(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2020 年 8 月 6 日 (06.08.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/155750 A1

(51) 国际专利分类号:
G06F 16/783 (2019.01)

(21) 国际申请号: PCT/CN2019/117261

(22) 国际申请日: 2019 年 11 月 11 日 (11.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910081793.7 2019年1月28日 (28.01.2019) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(72) 发明人: 杨雨晨 (YANG, Yvchen); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市世联合知识产权代理有限公司 (SL INTELLECTUAL PROPERTY CO., LTD.); 中国广东省深圳市福田区泰然六路泰然科技园205栋二层203-206, Guangdong 518000 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,

WO 2020/155750 A1

(54) **Title:** ARTIFICIAL INTELLIGENCE-BASED CORPUS COLLECTING METHOD, APPARATUS, DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 基于人工智能的语料收集方法、装置、设备及存储介质

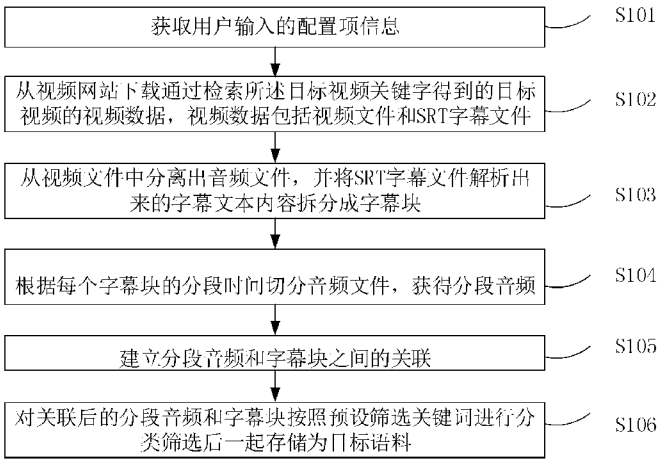


图 1

S101 Acquire configuration item information inputted by user
S102 Download from video website video data of target video produced by searching for target video keyword, video data comprising video file and SRT subtitle file
S103 Separate audio file from video file and split subtitle text content parsed from SRT subtitle file into subtitle blocks
S104 Segment audio file on the basis of segment time of each subtitle block to acquire segmented audios
S105 Establish associations between segmented audios and subtitle blocks
S106 Sort and screen associated segmented audios and subtitle blocks according to preset screening keyword and then jointly store as target corpus

(57) **Abstract:** An artificial intelligence-based corpus collecting method, apparatus, and device, and a storage medium, related to the technical field of natural language processing. The method comprises: acquiring configuration item information inputted by a user (S101); downloading from a video website video data of a target video produced by searching for a target video keyword, the video data comprising a video file and an SRT subtitle file (S102); separating an audio file from the video file, splitting a subtitle text content parsed from the SRT subtitle file into subtitle blocks (S103); segmenting the audio file on the basis of a segment time of each subtitle block

[见续页]

PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

to acquire segmented audios (S104); establishing associations between the segmented audios and the subtitle blocks (S105); sorting and screening the associated segmented audios and subtitle blocks according to a preset screening keyword and then jointly storing as a target corpus (S106). The method implements the goal of automatically and quickly collecting a corpus satisfying requirements of a certain type of scenarios and is highly efficient and inexpensive.

(57) 摘要：一种基于人工智能的语料收集方法、装置、设备及存储介质，属于自然语言处理技术领域。该方法包括获取用户输入的配置项信息（S101）；从视频网站下载通过检索所述目标视频关键字得到的目标视频的视频数据，视频数据包括视频文件和SRT字幕文件（S102）；从视频文件中分离出音频文件，并将SRT字幕文件解析出来的字幕文本内容拆分成字幕块（S103）；根据每个字幕块的分段时间切分音频文件，获得分段音频（S104）；建立分段音频和字幕块之间的关联（S105）；对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料（S106）。该方法能够实现自动快速收集到符合某一类场景需要的语料的目的，效率高而且成本低。

基于人工智能的语料收集方法、装置、设备及存储介质

本申请以 2019 年 1 月 28 日提交的申请号为 201910081793.7，名称为“基于人工智能的语料收集方法、装置、设备及存储介质”的中国发明专利申请为基础，并要求其优先权。

技术领域

本申请属于自然语言处理技术领域，涉及基于人工智能的语料收集方法、装置、设备及存储介质。

背景技术

人工智能（Artificial Intelligence，AI），是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。人工智能是计算机科学的一个分支，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器，该领域的研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等。

现实中，基于 AI 的自然语言处理过程中往往需要预先收集符合各类场景的语料，比如说在客服投诉自动处理平台应用中：需要关于表现“投诉”“不满”以及时间紧急程度的语料，以便于根据紧急、严重程度可以灵活调整工单接入的优先级以及处理对接人，以帮助快速处理投诉解决问题；又比如说，在幼儿早教以及幼儿趣味对话软件应用中，需要基于幼儿的声音而且情绪比较欢快活泼的语料。

现有收集某一场景的语料方式主要有：（1）通过免费的资源搜索获得某一场景的语料，这种方式获取到的语料非常有限，难以满足需求；（2）通过团队自己录音并进行标注而获得某一场景的语料,这种办法效率低，极其耗费人力；（3）通过渠道购买某一场景的语料，这种方式成本较高。

因此，现有语料收集的方法效率低而且成本很高，怎样快速收集到符合某一类场景需要的语料也成为迫切需要解决的问题。

发明内容

本申请实施例公开了一种能够快速收集到符合某一种场景的语料的基于人工智能的语料收集方法、装置、设备及存储介质。

本申请的一些实施例公开了一种基于人工智能的语料收集方法，包括：获取用户输入的配置项信息，所述配置项信息包括目标视频关键字和视频网站，所述视频网站为视频网站的

网址或视频网站的名称；从所述视频网站下载通过检索所述目标视频关键字得到的目标视频的视频数据，所述视频数据包括视频文件和 SRT 字幕文件；从视频文件中分离出音频文件，并将 SRT 字幕文件解析出来的字幕文本内容拆分成字幕块；根据每个字幕块的分段时间切分音频文件，获得分段音频；建立分段音频和字幕块之间的关联；对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料。

本申请的一些实施例还公开了一种基于人工智能的语料收集装置，包括：配置项信息获取模块，用于获取用户输入的配置项信息，所述配置项信息包括目标视频关键字和视频网站，所述视频网站为视频网站的网址或视频网站的名称；视频数据下载模块，用于从所述视频网站下载通过检索所述目标视频关键字得到的目标视频的视频数据，所述视频数据包括视频文件和 SRT 字幕文件；音频字幕处理模块，用于从视频文件中分离出音频文件，并将 SRT 字幕文件解析出来的字幕文本内容拆分成字幕块；音频切分模块，用于根据每个字幕块的分段时间切分音频文件，获得分段音频；音频字幕块关联模块，用于建立分段音频和字幕块之间的关联；筛选模块，用于对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料。

本申请的一些实施例还公开了一种计算机设备，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述处理器执行所述计算机可读指令时，实现上述基于人工智能的语料收集方法的步骤。

本申请的一些实施例还公开了一种非易失性可读存储介质，所述非易失性可读存储介质上存储有计算机可读指令，所述计算机可读指令被处理器执行时，实现上述基于人工智能的语料收集方法的步骤。

本申请的一个或多个实施例的细节在下面的附图和描述中提出，本申请的其他特征和优点将从说明书、附图以及权利要求变得明显。

附图说明

为了更清楚地说明本申请实施例的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其它的附图。

图 1 为本申请实施例提供的基于人工智能的语料收集方法的流程图；

图 2 为图 1 中步骤 S106 的第二种具体实施方式的流程图；

图 3 为图 1 中步骤 S106 的第三种具体实施方式的流程图；

图 4 为图 3 中步骤 S405 的一种具体实施方式的流程示意图；

图 5 为本申请实施例提供的基于人工智能的语料收集装置的示意图；

图 6 为图 5 中的音频字幕处理模块的示意图；

图 7 为图 5 中的筛选模块的第二实施例的结构示意图；

图 8 为图 5 中的筛选模块的第三实施例的结构示意图；

图 9 为图 8 中的语音状态参数分值计算模块的结构示意图；

图 10 本申请实施例中计算机设备 100 基本结构框图。

具体实施方式

为了便于理解本申请，下面将参照相关附图对本申请进行更全面的描述。附图中给出了本申请的较佳实施例。但是，本申请可以以许多不同的形式来实现，并不限于本文所描述的实施例。相反地，提供这些实施例的目的是使对本申请的公开内容的理解更加透彻全面。

除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同。本文中在本申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本申请。

本申请实施例提供了一种基于人工智能的语料收集方法。参阅图 1，为本申请实施例提供的基于人工智能的语料收集的示意图；如图 1 中所示意的，所述基于人工智能的语料收集的方法包括：

S101.获取用户输入的配置项信息，所述配置项信息包括目标视频关键字和视频网站，所述视频网站为视频网站的网址或视频网站的名称。

其中，目标视频关键字包括表示视频名称或者视频类型的关键字；视频网站可以为视频网站的名称例如爱奇艺、优酷，也可以为视频网站的网址例如爱奇艺、优酷的网址。

S102.从所述视频网站下载通过检索所述目标视频关键字得到的目标视频的视频数据，所述视频数据包括视频文件和 SRT 字幕文件。

具体的，本申请实施例中的视频数据包括有字幕文件和具有音视频信号的视频文件，视频数据可以是例如电影、电视剧、综艺、新闻、动画、歌曲等视频数据，也可以是涉及具体内容例如消费维权、投诉、点餐对话、具体的动画片内容等的视频数据。

