计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(4):300−307 [doi: 10.15888/j.cnki.csa.009043]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

不完全匹配的语音和文本语句级对齐①

徐 锴, 陶 冶, 李 辉

(青岛科技大学 信息科学技术学院, 青岛 266061)
通信作者: 陶 冶, E-mail: ye.tao@qust.edu.cn

摘 要: 语音文本自动对齐技术广泛应用于语音识别与合成、内容制作等领域, 其主要目的是将语音和相应的参考文本在语句、单词、音素等级别的单元进行对齐, 并获得语音与参考文本之间的时间对位信息. 最新的先进对齐方法大多基于语音识别, 一方面, 准确率受限于语音识别效果, 识别字错误率高时文语对齐精度明显下降, 识别字错误率对对齐精度影响较大; 另一方面, 这种对齐方法不能有效处理不完全匹配的长篇幅语音和文本的对齐. 该文提出一种基于锚点和韵律信息的文语对齐方法, 通过基于边界锚点加权的片段标注将语料划分为对齐段和未对齐段, 针对未对齐段使用双门限端点检测方法提取韵律信息, 并检测语句边界, 降低了基于语音识别的对齐方法对语音识别效果的依赖程度. 实验结果表明, 与目前先进的基于语音识别的文语对齐方法比较, 即使在识别字错误率为 0.52 时, 该文所提方法的对齐准确率仍能提升 45% 以上; 在音频文本不匹配程度为 0.5 时, 该文所提方法能提高 3%.

关键词: 语音文本对齐; 韵律信息; 锚点; 自动语音识别; 端点检测

引用格式: 徐锴,陶冶,李辉.不完全匹配的语音和文本语句级对齐.计算机系统应用,2023,32(4):300–307. http://www.c-s-a.org.cn/1003-3254/9043.html

Sentence Level Text-speech Alignment for Imperfect Transcriptions

XU Kai, TAO Ye, LI Hui

(School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: Automatic text-speech alignment technology is widely used in speech recognition and synthesis, content production, and other fields. Automatic text-speech alignment aims to align speech with text in sentence, word, and phoneme units and obtain the time alignment information. Most of the recent alignment methods are based on automatic speech recognition (ASR). On the one hand, the alignment accuracy is limited by the word error rate (WER) of ASR. On the other hand, such methods cannot effectively align imperfect transcriptions. This study proposes a text-speech alignment method based on anchor and prosodic information. Through fragment annotation based on boundary anchor weighting, speech is divided into aligned and unaligned fragments. For unaligned fragments, this study extracts their prosodic information by a dual-threshold endpoint detection method and detects the boundaries of sentences. This approach reduces the dependence of ASR-based text-speech alignment on the speech recognition effect. Compared with the current advanced ASR-based text-speech alignment methods, the proposed method can improve alignment accuracy by more than 45% when the WER is 0.52 and by at least 3% when the degree of incomplete matching is 0.5.

Key words: text-speech alignment; prosodic information; anchor; automatic speech recognition (ASR); endpoint detection

① 基金项目: 国家重点研发计划 (2018YFB1702902); 山东省高等学校青创科技支持计划 (2019KJN047)
收稿时间: 2022-09-07; 修改时间: 2022-10-21; 采用时间: 2022-10-27; csa 在线出版时间: 2022-12-23
CNKI 网络首发时间: 2022-12-27

文语对齐广泛应用在多媒体检索、广播电视媒体、有声读物制作等领域, 辅助完成新闻、演讲、会议等字幕生成以及为歌曲制作同步的歌词显示等任务. 目前文本语音对齐方法可以归纳为以下 3 类.

第 1 类是在强制对齐方法中使用隐马尔科夫模型 (hidden Markov models, HMM). 例如, Moreno 等[1] 提出一种基于自动语音识别 (automatic speech recognition, ASR) 和锚点的递归算法, 在连续语料中获取置信度较高的锚点, 并将语料分割成较小的能够使用 Viterbi 算法[2] 处理的语音片段. Gorman 等[3] 设计的 Prosodylab-Aligner 通过自动构建领域声学模型来进行强制对齐. McAuliffe 等[4] 基于 Kaldi[5] 实现了强制对齐系统 Montreal Forced Aligner, 支持基于预训练模型的音素或字级的对齐粒度. 这类方法主要应用在较短语料的音素对齐任务中, 当实验数据是长音频时, Viterbi 算法会产生大量的搜索树, 对齐效果不稳定[6].

第 2 类是基于语音合成 (text to speech, TTS) 和动态时间规整 (dynamic time warp, DTW) 算法. 例如, Boháč等[7] 通过 TTS 从文本生成音频特征序列, 使用 DTW 算法将不完全匹配的文本和语音进行时间对齐. Aeneas[8] 可完成在不同级别的对齐粒度上工作. Anguera 等[9] 将基于语音合成的文语对齐方法应用于有声读物制作之中. 这类方法对生成的音频质量有较高要求, 当音频质量低导致无法提取有效音频特征时, 对齐效果会明显下降[7]. 此外, 此类方法只有文本和语音内容相差不大, 才具有较高的可靠性、准确性和鲁棒性[7], 因此无法应对大量不完全匹配语料的对齐任务.

