



开放科学
(资源服务)
标识码
(OSID)

富媒体环境下语音和文本内容的对齐研究

韦向峰 袁毅 张全 池毓焕

中国科学院声学研究所 北京 100190

摘要: 语音媒体和文本媒体是富媒体中重要的两种媒体, 本文试图解决富媒体内容关联中语音和文本的自动对应和关联问题。本文通过自动语音识别、语音纠错处理、文本相似度计算等技术和方法, 把语音和文本的对齐问题转化为文本之间相似和对齐的问题, 通过从语音到文本、从文本到语音两种应用场景, 用实验证明了该方法的可行性和有效性。本文的语音文本对齐技术将是富媒体融合特别是语音媒体和文本媒体之间进行融合的重要支撑技术, 同时也为文本媒体和语音媒体的进一步关联重组和应用提供坚实的技术基础。

关键词: 富媒体; 语音识别; 锚点词; 编辑距离; 最长递增子序列

中图分类号: TP37, TP391.1 G35

Research on the Alignment of Speech and Text Contents in Rich Media Environment

WEI Xiangfeng YUAN Yi ZHANG Quan CHI Yuhuan

Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

Abstract: Speech medium and text medium are two important media in rich media environment. This paper tries to solve the problem of automatic correspondence and association between speech and text in rich media content association. This paper

基金项目: 富媒体数字出版内容组织与知识服务重点实验室开放基金项目 (ZD2018-07/03): “面向语音与文本内容的富媒体关联研究与技术实现”。

作者简介: 韦向峰 (1976-), 博士, 副研究员, 研究方向: 自然语言理解、语音处理、知识组织, E-mail: wxf@mail.ioa.ac.cn; 袁毅 (1967-), 高级工程师, 研究方向: 计算机软件、计算机网络; 张全 (1968-), 博士, 研究员, 研究方向: 自然语言理解、语言技术; 池毓焕 (1967-), 博士, 助理研究员, 研究方向: 自然语言理解、机器翻译、知识组织。

transforms the problem of alignment between speech and text into the problem of similarity and alignment between texts through automatic speech recognition, speech error correction and text similarity calculation. The feasibility and validity of this method are proved by experiments in two application scenarios, namely, from speech to text and from text to speech. The speech text alignment technology in this paper will be an important supporting technology for rich media integration, especially between speech media and text media. It will also provide a solid technical foundation for the further association, reorganization and application of text medium and speech medium.

Keywords: Rich media; speech recognition; anchor words; edit distance; maximum incremental subsequence

引言

文本和语音是常见的两种富媒体表现形式，它们一般在整体内容上保持一致。文本具有静态、可视、易划分、便于组合等特点，而语音则具有动态、连续、可听、划分困难、组合不易等特点，此外语音中还可能带有背景音或者噪音。因此，如何按照文本内容的划分单元对富媒体中的语音进行划分、对齐和组织，进一步按照知识组织体系进行关联和重组，成为富媒体融合的重要关键技术之一。目前，语音-文本自动对齐技术已经在诸多领域中得到应用，例如多媒体索引、训练基于大型语音语料库的语音识别和语音合成系统、语音电子书、视频字幕管理等^[1]。

语音-文本对齐技术是以自动语音识别处理为基础的，把参考文本与相应的语音按某种切分单位进行自动对齐的过程，其目的是获得语音信号与参考文本符号之间的映射关系，以期实现后面的语音信号或文字文本的知识组织^[2]。本文采用自动语音识别、语音纠错处理、文本对齐技术，把语音单元到文本内容的对齐和文本单元到语音内容的对齐问题，转化为文本和文本的对齐问题。首先，利用已

有的语音识别引擎系统，将语音媒体按端点检测分割后识别为文本，从而把语音-文本的对齐问题转化为文本与文本对齐的问题；第二，对语音识别的文本进行后处理，发现识别文本中的错误并纠错，得到准确度更高的语音识别文本结果；第三，使用有限状态机语言模型、编辑距离或词向量等文本相似度算法，根据文本划分单元的相似度将语音识别文本与参考原始文本对齐；最后，根据原始文本的划分单元得到与之对齐的语音识别文本，进而得到语音识别文本在语音媒体中的起止时间，利用语音媒体切分程序或工具即可得到与原始文本划分单元对齐的语音。

