用于大规模语料库建设的一种汉语语句切分方法

A Chinese sentence segmentation method for large-scale corpus construction

吴树兴，张秀琴（北京信息职业技术学院，北京 100015）
Wu Shuxing，Zhang Xiuqin（Beijing Information Technology College，Beijing 100015，China）

摘　要：
在语音识别和语音合成大规模语料库的构建中，需要把较长内容的语音文件切分成一定字数要求的语音数据文件和对应的文本文件。一种有效的自动切分方法是通过对单字占用时间的预测和元音主体数目的统计来评估切分点的位置，实现对语音数据的自动切分。实验表明，使用该方法进行切分的一次成功率可以达到92.8%，参数适当调整后的二次成功率为96.3%。整个切分过程中，进行人工调整的工作量很小，适合快速构建大规模语料库。

关键词：
元音主体；切分；语音合成；能熵比法
doi：10.12045/j.issn.1007-3043.2019.08.015
中图分类号：TN912.3
文献标识码：A
文章编号：1007-3043（2019）08-0070-04

Abstract：
In the construction of large-scale corpus of speech recognition and speech synthesis，it is necessary to divide the audio files of longer content into audio data files and corresponding text files with a certain number of words. An effective automatic segmentation method is to calculate the position of the segmentation point by counting the time occupied by the word and the statistics of the number of vowel subjects，so as to realize automatic segmentation of the audio data. Experiments using this method for segmentation show that the first success rate can reach 92.8%，and the secondary success rate after proper adjustment of parameters is 96.3%. During the entire segmentation process，the amount of manual adjustment is small，which is suitable for the rapid construction of large-scale corpus.

Keywords：
Vowel body；Segmentation；Speech synthesis；Energy entropy ratio method

引用格式：吴树兴，张秀琴. 用于大规模语料库建设的一种汉语语句切分方法[J]. 邮电设计技术，2019(8)：70-73.

0 前言

在语音识别和语音合成技术中，经常需要构建大规模训练语料库[1-6]。人工进行录制是建设语料库的常用方法，但这种方法建设周期长、投入的人力巨大[7-9]。近年来，许多学者尝试将语音识别技术引入到语料库建设中来[10-11]，其建设周期大幅缩短，同时减少了人力，但在语料库数据量非常大时，对错误进行人工调整也是非常耗时的。随着互联网的发展，音频资源越来越丰富，同时获取也更加方便、快捷，例如各种评书故事资源、各种讲故事栏目资料，可以充分利用这些语音资源和文本资料来构建语料库，但需要对这些资源重新进行加工处理。其中，将大段语料分成多个句子在构建语料库中是必不可少的，实现自动、准确的切分[12-13]，能够减小人工进行校正的工作量，缩短建设周期[14]。

在大规模语料库的构建中，需要把较长内容的语音文件分割成一定字数要求的语音数据文件和对应的文本文件，关键是除了分别将语音文件和对应文本进行正确切分外，还能够将切分的语音文件与文本内容准确无误地相对应。下面将具体描述本文所提出

基金项目：北京市教育委员会科技计划（KM201410857001）
收稿日期：2019-05-04

的一种比较有效的汉语语句自动切分方法。

1 汉语语句自动切分方法的总体结构

汉语语句自动切分方法的目标是将较长的语音文件和对应的文本文件分割为较短的多个语音文件和相应文本文件，并且每个语音文件的起始部分和结束部分需要具有一定的静音段。实现的主要思想是首先按照文本字数预估数据长度，在整个语音数据中预估数据段之后开始搜寻静音位置，确定静音位置后，在这段数据上采用能熵比法进行元音主体数目的判断，找到对应文本所包含字数的语音段，最后再确定精确的静音位置，完成句子的切分。

图1给出了该方法的总体框图结构，从图1中可以看出，该汉语语句自动切分方法的总体结构由语音数据初始处理模块、文本分句模块、噪声与信号门限估计模块、初始单字占用时间估计模块、数据截取模块、元音主体数目统计模块、单字占用时间预测模块、数据切分模块等组成。各个模块功能和作用如下：

a）语音数据初始处理模块：该模块负责读入要切分的语音数据，主要进行幅度归一化处理，迟滞处理，通过算术平均做数据平滑处理，得到与输入语音数据长度相等的数据Out_max_ave，最后将处理好的数据输出给噪声与信号门限估计模块、初始单字占用时间估计模块、数据截取模块、元音主体数目统计模块和数据切分模块。

b）文本分句模块：该模块将一个较长的汉语文本按照要求切分成若干句子，每个句子存储为一个文件，统计出整个汉语文本的字数输出给初始单字占用时间估计模块，同时统计每个句子的字数输出给数据截取模块和元音主体数目统计模块。

