双阈值的特定英语音频句子边界检测

刘秉权,徐 帅,李相前

(哈尔滨工业大学 计算机科学与技术学院,哈尔滨 150001,liubq@insun.hit.edu.cn)

摘 要: 为了提高英语音频句子切分的效果,提出了基于双阈值的句子边界检测方法.该方法针对VOA、BBC等特别适合英语学习者的音频所具有的波形规范、环境噪声小、速率通常比较稳定等特点,利用静音能量阈值和静音时延阈值来检测音频句子的边界,并辅以对照文本信息进行校正.针对VOA慢速英语的实验结果表明:单纯使用双阈值方法,音频切分的召回率超过96%,精确率超过94%;利用对照文本校正后,可进一步提高精确率.

关键词: 音频切分;边界检测;双阈值

中图分类号: TP37 **文献标志码:** A **文章编号:** 0367-6234(2010)02-0259-05

Boundary detection of special English audio sentence based on dual-threshold

LIU Bing-quan, XU Shuai, LI Xiang-qian

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, liubq@insun.hit.edu.cn)

Abstract: To improve the effect of sentence-evel English audio segmentation, a method of sentence boundary detection based on dual-threshold is proposed. With consideration of the characteristics of normative waveform, small noise and stable speed for English audio such as VOA and BBC, the method in this paper detects the sentence boundary of English audio via its quiet energy threshold and quiet delay threshold, and the corresponding audio text is used to revise the segmentation result. Experiments on special English of VOA show that the recall rate of segmentation exceeds 96% and the precision rate exceeds 94% by using the dual-threshold method only. After revision via the corresponding text, the precision rate can be improved further.

Key words: audio segmentation; boundary detection; dual-threshold

英语音频例句在学习过程中可有效提高英语学习者的听力水平和学习效率.目前互联网上存在大量的以篇章为单位的英语听力资源,通常辅以对照的英文文本.近年来,人们开展了一些针对不同应用的音频句子切割也即句子边界端点检测的研究.文献[1]结合N-ram语音模型提出一个基于韵律的句子切割算法.基于此,文献[2]把该算法扩展到多特征融合的特定情节.文献[3]应用句子切割技术来生成音频对话的摘要.文献[4]提出一个没有词信息的音频切割系统,它只利用静音特征来把音频内容分成不同的语义级别,包括句子分割.文献[5]除了静音持续时间外,又结合了能量、过零率、语速和韵律等特征,得到了一个相对满意的结果.另外,还有一些相关的研究,比如,Dagen Wang等[6]提出了多路径线形折叠算法来处理自发语音中的句子边界识别问题;Joanna Mrozinski等[7]为自动语音摘要系统提出一个自动句子分割方法,该方法基于word和class级别的统计信息来预测句子和非句子边界;Yang Liu等[8]为丰富语音识别系统,提出了自动检测句子边界和不流利现象的方法,以便语音识别能输出有结构的信息.为此本文提出了一种简单而有效的音频切分方法,是基于能量和静音时延双阈值的.

收稿日期: 2008-12-16.

基金项目: 国家自然科学基金资助项目(60673037);国家高技术研究发展计划资助项目(2006AA01Z197);黑龙江省自然科学基金资助项目(E200635).

作者简介: 刘秉权(1970—),男,博士,副教授.

1 边界检测方法

通过对大量样本的统计分析发现,不同语言单位(段、句、词等)的分割处,语音的某些特征有明显变化,比如几乎在每个语言单位之间,都会有一个停顿;在句子边界处,音频的能量特征就显著减少,这是用来识别句子边界的最显著特性. 于是,尝试捕捉这一特点来做边界检测,即通过分析音频能量特征的变化趋势来预测语言单位的边界.

1.1 短时分析原理

语音具有其特殊的产生机制,这导致语音信号是一种典型的非平稳信号. 不过,由于人所能听到的声音比起声音振动速度来讲要缓慢很多,因此现在的语音信号处理基本上都是基于这样的假设——语音信号具有短时平稳性,即语音信号特性随时间的变化是缓慢的,或语音信号在一个时间帧内是平稳信号. 这样,可以利用平稳信号的分析方法对其进行处理了.

