择的用户选择的语言类别的文本数据。

[0030] 在本实施例中，根据说话的时间间隔的长短，在对应的文本数据段中添加间隔标识符。由于通过语音识别得到文本数据段中包含了大量的标点符号，其中很多标点符号不符合上下文的语境，为了方便进一步校对文本数据段，可对语音识别得到文本数据段进行过滤，将文本数据段中标点符号所占字节转换成对应字节的间隔标识符。以方便人工校对时，修改成符合语境的标点符号。

[0031] 实施例二：

[0032] 图2示出了本发明实施例二提供的音视频字幕生成方法的流程图，详述如下：

[0033] 步骤S1，获取音视频数据，提取音视频数据中的音频数据。

[0034] 步骤S2，根据说话的时间间隔、及视频画面大小，对音频数据进行切分，得到符合说话方式、及适应视频画面大小的音频数据段，并记录音频数据段的时间信息。

[0035] 步骤S3，将音频数据段通过语音识别得到对应的文本数据段，并根据音频数据段的时间信息，得到对应的文本数据段的开始时间和结束时间，形成字幕文本。

[0036] 步骤S4，根据音频数据段的时间信息，将各音频数据段与其对应的文本数据段进行同步，时间信息包括各文本数据段的开始时间和结束时间。

[0037] 在本实施例中，为了提高字幕的同步精准度，将各音频数据段与其对应的文本数据段进行同步，可以是逐句进行同步将识别后的文本数据段依据开始时间和结束时间的时间戳生成字幕显示文本，按照一句时间戳加一句字幕的格式写入字幕文本。

[0038] 实施例三：

[0039] 图3示出了本发明实施例三提供的音视频字幕生成系统的结构示意图，为了便于描述，仅示出了与本发明实施例相关的部分，本发明实施例提供的音视频字幕生成系统，该系统包括：音频数据提取单元31，切分单元32，以及字幕文本形成单元33。

[0040] 具体地，音频数据提取单元31用于获取音视频数据，提取所述音视频数据中的音频数据；

[0041] 切分单元32用于根据说话的时间间隔、及视频画面大小，对所述音频数据进行切分，得到符合说话方式、及适应视频画面大小的音频数据段，并记录所述音频数据段的时间信息；以及

[0042] 字幕文本形成单元33用于将所述音频数据段通过语音识别得到对应的文本数据

段，并根据所述音频数据段的时间信息，得到对应的文本数据段的开始时间和结束时间，形成字幕文本。

[0043] 其中，字幕文本形成单元33包括：词库匹配单元331、及间隔标识符添加单元332。

[0044] 具体地，词库匹配单元331，用于将所述音频数据段与词库进行匹配，得到对应所述音频数据段的分类词库；根据所匹配的分类词库进行语音识别。

[0045] 该分类词库包括：两种以上的语种分类词库、及两种以上的专业学科分类词库。间隔标识符添加单元332，用于根据说话的时间间隔的长短，在对应的文本数据段中添加间隔标识符。

[0046] 实施例四：

[0047] 图4示出了本发明实施例四提供的音视频字幕生成系统的结构示意图，为了便于描述，仅示出了与本发明实施例相关的部分，本发明实施例提供的音视频字幕生成系统，该系统包括：音频数据提取单元31，切分单元32，字幕文本形成单元33，以及同步单元34。

[0048] 具体地，音频数据提取单元31用于获取音视频数据，提取所述音视频数据中的音频数据；

[0049] 切分单元32用于根据说话的时间间隔、及视频画面大小，对所述音频数据进行切分，得到符合说话方式、及适应视频画面大小的音频数据段，并记录所述音频数据段的时间信息；

[0050] 字幕文本形成单元33用于将所述音频数据段通过语音识别得到对应的文本数据段，并根据所述音频数据段的时间信息，得到对应的文本数据段的开始时间和结束时间，形成字幕文本；以及