c）噪声与信号门限估计模块：依据语音数据初始处理模块得到的数据来确定噪声与信号的门限值。为初始单字占用时间估计模块、数据截取模块、元音主体数目统计模块、数据切分模块4个模块提供噪声与信号的门限值T1。

d）初始单字占用时间估计模块：依据语音数据初始处理模块的输出数据、噪声与信号估计门限值以及文本分句模块统计的总字数来估算初始单字占用时间。

e）数据截取模块：将初始单字占用时间、Out_max_ave、T1，数据切分模块得到的前一句切分数据结束点和待切分句子字数作为输入数据，粗略估计待切分句子的切分点，输出到元音主体数目统计模块。

f）元音主体数目统计模块：该模块通过对元音主体数目的统计来对待切分句子的切分点进行语句级评估。模块中主要使用能熵比法进行端点检测和元音主体数目估计，具体实现请参见文献[15]和[16]。

g）数据切分模块：依据Out_max_ave、T1和元音主体数目统计模块输出的静音位置点，计算出待切分句子的精确切分位置，然后从整个原始数据中截取数据，存储为语音数据文件。

2 方法具体实现

该方法具体实现时，首先，需要准备好待切分的语音数据和对应的汉语文本文件，对这些文件进行统一编号和命名，放在指定文件夹里，并建立存储切分文件的文件夹。使用C语言、Matlab语言、Python语言等易于编程的语言来实现该切分方法，根据文件数量设置循环次数。下面是一个数据的切分处理过程，可以分成以下8个步骤来实现。

第1步，首先处理文本文件，将一个汉语文本按照

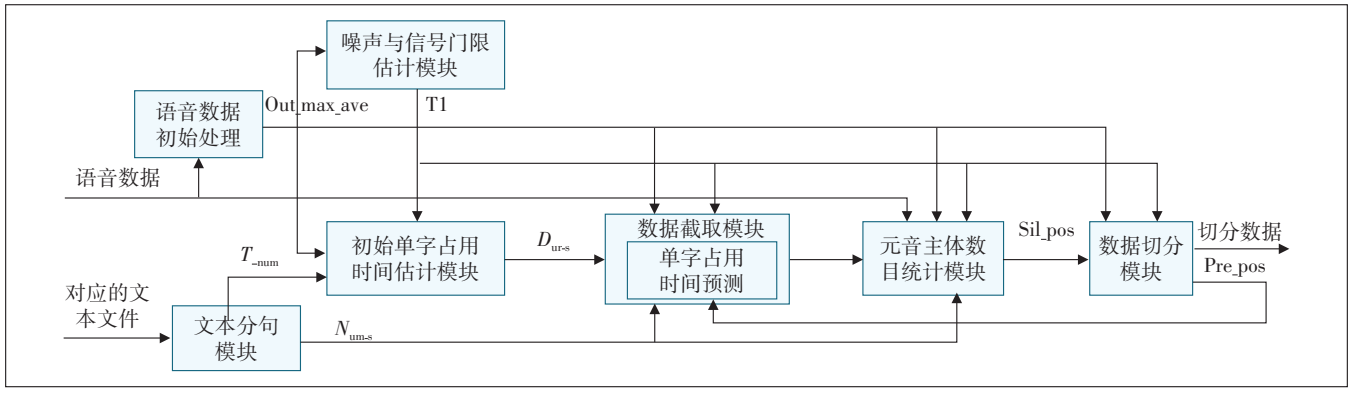

图1　一种汉语语句自动切分方法总体框图结构

要求切分成若干句子，每个句子存储为一个文件。例如要求切分的每个句子要大于等于5个汉字，遇到句子结束标点符号时则不限制字数，这步操作由文本分句模块来实现。然后将要切分的语音数据输入到语音数据初始处理模块中，做幅度归一化处理和迟滞处理，再将前后若干个数据取算术平均值作为该点的输出数据记为Out_max_ave。

第2步，将Out_max_ave输入到噪声与信号门限估计模块中，采用试探法来确定门限值，设门限值为k×0.001，k为正整数，将k由小至大设定不同的值，得到不同的门限值，对整个语音文件的语音部分采样点数使用计数器进行统计，计算不同门限值下语音采样点数变化率，当这个变化率较小时，说明门限值是稳定的，也就说明对应的k值是合适的，此时的k×0.001就是要寻找的噪声与信号估计门限T1。