短时平均过零率、短时平均能量和平均幅度等,都是在这种短时平稳假定下从时域来分析的一些物理量. 这种时间依赖变量处理的基本手段,一般是用一个长度有限的窗序列($w(m)$)截取一段语音信号来进行分析,并让这个窗滑动以便分析任一时刻附近的信号.

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)] \cdot w(n-m). \tag{1}$$

式中: $T[\cdot]$ 为要对输入信号所作的运算, $x(m)$ 为输入信号序列, $\cdot$ 为点积运算.

1.2 能量阈值

在计算能量值之前,需要对音频信号序列加窗. 通常使用的窗函数包括矩形窗、海明窗、汉宁窗等等.

1)矩形窗:

$$w(n) = \begin{cases} 1, 0 \leqslant n \leqslant N-1 \\ 0, \text{其他}. \end{cases} \tag{2}$$

2)汉宁窗:

$$w(n) = 0.5\left(1 - \cos\left(2\pi \frac{k}{N+1}\right)\right), \quad n = 1,2,\cdots,N. \tag{3}$$

3)海明窗:

$$w(n) = 0.54 - 0.46\cos\left(2\pi \frac{k}{N-1}\right), \quad k = 1,2,\cdots,N. \tag{4}$$

式中: N 为窗长.

本文使用了简单而有效的海明窗. 窗长为20 ms,窗是没有搭接的.

加窗之后,可以得到 m 帧的短时能量 E.

$$E_n = \sum_{m=0}^{n} [x(m)\, w(n-m)]^2. \tag{5}$$

式中: $x(m)$ 为原始的音频信号样本序列, $w(n-m)$ 为窗函数.

不同的声音段有不同的能量,通常来说,停顿段的能量要比平均能量小很多. 因此可以估计一个能量阈值,如果一帧的能量小于这个阈值,则认为它在一个停顿段里. 有两种方法来估计这个阈值:

1) $E_t = K * \text{Energy\_Average\_Value}, (K \ll 1)$. 这种方法阈值是可以变化的,以适应不同种类的音频.

2)可以针对不同类型音频手动计算大量的阈值. 这种方法有着更好的效果,只是需要预先统计工作,并且阈值是固定不变的.

1.3 静音时延阈值

利用静音能量阈值可以很好地检测到语音的边界,但是并不是只有句子边界处才有语音的消减,不同的语言单位之间都存在这种现象,比如段与段之间,分句与分句之间,甚至词与词之间都可能被判为检测点. 不同语言单元间的确都存在波形幅度衰减的现象,除了衰减的幅度不同外,还有一个明显的特征即衰减持续的时间不同. 段与段之间,衰减最明显并且持续时间最长,句与句之间,衰减也比较明显,只是持续时间略短,分句与分句之间,持续时间更短,而词与词之间的衰减幅度和持续时间都不甚明显. 鉴于这一特征,利用静音时延阈值来区分. 对每类音频的不同语言单元之间尤其是句与句之间的静音段进行分析,统计出它们的平均段长和最短段长,然后就可以采取某种策略得到静音时延阈值. 比如,利用平均段长乘以一个小于1的系数或直接利用最短段长,如果加的窗是无重叠的,只需将选取的段长除以窗长,即是静音时延阈值.

1.4 双阈值方法

利用能量阈值和静音时延阈值,可以识别出音频句子边界. 如果能量低于能量阈值并且持续时间超过时延阈值,就可以认为是一个句子的边界. 但是这种方法不能处理一些特殊情况. 当两种检测方法都满足,但不是句子边界的时候,就不能处理了. 在本文中,使用了和音频相关的文本来对这种方法进行校正.

2 文本的作用

几乎所有的广播电台在播出新的音频时,都

有针对的文本发布. 这些文本不同于文语转换中得到的文本,它们是经过编辑的,有标点符号的规范的准确的文本. 同特定音频一样,可以方便地从网络上获得. 而且从国内外的研究现状可以看出,文本句子边界检测已经取得了非常高的准确率. 因此如果音频有对应的文本,可以用它来解决错误切分的校正问题. 首先,检测出文本的句子边界,然后根据取得的信息,来消除错误的切分.