第 3 类是基于 ASR 和文本匹配的方法. 例如, SailAlign[10] 文语对齐算法利用对齐的语音和文本重新训练声学模型, 接着使用声学模型对未对齐的语音和文本进行对齐, 在加入一定噪音的 TIMIT 语料库进行实验, 该算法的准确率达到 80% 以上. Mocanu 等[11] 实现了一个面向听障人士的字幕同步和定位系统, 能够自动完成字幕和视频画面对齐并动态调整字幕显示时间. González-Carrasco 等[12]基于 ASR 设计了一个用于西班牙语直播节目的字幕同步框架. Deep-Sync[13] 利用 ASR 进行语音转录并结合 BERT 模型在语义层面进行完成对齐. 这类方法采用了成熟的 ASR 识别技术, 一定程度上降低了标注数据要求, 但对齐的准确率则容易受到 ASR 识别效果的影响, 当 ASR 识别准确率偏低时, 对齐效果会明显下降[14].

上述 3 类方法在一定程度上能够处理完全匹配的音频和文本的对齐, 但均难以处理非完全匹配的语料. 例如, 当语料中出现错音、添音和吞音, 以及“音对字错”和“音错字错”[15] 等情形时, 算法的对齐精度均会受到不同程度的影响.

针对大量连续非完全匹配的语料对齐问题, 本文提出一种基于锚点和韵律信息的文语对齐方法, 在目前先进的基于 ASR 的文语对齐方法基础上结合基于韵律信息的语句边界检测, 能克服不同类型的语音识别错误导致的语句失配问题[16], 提升了在 ASR 识别出现错误时的文语匹配准确率.

本文的主要贡献包括:

(1) 提出基于锚点和韵律信息的文本语音匹配框架, 弥补了目前基于 ASR 的对齐算法受识别效果影响大以及无法有效对齐不完全匹配语料的不足.

(2) 设计了基于边界锚点加权的片段标注和基于韵律信息的语句边界检测算法, 降低了基于语音识别的对齐方法对语音识别效果的依赖程度.

(3) 提供了一套中文语句级的文语对齐数据集, 包含衍生的删除、插入、替换错误导致语句失配的不完全匹配数据集. 为基于语音识别的语句级文语对齐算法在准确度、以及算法的对齐鲁棒性等方面的检验提供了参考.

1 问题定义及方法流程

假设第i句的参考文本序列为$sentence^i = \{S_1, S_2, \cdots, S_\alpha\}$, S_p为参考文本序列中的第p个字词, α为参考文本序列长度, $1 \leqslant p \leqslant \alpha$. 音频的语音识别结果序列为$A = \{\alpha_1, \alpha_2, \cdots, \alpha_\beta\}$, β为语音识别结果序列的长度, α_j为音频语音识别结果序列中的第j个结果, $1 \leqslant j \leqslant \beta$. α_j是一个标记了在音频中发音开始时间和结束时间的字词. ST和ET分别表示开始时间和结束时间, 用ST^i和ET^i表示第i句参考文本的开始时间和结束时间.

文本语音对齐的目标是建立文本和语音的时间对位关系, 即在A中找到$\{\alpha_j, \alpha_k\}$, 且$1 \leqslant j \leqslant k \leqslant \beta$. α_j和α_k分别对应第i句参考文本的开头和结尾的字词, ST_{α_j}表示α_j的开始时间, ET_{α_k}表示α_k的结束时间. 使得第i句的开始时间ST^i由α_j的开始时间ST_{α_j}决定, 第i句的结束时间ET^i由α_k的结束时间ET_{α_k}决定, 即, $ST^i = ST_{\alpha_j}$, $ET^i = ET_{\alpha_k}$.

如图 1 所示, 预处理阶段首先将$sentence^i$和A转为音节[9], 这样可以解决一部分“音对字错”类型的语音识别错误导致的文本语音失配问题. 然后使用动态规划

搜索锚点 (anchor words, AW), 锚点是用于定位语音识别文本与参考文本的字词, 将 ASR 识别结果与参考文本进行对照, 利用字符串匹配算法, 从候选片段中选出在位置和内容上一致的字词即称为锚点.