网络爬虫（又被称为网页蜘蛛、网络机器人），是一种按照一定的规则，自动地抓取万维网信息的计算机可读指令或者脚本。具体的，下载视频数据的实施方式可以包括：

第一种方式，网络爬虫根据用户输入的视频网址找到该视频网址，打开包含该目标视频的网页，自动下载目标视频。例如，当需要下载电影《当幸福来敲门》，用户可以预先设置包含有该电影的视频网址，网络爬虫根据用户输入的视频网址找到该视频网址，打开包含该

目标视频的网页，自动下载目标视频。

第二种方式，获取用户输入的视频网站，视频网站可以为视频网站的名称例如爱奇艺、优酷，也可以为视频网站的网址例如爱奇艺、优酷的网址。当获取的视频网站为视频网站的名称例如爱奇艺、优酷等，网络爬虫在预设的搜索引擎像百度等网页中输入视频网站的名称检索到视频网站的网址，打开视频网站例如爱奇艺，在爱奇艺、优酷等视频网站的搜索框中输入目标视频关键字搜索到目标视频，网络爬虫根据搜索的结果，依次打开搜索出来的网页将所有的视频均下载下来。当获取的视频网站为视频网站的网址例如爱奇艺、优酷的网址时，网络爬虫打开对应的视频网站，并在视频网站的搜索框中输入目标视频关键字搜索到目标视频，网络爬虫根据搜索的结果，依次打开搜索出来的网页将所有的视频均下载下来。其中的目标视频关键字可以为例如“熊出没”等动画片的名称，也可以为例如“做饭”等表示视频内容的关键字。

目标视频关键字的设置：现实中，会有客服自动投诉处理平台，当我们知道某些电影属于情绪比较愤慨的电影，可以预先设置关键字为该电影的名称例如“XXX”，也可以设置用于筛选需要下载的视频资源的类型的关键字，比如说，我们知道某一类节目（像调解节目，关于售后维权等节目）里面有很多关于抱怨、气愤、不满的内容，可以设置该类节目的名称例如“消费主张”为目标视频关键字；在一些场景中，气氛是比较欢快活泼的，例如说关于幼儿早教方面也涉及很多语音识别等技术，可以设置用于筛选需要下载的视频资源的类型的关键字，比如说，我们知道某一类节目像动画节目里面多是属于幼儿看的节目，是很适合早教的内容，可以设置关键字为“动画+幼儿”。

进一步的，为了表明需要的是视频格式的资源可以在目标视频关键字中加上“视频”，例如“消费主张+视频”、“动画+幼儿+视频”等限定搜索的为视频资源。

以上只是列举，不用于限制本申请。

S103.从视频文件中分离出音频文件，并将 SRT 字幕文件解析出来的字幕文本内容切分成字幕块。

具体的，步骤 S103 中包括两个分步骤：从视频文件中分离出音频文件，具体的，通过视频音频分离技术将视频文件中的音频分离出来，得到单独的音频文件。将 SRT 字幕文件解析出来的字幕文本内容切分成字幕块。

步骤 S103 中的两个分步骤属于并列关系,不分时间先后。

具体的，在本实施例中，解析 SRT 字幕文件可以得到如下所示的字幕文本内容；

1

00:00:00,162 --> 00:00:01,875

从现在开始

2

00:00:02,800 --> 00:00:03,000

我只疼你，宠你，不会骗你

3

00:00:06,560 --> 00:00:11,520

没有人能打你，骂你，欺负你，有人欺负你我会第一时间出来帮你

其中，“1”、“2”、 “3”表示字幕的序号，例如“1”代表该音频信号中出现的第 1 个字幕，“2” 代表该音频信号中出现的第 2 个字幕，“3” 代表该音频信号中出现的第 3 个字幕；音频信号中主要包括有字幕的部分和没有字幕的空白部分，每一个字幕都对应两个时间，第一个“时间”(“-->”左边的时间)表示音频信号中字幕出现的开始时间，第二个“时间”(“-->”右边的时间）表示字幕结束的时间，从开始时间到结束时间为字幕的播放时间。例如“00:00:00,162”为音频信号中第 1 个字幕出现的开始时间，“00:00:01,875”为第 1 个字幕结束的时间，“00:00:00,162--〉00:00:01,875”为第 1 个字幕的字幕内容“从现在开始”的播放时间。“从现在开始”为第 1 个字幕的字幕内容，“我只疼你，宠你，不会骗你” 为第 2 个字幕的字幕内容，“没有人能打你，骂你，欺负你，有人欺负你我会第一时间出来帮你”为第 3 个字幕的字幕内容。

具体的，本实施例中，结合播放时间和断句符对字幕文本内容进行分块，得到字幕块；例如“从现在开始”拆分为一个字幕块，“我只疼你，宠你，不会骗你”拆分为一个字幕块，“没有人能打你，骂你，欺负你，有人欺负你我会第一时间出来帮你” 拆分为一个字幕块。

S104. 根据每个字幕块的分段时间切分音频文件，获得分段音频。

在解析得到的字幕文本中，每一个字幕都对应两个时间，第一个“时间”表示音频信号中字幕出现的开始时间，第二个“时间”表示字幕结束的时间，从开始时间到结束时间为字幕的播放时间。由于字幕块是按照字幕的播放时间拆分的，因此由字幕的播放时间可以得到每个字幕块的开始时间到结束时间，再按照每一个字幕块的开始时间到结束时间将音频文件进行切分，比如说切分成“00:00:00,162--〉00:00:01,875”、“00:00:02,800--〉00:00:03,000”、“00:00:06,560--〉00:00:11,520”等等一段一段的分段音频，最后拆分出来的分段音频与字幕块一一对应。

S105. 建立分段音频和字幕块之间的关联。

将分段音频和字幕块进行关联，例如时间段为“00:00:00,162--〉00:00:01,875”的分段音频与分段字幕“从现在开始”关联起来。关联起来的分段音频和字幕块可以存储在一个

指定的文件夹地址也可以分开存储，但是两者的文件名称需要一致。

S106.对关联后的字幕块分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料。

在本申请的实施例中，获取用户输入的配置项信息，从所述视频网站下载目标视频的视频数据；然后处理所述视频数据，从视频文件中分离出音频文件，并将 SRT 字幕文件解析出来的字幕文本内容切分成字幕块；根据每个字幕块的分段时间切分音频；将分段音频和字幕块关联起来；对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料，实现了快速自动收集到所需要的符合某一类场景例如符合预设筛选关键词的语料的目的，效率高而且成本低。

具体的，步骤 S106 的第一种具体实施方式中，所述对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料的步骤包括:分析每个字幕块是否包含有与预设筛选关键词相匹配的文本；将包含有相匹配的文本的字幕块连同与该字幕块相关联的分段音频一起存储到指定的第一位置。

具体的，本申请实施例中通过预设筛选关键词分类筛选的方式帮助筛选需要的语料。通常人在愤怒的状态下可能会使用更多的侮辱性的词汇；在高兴的时候，可能会使用一些积极向上的词汇。因此，如果需要收集愤怒情绪的语料，预设筛选关键词可以为“太过分”、“我很生气”或者表示骂人的“贱人”、“傻子”等等，如果需要收集积极向上情绪的语料，预设筛选关键词可以为“天天向上”、“奋斗”、“加油”等等。

抓取每个字幕块的文本与预设的筛选关键词进行比对，确认所述字幕块是否包含有与预设筛选关键词相匹配的文本，其中，筛选关键词匹配的方式可以为模糊匹配。存储下来的相关联的分段音频和字幕块为目标语料。

请参阅图 2，图 2 为图 1 中步骤 S106 的第二种具体实施方式的流程图；

具体的，在一些实施例中，所述对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料的步骤包括:S301.分析每个字幕块是否包含有与预设筛选关键词相匹配的文本；S302.将包含有相匹配的文本的字幕块连同与该字幕块相关联的分段音频一起存储到指定的第一位置。

具体的，本申请实施例中通过预设筛选关键词分类筛选的方式帮助筛选需要的语料。通常人在愤怒的状态下可能会使用更多的侮辱性的词汇；在高兴的时候，可能会使用一些积极向上的词汇。因此，如果需要收集愤怒情绪的语料，预设筛选关键词可以为“太过分”、“我很生气”或者表示骂人的“贱人”、“傻子”等等，如果需要收集积极向上情绪的语料，预设筛选关键词可以为“天天向上”、“奋斗”、“加油”等等。

抓取每个字幕块的文本与预设的筛选关键词进行比对，确认所述字幕块是否包含有与预设筛选关键词相匹配的文本，其中，筛选关键词匹配的方式可以为模糊匹配。存储下来的相关联的分段音频和字幕块为目标语料。

S303.判断前述存储在第一位置的各分段音频的每一个语音状态参数是否在预设标准区间；S304.挑选出所有的语音状态参数均在预设标准区间的分段音频连同与该分段音频相关联的字幕块一起存储到指定的第二位置。

具体的，本实施例中的筛选配置项信息除了包括用于筛选的筛选关键词之外，还包括用于辅助分析分段音频的情绪类别的语音状态参数，语音状态参数可以包括音量、频率、振幅、语速以及语调。

所述挑选出所有的语音状态参数均在预设标准区间的分段音频连同与该分段音频相关联的字幕块一起存储到指定的第二位置的步骤之前，还包括：预先设置每一个语音状态参数的标准区间。