1 国内外研究概述

国外对语音-文本对齐的研究起步很早，已经出现了针对英语法语的语音-文本对齐技术，并应用到实际的系统中。例如，美国普渡大学的研究人员利用文本语音对齐技术来生成视频字幕，该算法运用自动语音识别的结果，使其与视频同步，同时将其与通过人工标注的字幕进行校对，通过这种方式来产生比较准确的离线字幕^[2-3]；Brousseau 等设计并开发了一

个专门为法语电视节目提供实时字幕的系统,电视音频经由 *re-speaker* 重新发音后由语音识别引擎识别为文本,这个字幕系统于 2004 年在加拿大上线并为加拿大的法语地区广播电视系统提供服务^[2];美国南加州大学的 Katsamanis 等在多媒体信息检索中应用了词汇和音素级别的语音 - 文本对齐技术,并开发出了相应的 SailAlign 开源软件工具包^[4]。

国内对语音 - 文本的对齐研究主要集中在语音识别领域,例如中国科学院声学研究所申请的专利“一种在线语音文本对齐系统及方法”^[5],该专利发明具有自动错误检测文本和语音中的不匹配段落并跳过这些段落的功能,可以实时的产生在线输入语音流和对应的文本的对齐结果,并能够处理带错误的长文本^[6]。张巍等人提出了一种利用开放识别引擎和基于有限状态自动机的语言模型来得到语音与文本一一对齐数据的算法,来摆脱对于大量标注数据的依赖,可用于语音识别声学模型的训练,并对文本和语音进行迭代对齐,实现自适应的文语对齐^[7]。山东大学的陶冶在“文本语音匹配的研究和应用”中提出了一种改进的基于扩展匹配网络的容错对准算法和一种适用于大量连续语料和非完全匹配语料的对准算法,并实现了文本匹配引擎,取得了良好的对齐效果^[2-3]。在语音识别文本的纠错处理方面,中国科学院声学研究所韦向峰等人利用语音混淆矩阵和语义概念关联关系,对语句层面的语音纠错进行了深入研究,这一语音处理的纠错方案为建立形成语义内容层面的语音 - 文本关联技术奠定了坚实的基础^[8]。高红坤等在 SailAlign 的基础上进行了中文的文语对齐研究,添加了中文的

语言模型和声学模型,实现了中文语音识别中的自动文语对齐^[9]。北京邮电大学的何可嘉基于语音识别和动态规划锚点(可信对齐区域)的递归对齐算法,结合端点检测、语音检测以及说话人分割等音频分割技术,完成了广播语音的自动标注系统的搭建工作,对广播语音实现内容简介层、说话人身份层、说话内容层三层信息的自动标注^[10]。目前我国的中文语音识别和语音合成水平已处于世界领先地位,在技术研发和技术积累方面,也有了丰富的积累和坚实的基础,这也为中文的语音 - 文本对齐技术的发展提供了基础。

在自动语音识别或者语音语料标注系统中,传统上是通过基于 Viterbi 算法的强制对齐技术(Force-alignment)来实现的^[1]。然而,在语音媒体包含噪声或者语音时长较长和转写文本正确率不高的情况下,采用这样的方法是无法胜任的。因为对于较长语音段,强制对齐需要很长的时间且准确率大大下降^[9]。虽然语音 - 文本对齐的研究已经取得了很大的进展,但在应用在实际的过程中还有不少困难,主要原因有以下几个方面^[9,11]。(1) 语音的质量:目前绝大部分用于进行语音识别系统训练的语音数据都是处在良好的实验室环境下录制的,而人们在实际生活中保存的语音数据质量往往比较差,所以语音识别系统词错误率一般较高。语音 - 文本对齐技术以语音识别的文本结果为前提,所以语音质量的好坏将影响语音和文本进行对齐的正确率。(2) 长语音的处理^[2]:在语音的识别过程中,为了完成强制对齐,必须按照原始文本与其相对应的声学模型得到匹配的网格。但是要想产生这样的匹配网格,一般大约需要

成千上万个状态,并且搜索路径将随着状态数量的增加而变得特别复杂。在进行多媒体文档检索过程中,语音的长度至少十几分钟,有的达到几个小时,这样的语音使得传统的维特比算法进行解码的过程所需要的时间和空间飞快增加。(3)不完整的原始文本:在实际生活中,原始文本很难与对应的语音达到完全对齐,而相对于完整的原始文本,不完整的原始文本通常是容易获得的(例如,教师的讲课内容一般以演讲稿为参考;电影和电视剧的角色之间的对白剧本等)。这些原始文本中都包含有不少信息,但是其中的内容可能不完整、缺失和冗余,需要研究如何利用这些不完整的原始文本得到准确的对齐信息。(4)缺乏语义线索和语义处理能力:维特比算法并没有考虑语音和文本的语义信息,只是通过训练语料的形式特征与语音信号的频谱特征进行了对应。由于语音媒体和文本媒体除了形式特征还有语义方面的特征,所以如何利用语音和文本中的语义信息进行对齐,以及如何利用语义分析和语义特征提升语音识别以及语音与文本的语义关联关系发现,是一个值得研究的重要问题。