2.1　文本句子边界检测

在对文语转换的文本进行处理时,更倾向于利用复杂的模型来取得满意的结果,不过这样计算量偏大,不适合用于实际的系统中. 其实音频对应的文本如同音频本身一样,具有严格的规范性:整洁、标点等重要信息正确. 在文本的句子末尾,通常有一个表示句子结束的标点符号,比如"."、"?"、"!"等. 但是,这些标点符号并不是专有的,有时它们可能出现在句子中间. 比如说,在"Mr. Smith"中,"."就不是句子的结尾符. 因为音频相关的文本是很规则的,因此可以建立一些规则来处理这些特殊的情况. 首先需要建立一个前缀词典,包括"Ms."、"Mr."、"U. S."等等. 规则定义为(结果 = T:如果标点是句子结尾符;结果 = F:如果标点不是句子结尾符):

1)如果下一个字符不是空格,那么结果 = F.

2)如果下一个字符是空格,并且当前标点是"?"或"!",则结果 = T.

3)如果下一个字符是空格,并且当前标点是".",如果空格后是个小写字母,则结果 = F.

4)如果下一个字符是空格,并且当前标点是".",空格后是个大写字母,如果标点前字符不在前缀词典里,则结果 = T,否则结果 = F.

5)如果下一个字符是空格,并且当前标点是".",如果空格后是其他字符(不是大写也不是小写字母),结果 = T.

用这些规则来检测句子边界,最终可以得到句子数. 然后和用双阈值切分出来的结果比较. 根据差别,可以判断音频切分是否有误检,并可以估计误检的数量.

2.2　双阈值校正

利用被误切的句子所含的字符数和音频占有的时长都会变短的事实. 初步考虑是利用单句长阈值. 所谓单句长阈值就是标识一句话长度的阈值,如果某句子的长度小于该阈值,就认为该句子是被误切的. 但这样存在一个明显的问题,就是如果音频中最短句子的句长(记为 L_1)小于被误切句子的较短的部分(记为 L_2)时,处理起来就比较麻烦. 比如,如果单句长阈值过小($< L_1$),由于 $L_2 > L_1 >$ 单句长阈值,所以检测不到 L_2,也即检测不到误切点;如果单句长阈值过大($> L_2$),这样虽能检测到 L_2,因为单句长阈值 $> L_2 > L_1$,所以同时也会检测到 L_1,即把最短句当作误切句而删掉. 因此,这种情况下确定一个合适的单句长阈值是比较困难的.

不过,可以借助短句(小于单句长阈值的句子称为短句)前后两句的力量来加以判断. 对于误切句,其前后两句中必定有一句是被误切的另一部分,因此,把误切句和其前后两句加起来其实是音频中实际的两句话. 而对于非误切句,加起来后就是实际的三句话. 从概率角度讲,实际文本或音频中,三句话的长度大于两句话的长度是个大概率事件. 基于此,可以利用三句长阈值,即短句加上其前后两句的长度的阈值,结合单句长阈值一块来解决上述问题. 单句长阈值和三句长阈值为

$$\text{singleThreshold} = k \times \frac{\text{TimeLength}}{\text{SentenceNumber}},\quad 0 < k < 1. \tag{6}$$

$$\text{ThreeThreshold} = 3 \times k \times \frac{\text{TimeLength}}{\text{SentenceNumber}},\quad 0 < k < 1. \tag{7}$$

式中:TimeLength 为音频的时间长度,可以从解码后的 WAV 头文件里的信息计算而得,SentenceNumber 为文本切分后得到的句子数,文本切分完成后保存有这一信息. 这样,TimeLength/SentenceNumber 得到的就是单句的平均时间句长. K 为正常量($k < 1$).

3　实验结果分析

针对不同种类的特定英语音频,都有相似的处理方法和相似的特性. 本文的实验以 VOA Special English 为例来进行结果分析. 实验选取 28 组时长为 4 min 左右的 VOA Special 音频,它们全部来自中国日报网站的英语点津子站(http://www.chinadaily.com.cn/language_tips/index.html). 28 组音频颇具代表性,内容上涉及经济学、教育、健康、发展、农业、新闻等热门领域,时间上主要有 2006 年 8、9 月份和 2007 年 5、6 月份的报道组成,播音来自常见的 6 个播音员,其中:3 个男声:Steve Ember、Doug Johnson 和 Mario Ritter,3 个女声:Barbara Klein、Katherine Cole 和 Faith Lapidus 等.