具体的，预先设置每一个语音状态参数的标准区间的步骤包括：获取标记有目标情绪类别的语料样本进行统计分析，得到目标情绪类别下每一个语音状态参数所占的概率均大于预设值的语音状态参数的范围；从所述范围内提取一个包含于所述范围内的区间作为预设标准区间。其中，语料样本可以由人工采集好自己认为符合自己想要的某一类情绪的样本，也可以为获取到的已有的由其他方式收集的样本；包含于所述范围内的区间可以为与所述范围相同的区间也可以为所述范围以内的区间，例如所述范围为 50～70，包含于所述范围内的区间可以为 50～70，也可以为 50～60、55～65、60～70 等等。

更具体的，在本实施例中，例如说关于频率这个语音状态参数我们找到一个标记有目标情绪类别（例如愤怒）的语料样本库，测试每一个语料样本的频率值，并画出频率的概率正态分布图，发现所有的语料样本中频率在范围 50～70Hz 内的样本在所有的语料样本所占的概率均大于预设值（例如 97%），就可以得到目标情绪类别下频率所占的概率均大于预设值的语音状态参数的范围，同样的，采用该方法可以得到目标情绪类别下每一个语音状态参数所占的概率均大于预设值的语音状态参数的范围；可以将该范围区间作为预设标准区间，也可以在区间 50～70Hz 内挑选一个小区间例如 50～60 Hz、55～65 Hz、60～70 Hz 作为预设标准区间，其他的语音状态参数例如音量、振幅、语速以及语调同样处理。

所有的语音状态参数均在预设标准区间的分段音频，即分段音频的五个语音状态参数均在各自对应的预设标准区间。

请参阅图 3，图 3 为图 1 中步骤 S106 的第三种具体实施方式的流程图；

步骤 S302 将包含有相匹配的文本的字幕块连同与所述字幕块相关联的分段音频一起存

储到指定的第一位置之后，还包括：S405.计算出存储在所述第一位置的各分段音频的每一个语音状态参数的分值。S406.将同一分段音频中所有的语音状态参数的分值进行求和运算，确认总分值是否达到预设的阈值。其中，预设阈值可以由经验或者需求自行设置，例如说 80 分、90 分等。S407.将总分值达到预设阈值的分段音频连同与该分段音频相关联的字幕块一起存储到指定的第三位置。

具体的，本实施例中的筛选配置项信息除了包括用于筛选的筛选关键词之外，还包括用于辅助分析分段音频的情绪的语音状态参数，语音状态参数包括音量、频率、振幅、语速以及语调。

接着请参阅图 4，图 4 为图 3 中步骤 S405 的一种具体实施方式的流程示意图；更具体的，计算出存储在所述第一位置的各分段音频的每一个语音状态参数的分值的步骤包括：

S501. 获取标记有目标情绪类别的语料样本进行统计分析，得到目标情绪类别下每一个语音状态参数所占的概率均大于预设值的语音状态参数的范围。

其中，语料样本可以由人工采集好自己认为符合自己想要的某一类情绪的样本，也可以为获取到的已有的由其他方式收集的语料样本。在本实施例中，例如说关于频率这个语音状态参数我们找到一个标记有目标情绪类别（例如愤怒）的语料样本库，测试每一个语料样本的频率值，并画出频率的概率正态分布图，发现频率在范围 50～70Hz 内的样本在所有的语料样本中所占的概率均大于预设值（例如 97%），就可以得到目标情绪类别下频率所占的概率均大于预设值的语音状态参数的范围，同样的，采用该方法就可以得到目标情绪类别下每一个语音状态参数所占的概率均大于预设值的语音状态参数的范围。

S502.在该范围内挑选一个数值例如中值作为预设的语音状态参数标准值。

其中，频率标准值用 $W_{频率}$ 表示，$W_{音量}$、$W_{振幅}$、$W_{语速}$ 和 $W_{语调}$ 分别表示其他语音状态参数的预设标准值。

S503. 测试所述存储在第一位置的各分段音频的每一个语音状态参数值。

S504.基于预设的语音状态参数标准值、测试的语音状态参数值以及接收到的权重值，按以下公式计算每一个语音状态参数的分值：

$M_i=100*S_i*(X_i/W_i)$；其中，M_i 为每一个语音状态参数的分值，S_i 为每一个语音状态参数的权重值，X_i 为测试的语音状态参数值，W_i 为预设的语音状态参数标准值，$_i$ 代表语音状态参数，具体可以为音量、频率、振幅、语速和语调。

具体的，将所述存储在第一位置的各分段音频实际测试出来的语音状态参数具体值 X_i 与预设的语音状态参数标准值 W_i 进行比较，得出的数值称为相似度 P_i，即 $P_i=X_i/W_i$；

例如，将所述存储在第一位置的各分段音频实际测试出来的频率具体值与预设的频率标

准值$W_{频率}$进行比较，得出频率相似度$P_{频率}$，频率具体值用$X_{频率}$表示，$X_{音量}$、$X_{振幅}$、$X_{语速}$、$X_{语调}$分别表示其他语音状态参数的具体值，$P_{音量}$、$P_{振幅}$、$P_{语速}$和$P_{语调}$分别表示其他语音状态参数的相似度，具体的公式如下：

$P_{音量}=X_{音量}/W_{音量}$，$P_{频率}=X_{频率}/W_{频率}$，$P_{振幅}=X_{振幅}/W_{振幅}$，$P_{语速}=X_{语速}/W_{语速}$，$P_{语调}=X_{语调}/W_{语调}$。

接收预设的每一个语音状态参数的权重值。权重值用S_i表示，每个语音状态参数的权重值分别为$S_{音量}$、$S_{频率}$、$S_{振幅}$、$S_{语速}$和$S_{语调}$；预先给每一个语音状态参数设置一个权重值，例如人在愤怒的时候，声音明显会响亮很多，因此音量的权重就比较大，可以设为60%。

由$M_i=100*S_i*P_i$进一步得出公式$M_i=100*S_i*(X_i/W_i)$，参考公式$M_i=100*S_i*(X_i/W_i)$计算每一个语音状态参数的分值。具体的，参考如下公式：

$M_{音量}=100*S_{音量}*(X_{音量}/W_{音量})$，$M_{频率}=100*S_{频率}*(X_{频率}/W_{频率})$，$M_{振幅}=100*S_{振幅}*(X_{振幅}/W_{振幅})$，$M_{语速}=100*S_{语速}*(X_{语速}/W_{语速})$，$M_{语调}=100*S_{语调}*(X_{语调}/W_{语调})$

其中，$M_{音量}$、$M_{频率}$、$M_{振幅}$、$M_{语速}$以及$M_{语调}$分别表示每一个语音状态参数的分值；$S_{音量}$、$S_{频率}$、$S_{振幅}$、$S_{语速}$和$S_{语调}$分别表示每一个语音状态参数的权重值；

因此，可以计算出每一个语音状态参数的分值。

反过来继续参阅图3，S406.将同一分段音频中所有的语音状态参数的分值进行求和运算，确认总分值是否达到预设的阈值。具体的，用M表示分段音频的总分值，根据公式$M=M_{音量}+M_{频率}+M_{振幅}+M_{语速}+M_{语调}$得出同一分段音频的总分值。

将同一分段音频的总分值与预设的阈值进行比较，确认总分值是否达到预设的阈值。

S407.将总分值达到预设阈值的分段音频连同与该分段音频相关联的字幕块一起存储到指定的第三位置。具体的，如果总分值大于或者等于预设阈值，则将分段音频连同与该分段音频相关联的字幕块一起存储到指定的第三位置。

本申请实施例带来的好处是实现了自动快速收集到所需要的符合某一类场景的语料的目的，效率高而且成本低，设置了多种语音状态参数，通过对获取标记有情绪类别的语料样本进行统计分析，得出一个范围，选出范围内的区间作为预设标准区间或者选出范围内的某一个具体值作为预设标准值，对分段音频进行测试、计算打分，挑选出的目标语料的情绪更加符合标准。

本申请实施例提供了一种基于人工智能的语料收集装置，参阅图 5，图 5 为本申请所述基于人工智能的语料收集装置的第一种实施例的结构示意图；

所述基于人工智能的语料收集装置包括：配置项信息获取模块 51、视频数据下载模块 52、

音频字幕处理模块 53、音频切分模块 54、音频字幕块关联模块 55 和筛选模块 56。配置项信息获取模块 51，用于获取用户输入的配置项信息；其中，所述配置项信息包括目标视频关键字和视频网站，所述视频网站为视频网站的网址或视频网站的名称。视频数据下载模块 52，用于从所述视频网站下载通过检索所述目标视频关键字得到的目标视频的视频数据，所述视频数据包括视频文件和 SRT 字幕文件；音频字幕处理模块 53，用于从视频文件中分离出音频文件，并将 SRT 字幕文件解析出来的字幕文本内容拆分成字幕块；音频切分模块 54，用于根据每个字幕块的分段时间切分音频文件，获得分段音频；音频字幕块关联模块 55，用于建立分段音频和字幕块之间的关联；筛选模块 56，用于对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料。

参阅图 6，图 6 为图 5 中的音频字幕处理模块的结构示意图；具体的，在本申请实施例中，音频字幕处理模块 53 包括：字幕拆分模块 531，用于解析 SRT 字幕文件得到字幕文本内容；结合播放时间和断句符对字幕文本内容进行分块，得到字幕块；音频视频分离模块 532，用于从视频文件中分离出音频文件。