因此,本文将探索利用自然语言处理和语音处理技术,解决长语音与文本划分单元的对齐问题,特别是原始文本不完整、与语音音频不能完全一一对应情形下的文本划分单元与语音单元的对齐。通过自动语音识别将语音媒体转换为识别文本,建立纠错容错处理机制,通过文本与文本之间的相似度计算、利用语音媒体的自动切分及其与识别文本的对应关系,实现文本媒体和语音媒体的自动对齐。这些技术是富媒体融合特别是语音媒体和文本媒体之间

进行融合的重要支撑技术,同时也为文本媒体和语音媒体的进一步关联重组和应用提供坚实的技术基础。

2 语音媒体识别为文本

2.1 自动语音识别

在语音-文本自动对齐的过程中,第一步是利用语音识别技术把语音内容自动识别为文字文本。随着语音识别技术的发展和深度神经网络的引入,语音识别技术取得了重大进步,各种免费开放的语音识别云端平台也提供了语音识别的API接口^[12]。本文选择使用第三方的语音识别技术和平台,把音视频文件中的语音内容转换为文字文本。由于原始的音视频媒体文件时长较长,格式也不统一,为了对接第三方的语音识别系统,需要对音视频文件进行预处理和语音内容切分。在预处理阶段,对需要处理的音视频资料(如wav、mp3、mp4文件)进行归一化预处理,利用音视频转换软件,统一转换为采样率为16000Hz/秒单声道音频文件。在语音内容切分阶段,对经过归一化预处理的音频文件,采用端点检测技术对文件进行切分。端点检测算法采用的是基于能量的特征检测方法,即首先检测信号的强度,并且假设语音能量大于背景噪声能量;然后对于能量大于某一门限的语音内容段,可以认为是有人类语音存在的内容段。这样,经过预处理和根据端点检测切分后,语音媒体文件被切分为多个音频文件,每个音频文件可以认为是只包含一个段落、句子或短语的音频文件,并保留音频文件在原始语音媒体中的起止时间,如音

频文件可命名为：原文件名_起始时间-结束时间.wav，其中时间的单位为毫秒 ms。而这些多个文件之间的时间间隔的语音媒体内容被认为是含人类语音内容的无语音信号，因此不保存其语音内容。

经过预处理和语音内容切分后得到的语音内容音频文件，就可以调用第三方的语音识别系统对切分好的音频文件进行语音识别。识别后得到的文字文本内容保存在.txt文件中（文件名与音频文件名称相同），如图1所示。

20180905_0-2150	2018/11/20 15:04	文本文档	1 KB
20180905_0-2150	2018/11/20 11:54	WAV 文件	68 KB
20180905_2550-6050	2018/11/20 15:20	文本文档	1 KB
20180905_2550-6050	2018/11/20 11:54	WAV 文件	110 KB

图1 音频文件及其语音识别结果文件

语音识别结果文件中的文字文本内容例子如图2所示。



图2 语音识别结果文件中的内容（示例）

2.2 语音识别结果的纠错处理

虽然可以通过自动语音识别系统把语音媒体识别为文字文本，但是由于自动语音识别的技术复杂性和各种影响因素如背景噪音、声学信号处理和声学模型、语义处理和语言模型等，都会使语音识别的文字结果中很容易出现识别错误的字词甚至句子，因此还需要对识别文本进行纠错处理。

语音识别错误的原因从音节的发音角度看有错音、添音和吞音，从音字转换的错误类型上可以分为“音对字错”和“音错字错”^[13]。根据错误数据的统计，发现错误主要集中在错音和“音错字错”。要实现对错音的纠错处理，

首先构筑声母和韵母的混淆矩阵，根据声韵组合可以构造出每个音节的混淆音；然后验证音节混淆音是否覆盖正确的结果，并根据验证情况对一些音节进行调整；最终使得根据错误音节的混淆音给出的正确音候选集前十名包括的正确音大于95%^[14]。