首先,系统对从网络上下载的 MP3 音频进行解码操作. 这样就得到 WAV 波形文件. 然后,对

WAV 文件进行加窗并提取音频特征,窗函数选海明(Hamming)窗,窗长为 20 ms,窗之间无重叠. 最后,利用基于双阈值的方法对其进行切割. 应注意的是,系统也同时对对应的文本信息进行按句切割,其中文本信息是通过将音频相应的网页文字拷贝到. txt 文件而得到. 切割时,28 组音频都采用默认的阈值,即统一的静音能量阈值和统一的静音时延阈值. 实验结束得到三部分数据:MP3 解码数据、文本切分数据和基于双阈值的音频切分数据. 为了检测辅助文本在音频切分中的作用,简单调整系统,可分别得到校正前的音频切分数据和校正后的音频切分数据.

实验结果的几个评价参数为:

1) 召回率. 召回率 = 检测出的正确切分点数目/真实的切分点数目.

2) 精确率. 精确率 = 检测出的正确切分点数目/检测出的所有切分点数目.

3) 误检率. 误检率 = 检测出的错误切分点数目/检测出的所有切分点数目.

4) 漏检率. 漏检率 = 未检测出的正确切分点数目/检测出的所有切分点数目.

对文本切分数据和校正前后的音频切分数据进行详细分析,分析的具体方法是对切分后得到的数据手工逐个检查.

3.1 文本切分结果分析

文本切分采用的是基于规则的切分方法. 切分对象是 28 组音频对应的 28 篇文本信息. 实验结果如表 1 所示.

表 1 文本切分结果

切分得到的句子数	实际句子数	召回率/%	精确率/%
836	835	1	99.9

其中唯一一处误检是由于对人名 John F. Kennedy 进行检测时,将 F 后的“.”当作了句子边界. 原因是由于缩略语词典里没有收录“F”,无法将“F.”识别为前缀缩写. 总体来看,文本切分的结果是非常令人满意的. 这很好地证实了特定音频对应的文本具有规范、整洁的特点. 从而,文本切分工作能保证为音频提供很准确的对照文字,并为基于此进行的校正工作提供了可靠的基础.

3.2 基于双阈值的音频切分结果分析

基于双阈值的音频切分算法对 28 组音频文件处理后得到结果数据如表 2 所示.

在得到的 856 个句子边界中,存在两种类型的错误:1)错误检测点,即把非句点判作了句子边界,这类错误共 52 个. 2)漏检测点,即实际的句子终结符号没有被判作句子边界,这类错误共 30 个.

$$\text{切分得到句子数} = \text{实际句子数} - \text{漏切割点} + \text{错误检测点}. \tag{8}$$

表 2 双阈值音频切分结果

切分得到的句子数	实际句子数	召回率/%	精确率/%
856	835	96.41	94.04

错误的参数分析如表 3 所示.

表 3 错误分析表

切分得到的句子数	实际句子数	召回率/%	精确率/%
52	30	6.07	3.50

在形式上,错误多来自两种错误的切割方式:

1)一分为二型. 典型的一分为二型即把一句完整的句子从中间切开,分成了两句话. 这是造成错误检测点的主要原因. 在分析过程中,也发现有把一句话切分成三句话的情况. 不过这种情况比较少见,一共只出现 1 次.

分析发现,切割点多出现在特殊标点符号处,比如“- -”和“;”等,或语义上具有并列、转承和举例等的地方,或短语和播音员特别强调的地方. 比如(斜体边界即为切割处):

1) The other finding was that the reduction happened mainly among older women - - the main users of the therapy.

2) The protests resulted in the choice of a different president who is more popular with the students, Robert Davila.

3) Rajat Nag is managing director general of the bank; we called him in Manila.

4) In 2004, Thomas Matrka was a graduate student in engineering.

文本辅助的音频切分算法,就是尝试利用文本信息来校正这种错误情况,以减少错误检测点.