参阅图 7，图 7 为图 5 中的筛选模块的第一实施例的结构示意图；具体的，在一些实施例中，筛选模块 56 包括：关键词匹配模块 561，用于分析每个字幕块是否包含有与预设筛选关键词相匹配的文本；第一存储模块 562，用于将包含有相匹配的文本的字幕块连同与改字幕块相关联的分段音频一起存储到指定的第一位置。

进一步的，在另一些实施例中，筛选模块 56 除了包括关键词匹配模块 561 和第一存储模块 562 之外还包括：语音状态参数判断模块 563，用于判断所述存储在第一位置的分段音频的每一个语音状态参数是否在预设的标准区间；其中，所述语音状态参数包含于预设的筛选配置项信息，用于辅助分析分段音频的情绪；第二存储模块 564，用于挑选出所有的语音状态参数均在预设的标准区间的分段音频连同与该分段音频相关联的字幕块一起存储到指定的第二位置。

参阅图 8，图 8 为图 5 中的筛选模块的第二实施例的结构示意图；具体的，在另一些实施例中，筛选模块 56 除了包括关键词匹配模块 561 和第一存储模块 562 之外还包括：语音状态参数分值计算模块 565，用于计算出存储在所述第一位置的各分段音频的每个语音状态参数的分值；总分值计算判断模块 566，用于将同一分段音频中所有的语音状态参数的分值进行求和运算，并确认总分值是否达到预设的阈值；第三存储模块 567，用于将总分值达到预设阈值的分段音频连同与该分段音频相关联的字幕块一起存储到指定的第三位置。

参阅图 9，图 9 为图 8 中的语音状态参数分值计算模块的结构示意图；具体的，语音状态参数分值计算模块 565 包括：范围分析模块 5651，用于获取标记有目标情绪类别的语料样

本进行统计分析，得到目标情绪类别下每一个语音状态参数所占的概率均大于预设值的语音状态参数的范围；其中，语料样本可以由人工采集好自己认为符合自己想要的某一类目标情绪类别的样本，也可以为获取到的已有的由其他方式收集的样本。在本实施例中，例如说关于频率这个语音状态参数我们找到一个标记有目标情绪类别（例如愤怒）的语料样本库，测试每一个语料样本的频率值，并画出频率的概率正态分布图，发现频率在范围 50～70Hz 内的样本在所有的语料样本中所占的概率均大于预设值（例如 97%），就可以得到目标情绪类别下频率所占的概率均大于预设值的语音状态参数的范围，同样的，采用该方法就可以得到目标情绪类别下每一个语音状态参数所占的概率均大于预设值的语音状态参数的范围。标准值设置模块 5652，用于在所述范围内挑选一个数值作为预设的语音状态参数标准值；其中，频率标准值用 $W_{频率}$ 表示， $W_{音量}$ 、 $W_{振幅}$ 、 $W_{语速}$ 和 $W_{语调}$ 分别表示其他语音状态参数的预设标准值，作为预设标准值所挑选的数值可以为中值或者范围内任何一个数值,例如,在本实施例中,挑选频率中值 60Hz 作为频率的预设标准值。测试值模块 5653，用于测试所述存储在第一位置的各分段音频的每一个语音状态参数值。分值计算模块 5654，用于基于预设的语音状态参数标准值、和测试的语音状态参数值以及接收到的权重值，按以下公式计算每一个语音状态参数的分值：

$M_i=100*S_i*(X_i/W_i)$；其中，M_i 为每一个语音状态参数的分值，S_i 为每一个语音状态参数的权重值，X_i 为测试的语音状态参数值，W_i 为预设的语音状态参数标准值，i 代表语音状态参数，具体可以为音量、频率、振幅、语速和语调。

本申请的一实施例公开了一种计算机设备。具体请参阅图 10，为本申请的一实施例中计算机设备 100 基本结构框图。如图 10 中所示意的，所述计算机设备 100 包括通过系统总线相互通信连接存储器 101、处理器 102、网络接口 103。需要指出的是，图 10 中仅示出了具有组件 101-103 的计算机设备 100，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。其中，本技术领域技术人员可以理解，这里的计算机设备是一种能够按照事先设定或存储的指令，自动进行数值计算和/或信息处理的设备，其硬件包括但不限于微处理器、专用集成电路(Application Specific Integrated Circuit，ASIC)、可编程门阵列(Field－Programmable Gate Array，FPGA)、数字处理器 (Digital Signal Processor，DSP)、嵌入式设备等。

所述计算机设备可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述计算机设备可以与用户通过键盘、鼠标、遥控器、触摸板或声控设备等方式进行人机交互。

所述存储器 101 至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、

多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、随机访问存储器（RAM）、静态随机访问存储器（SRAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、可编程只读存储器（PROM）、磁性存储器、磁盘、光盘等。在一些实施例中，所述存储器 101 可以是所述计算机设备 100 的内部存储单元，例如该计算机设备 100 的硬盘或内存。在另一些实施例中，所述存储器 101 也可以是所述计算机设备 100 的外部存储设备，例如该计算机设备 100 上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。当然，所述存储器 101 还可以既包括所述计算机设备 100 的内部存储单元也包括其外部存储设备。本实施例中，所述存储器 101 通常用于存储安装于所述计算机设备 100 的操作系统和各类应用软件，例如上述基于人工智能的语料收集方法等。此外，所述存储器 101 还可以用于暂时地存储已经输出或者将要输出的各类数据。

所述处理器 102 在一些实施例中可以是中央处理器（Central Processing Unit，CPU）、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器 102 通常用于控制所述计算机设备 100 的总体操作。本实施例中，所述处理器 102 用于运行所述存储器 101 中存储的计算机可读指令或者处理数据，例如运行上述基于人工智能的语料收集方法的计算机可读指令。

所述网络接口 103 可包括无线网络接口或有线网络接口，该网络接口 103 通常用于在所述计算机设备 100 与其他电子设备之间建立通信连接。

本申请还提供了另一种实施方式，即提供一种非易失性可读存储介质，所述非易失性可读存储介质存储有单据信息录入流程，所述单据信息录入流程可被至少一个处理器执行，以使所述至少一个处理器执行上述任意一种基于人工智能的语料收集方法的步骤。

最后应说明的是，显然以上所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例，附图中给出了本申请的较佳实施例，但并不限制本申请的专利范围。本申请可以以许多不同的形式来实现，相反地，提供这些实施例的目的是使对本申请的公开内容的理解更加透彻全面。尽管参照前述实施例对本申请进行了详细的说明，对于本领域的技术人员来而言，其依然可以对前述各具体实施方式所记载的技术方案进行修改，或者对其中部分技术特征进行等效替换。凡是利用本申请说明书及附图内容所做的等效结构，直接或间接运用在其他相关的技术领域，均同理在本申请专利保护范围之内。

权 利 要 求 书

1. 一种基于人工智能的语料收集方法，其特征在于，包括：

获取用户输入的配置项信息，所述配置项信息包括目标视频关键字和视频网站，所述视频网站为视频网站的网址或视频网站的名称；

从所述视频网站下载通过检索所述目标视频关键字得到的目标视频的视频数据，所述视频数据包括视频文件和 SRT 字幕文件；

从所述视频文件中分离出音频文件，并将所述 SRT 字幕文件解析出来的字幕文本内容拆分成字幕块；

根据每个所述字幕块的分段时间切分所述音频文件，获得分段音频；

建立所述分段音频和所述字幕块之间的关联；

对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料。

2. 根据权利要求 1 所述的基于人工智能的语料收集方法，其特征在于，所述对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料的步骤具体包括：

分析每个字幕块是否包含有与预设筛选关键词相匹配的文本；

将包含有相匹配的文本的字幕块连同与所述字幕块相关联的分段音频一起存储到指定的第一位置。

3. 根据权利要求 2 所述的基于人工智能的语料收集方法，其特征在于，所述将包含有相匹配的文本的字幕块连同与所述字幕块相关联的分段音频一起存储到指定的第一位置的步骤之后，还包括：

判断存储在第一位置的各分段音频的每一个语音状态参数是否在预设的标准区间；

挑选出所有的语音状态参数均在预设标准区间的分段音频连同与所述分段音频相关联的字幕块一起存储到指定的第二位置。

4. 根据权利要求 3 所述的基于人工智能的语料收集方法，其特征在于，所述预设标准区间的设置方法，具体包括：

获取标记有目标情绪类别的语料样本进行统计分析，得到目标情绪类别下每一个语音状态参数所占的概率均大于预设值的语音状态参数的范围；

从所述范围内提取一个包含于所述参考范围内的区间作为预设标准区间。

5. 根据权利要求 2 所述的基于人工智能的语料收集方法，其特征在于，所述将包含有相

匹配的文本的字幕块连同与所述字幕块相关联的分段音频一起存储到指定的第一位置的步骤之后还包括：

计算出存储在所述第一位置的各分段音频的每一个语音状态参数的分值；

将同一分段音频中所有的语音状态参数的分值进行求和运算，确认总分值是否达到预设的阈值；

将总分值达到预设阈值的分段音频连同与所述分段音频相关联的字幕块一起存储到指定的第三位置。

6. 根据权利要求 5 所述的基于人工智能的语料收集方法，其特征在于，所述计算出存储在所述第一位置的各分段音频的每一个语音状态参数的分值的步骤具体包括：

获取标记有目标情绪类别的语料样本进行统计分析，得到目标情绪类别下每一个语音状态参数所占的概率均大于预设值的语音状态参数的范围；

在所述范围内挑选一个数值作为预设的语音状态参数标准值；

测试所述存储在第一位置的各分段音频的每一个语音状态参数值；

基于预设的语音状态参数标准值、测试的语音状态参数值以及接收到的权重值，按以下公式计算每一个语音状态参数的分值：

$M_i=100*S_i*(X_i/W_i)$；其中，M_i为每一个语音状态参数的分值，S_i为每一个语音状态参数的权重值，X_i为测试的语音状态参数值，W_i为预设的语音状态参数标准值，i代表语音状态参数。