从句法语义的角度看，语音识别错误的主要表现是语句中出现了“孤魂”（与前后文字没有联系的错误字词）。“孤魂”就是错误可能出现的地方，因此只要在认为有错误的位置上给出可替换的词语候选集，纠错实际上就转化为从词语候选集中选出正确的词语^[15]。通过尝试候选集中的所有词语，依据句类分析技术就可以找到合乎语句概念关联脉络和语义约束要求的词语，完成纠错处理。下面是一个具体的语音识别纠错处理的例子，语音识别系统输出的句子是“我国今后每年都要今后许多小麦”（其中第二个“今后”识别结果错误，应为“进口”）^[8,15]。

1) 预处理：根据词语概念的关联性，进入孤魂可疑集合的词语是：

我国，今后（第二个）

2) 句类分析：句子没有能够假设为全局特征语义块的动态概念，同时也没有迹象表明句子属于无特征语义块的句类。

3) 孤魂发现：由于“今后（第二个）”前面有时间概念（“今后每年”），同时有特征语义块的逻辑说明概念（“都要”），这两类概念合起来经常充当特征语义块要素的前修饰部分。表明这个位置应当出现一个动态概念。“今后（第二个）”更可能是孤魂。“我国”是孤魂的可能性要小于“今后（第二个）”。孤魂

发现只是根据可疑集合和数据黑板的数据判断孤魂的可能性，不是作定案的判断。孤魂消除如果失败，还可以再回到这里，根据现场的数据重新进行孤魂发现。

4) 孤魂消除：根据孤魂发现给出的孤魂记录结果，以及对该位置上的概念预期情况，利用混淆音数据寻找在这个位置上的候选词语集合，得到“浸透”“进口”两个动态概念。用它们对应的句类的句类知识对句子进行检验，否定掉“浸透”，认可“进口”为特征语义块。同时，确定句子属于物转移句句类，有孤魂嫌疑的“我国”充当了物转移动作的发出者，表明与它相关联的词语是隔开的，消除了它的孤魂嫌疑。

5) 完成纠错处理，给出建议的正确结果。

由此，通过对文本中的句子进行文本的语义分析，依据概念层次网络理论对句子的语义结构和词语概念的关联性找到在语义上缺乏关联的词语，根据语义概念预期和语音的混淆音数据候选集可以检验得到正确的词语文本，从而实现语音识别文本的错误发现和错误纠正处理，提高语音识别文本的正确率，为后续文本对齐提供更好的准确文字基础。

3 语音 - 文本的相互对齐

3.1 从语音到文本的对齐

在某些应用场景中，已知整体语音与整体文本，但需要从某个语音片段找到其在整体文本中的对应文本（即整体文本的一部分），称为从语音到文本的对齐。例如，在有声小说的语音文件到文本文件对齐的应用场景中，语

音片段是有声小说的一部分（比如一集），已知的原始文本文件是有声小说的整体文本。语音到文本的对齐过程需要把语音片段识别为文本，通过语音识别文本找到原始整体文本中与之最为相似的部分文本（即与该集语音片段一一对应的文本）。

利用前述语音识别和纠错处理技术，可以得到语音文件按端点检测切割后的多个语音文件及其语音识别文本，但是语音识别即使经过纠错处理仍然可能没有文本结果或者识别文本中的文字错误较多，所以无法使用字符串匹配搜索的方法找到原始文件中的对应文本。为此，本文提出一种基于“锚点词”和编辑距离的语音到文本对齐方法，步骤如下：1) 对语音识别文本进行自动分词处理，根据词语的长度、词性、词频等特征确定得到锚点词；2) 对原始文本按句子进行切分，得到锚点词在原始文本文件中所在句子位置信息；3) 按照锚点词在语音识别文本中的位置和原始文本中的位置，去除出现位置交叉或重叠的“伪锚点”，对于顺序错乱的锚点位置找出最长递增子序列；4) 根据锚点词在语音识别文本中所在的句子，计算其与锚点词在原始文本文件中的同样窗口长度句子的编辑距离，根据编辑距离的大小（小于一定阈值）确定得到语音识别句子在原始文本文件中的位置；5) 对于无法在原始文本中找到锚点词或者无法找到对应文本的情况，可以适当扩大锚点词的范围或者人工制定锚点词集合，再执行步骤2)到步骤4)。