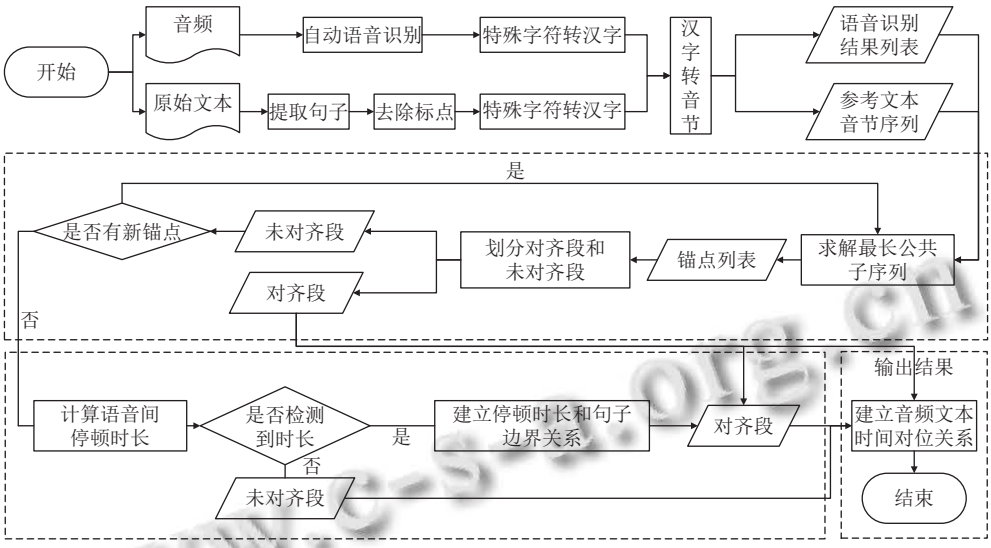

图 1 文语对齐算法流程图

进而根据锚点位置进行片段标注, 将语句划分为对齐段和未对齐段, 接下来, 针对语音识别错误导致的未对齐段循环搜索锚点和片段标注直到没有新的锚点出现为止.

最后, 针对剩余未对齐段, 这部分未对齐段的语音无法正确识别, 因此借助语音端点检测提取韵律信息进行语音边界检测完成未对齐段语句对齐, 如图 2(c) 所示, 这一步可以解决一部分“音错字错”类型的语音识别错误导致的失配问题.

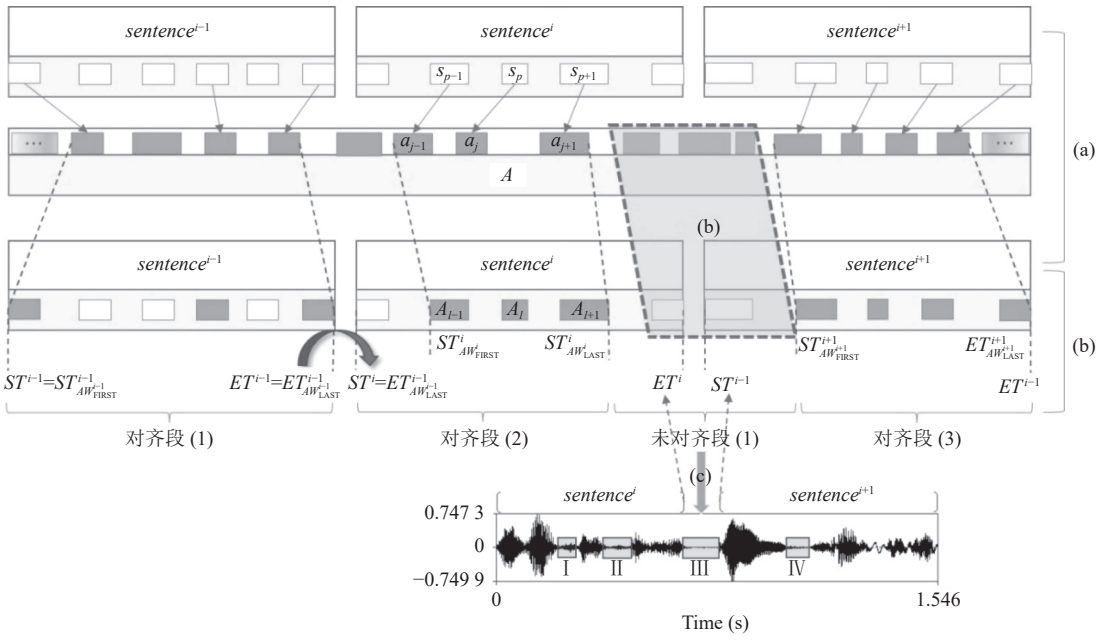

图 2 文语对齐算法实例图

2 基于锚点和韵律信息的文语对齐框架

本文方法主要由以下 3 部分组成.

(a) 基于动态规划的锚点搜索方法. 在原始文本与识别文本的音节序列中, 通过动态规划寻找锚点, 如图 2 中 (a) 内容A_l所示.

(b) 基于边界锚点加权的片段标注方法. 根据锚点

在参考文本语句中的位置关系, 在语料中划分对齐段和未对齐段, 并对未对齐段重复执行 (a) 和 (b), 直到没有新锚点出现为止, 如图 2 中 (b) 内容所示.

(c) 基于韵律信息的边界检测方法. 结合韵律信息检测未对齐段语句边界, 如图 2 中 (c) 内容所示.

2.1 基于动态规划的锚点搜索方法

在寻找锚点之前本文对语音识别结果和原文本进行预处理, 图 3 是一则新闻文本片段预处理前后的变化. 对原始文本进行分句、数字和英文等符号转为汉字、汉字转音节、删除标点符号的处理, 如“2022”和“30.37”等黑体标注的词汇, 将其转换为“二零二二”和“三十点三七”.