[0051] 同步单元34用于根据音频数据段的时间信息，将各音频数据段与其对应的文本数据段进行同步，所述时间信息包括各文本数据段的开始时间和结束时间。

[0052] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，所述的程序可以存储于一计算机可读取存储介质中，所述的存储介质，如ROM/RAM、磁盘、光盘等。

[0053] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本发明的保护范围之内。

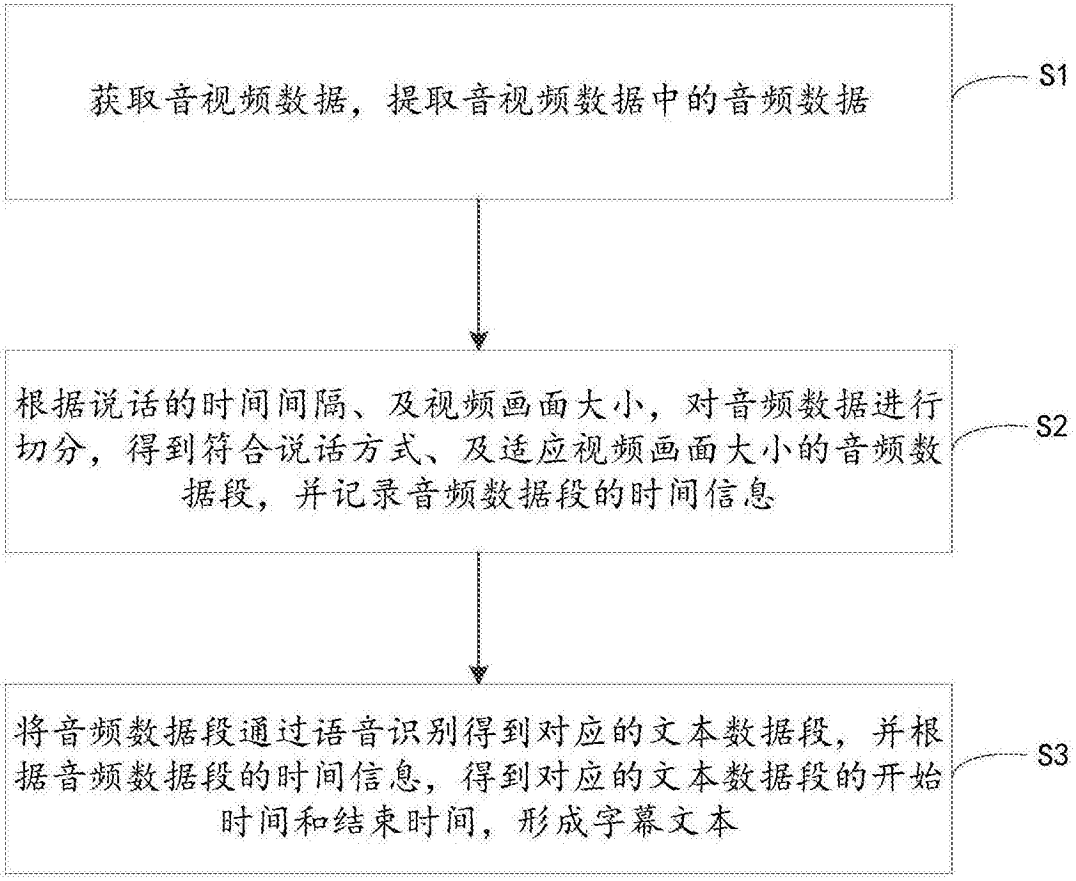

获取音视频数据，提取音视频数据中的音频数据
S1
根据说话的时间间隔、及视频画面大小，对音频数据进行切分，得到符合说话方式、及适应视频画面大小的音频数据段，并记录音频数据段的时间信息
S2
将音频数据段通过语音识别得到对应的文本数据段，并根据音频数据段的时间信息，得到对应的文本数据段的开始时间和结束时间，形成字幕文本
S3