第3步，使用语音数据Out_max_ave、噪声与信号估计门限T1和文本分句模块输出的总字数$T_{_num}$，计算初始单字占用时间。将语音数据Out_max_ave所有的点，只要大于T1就认为是语音数据部分，使用计数器统计语音部分的采样点数M，使用公式$D_{ur_s}=M/f_s/T_{_num}$，得到初始单字占用时间，其中，f_s是采样率。

第4步，将语音数据Out_max_ave、T1、上一步得到的初始单字占用时间D_{ur_s}和待切分句子字数N_{um_s}输入至数据截取模块，获得待切分句子的切分点。其方法是对数据Out_max_ave从头向后搜索，当大于门限值T1的采样点总数大于等于$N_{um_s}\times D_{ur_s}\times 2\times f_s$时，该采样点为待切分句子的结束点，记为pos_end，跳转至第6步。

第5步，将语音数据Out_max_ave、T1、前一句语音数据、前一句切分数据结束点Pre_pos和待切分句子字数输入至数据截取模块，获得待切分句子的切分点。

a）统计前一句语音数据的语音采样点数（值大于T1），记为$v_{_counter_s}$，用$t_1=v_{_counter_s}/f_s/N_{um_s}$得到前一句的平均单字占用时间，这里字数$N_{um_s}$是由文本分句模块计算的前一句的字数。

b）然后根据计算得到的平均单字占用时间来预测当前句所占用时间。如果当前句之前的第3句占用时间为t_3，当前句之前的第2句所占用时间为t_2，当前句之前的第1句所占用时间为t_1，预测当前句单字占用时间为$t=\frac{1}{4}\times t_3+\frac{1}{4}\times t_2+\frac{1}{2}\times t_1$（如果当前句是第2句，则$t=t_1$，如果当前句是第3句，则$t=\frac{1}{4}\times t_2+\frac{3}{4}\times t_1$），因此，得到当前句所占用时间为$t_{_cur}=t\times N_{um_s}\times 1.3$，其中，$N_{um_s}$为当前句字数，系数1.3是为了保证能够取到足够的数据量。这一部分对应着数据截取模块中的单字占用时间预测。

c）从数据切分模块输出的前一句语音数据的结束点Pre_pos开始向后搜索Out_max_ave，使用T1作为门限，找到待切分句子的结束点（当大于T1的采样点总数大于等于$t_{_cur}$时），记为Pos_end。

第6步，从上一步得到的pos_end位置开始向采样点索引减小方向对语音数据Out_max_ave进行搜索，如果搜索到连续1 s Out_max_ave的值小于T1，停止搜索，该位置即为静音段，记为v_end，将前一句语音数据的结束点Pre_pos作为起始点，v_end作为结束点截取数据，输入到元音主体数目统计模块中，得到这段数据的元音主体个数，由于音节数目与元音主体数目是一致的，也就得到了这一段数据的字数，将这一字数与文本统计出的字数进行比较，如果这一字数大于文本统计出的字数，继续向采样点索引减小方向搜索，先搜索语音段，如果搜索到连续0.05 s Out_max_ave的值大于T1，该位置即为语音段，继续向采样点索引减小方向搜索静音段，如果搜索到连续0.3 s Out_max_ave的值小于T1，停止搜索，该位置即为静音段，将该位置记为Sil_pos，去掉该点后边的数据，得到的数据段，输入到元音主体数目统计模块中，得到这段数据的元音主体个数，继续将这一字数与文本统计出的字数进行比较，如果这一字数大于文本统计出的字数，则重复这一过程，如果小于，则停止。找到与文本统计出的字数最接近的元音主体数目，相应的静音段即为所寻找的静音段，更新Sil_pos值。

第7步，将第6步得到的静音段的位置点Sil_pos，输入到数据切分模块。在数据切分模块中，由位置点Sil_pos开始分别向采样点索引减小方向和采样点索引增大方向搜索，如果搜索到连续0.1 s Out_max_ave的值大于T1，则认为搜索到语音段边界，较小索引值记为begin_s，较大索引值记为end_s，中间索引值（begin_s+end_s）/2向下取整作为当前句语音数据的结束点，将前一句语音数据的结束点Pre_pos作为当前句语音数据的起始点（如果当前句是首句，则当前句语音数据的起始点即为原始数据的起始点），从整个原始数据中截取数据，存储为当前句的语音数据文件。最后使用当前句语音数据的结束点来更新Pre_pos。

第8步，如果整个数据都被切分完成，则结束，否

者重复执行第5步至第7步。

3 切分效果评估

为了验证前面所提出的汉语语句自动切分方法，对单田芳的评书资源《白眉大侠》进行切分，该评书一共有320集，每集对应一个语音文件和一个文本文件，每集大约20多分钟，共100多小时的语音资料。