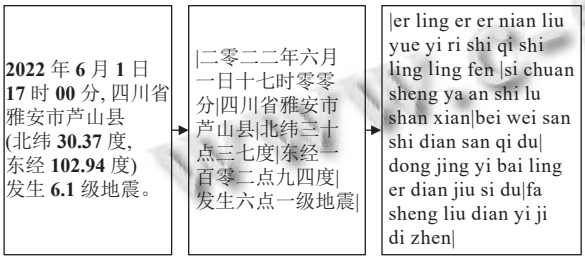

图 3 文本预处理示意图

锚点的搜索问题就是在$sentence^i$和A两个序列中寻找最长公共子序列$Anchors=\{A_1,A_2,\cdots,A_l\}$的问题求解过程, $Anchors$就是寻找的锚点集合.

使用动态规划求解该问题, 用一个二维矩阵存储迭代过程中当前的最长公共子序列长度, 其中, $D[\alpha][\beta]$记录$sentence^i$和A的最长公共子序列的长度, 当$\alpha=0$或$\beta=0$时, 空序列是$sentence^i$和A的最长公共子序列, 故$D[\alpha][\beta]=0$, 其他情况下, 可建立递推关系为:

$$D[\alpha][\beta]=\begin{cases}0, & \alpha=0\text{或}\beta=0\\ D[\alpha-1][\beta-1]+1, & sentence^i[\alpha]=A[\beta]\\ \max(D[\alpha-1][\beta],D[\alpha][\beta-1]), & sentence^i[\alpha]\neq A[\beta]\end{cases} \tag{1}$$

由于$D[\alpha][\beta]$是两个序列的最长公共子序列长度, 而本文最终是希望求解最长公共子序列, 因此根据$D[\alpha][\beta]$回溯即可得到所求锚点序列, 获得的锚点在参考文本的语句中标记后, 时间戳就可以从 ASR 结果和参考文本的部分单词建立联系, 参考文本语句中标识的锚点将继承 ASR 输出中相应单词的时间戳.

2.2 基于边界锚点加权的片段标注方法

根据语句中包含的锚点时间戳确定语句的发音时刻, 不需要进一步计算与每个单词关联的时间戳[17], 对齐段与未对齐段的划分具体分为 3 种情况讨论.

对于第i句的开头和结尾存在锚点的情况, 能够直接完成语句对齐; 对于开头和结尾不存在锚点但是相邻语句的开头结尾存在锚点, 可以借助相邻语句的锚点, 间接完成语句对齐. 对于不能完成对齐的语句根据句中的锚点位置划分对齐段和未对齐段, 具体定义如下对齐规则.

(1) 如图 2 对齐段 (1) 所示, 对于一个包含多个锚点的语句$sentence^{i-1}$, 我们遍历其中的元素, 第 1 个锚点AW^{i-1}_{FIRST}如果是位于语句开头, 那么第$i-1$句的开始时间ST^{i-1}就由第 1 个锚点来确定, 即, $ST^{i-1}=ST^{i-1}_{AW^{i-1}_{\text{FIRST}}}$, 如式 (2), 同理, 最后一个锚点$AW^{i-1}_{\text{LAST}}$若位于语句结尾, 那么该句的结束时间$ET^{i-1}$就由最后一个锚点的结束时间确定, 即, $ET^{i-1}=ET^{i-1}_{AW^{i-1}_{\text{LAST}}}$, 如式 (3):

$$\exists AW^{i-1}_{\text{FIRST}}\rightarrow ST^{i-1}=ST^{i-1}_{AW^{i-1}_{\text{FIRST}}} \tag{2}$$

$$\exists AW^{i-1}_{\text{LAST}}\rightarrow ET^{i-1}=ET^{i-1}_{AW^{i-1}_{\text{LAST}}} \tag{3}$$

(2) 如图 2 对齐段 (2) 所示, 而当第i句开头或者结尾没有锚点时, 我们可以借助第$i-1$句的最后一个锚点的结束时间$ET^{i-1}_{AW^{i-1}_{\text{LAST}}}$和第$i+1$句的第一个锚点的开始时间$ST^{i+1}_{AW^{i+1}_{\text{FIRST}}}$来确定第$i$句的时间范围, 如式 (4) 和式 (5) 所示:

$$\exists AW^{i-1}_{\text{LAST}}\rightarrow ST^{i}=ET^{i-1}_{AW^{i-1}_{\text{LAST}}} \tag{4}$$

$$\exists AW^{i+1}_{\text{FIRST}}\rightarrow ET^{i}=ST^{i+1}_{AW^{i+1}_{\text{FIRST}}} \tag{5}$$

(3) 如图 2 未对齐段 (1) 和对齐段 (3) 所示, 对未对齐的语句划分对齐段和未对齐段, 这部分语句的特征是相邻的语句中包含锚点但是开头和结尾没有锚点, 因此使用以下过程划分对齐段和未对齐段. 以第i句和第$i+1$句为例, 遍历这两句中的锚点, 确定第i句的最后一个锚点AW^{i}_{LAST}和第$i+1$的第 1 个锚点AW^{i+1}_{FIRST}, 根据锚点位置将语句划分为对齐段 (3) 和未对齐段 (1).