7. 根据权利要求 4 或 6 所述的基于人工智能的语料收集方法，其特征在于，所述获取标记有目标情绪类别的语料样本进行统计分析的步骤包括：

获取标记有目标情绪类别的语料样本库；

测试所述语料样本库中每一个语料样本的频率值，以获取关于所述频率值的概率正太分布图；

基于所述概率正太分布图进行统计分析。

8. 根据权利要求 1 至 7 任一项所述的基于人工智能的语料收集方法，其特征在于，所述将 SRT 字幕文件解析出来的字幕文本内容拆分成字幕块的步骤具体包括：

解析 SRT 字幕文件得到字幕文本内容；

结合播放时间和断句符对字幕文本内容进行分块，得到字幕块。

9. 一种基于人工智能的语料收集装置，其特征在于，包括：

配置项信息获取模块，用于获取用户输入的配置项信息，所述配置项信息包括目标视频关键字和视频网站，所述视频网站为视频网站的网址或视频网站的名称；

视频数据下载模块，用于从所述视频网站下载通过检索所述目标视频关键字得到的目标视频的视频数据，所述视频数据包括视频文件和 SRT 字幕文件；

音频字幕处理模块，用于从视频文件中分离出音频文件，并将 SRT 字幕文件解析出来的字幕文本内容拆分成字幕块；

音频切分模块，用于根据每个字幕块的分段时间切分音频文件，获得分段音频；

音频字幕块关联模块，用于建立分段音频和字幕块之间的关联；

筛选模块，用于对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料。

10. 根据权利要求 9 所述的基于人工智能的语料收集装置，其特征在于，所述音频字幕处理模块包括：

字幕拆分模块，用于解析 SRT 字幕文件得到字幕文本内容；结合播放时间和断句符对字幕文本内容进行分块，得到字幕块；音频视频分离模块 532，用于从视频文件中分离出音频文件。

11. 一种计算机设备，包括存储器和处理器，其特征在于，所述存储器中存储有计算机可读指令，所述处理器执行所述计算机可读指令时，实现如下基于人工智能的语料收集方法的步骤：

获取用户输入的配置项信息，所述配置项信息包括目标视频关键字和视频网站，所述视频网站为视频网站的网址或视频网站的名称；

从所述视频网站下载通过检索所述目标视频关键字得到的目标视频的视频数据，所述视频数据包括视频文件和 SRT 字幕文件；

从所述视频文件中分离出音频文件，并将所述 SRT 字幕文件解析出来的字幕文本内容拆分成字幕块；

根据每个所述字幕块的分段时间切分所述音频文件，获得分段音频；

建立所述分段音频和所述字幕块之间的关联；

对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料。

12. 根据权利要求 11 所述的计算机设备，其特征在于，所述对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料的步骤具体包括：

分析每个字幕块是否包含有与预设筛选关键词相匹配的文本；

将包含有相匹配的文本的字幕块连同与所述字幕块相关联的分段音频一起存储到指定的第一位置。

13. 根据权利要求 12 所述的计算机设备，其特征在于，所述将包含有相匹配的文本的字幕块连同与所述字幕块相关联的分段音频一起存储到指定的第一位置的步骤之后，还包括：

判断存储在第一位置的各分段音频的每一个语音状态参数是否在预设的标准区间；

挑选出所有的语音状态参数均在预设标准区间的分段音频连同与所述分段音频相关联的字幕块一起存储到指定的第二位置。

14. 根据权利要求 13 所述的计算机设备，其特征在于，所述预设标准区间的设置方法，具体包括：获取标记有目标情绪类别的语料样本进行统计分析，得到目标情绪类别下每一个语音状态参数所占的概率均大于预设值的语音状态参数的范围；

从所述范围内提取一个包含于所述参考范围内的区间作为预设标准区间。

15. 根据权利要求 12 所述的计算机设备，其特征在于，所述将包含有相匹配的文本的字幕块连同与所述字幕块相关联的分段音频一起存储到指定的第一位置的步骤之后还包括：

计算出存储在所述第一位置的各分段音频的每一个语音状态参数的分值；

将同一分段音频中所有的语音状态参数的分值进行求和运算，确认总分值是否达到预设的阈值；

将总分值达到预设阈值的分段音频连同与所述分段音频相关联的字幕块一起存储到指定的第三位置。

16. 一个或多个非易失性可读存储介质，其特征在于，所述非易失性可读存储介质上存储有计算机可读指令，所述计算机可读指令被一种处理器执行时，实现如下基于人工智能的语料收集方法的步骤：获取用户输入的配置项信息，所述配置项信息包括目标视频关键字和视频网站，所述视频网站为视频网站的网址或视频网站的名称；

从所述视频网站下载通过检索所述目标视频关键字得到的目标视频的视频数据，所述视频数据包括视频文件和 SRT 字幕文件；

从所述视频文件中分离出音频文件，并将所述 SRT 字幕文件解析出来的字幕文本内容拆分成字幕块；

根据每个所述字幕块的分段时间切分所述音频文件，获得分段音频；

建立所述分段音频和所述字幕块之间的关联；

对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料。

17. 根据权利要求 16 所述的非易失性可读存储介质，其特征在于，所述对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料的步骤具体包括：

分析每个字幕块是否包含有与预设筛选关键词相匹配的文本；

将包含有相匹配的文本的字幕块连同与所述字幕块相关联的分段音频一起存储到指定的第一位置。

18. 根据权利要求 17 所述的非易失性可读存储介质，其特征在于，所述将包含有相匹配的文本的字幕块连同与所述字幕块相关联的分段音频一起存储到指定的第一位置的步骤之后，还包括：

判断存储在第一位置的各分段音频的每一个语音状态参数是否在预设的标准区间；

挑选出所有的语音状态参数均在预设标准区间的分段音频连同与所述分段音频相关联的字幕块一起存储到指定的第二位置。

19. 根据权利要求 18 所述的非易失性可读存储介质，其特征在于，所述预设标准区间的设置方法，具体包括：获取标记有目标情绪类别的语料样本进行统计分析，得到目标情绪类别下每一个语音状态参数所占的概率均大于预设值的语音状态参数的范围；

从所述范围内提取一个包含于所述参考范围内的区间作为预设标准区间。

20. 根据权利要求 17 所述的非易失性可读存储介质，其特征在于，所述将包含有相匹配的文本的字幕块连同与所述字幕块相关联的分段音频一起存储到指定的第一位置的步骤之后还包括：计算出存储在所述第一位置的各分段音频的每一个语音状态参数的分值；

将同一分段音频中所有的语音状态参数的分值进行求和运算，确认总分值是否达到预设的阈值；

将总分值达到预设阈值的分段音频连同与所述分段音频相关联的字幕块一起存储到指定的第三位置。

21. 根据权利要求 19 所述的非易失性可读存储介质，其特征在于，所述计算出存储在所述第一位置的各分段音频的每一个语音状态参数的分值的步骤具体包括：

获取标记有目标情绪类别的语料样本进行统计分析，得到目标情绪类别下每一个语音状态参数所占的概率均大于预设值的语音状态参数的范围；

在所述范围内挑选一个数值作为预设的语音状态参数标准值；

测试所述存储在第一位置的各分段音频的每一个语音状态参数值；

基于预设的语音状态参数标准值、测试的语音状态参数值以及接收到的权重值，按以下公式计算每一个语音状态参数的分值：

$M_i=100*S_i*(X_i/W_i)$；其中，M_i 为每一个语音状态参数的分值，S_i 为每一个语音状态参数的权重值，X_i 为测试的语音状态参数值，W_i 为预设的语音状态参数标准值，i 代表语音状态参数。