3.1.1 锚点词

锚点词是用于定位语音识别文本与原始文本，具有特征明显、辨识度高等特点的词语^[16-17]，

例如“无期徒刑”、“假面具”、“炉台”等。一般来说,锚点词主要是专有名词、术语,以及人名、地名、时间、日期等实体名词。虽然锚点词的数量可能并不多,但是要在原始文本(如整本小说)中多次按字符串匹配寻找锚点词,其效率是极其低下的。为了提高锚点词的查找和匹配效率,引入了有限状态自动机(DFA)的算法。首先,把所有锚点词以字为单位,按树的形式存储,例如“假面具”和“假面人”在存储时的第一和第二个字都是相同的,只是在第三个字上分叉成树。其次,在匹配的时候,以原始文本的字为单位,如果某个字是锚点词的第一个字,则只需循环判断下一个字或者后面的字是否是锚点词词库中的最后一个字即可;第三,如果匹配到多个锚点词,则以最大匹配规则为准,返回长度最长的锚点词。

3.1.2 最长递增子序列算法

一个锚点词可以在原始文本中多次出现,多个锚点词出现的多个位置就会形成很多个先后位置顺序序列。为了得到最能反映锚点词在语音识别文本中的正确顺序的问题,采用最长递增子序列的算法来获得锚点的正确位置,即锚点词最多且先后顺序正确的位置序列才是最能反映锚点词正确顺序的子序列。最长递增子序列的问题描述如下:如果 $L=\{a_1, a_2, a_3, \dots, a_n\}$ 是一个整数位置序列,需要从中找出多个子序列 $Lin=\{a_{k_1}, a_{k_2}, a_{k_3}, \dots, a_{k_m}\}$, 满足 $k_1 < k_2 < k_3 < \dots < k_n$, 且 $a_{k_1} < a_{k_2} < a_{k_3} < \dots < a_{k_m}$, 当 m 最大时得到的就是最长递增子序列。

Patience Sort 算法^[10]可以解决这个问题,类似于接龙游戏,首先从数据序列中取出第一个元素,作为第一堆的数据,后续的数据只要

是不比它大的,便可以放上去,否则就必须在右边另起一堆。之后的数据也是一样,只要有任何一堆可以放的就放,否则也要另起一堆。只要遵循将数据尽可能放到最左边的那一堆的贪心算法(Greedy Method),则所需要的堆数会是最少。当所有数据都分堆完毕,最后再将各个堆组合在一起成为排序好的数据^[10]。例如如图3所示,Text表示锚点词在语音识别文本的句子编号,Raw Anchor表示的是锚点词在原始文本中可能的出现顺序。

Text	1	2	3	4	5	6	7	8	9
Raw Anchor	6	2	3	5	1	7	8	9	4

图3 备选锚点的位置序列

Patience Sort 算法的堆放顺序如图4所示,先放数字6,再放数字2;因为数字3比数字2大,所以只能另起一堆放;以此类推,可得到图4所示的堆放结果。

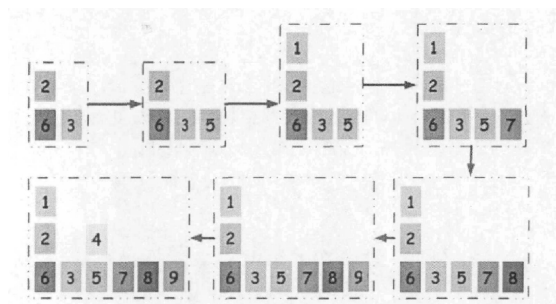


图4 锚点堆放算法的过程

每次堆放时都记录下锚点的路径,当所有的锚点都计算完后,从最后一个堆的位置往回溯,就得到了最长递增子序列,也就是最可能的锚点序列,如图5所示。

2	3	5	7	8	9
---	---	---	---	---	---

图5 锚点最长递增子序列结果

3.1.3 编辑距离的算法

相同的锚点词在语音识别文本和在原始文本文件中的文本有可能内容不同,即使文本长度也相同,所以还需要计算语音识别后锚点词所在的文字文本与锚点词所在的原始文本的相似程度,相似度大或者完全相同的原始文本中的文本就可以作为与语音内容对应的文本候选^[18-19]。文本的相似度可以用动态规划极小化和 Levenshtein 距离(编辑距离)得到^[2]。编辑距离,又称 Levenshtein 距离,是指两个字串之间,由一个转成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符,插入一个字符,删除一个字符^[20]。例如,将“kitten”一词转成“sitting”可以经过以下编辑操作:sitten ($k \rightarrow s$)、sittin ($e \rightarrow i$)、sitting ($\rightarrow g$)。由此得到两个字符串 s 和 t 之间的编辑距离的具体算法步骤如下:

1) 构造行数为 $m+1$ 列数为 $n+1$ 的矩阵 d ,用来保存完成某个转换需要执行的操作的次数,将字符串 $s[1..n]$ 转换到字符串 $t[1..m]$ 所需要执行的操作次数为 $d[n][m]$ 的值。

2) 初始化矩阵 d 的第一行为 0 到 n , 第一列为 0 到 m 。

3) 检查每个从 1 到 n 的 $s[i]$ 字符。

4) 检查每个从 1 到 m 的 $t[j]$ 字符。

5) 将串 s 和串 t 的每一个字符进行两两比较,如果相等,则让 $cost$ 为 0,如果不等,则让 $cost$ 为 1。

6-1) 如果可以在 k 个操作里面将 $s[1..i-1]$ 转换为 $t[1..j]$,那么就可以将 $s[i]$ 移除,然后再做这 k 个操作,所以总共需要 $k+1$ 个操作。

6-2) 如果可以在 k 个操作内将 $s[1..i]$ 转换为 $t[1..j-1]$,也就是说 $d[i,j-1]=k$,那么就可以将 $t[j]$ 加上 $s[1..i]$,这样总共就需要 $k+1$ 个操作。

6-3) 如果可以在 k 个步骤里面将 $s[1..i-1]$ 转换为 $t[1..j-1]$,那么就可以将 $s[i]$ 转换为 $t[j]$,使得满足 $s[1..i]=t[1..j]$,这样总共也需要 $k+1+cost$ 个操作。(这里加上 $cost$,因为如果 $s[i]$ 刚好等于 $t[j]$,那么就不需要再做替换操作,即可满足,如果不等,则需要再做一次替换操作,那么就需要 $k+1$ 次操作)。

7) 比较 6-1)、6-2)、6-3) 所得的操作个数,取最小值作为 $d[i,j]$ 的值。

8) 重复执行步骤 3) 到步骤 7), 最后的编辑距离结果就存在 $d[n,m]$ 中。

这样,通过把语音媒体文件切分为多个语音音频文件及其语音识别文本,从语音识别文本得到锚点词,再根据锚点词和语音识别文本得到原始文本中的候选文本,计算语音识别文本与候选文本的编辑距离后可以找到语音识别文本在原始文本中的位置,即得到语音文件对应的原始文本部分。

3.2 从文本到语音的对齐

在某些应用场景中,已知整体文本与整体语音,但是需要根据文本片段的准确内容,在整体语音中找出并得到与文本片段相对应的语音片段,称之为从文本到语音的对齐。在文本到语音的对齐场景中,例如已知 PPT 讲者的正式演讲稿(按幻灯片组织的多个文本片段),根据一个幻灯片的文本片段到 PPT 演讲录音文件中找到该幻灯片的演讲起止时间,并得到一个演讲录音文件。

与语音到文本的对应方法步骤类似,可以先把一个幻灯片的文本片段进行自动分词,然后根据词语的长度、词性和词频特征得到锚点词,利用锚点词到整个 PPT 演讲录音文件的语音识别文本(多个且有对应的语音文件)中进行锚点定位,根据锚点上下文文本的编辑距离得到对应的语音识别文本(即得到相应语音文件),最后把对应的多个语音文件合成一个语音文件,就是该幻灯片文本对应的演讲录音。

与文本到语音的对应所不同的是:1)如果该幻灯片的前一幻灯片文本或者后一幻灯片文本已经对齐(与语音片段对应),那么就可以把已对应的前后幻灯片的语音片段作为要对应的语音文件的前后边界,这将缩小检索计算空间,提高文本到语音的对应效率;2)由于幻灯片的文本片段是精确的,因此可以只使用文本片段的第一句和文本片段的最后一句作为代表文本进行文本到语音的对应,在无法对应时再用文本片段内部的文本进行对应,这样也可以提高对应效率;3)得到文本片段对应的多个语音文件(语音识别处理后的多个文件),必须把这些多个语音文件合成一个语音文件,或者从整体的 PPT 演讲录音文件(一个语音文件)中截取出文本片段所对应的语音片段。由于在语音识别处理过程中保留了这些多个语音文件在 PPT 演讲录音文件中的起止位置,所以采用从一个整体语音文件中截取起止时间(根据多个语音文件得到)的方法获取文本片段对应的一个语音文件。