针对未对齐段和对应 ASR 结果, 如图 2(b) 所示, 递归执行第 1 步寻找锚点和第 2 步划分对齐段和未对齐段, 直到没有新的锚点出现为止. 这样通过缩小未对齐问题范围, 一方面可以再次寻找到锚点, 另一方面也可以阻止错误的传播, 保证对准算法的鲁棒性.

2.3 基于韵律信息的语句边界检测方法

针对仍然存在的未对齐段, 结合韵律信息检测语

句边界, 通过统计数据集中语句边界和非边界处停顿时长的分布发现, 语句边界处的停顿时间通常比短语处的时间要长, 说话人用较长时间的停顿作为语句的结束, 因此, 如果当前词和下一个词之间的停顿持续时间较长, 那么当前词可能对应语句边界[18].

语音端点检测 (voice activity detection, VAD) 是在一段包含语音的信号中分离出语音信号和非语音信号, 并确定语音信号的起始端点和终止端点[19]. 借助语音端点检测可以帮助提取音频的发音时长韵律信息, 有声书场景的语音语速适中, 节奏明显[20], 并且在专业录音棚中录制噪声较小. 因此, 本文根据有声书场景的音频特点, 采用了基于短时能量和过零率的双门限端点检测. 声音中包含浊音部分和静音部分, 静音包括清音、噪音和无声, 清音属于声音中的辅音, 能量小, 过零率高.

发音时清音和浊音之间的能量差别明显[21], 因此先利用短时能量提取浊音, 短时平均能量是语音信号平方经过一个窗函数的滤波输出所得到的信号:

$$E_n = \sum_{m=0}^{N-1} x_n^2(m) \tag{6}$$

接下来用过零率提取清音, 未对齐段第n帧语音信号$x_n(m)$的短时过零率z_n:

$$Z_n = \frac{1}{2}\sum_{m=0}^{N-1} |sgn[x_n(m)] - sgn[x_n(m-1)]| \tag{7}$$

根据获得的短时平均能量和短时过零率, 采用双门限端点检测方法进行未对齐段的语音端点检测, 结合 VAD 结果, 计算得到语音段的开始和结束时间以及各静音段的时长, 对语句边界做进一步估计, 从而解决 ASR 识别“音错字错”导致的语句失配问题, 如图 2(c) 所示, 静音段时长关系为Ⅲ> Ⅱ >Ⅳ> Ⅰ, 因此选取静音段Ⅲ作为未对齐段 (1) 的语句分界点.

3 实验分析

使用有声读物内容创作过程中的语音与文本资源构建了数据集, 通过对比实验检验了所提方法在语句级别的对齐任务上的准确度、鲁棒性和性能等.

3.1 评价指标

为了客观评价所提出的对准算法在语句级别的对齐准确率 (*Accuracy*), 准确率评分指标定义如式 (8):

$$Accuracy = \frac{TP}{TP+FP} \tag{8}$$

其中, FP表示带有错误时间戳的参考文本语句, TP表示正确对准的参考文本语句.

如果参考文本语句的时间戳 (ST_f^i和ET_f^i) 与对应的手工标注的开始和结束时间小于指定的误差阈值范围, 则认为是对齐的, 如式 (9):

$$TP = \sum_{i=1}^{M} sgn(\max(\left|ST_f^i - ST_{GT}^i\right|, \left|ET_f^i - ET_{GT}^i\right|) \leqslant Th) \tag{9}$$

其中, $sgn(\cdot)$是符号函数, 如果内部条件为真返回 1, 否则返回 0, ST_{GT}和ET_{GT}代表这个词的真实的开始时间和结束时间, M是语句总数, Th是容差阈值.

3.2 数据集

将有声读物内容创作过程中的语音和对应文本进行预处理并标记每句话在音频中对应的发音时间, 构建中文语句级的文语对齐数据集, 下载链接: https://github.com/xukai98/text_speech_alignment/blob/main/datasets.csv. 该基础数据集经过人工校准, 参考文本内容和音频的内容相符, 总数据集约有 10 h 的音频和 5 587 条参考文本的语句.

此外, 我们在基础数据集基础上, 将参考文本进行不同程度的插入、删除、替换, 构建不完全匹配的数据集, 对方法在语料不完全匹配情况下的表现以及不同类型的错误对对准方法的影响程度进一步评测.

除上述构建的有声读物数据集以外, 还使用 TIMIT 数据集验证本文的方法在经典语音识别数据集上的效果. TIMIT 数据集一共包含 6 300 个句子, 所有的句子都在音素级别上进行了手动分割和标记.

3.3 不同方法的语句对齐准确率分析

不同的 ASR 算法的识别准确率不同, 采用字错误率 (word error rate, *WER*) 指标[19] 进行评价:

$$WER = \frac{S+D+I}{S+D+H} \tag{10}$$

其中, S为替换的字数, D为删除的字数, I为插入的字数, H为正确的字数.

本节将基线方法[11] 与基于 DTW 的文语对齐方法[8] 和本文方法进行对比. 实验中, 使用 WER=0.1055 的 ASR 工具实现基线方法和本文方法.