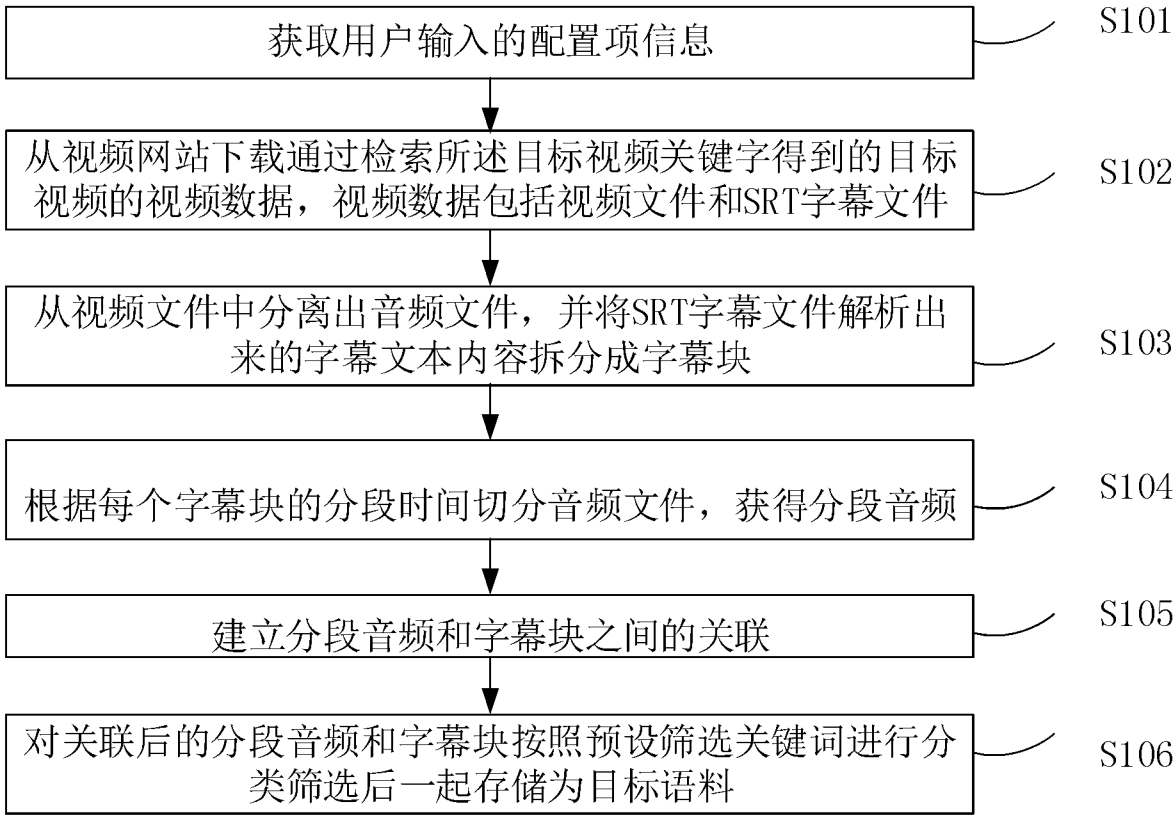

获取用户输入的配置项信息
S101
从视频网站下载通过检索所述目标视频关键字得到的目标视频的视频数据，视频数据包括视频文件和SRT字幕文件
S102
从视频文件中分离出音频文件，并将SRT字幕文件解析出来的字幕文本内容拆分成字幕块
S103
根据每个字幕块的分段时间切分音频文件，获得分段音频
S104
建立分段音频和字幕块之间的关联
S105
对关联后的分段音频和字幕块按照预设筛选关键词进行分类筛选后一起存储为目标语料
S106

图 1

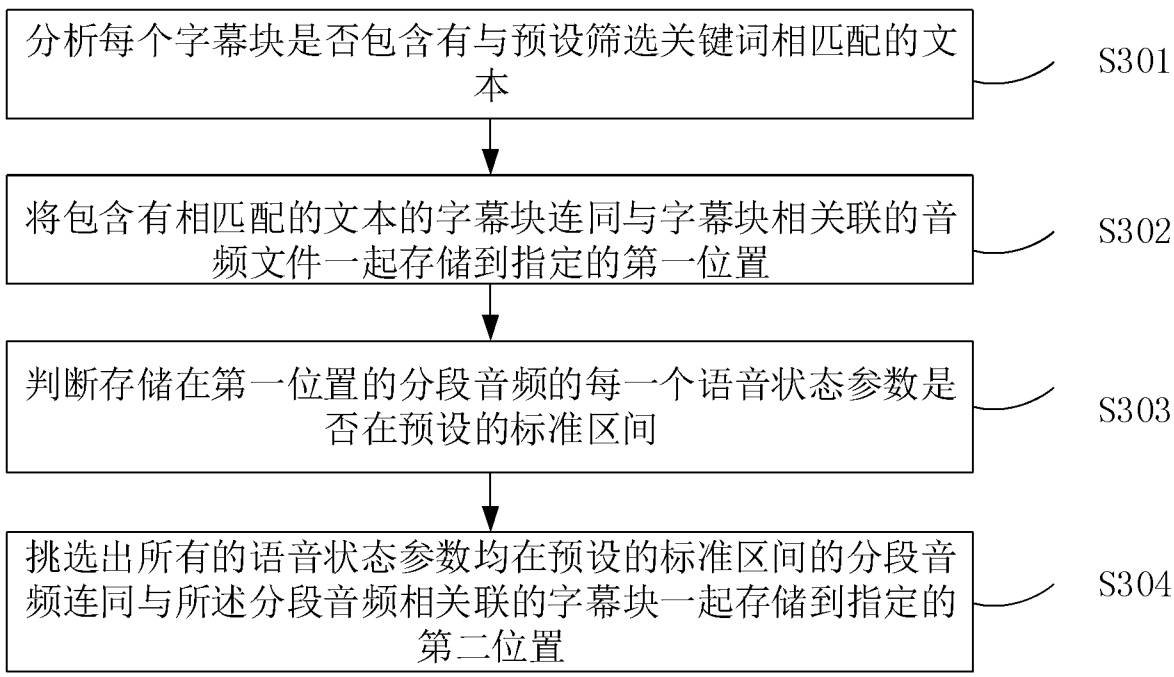

分析每个字幕块是否包含有与预设筛选关键词相匹配的文本
S301
将包含有相匹配的文本的字幕块连同与字幕块相关联的音频文件一起存储到指定的第一位置
S302
判断存储在第一位置的分段音频的每一个语音状态参数是否在预设的标准区间
S303
挑选出所有的语音状态参数均在预设的标准区间的分段音频连同与所述分段音频相关联的字幕块一起存储到指定的第二位置
S304

图 2

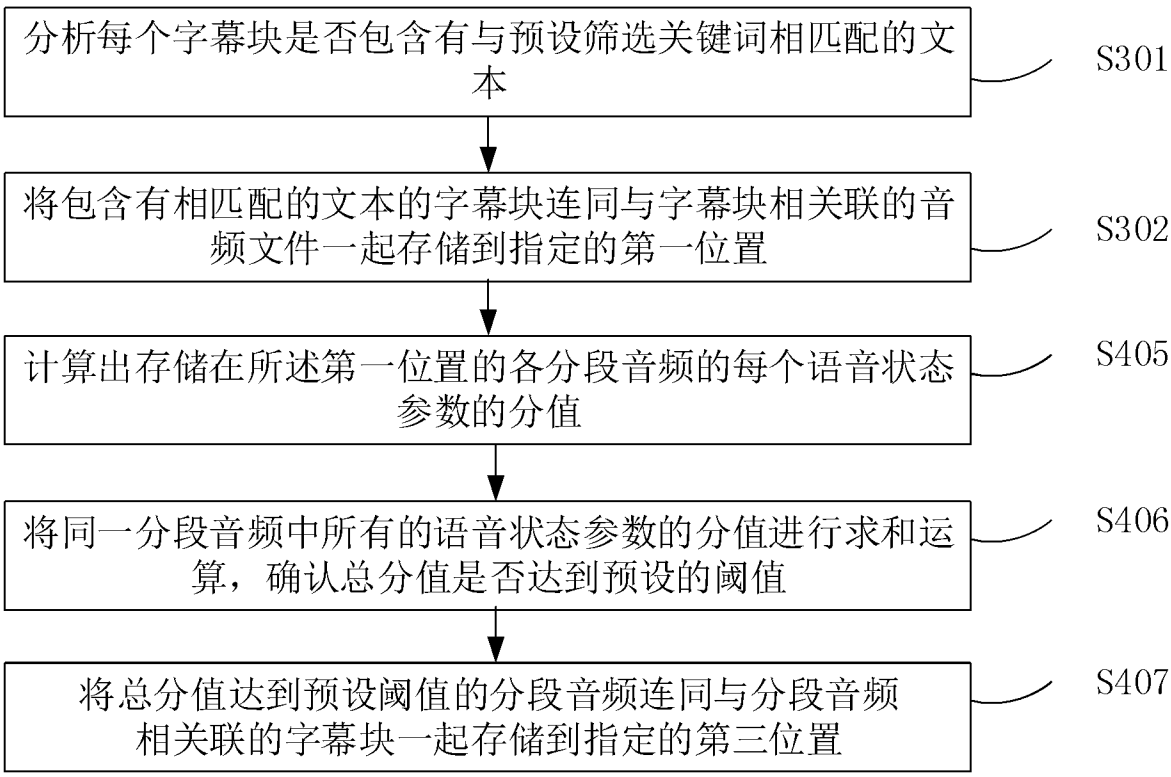

分析每个字幕块是否包含有与预设筛选关键词相匹配的文本
S301
将包含有相匹配的文本的字幕块连同与字幕块相关联的音频文件一起存储到指定的第一位置
S302
计算出存储在所述第一位置的各分段音频的每个语音状态参数的分值
S405
将同一分段音频中所有的语音状态参数的分值进行求和运算，确认总分值是否达到预设的阈值
S406
将总分值达到预设阈值的分段音频连同与分段音频相关联的字幕块一起存储到指定的第三位置
S407