4 实验

语音与文本的对应关系可以分为两种:一

种是一一对应关系,另一种是部分对应关系。在一一对应关系中,语音所读的文字与要对应的文本中的文字是一字一句完全相同的,例如有声读物或有声小说,朗读者的语音可以一字一句与文本文字完全对应。但是,由于文本存储的形式与语音(声音)的存储形式不完全一致,造成了语音媒体与文本媒体之间需要进一步加工对齐的情形。例如,有声小说的文本内容可能以整本小说为存储单位,存放在一个文本文件中;而声音文件因为时长、播放形式等原因,往往是以“集”(章、回、节等)为单位,分成多个声音文件存储,在出版、播放或展现单个声音文件时就需要自动找到单个声音文件所对应的文本文字,而不是整本小说的所有文字。同时,一一对应的语音文本富媒体语料,经过顺序打乱、内容修改等处理,还可以变成部分对应关系的语音文本富媒体语料,以便对语音-文本对齐技术进行测试和评价。在部分对应关系中,语音所讲文字与要对应的文本文字并不是一一对应关系,只有部分文字是相同的,而相同文字部分出现的先后顺序也可以不同。例如,在 PPT(PowerPoint)的讲演过程中,讲述者的语音与 PPT 中的文字往往不是一一对应的,但是讲述者的某段语音与 PPT 中的某页幻灯片在人听来却是可以一一对应的,在播放某段语音时希望自动播放对应的幻灯片页,或者播放某个幻灯片页时播放对应的语音,在这种场景下就需要用到语音-文本的自动对齐技术。

本文选择了两本有声小说作为从语音到文本的对齐实验样本,一本是《芙蓉镇》,另一本是《悲惨世界》。《芙蓉镇》的语音内容共 28 集,每集 28 分钟左右(男女声),选择了

4集作为研究样本,分别是3、10、17、24;文本内容为《芙蓉镇》小说的整本内容。《悲惨世界》的语音内容共39集,每集25分钟左右(男女声并有音乐和背景渲染声),选择了4集作为研究样本,分别是36、37、38、39。此外,还选择了一个“未来学校”的PPT演讲稿和演讲音视频作为从文本到语音的对齐样本,PPT演讲稿的文本内容与语音的文本内容并不是一一对应关系,而是部分对应关系。由于原始媒体资源为视频演讲,所以采用人工转写的方法得到其中的10个幻灯片演讲文本,视频对应的音频文件由转换工具自动生成。

在实验中《悲惨世界》的第36集被切分为429个语音识别文件结果,从语音识别的结果来看识别出的句子多为短句或词语,也有部分是长句子,总体而言识别的准确率不高。利用这429个语音识别文件的文字结果,我们获得了58个锚点词,通过这些锚点词定位得到语音识别文本在整本小说文本中的位置,但是开始部分和结束部分有误差。这是因为语音识别的文本并不准确,因此在相似度计算时,无法准确得到第一句和最后一句的对应结果,但是从整体语音识别句子文本的对齐总数计算,其准确率达到了86.37%。在从文本到语音的对齐实验中,视频演讲的语音识别结果有很多是较长的句子或者多个句子,但是其识别准确率也不高,结果多数为常用词语,涉及专业词汇或者专有名词则出现识别错误,有的句子一错就错到底(整个句子都识别错误)。由于文本内容是准确的,因此策略为先寻找头句和尾句的对应语音识别结果文字,但由于语音识别结果错误率较高,往往出现找不到对应识别句子文本

的情况。所以最终找到的对应语音片段也有一定范围的误差,从语音时长的角度计算,其错误率平均在20%以下。分析出现错误的句子,除了语音识别结果文本的影响外,还有文本相似度计算出现的错误,即锚点词定位后的句子文本之间的对应关系也出现了一定的错误,这在未来的研究中需要加以改进。

5 结语

本文研究了在富媒体环境中语音媒体与文本媒体的相互对齐问题,把语音到文本或者文本到语音的对齐问题转化为文本之间的对齐问题。对于从语音到文本的对齐场景,通过语音识别可以得到语音对应的文本文字,由于自动语音识别技术的限制,其中包含了大量错误的识别文字结果,因此需要一定的纠错处理;通过语音识别结果可以得到其中的“锚点词”,通过锚点词可以找到对应的候选句子文本,再通过计算句子文本的相似度判断是否是对应文本。对于从文本到语音的对齐场景,从文本片段可以得到准确的首尾句,根据首尾句得到锚点词或关键词语,再到语音识别结果文本中得到文本候选,通过文本相似度计算判断得到对应的语音识别文本句子,根据对应的语音识别文本句子所在的语音文件起止时间,得到文本片段的对应语音片段。该方法在实验环境下具有一定的准确率,但是由于语音识别结果是基础和前提,而语音识别正确率和纠错处理技术目前还有很多问题需要攻克和解决,因此还需要进一步研究如何提高文本语音对齐的性能和效果。总之,本文探索了富媒体中的语音媒体