表 1 给出了使用完全匹配的数据集, 目前先进的基于 ASR 的基线方法、基于 DTW 的方法 (DTW) 以及本文方法在不同容差水平下的对齐结果. 研究表明,

有声读物发音的节奏时间为 1 s 时不会影响听感[20], 因此选择 *Th*=1.0 s 来评测对齐准确率是合理的.

表 1 不同方法的语句对齐准确率

Th (s)	本文方法 (%)	基线方法 (%)	DTW (%)
0.1	4.81	3.60	5.75
0.2	7.82	6.21	21.83
0.3	7.82	6.21	22.98
0.4	7.82	6.21	22.98
0.5	12.62	11.02	25.28
0.6	64.53	62.12	27.58
0.7	98.19	95.79	31.03
0.8	99.59	97.19	41.37
0.9	100	97.59	54.02
1.0	**100**	**97.59**	**60.92**

表 1 结果显示, *Th*=1.0 s 时, 目前先进的基于 ASR 的基线方法语句对齐准确率高于 97%, 本文提出的基于锚点和韵律信息的文语对齐方法能达到高于 99% 的准确率, 而基于 DTW 的方法仅为 60.92%.

3.4 本文方法在 TIMIT 数据集上的对齐准确率分析

为了验证本文方法在经典语音识别数据集上的对齐效果, 将本文方法在 TIMIT 数据集上进行文语对齐实验. 图 4 给出了本文方法在不同容差阈值水平下的对齐结果.

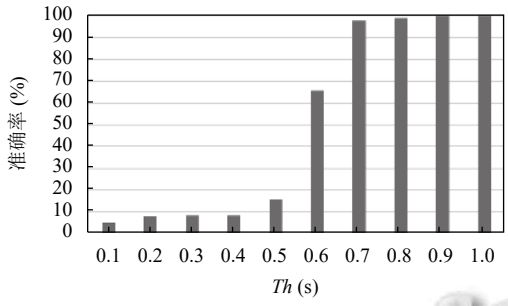

图 4 本文方法在 TIMIT 数据集的对齐准确率

实验结果显示, *Th*=1 s 时, 本文提出的基于锚点和韵律信息的文语对齐方法能达到高于 99% 的准确率.因为 TIMIT 语料库的音频和文本是完全匹配的, 并且语速均匀, 有利于本文方法中锚点搜索以及韵律信息检测, 所以在 *Th*=1 s 时能够有较高的准确率.该语料库在音素级别上进行了开始和结束时间的标注, 语句的起始时间更为精确, 当 *Th*≤0.5 s 时对齐效果并不理想, 当 *Th*=0.5 s 时, 仅有约 15% 的对齐准确率.

为了进一步验证本文提出的基于动态规划的锚点搜索方法在迭代多次时其结果是否可以趋于稳定, 我们在 TIMIT 数据集上设置该方法为多次迭代, 并记录迭代次数和对齐准确率的关系, 如图 5 所示. 本文所提的基于动态规划的锚点搜索方法是针对求解锚点和缩短未对齐段进行的循环迭代. 随着迭代次数的增加未对齐段范围不断缩小, 实验结果显示, 第 2 次迭代相较于第 1 次会有较大的提升, 而第 2 次到第 3 次迭代仅提升约 2%, 当该算法迭代次数超过 3 次时, 对齐准确率约为 97%, 对齐效果趋于稳定.

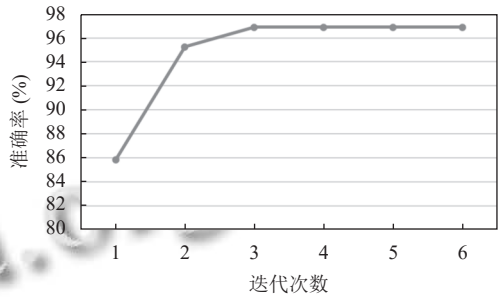

图 5 迭代次数对对齐准确率的影响

3.5 语音识别效果变差时对齐鲁棒性分析

为了验证在语音识别效果变差时基于语音识别的对齐方法鲁棒性差异, 本文在完全匹配的数据集上, 分别使用 *WER*=0.1055、*WER*=0.2533、*WER*=0.4083、*WER*=0.5223 的 ASR 引擎进行实验, 结果如图 6 所示. 用 P1 表示基于边界锚点加权的片段标注方法, P2 表示基于韵律信息的语句边界检测方法. ASR 表示目前先进的基于语音识别的基线方法, ASR+P1 表示在基线方法基础上添加了本文提出的基于边界锚点加权的片段标注方法, ASR+P1+P2 表示在基线方法基础上添加了基于边界锚点加权的片段标注方法和基于韵律信息的语句边界检测方法.