图 3

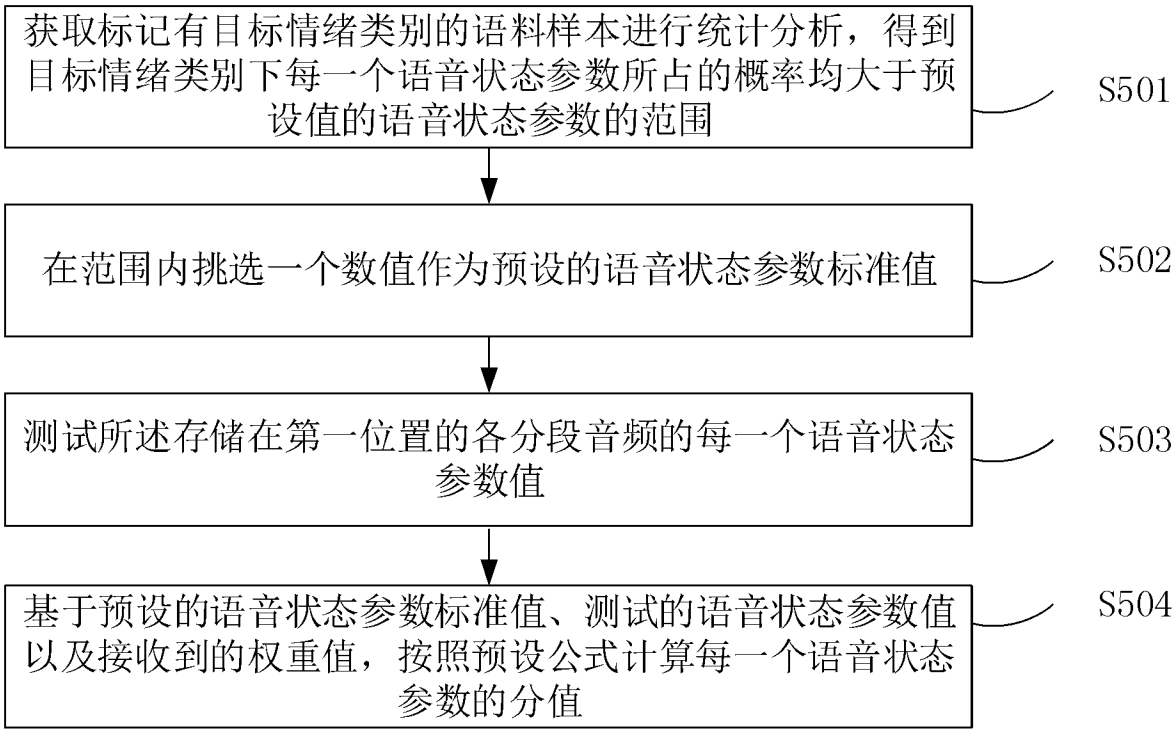

获取标记有目标情绪类别的语料样本进行统计分析，得到目标情绪类别下每一个语音状态参数所占的概率均大于预设值的语音状态参数的范围
S501
在范围内挑选一个数值作为预设的语音状态参数标准值
S502
测试所述存储在第一位置的各分段音频的每一个语音状态参数值
S503
基于预设的语音状态参数标准值、测试的语音状态参数值以及接收到的权重值，按照预设公式计算每一个语音状态参数的分值
S504