与文本媒体的相互关联方法,通过把语音媒体转化为文本媒体的方法,找到语音媒体与文本媒体的内容联系和时间联系,利用文本之间的相似度把语音-文本的相互对应关系转变为文本-文本的对应关系,实现了不同富媒体之间的关联计算。本文的语音-文本对应方法可以应用于有声读物的文本自动编辑和管理、文本讲稿的语音片段自动获取,以及音视频的字幕编辑管理等方面。随着富媒体资源和应用场景的不断丰富和涌现,本文所形成的富媒体关联方法将会得到更多更大的应用。

参考文献

- [1] 张志楠. 语音 Corpus 的自动构建和语音最小化标注的研究 [D]. 青岛: 中国海洋大学, 2013.
- [2] 周陈. 长音频与文本的对齐研究 [D]. 湘潭: 湘潭大学, 2013.
- [3] 陶冶. 文本语音匹配的研究和应用 [D]. 济南: 山东大学, 2009.
- [4] Katsamanis A, Black M P, Georgiou P G, Goldstein L, Narayanan S. SailAlign: Robust long speech-text alignment[C]. Proceedings of Workshop on New Tools and Methods for Very Large Scale Research in Phonetic Sciences, January 28-31, 2011.
- [5] 颜永红, 高杰, 赵庆卫, 等. 一种在线语音文本对齐系统及方法 [P]. 中国专利: H04N5/278GK101651788SQ20081022479, 2010-02-17.
- [6] 李传越. 基于通用识别器的文语对齐技术的研究 [D]. 青岛: 中国海洋大学, 2014.
- [7] 张巍, 王永远, 贾晓茹, 等. 无标注的含噪中文长篇幅语音文本的文语对齐研究 [J]. 中国海洋大学学报 (自然科学版), 2015, 45(10):121-126.
- [8] 韦向峰, 张全, 熊亮. 一种基于语义分析的汉语语音识别纠错方法 [J]. 计算机科学, 2006(10):152-155.
- [9] 高红坤. 基于 SilAlign 的中文语音文语对齐的研究 [D]. 青岛: 中国海洋大学, 2015.
- [10] 何可嘉. 广播语音的自动标注系统 [D]. 北京: 北京邮电大学, 2010.
- [11] A. Haubold, R.J. Kender. "Alignment of Speech to Highly imperfect Text Transcriptions", in 2007 IEEE international conference on Multimedia and Expo, 2007: 224-227.
- [12] 姬艳. 发音器官在生理识别中的贡献率及相互补偿的研究 [D]. 天津: 天津大学, 2017.
- [13] 李晶皎, 张王利, 姚天顺. 汉语语音理解中自动纠错系统的研究 [J]. 软件学报, 1999(4):42-46.
- [14] 王兴建. 语音识别后文本处理系统中文本语音信息评价算法研究 [D]. 北京: 北京邮电大学, 2010.
- [15] 张全, 张倪, 韦向峰. 汉语语音识别的纠错处理 [C]. 自然语言理解与机器翻译——全国第六届计算语言学联合学术会议论文集. 2001.
- [16] 塞麦提·麦麦提敏, 侯敏, 吐尔根·伊布拉音. 基于锚点句对的汉维句子对齐方法 [J]. 计算机工程, 2015, 41(4):166-170.
- [17] 王斌. 汉英双语语料库自动对齐研究 [D]. 北京: 中国科学院计算技术研究所, 1999.
- [18] 徐瑛. 一种综合加权的词语语义相似度计算研究 [D]. 青岛: 青岛理工大学, 2011.
- [19] 薛慧芳. 句子相似度计算理论及应用研究 [D]. 西安: 西北大学, 2011.
- [20] 百度百科. 编辑距离 — 百度百科 [EB/OL]. [2019-03-20]. <https://baike.baidu.com/item/%E7%BC%96%E8%BE%91%E8%B7%9D%E7%A6%BB/8010193?fr=aladdin>.