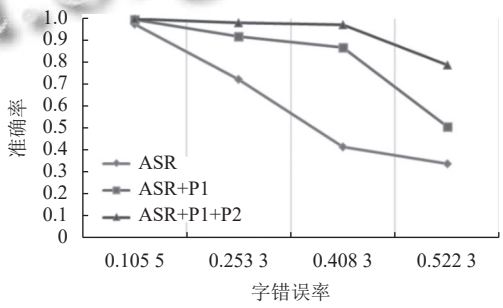

图 6 *WER* 对基于 ASR 的文语对齐方法的影响

可以看出一方面 ASR 的识别准确率对所有基于 ASR 的对准方法都是有影响的, ASR 识别正确率越高, 对准方法精度越高; 另一方面, 即使在一个识别准确率相对低的 ASR 引擎上实验, 本文方法依然比目前先进的基于 ASR 方法的对齐准确率更高, 并且受 ASR 识别率降低所带来的对准精度下降的影响更小.

目前先进的基于 ASR 的对准方法, 词错误率由 0.105 5 增大到 0.522 3 时, 准确率下降了约 50%, 而本文提出的对齐方法仅下降了约 20%.

产生这种结果的原因是针对“音对字错”的错误本文使用音节序列进行对准并且对位于语句边界的锚点在语句片段标注过程中赋予了更高的权重, 当语句中可用锚点因为 ASR 识别准确率下降减少时, 这些位于语句边界的锚点在对准过程中能够发挥更关键的作用. 虽然词错误率升高语音识别效果变差, 基于动态规划的锚点搜索方法无法求解出充足的锚点, 导致第一次对齐不能产生良好的对齐效果, 语句中存在大量未对齐段. 但是, 第一次对齐剩余的未对齐段就会进入到基于韵律信息的边界检测方法中, 检测未对齐段的韵律信息, 将未对齐段语句中的较长停顿作为语句的边界. 针对未对齐段的基于韵律信息的边界检测能够有效解决由 ASR 字错误率增高带来的语句失配问题, 对于音频和文本不完全匹配导致的第一次无法有效对齐的情况也能有效对齐.

3.6 数据集不完全匹配时对齐鲁棒性分析

进一步验证本文方法在不完全匹配数据集上执行对准任务的表现, 并分析不同的错误类型对不同方法的对齐准确率的影响. 为了评估算法在语料不完全匹配情况下的表现, 本文将先进的基于 ASR 的对准方法作为基线和基于 DTW 的方法以及本文方法在基础数据集上对参考文本语句进行不同比例的破坏, 破坏操作包括删除 (Del)、插入 (Ins) 和替换 (Sub).

表 2 中不同方法的对比结果显示, 删除错误对各种方法的对齐性能影响最大, 即使先进的基于 ASR 的方法的对齐准确率也会受到插入和替换错误的明显影响, 而本文所提方法可以有效降低插入和替换错误所导致的影响. 基于 DTW 的方法虽然能够降低插入和删除错误的影响, 但是该方法的对准精度相对于本文所提方法相差 20%.

表 2 数据集不完全匹配时对齐鲁棒性分析

破坏比例	DTW (%)			基线方法 (%)			本文方法 (%)		
	Del	Sub	Ins	Del	Sub	Ins	Del	Sub	Ins
0.1	58.62	63.21	63.21	93.38	88.17	92.38	98.39	99.59	99.39
0.2	58.62	58.62	62.06	65.33	72.34	84.97	76.95	98.59	99.59
0.3	57.47	57.47	59.77	40.28	54.11	71.34	57.72	98.39	98.59
0.4	43.67	57.47	59.77	13.83	41.48	55.91	23.25	97.39	98.59
0.5	**40.22**	56.32	58.62	**2.4**	33.66	49.49	**5.01**	97.39	98.38

3.7 文语对齐方法的时间开销分析

最后对算法执行效率进行评测, 我们在相同物理条件的机器上将本文所提方法与目前先进的基于 ASR 的方法和基于 DTW 的方法进行比较, 如图 7 所示.

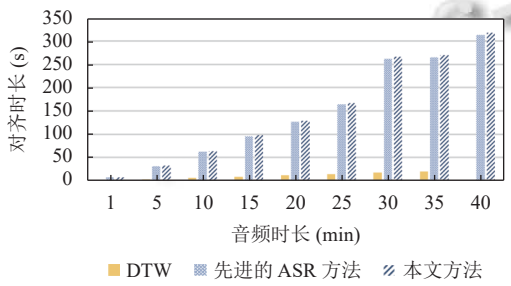

图 7 各方法执行效率对比

随着音频时长从 1 min 增加到 35 min, 基于 DTW 的对齐方法始终稳定在 50 s 之内, 但是当音频时长增加到 40 min 时, 基于 DTW 的对齐方法由于巨大的内存开销在此机器上无法完成对齐. 而基于 ASR 的方法由于模型的加载等需要耗费大量时间, 随着音频时长增加, 对齐时长也由 50 s 增加到 300 s, 与 DTW 的方法相比要慢很多, 而本文所提方法和目前先进的基于 ASR 的对准方法相比并没有因为附加的片段标注阶段和语句边界检测阶段导致额外明显的时间开销.

4 结束语

本文针对不完全匹配的长篇幅音频和文本对齐问题提出了一种基于锚点和韵律信息的语句级文语对齐方法框架, 改进了目前先进的基于 ASR 的对齐方法. 针对基于 ASR 的对齐方法中可能存在的“音对字错”和“音错字错”的识别错误, 通过制定赋予边界锚点更高权重的语句对准规则和针对未对齐段结合语音的韵律信息检测语句边界来提高基于 ASR 对准方法的准确率和鲁棒性.

