一种通过语音识别和音频分割来自动生成字幕的解决方案

胡中毓

摘要：随着网络科技的发展，人人都有机会在网络上发布自己的原创视频。但是，视频字幕的制作通常比较繁琐，而当前并没有太多的自动字幕生成方案，故而本论文阐述了一种自动字幕生成的方案。此解决方案整合了现有的语音识别和音频分割技术来进行自动字幕生成。实验结果证明，该方案可以有效的协助使用者进行字幕制作。

关键词：视频制作，视频转换，音频分割，语音识别，字幕生成

A Solution For Automatically Generating Subtitles By Speech Recognition And Audio Segmentation

HU Zhongyu

Abstract: With the development of internet technology, everyone has the chance to release original videos onto the internet. However, the production of subtitle is usually tedious, and there is not much available solution for this problem. To solve it, this paper describes a new scheme which integrates existing speech recognition and audio segmentation techniques for automatic caption generation. The experimental results show that the solution can effectively assist users in subtitle production.

Keywords: video production; video conversion; audio segmentation; speech recognition; subtitle generation.

正文

1. 引言

随着网络科技的发展，人人都有机会在网络上发布自己的原创视频。字幕的存在可以让观众更轻松地获取视频的信息。两种字幕形式较为常见：直接集成在视频图像中的内嵌型，和以srt文本文件为格式临时导入视频的外挂型[1]。视频字幕的制作通常比较繁琐，现有的一些方法需要使用者进行较为复杂的操作：创作者可以通过专业软件，在与视频内容进行对比的同时将字幕同步时间轴内嵌[2]；或者在任意文本编辑器中分别输入每一段字幕的起始时间，结束时间以及字幕文字，然后直接改后缀名使其成为srt外挂字幕文件[3]。

为了简化字幕制作的过程，笔者对现有的语音识别和音频分割技术进行整合，在调取语音识别库对视频语音进行转换以及运用k-means聚类算法来进行较为精准的语音片段分割之后，整合生成srt格式外挂字幕文件，可以直接在视频文件中加载，方便使用。

1. 软件实现
   1. 视频预处理

Ffmpeg是一个可以通过命令行即可运行的免费音视频处理工具[4]。笔者用Python的subprocess库调用ffmpeg，将待处理视频（如mp4）转换成合适的音频格式（wav）。

* 1. 语音分割

自动字幕生成方案很重要的一点就是对语句的精确分割。成功分割后可以用现成的语音识别技术将每句话单独识别即可。语句的分割，一般可用静音间断来作为分割依据，但有一段较为低分贝的音频时，可以认为是一句话与另一句话的分割处。



音频分割处

从上图可以看出，静音段可能是句与句之间的间断，但也可能只是两个词之间的间断。将静音段长度设为L秒，将该段是句与句之间分割段的概率设为α，则两者有以下关系：

α = b · L (b>0)

也就是说，一个静音段越长，越可能是两句话之间的分割。

由于计算机中音量概念与物理学概念不同，默认当音量低于-16分贝时，此音频段为静音段。

初步设定当静音片段长度达到700毫秒时，程序将对音频语句作出分割。介于人与人的说话方式都有不同，用户能够自行对此数值进行修改。

通过k-means聚类算法来分析音频中静音段的平均时长从而来确定分割的基本标准[5]。

上图展示了一段音频中静音片段的长度。我们可以发现静音片段的长度大概聚集在0.5秒～1.0秒之间，只有部分处于10秒左右。使用k-means算法，排除特例之后，将主要的静音片段判断标准定为0.6秒左右

* 1. 语音识别

使用百度、讯飞等开源语音识别库将语音转化为文字

语音文字识别技术在当下已经非常成熟，故而本方案直接采用了百度和讯飞的双重语音文字识别引擎，双引擎增加可靠性，同时互相对比可以增加识别准确率。

将分割后的音频逐个上传到百度和讯飞的服务器[6]，并将返回的结果比对，如有不同，则可以标注该段，让用户自行选择。

* 1. 处理过长字幕

设l为最适合一段字幕的字符数（l=20）。当一段字幕大于l，通过分割来让其接近l。我们通过python中的jieba库进行中文分词使单独的词语不作为分割点[7]，从而保证语意不变。对于一段字幕，从长度l的地方开始向前进行索引直到jieba分词的分词点并在此处进行分割。

* 1. 合并时间轴与字幕

将语音识别后的字幕段落与时间轴合并，输出成srt文件。

1. 实验结果和分析

选取语句较为清晰，背景声音不嘈杂的视频测试。将运行后得到的字幕结果与

工听写所得的字幕进行对比。

|  |  |  |  |
| --- | --- | --- | --- |
|  | 分割准确率（正确分割语句数/总语句数） | 单个片段字幕长度平均值 | 语音识别正确率（正确字数/总字数） |
| 样本1 | 32/38（84.2%） | 18.2 | 69.5% |
| 样本2 | 48/54（89.9%） | 16.4 | 65.2% |
| 样本3 | 29/35（82.9%） | 17.8 | 70.2% |
| 样本4 | 31/36（86.1%） | 18.4 | 55.4% |
| 平均值 | 85.8% | 17.7 | 65.1% |

通过对吐字清晰的视频进行测试，结果令人满意。语音分割准确率较高，平均值约85%；单个片段的字符数稳定在20字以内，使每段文字不过长也不过短；但是受制于现有语音识别技术，字幕正确率仅有半成。

1. 总结语

字幕制作有许多方式，但使用起来大多费时费力。本人通过整合现有语音识别和音

分割技术制作的自动字幕工具在分割方面有较好的效果，但是语音识别正确率不是很高。不过由于能够自动划分时间片段并生成srt文件，使用者可以在生成之后进行手动的简单修改即可达到较好的效果。

参考文献

[1]千江有水,loulii.玩转字幕 从Sub到Srt[J].电脑迷,2006(04):14-15.

[2]谢小粮.用Premiere Pro制作影片的滚动字幕[J].电脑知识与技术,2013,9(28):6421-6422.

[3]张海兵,刘阁.视频公开课字幕文件的制作技术探索[J].工业技术与职业教育,2013,11(03):9-10+26.

[4]程会敏,周恕义,李小勇,陈圣奇.基于FFmpeg开源组件的视频自动转码[J].中国现代教育装备,2014(07):23-25.

[5]李钊,袁文浩,任崇广.基于搜索空间划分与Canopy K-means聚类的种群初始化方法[J/OL].控制与决策:1-7[2019-08-07].https://doi.org/10.13195/j.kzyjc.2019.0358.

[6]杨焕峥,杨国华,徐玲,孙慧敏.云端AI与本地相结合的嵌入式语音识别系统[J].宁波职业技术学院学报,2019,23(01):86-89.

[7]何南思,杨霁.《Python高级编程》课程信息化教学设计初探——以“Jieba库应用”课程单元为例[J].信息与电脑(理论版),2019(09):239-240+243.