图1

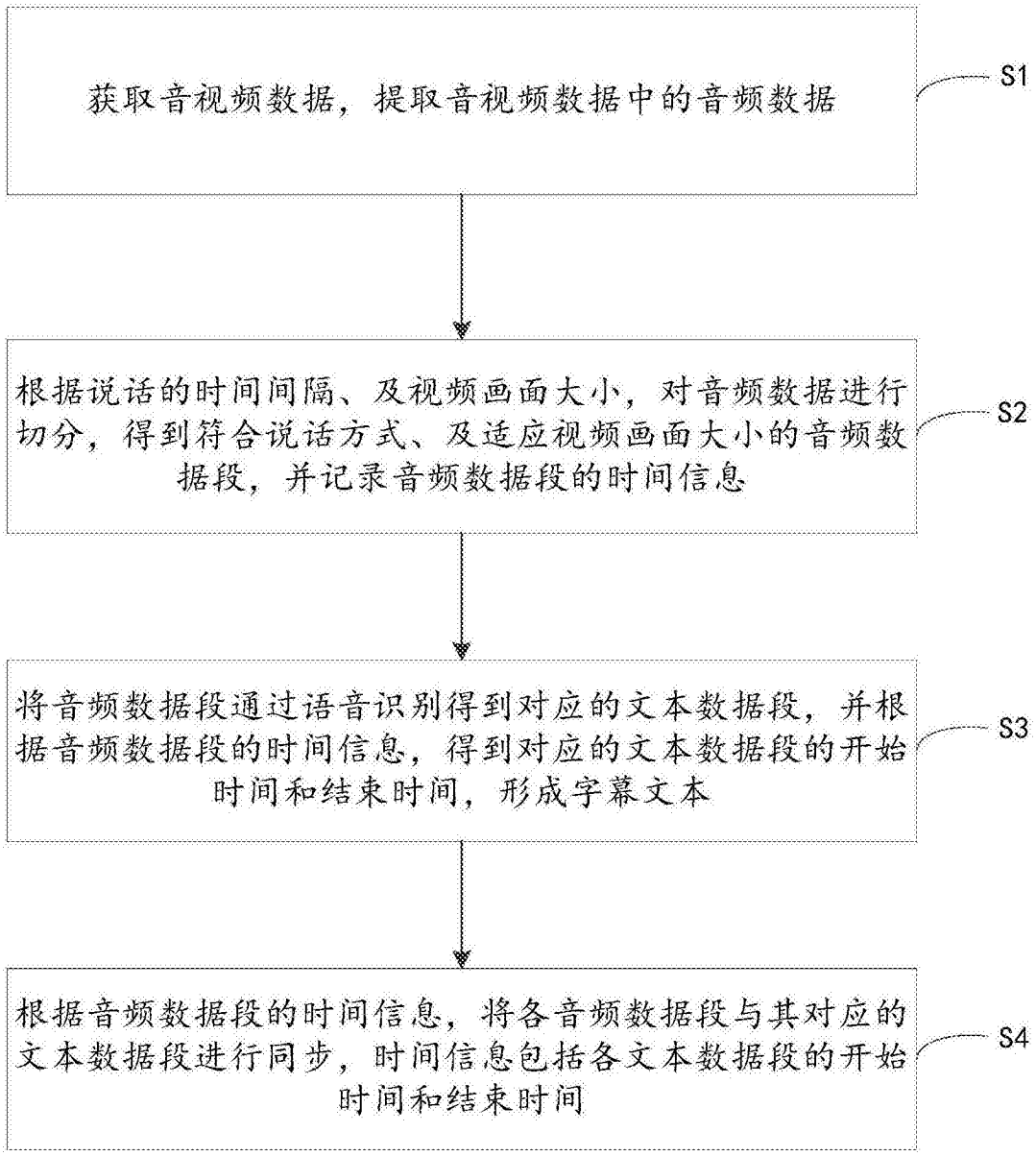

获取音视频数据，提取音视频数据中的音频数据
S1
根据说话的时间间隔、及视频画面大小，对音频数据进行切分，得到符合说话方式、及适应视频画面大小的音频数据段，并记录音频数据段的时间信息
S2
将音频数据段通过语音识别得到对应的文本数据段，并根据音频数据段的时间信息，得到对应的文本数据段的开始时间和结束时间，形成字幕文本
S3