图 4

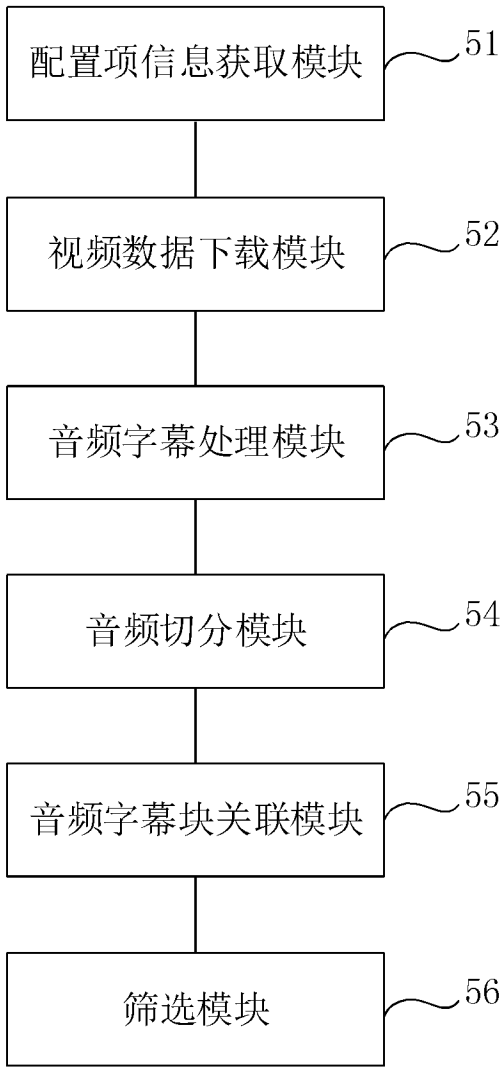

配置项信息获取模块
51
视频数据下载模块
52
音频字幕处理模块
53
音频切分模块
54
音频字幕块关联模块
55
筛选模块
56

图 5

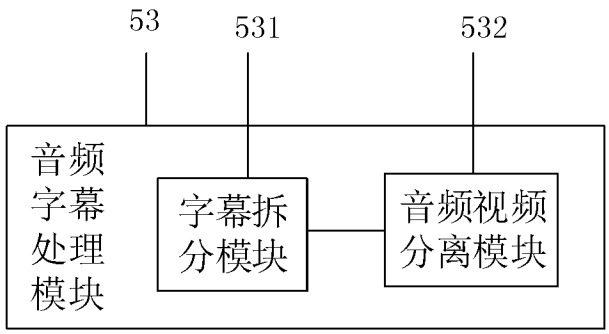

53
531
532
音频字幕处理模块
字幕拆分模块
音频视频分离模块

图 6

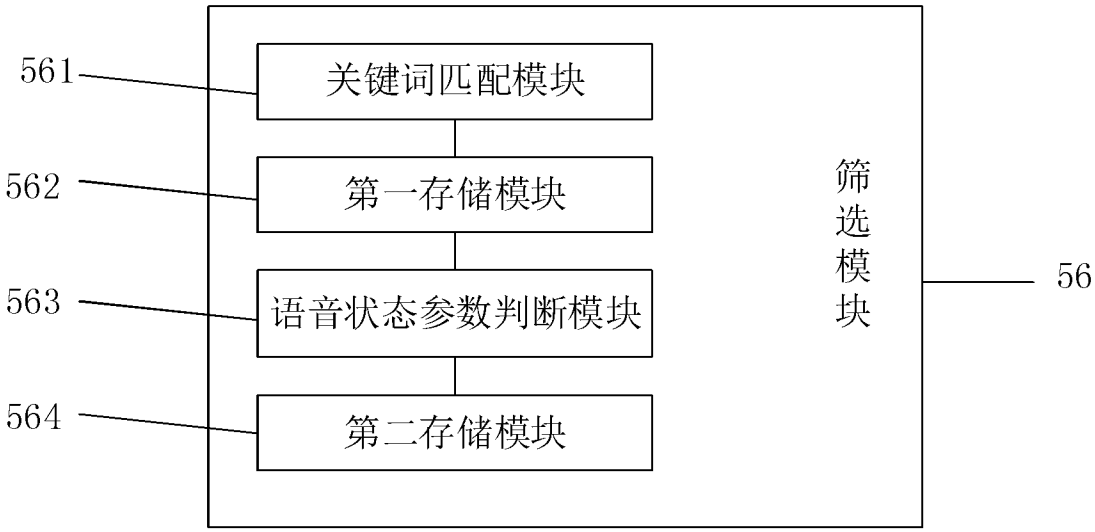

561
关键词匹配模块
562
第一存储模块
563
语音状态参数判断模块
564
第二存储模块
筛选模块
56

图 7

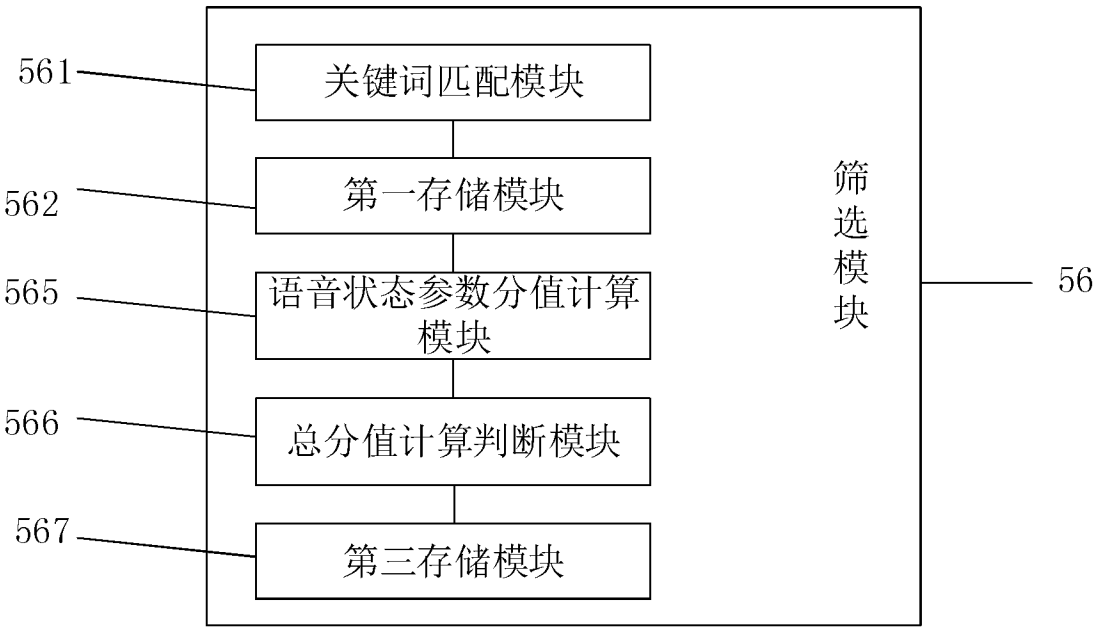

561
关键词匹配模块
562
第一存储模块
565
语音状态参数分值计算模块
566
总分值计算判断模块
567
第三存储模块
筛选模块
56

图 8

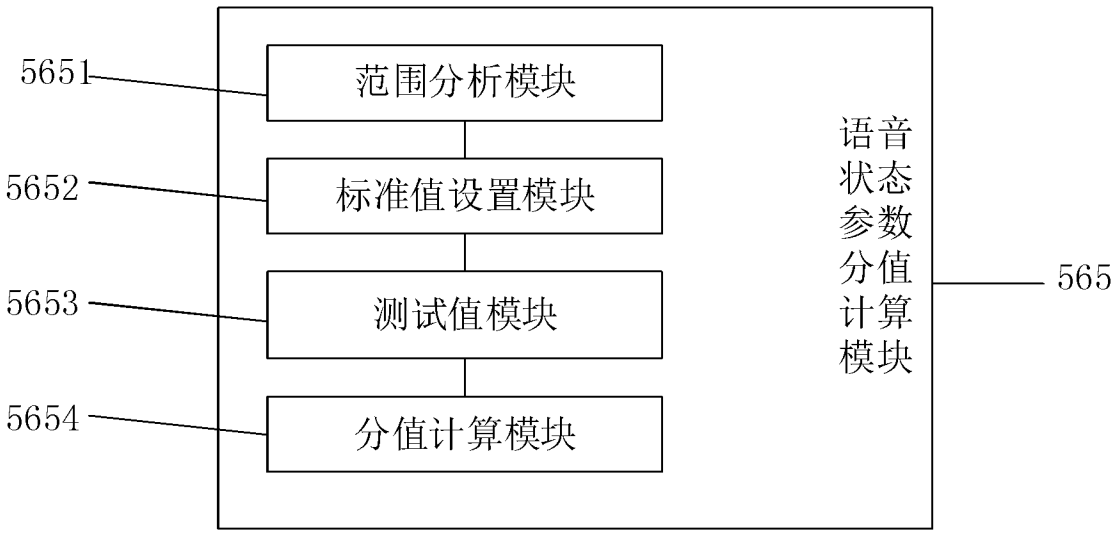

5651
范围分析模块
5652
标准值设置模块
5653
测试值模块
5654
分值计算模块
语音状态参数分值计算模块
565

图 9

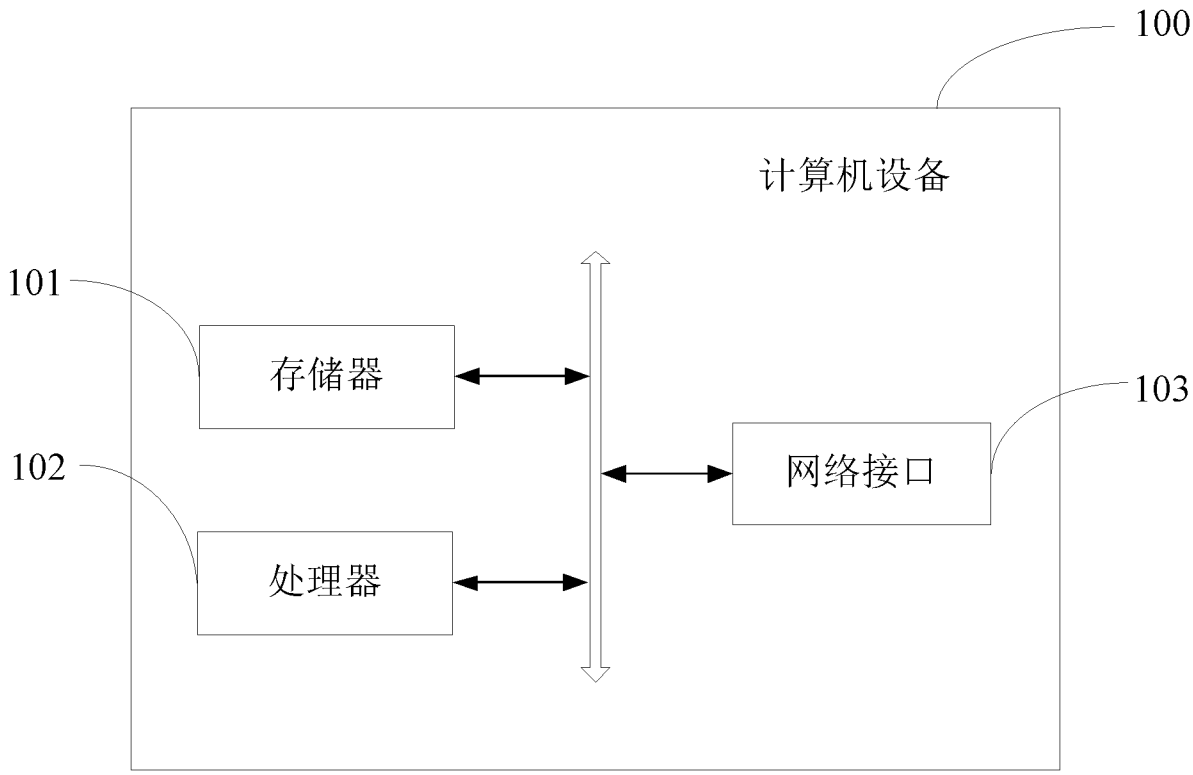

100
计算机设备
101
存储器
102
处理器
103
网络接口

图 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/117261

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/783(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNKI; CNPAT; WPI; EPODOC; IEEE: 音频, 字幕, 关键词, 视频, 切分, 分段, 分块, 语料, 情绪, audio, subtitle, key word, video, segment, corpus, emotion

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110008378 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 12 July 2019 (2019-07-12) claims 1-10, and description, paragraph [0124]	1-21
Y	CN 103324685 A (DALIAN UNIVERSITY OF TECHNOLOGY) 25 September 2013 (2013-09-25) description, paragraphs [0005]-[0036]	1-4, 7-14, 16-19
Y	CN 104978961 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 14 October 2015 (2015-10-14) description, paragraphs [0036]-[0048]	1-4, 7-14, 16-19
Y	CN 105047203 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 11 November 2015 (2015-11-11) description, paragraphs [0035]-[0047]	1-4, 7-14, 16-19
Y	CN 108122552 A (SHANGHAI ZHIZHEN INTELLIGENT NETWORK TECHNOLOGY CO., LTD.) 05 June 2018 (2018-06-05) description, paragraphs [0026]-[0047]	3-4, 7, 13-14, 18-19
A	CN 108268539 A (SHANGHAI JIAO TONG UNIVERSITY) 10 July 2018 (2018-07-10) entire document	1-21

☑ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 January 2020	**31 January 2020**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration** **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/117261

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2007253678 A1 (SARUKKAI, R.R.) 01 November 2007 (2007-11-01) entire document	1-21

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/117261

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110008378	A	12 July 2019	None			
CN	103324685	A	25 September 2013	None			
CN	104978961	A	14 October 2015	WO	2016188329	A1	01 December 2016
				IN	201727044950	A	20 April 2018
				US	2018158469	A1	07 June 2018
				EP	3340238	A1	27 June 2018
				JP	2018522288	A	09 August 2018
CN	105047203	A	11 November 2015	None			
CN	108122552	A	05 June 2018	HK	1256390	A0	20 September 2019
CN	108268539	A	10 July 2018	None			
US	2007253678	A1	01 November 2007	WO	2007130799	A1	15 November 2007
				US	2011211814	A1	01 September 2011

国际检索报告

国际申请号
PCT/CN2019/117261

A. 主题的分类

G06F 16/783(2019.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNKI;CNPAT;WPI;EPODOC;IEEE: 音频，字幕，关键词，视频，切分，分段，分块，语料，情绪，audio，subtitle，key word，video，segment，corpus，emotion

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 110008378 A (平安科技深圳有限公司) 2019年 7月 12日 (2019-07-12) 权利要求1-10，说明书第[0124]段	1-21
Y	CN 103324685 A (大连理工大学) 2013年 9月 25日 (2013-09-25) 说明书第[0005]-[0036]段	1-4, 7-14, 16-19
Y	CN 104978961 A (腾讯科技深圳有限公司) 2015年 10月 14日 (2015-10-14) 说明书第[0036]-[0048]段	1-4, 7-14, 16-19
Y	CN 105047203 A (腾讯科技深圳有限公司) 2015年 11月 11日 (2015-11-11) 说明书第[0035]-[0047]段	1-4, 7-14, 16-19
Y	CN 108122552 A (上海智臻智能网络科技股份有限公司) 2018年 6月 5日 (2018-06-05) 说明书第[0026]-[0047]段	3-4, 7, 13-14, 18-19
A	CN 108268539 A (上海交通大学) 2018年 7月 10日 (2018-07-10) 全文	1-21
A	US 2007253678 A1 (SARUKKAI, Ramesh R.) 2007年 11月 1日 (2007-11-01) 全文	1-21

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2020年 1月 15日	2020年 1月 31日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 穆滢 电话号码 86-(10)-53961531

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/117261

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	110008378	A	2019年 7月 12日	无			
CN	103324685	A	2013年 9月 25日	无			
CN	104978961	A	2015年 10月 14日	WO	2016188329	A1	2016年 12月 1日
				IN	201727044950	A	2018年 4月 20日
				US	2018158469	A1	2018年 6月 7日
				EP	3340238	A1	2018年 6月 27日
				JP	2018522288	A	2018年 8月 9日
CN	105047203	A	2015年 11月 11日	无			
CN	108122552	A	2018年 6月 5日	HK	1256390	A0	2019年 9月 20日
CN	108268539	A	2018年 7月 10日	无			
US	2007253678	A1	2007年 11月 1日	WO	2007130799	A1	2007年 11月 15日
				US	2011211814	A1	2011年 9月 1日