本文构建了一个语句级中文的文语对齐数据集, 并在此基础上构建多个不完全匹配的文语对齐数据集, 为基于语音识别的语句级文语对齐算法在准确度、以及算法的对齐鲁棒性等方面的检验提供了参考.

本文仅采用了单一的韵律信息进行语句边界检测, 因此后续将考虑使用包含重音、基频等其他韵律特征进行辅助. 此外在低信噪比情况下, VAD 检查也会存在偏差, 因此后续仍需考虑在低信噪比情况下的语句边界检测.

参考文献

1 Moreno PJ, Joerg C, Van Thong JM, *et al*. A recursive algorithm for the forced alignment of very long audio segments. Proceedings of the 5th International Conference on Spoken Language Processing. Sydney: ISCA, 1998. 2711–2714.

2 Jo J, Kim HG, Park IC, *et al*. Modified viterbi scoring for HMM-based speech recognition. Intelligent Automation and Soft Computing, 2019, 25(2): 351–358.

3 Gorman K, Howell J, Wagner M. Prosodylab-aligner: A tool for forced alignment of laboratory speech. Canadian Acoustics, 2011, 39(3): 192–193.

4 McAuliffe M, Socolof M, Mihuc S, *et al*. Montreal forced aligner: Trainable text-speech alignment using kaldi. Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm: ISCA, 2017. 498–502.

5 Ravanelli M, Parcollet T, Bengio Y. The PyTorch-Kaldi speech recognition toolkit. Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019. 6465–6469.

6 牛米佳, 飞龙, 高光来. 蒙古语长音频语音文本自动对齐的研究. 中文信息学报, 2020, 34(1): 51–57. [doi: 10.3969/j.issn.1003-0077.2020.01.007]

7 Boháč M, Blavka K. Text-to-speech alignment for imperfect transcriptions. Proceedings of the International Conference on Text, Speech and Dialogue. Berlin, Heidelberg: Springer, 2013.536–543.

8 Readbeyond. Aeneas. https://www.readbeyond.it/aeneas/

9 Anguera X, Perez N, Urruela A, *et al*. Automatic synchronization of electronic and audio books via TTS alignment and silence filtering. Proceedings of the 2011 IEEE International Conference on Multimedia and Expo. Barcelona: IEEE, 2011. 1–6.

10 高红坤. 基于 SailAlign 的中文语音文语对齐的研究 [硕士学位论文]. 青岛: 中国海洋大学, 2015.

11 Mocanu B, Tapu R. Automatic subtitle synchronization and positioning system dedicated to deaf and hearing impaired people. IEEE Access, 2021, 9: 139544–139555. [doi: 10.1109/ACCESS.2021.3119201]

12 González-Carrasco I, Puente L, Ruiz-Mezcua B, *et al*. Sub-sync: Automatic synchronization of subtitles in the broadcasting of true live programs in Spanish. IEEE Access, 2019, 7: 60968–60983. [doi: 10.1109/ACCESS.2019.2915581]

13 Martín A, González-Carrasco I, Rodriguez-Fernandez V, *et al*. Deep-Sync: A novel deep learning-based tool for semantic-aware subtitling synchronisation. Neural Computing and Applications, 2021: 1–15.

14 Bordel G, Penagarikano M, Rodríguez-Fuentes LJ, *et al*. Probabilistic kernels for improved text-to-speech alignment in long audio tracks. IEEE Signal Processing Letters, 2016, 23(1): 126–129. [doi: 10.1109/LSP.2015.2505140]

15 韦向峰, 袁毅, 张全, 等. 富媒体环境下语音和文本内容的对齐研究. 情报工程, 2019, 5(2): 17–27.

16 Ahmed I, Kopparapu SK. Technique for automatic sentence level alignment of long speech and transcripts. Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon: ISCA, 2013. 1516–1519.

17 Teytaut Y, Roebel A. Phoneme-to-audio alignment with recurrent neural networks for speaking and singing voice. Proceedings of the 22nd Annual Conference of the International Speech Communication Association. Brno: ISCA, 2021. 61–65.

18 刘聪, 万根顺, 高建清, 等. 基于韵律特征辅助的端到端语音识别方法. 计算机应用, 2022: 1–6. [doi: 10.11772/j.issn.1001-9081.2022010009]

19 Sharma S, Sharma A, Malhotra R, *et al*. Voice activity detection using windowing and updated K-means clustering algorithm. Proceedings of the 2nd International Conference on Intelligent Engineering and Management (ICIEM). London: IEEE, 2021. 114–118.

20 Axtell B, Munteanu C, Demmans Epp C, *et al*. Touch-supported voice recording to facilitate forced alignment of text and speech in an E-reading interface. Proceedings of the 23rd International Conference on Intelligent User Interfaces. Tokyo: ACM, 2018. 129–140.

21 Sharma B, Gupta C, Li HZ, *et al*. Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models. Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019. 396–400.

(校对责编: 牛欣悦)