根据音频数据段的时间信息，将各音频数据段与其对应的文本数据段进行同步，时间信息包括各文本数据段的开始时间和结束时间
S4

图2

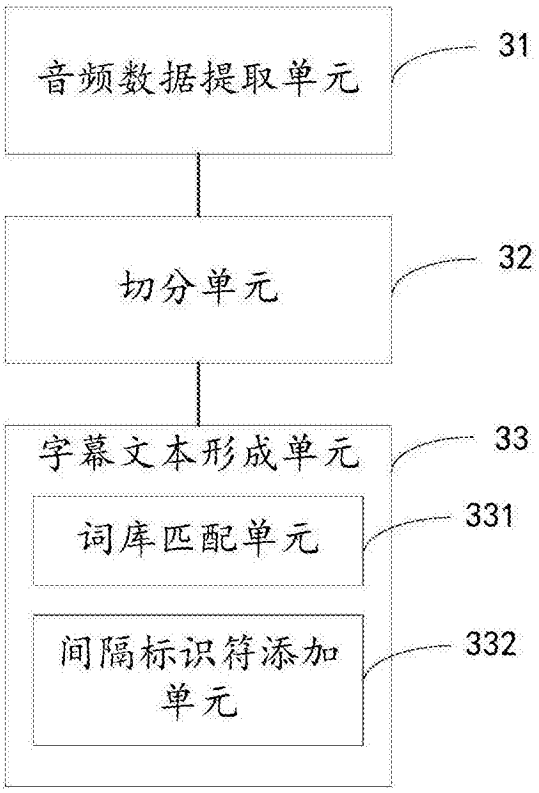

音频数据提取单元
31
切分单元
32
字幕文本形成单元
33
词库匹配单元
331
间隔标识符添加单元
332

图3

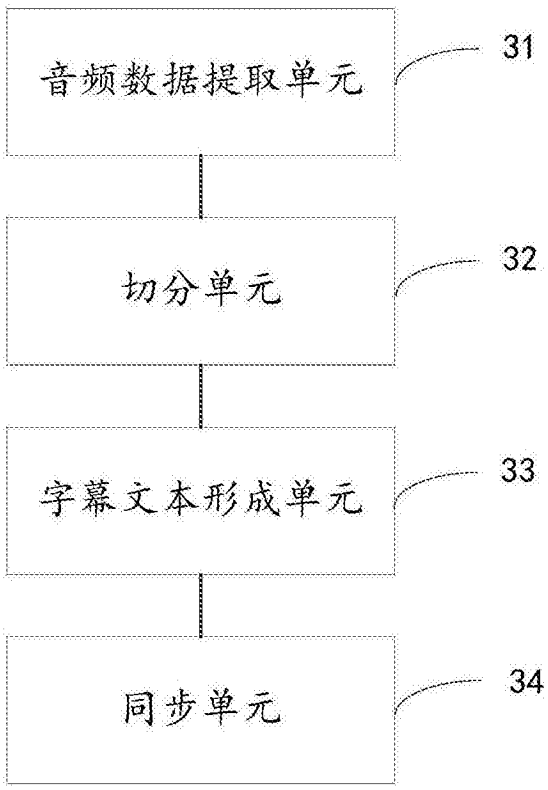

音频数据提取单元
31
切分单元
32
字幕文本形成单元
33
同步单元
34

图4