2)合二为一型. 典型的合二为一型是指把两句完整的话连到了一起,当作一句话了. 这样每出现一次这种情况,就造成了一个漏检测点. 还有一种情况就是一个完整的句子和邻近的半句连在了一起,这样不仅会漏掉一个检测点还会增加一个错误的检测点.

对合二为一型的检测结果进行分析,发现更具规律性. 它们大都具有语义上的联系,比如存在并列、承接、转折、因果等关系. 例如:

1) You might suddenly feel lightheaded. You

might also feel sick to your stomach.

2) Wearing safety protection like elbow pads and leg guards during activities is a good idea. If you think these might be restrictive, try a cast.

3) Now the wool is orange or yellow. Or at least it should be.

4) The WTO tried to launch a ninth round in Seattle in 1999. But trade ministers argued and free trade opponents rioted.

这种句子连在一起通常能表达一种更加清晰的意思,所以这种错误的切分方式,只要不是把一个完整的句子和一个半句连在一起,通常不影响在实用系统中使用. 比如在英语学习系统中,作为例句来解释句中的某个单词时,就不会降低使用的效果.

3.3　文本校正的音频切分结果分析

文本除了能给切割后得到的音频句子提供对应的文字,丰富资源库,提高学习效率,还应该具有更广泛的应用. 本文尝试利用文本信息来提高音频切割算法. 主要是利用文本切分后得到的句子数,来判断音频切割后检测到的序列中是否含有一分为二切割方式造成的错误检测点. 实验结果数据及错误分析如表 4 和表 5 所示.

表 4　文本校正后的音频切分结果

切分得到的句子数	实际句子数	召回率/%	精确率/%
851	835	96.29	94.48

表 5　错误分析表

切分得到的句子数	实际句子数	召回率/%	精确率/%
48	31	5.64	3.64

本实验中,双阈值文本校正算法一共校正了 4 处错误检测点,但同时也带来了 1 处漏检测点(将两句完整的话连到一起了). 可见,基于单句长阈值和三句长阈值的文本校正算法能在一定程度上增加切分结果的准确性.

4　结　论

1) 基于规范音频很少有环境噪声、速率通常比较稳定等特性,本文提出一种双阈值方法进行英语音频句子切分,该方法利用静音能量阈值和静音时延阈值来检测音频句子的边界. 同时介绍了一种基于规则的英文文本句子切分方法,并利用切分好的文本信息对音频切分结果进行校正.

2) 针对 VOA 慢速英语的实验结果表明:单纯使用双阈值方法,音频切分的召回率超过 96%,精确率超过 94%;相对而言,文本切分的精确率和召回率接近 100%,其可靠性明显高于音频切分;利用切分后的对照文本校正后,可进一步提高音频切分的精确率.

参考文献:

[1] SHRIBERG E, STOLCKE A. Prosody-based automatic segmentation of speech into sentences and topics[J]. Speech Communication, 2000, 32(1/2): 127 – 154.

[2] SHRIBERG E, STOLCKE A, BARON D. Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, disfluencies, and overlapping speech[C]//Proc ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding. [s. n.]: [s. l.], 2002: 139 – 146.

[3] ZECHNER K. Automatic generation of concise summaries of spoken dialogues in unrestricted domains[C]//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2001: 199 – 207.

[4] PFEIFFER S. Pause concepts for audio segmentation at different semantic levels[C]//Proceedings of the Ninth ACM International Conference on Multimedia. New York: ACM, 2001: 187 – 193.

[5] WANG D, LU L, ZHANG H J. Speech segmentation without speech recognition[C]//Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal. Washington: IEEE Xplore Digital Library, 2003: I-468 – I-471.

[6] WANG D G, SHRIKANTH S. A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, 2004 Proceedings. N. J.: IEEE Piscataway, 2004: 525 – 528.

[7] MROZINSKI J, WHITTAKER E W D, CHATAIN P, *et al.* Automatic sentence segmentation of speech for automatic summarization[C]//Proceedings 2006 IEEE International Conference on Acoustics, Speech and Signal. Washington: IEEE Xplore Digital Library, 2006: I – 981 – I – 984.

[8] LIU Y, SHRIBERG E, STOLCKE A, *et al.* Enriching speech recognition with automatic detection of sentence boundaries and disfluencies[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(5): 1526 – 1540.

(编辑　张　红)