要求切分的每个句子要大于等于5个字，遇到句子结束标点符号时则不限制字数，按照此要求，每一集只有把全部的句子正确切分出来才算正确切分，首先对320集的语音文件和文本文件使用相同的参数，进行切分，能够成功完成切分297集，23集发生错误，成功率为92.8%，称为一次成功率，我们再对发生错误的23集，每一集调整合适的参数，使得能够进行正确切分，最终可以成功完成切分11集，12集无法自动进行切分，成功率为96.3%，称为二次成功率。

将这12集进行手动切分，使用Matlab工具将语音数据读入，找到位于文件中间段落间停顿（也称为静音段）较大的位置，取该静音段的中间位置作为切分点，将该文件一分为二，然后再将这个静音段之前一句和之后的一句话听辨出来，根据所听辨出的文字，在对应的文本文件中找到该静音段所对应的位置，将对应的文本文件按照这个位置一分为二，这样就将语音文件和对应的文本文件各自分成了2个文件，并进行存储，无法自动切分成功的12集语音数据和对应文本数据，按照此方法进行切分，共形成24个语音文件和24个文本文件，然后再使用自动切分算法进行自动切分，仍然有5个文件发生错误，对这5个文件再进行人工切分，每个文件仍然一分为二，形成10个语音文件，再对文本文件进行切分，也形成10个对应文本文件，再进行自动切分，…，直至完成所有句子的切分。

最后共切分出125 928个句子，整个过程共耗时45 h，时间主要用在参数调整和手动切分上。在手动切分过程中发现最终有32个句子仍然无法切分出正确结果，这些语音文件和文本不采用，产生这些错误的原因是语音中存在其他干扰，因此无法进行切分。

4 结束语

本文提出了一种用于大规模语料库构建中的汉语语句自动切分方法。它可以把较长内容的语音文件切分成一定字数要求的多个语音数据文件和对应的文本文件。使用该方法，对100多小时的语音文件和文本文件进行了切分，实验结果表明，这里所提出的语句自动切分方法相比于传统的录音方案和语音识别再进行人工调整的方案，准确率高，并且人工参与的工作量非常小，适合快速构建语料库。另外，该方法在参数的适应性方面仍然存在改进的空间。

参考文献：

[1] 蔡莲红，赵世霞. 汉语语音合成语料库的研究与建立[J]. 语言文字应用，1999(3)：97-102.
[2] 袁家宏. 大规模语音语料库的采集、处理和研究[J]. 语言学研究，2017(1)：34-42.
[3] 曲维光，唐旭日，俞敬松. 超大规模语料库精加工技术研究[J]. 当代语言学，2009(4)：136-146+190.
[4] 才让加. 面向自然语言处理的大规模汉藏(藏汉)双语语料库构建技术研究[J]. 中文信息学报，2011(11)：157-161.
[5] 凌震华. 基于统计声学建模的语音合成技术[D]. 合肥：中国科学技术大学，2008.
[6] BARRA-CHICOTE R，YAMAGISHI J，KING S，et al. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech[J]. Speech Communication，2010，52(5)：394-404.
[7] 孙岭，胡郁，王仁华. 中文语音合成系统中的语料库设计[C]// 第6届全国人机语音通讯学术会议.2001.
[8] ZHU W. Corpus building for data-driven TTS systems[C]// IEEE Workshop on Speech Synthesis. 2002.
[9] 俞炯，贺琳. 言语合成语料库的制作[C]// 第7届全国人机语音通讯学术会议，厦门，2003：292-295.
[10] 汤胜良，张士礼，张志平，等. 基于新闻联播语料库的语音合成系统[C]// 第8届全国人机语音通讯学术会议. 北京，2005：335-338.
[11] 何彬. 基于语音识别和语音合成的汉语语音转换技术研究[D]. 昆明：云南大学，2013.
[12] 陈肖霞. 连续话语语料库的语音切分和标记[J]. 语言文字应用，2000(5)：78-82.
[13] 苗玺. 中文语料库切分不一致字串分类校验方法研究[D]. 太原：山西大学，2006.
[14] 张志楠，李琳琳，张巍. 高准确度无标注的句子切分算法的研究[C]// 第十二届全国人机语音通讯学术会议，2013.
[15] 易克初，田斌，付强. 语音信号处理[M]. 北京：国防工业出版社，2000：51-67.
[16] 宋之用. Matlab在语音信号分析与合成中的应用[M]. 北京：北京航空航天大学出版社，2013：203-204.

作者简介：
吴树兴，副教授，博士，主要从事信号与信息处理技术、移动通信技术的教学和研究工作；张秀琴，讲师，硕士，主要从事汉语语言学和语用学的教学和科研工作。