汉语语音识别中声学界标点引导的随机段模型解码算法

晁 浩[1,2] 杨占磊[2] 刘文举[2]
(河南理工大学计算机科学与技术学院 焦作 454000)[1]
(中国科学院自动化研究所模式识别国家重点实验室 北京 100190)[2]

摘 要 提出了一种随机段模型的解码优化算法。检测出具有语音学意义的界标点,根据这些界标点分析临近语音段的边界信息和声韵母类别信息,最后将这些边界信息和类别信息用于指导随机段模型的搜索过程。实验中,两种类型的界标点能较为准确地被检测出来,并用于指导随机段模型的解码,在"863-test"测试集上进行的汉语连续语音识别实验显示,在正确率只有轻微下降的同时,解码时间下降了12.92%,这表明了将语音学知识引入语音识别系统的有效性。

关键词 语音识别,随机段模型,解码,界标点
中图法分类号 TP391 **文献标识码** A

Landmark Guided Segmental Speech Decoding Algorithm for Continuous Mandarin Speech Recognition

CHAO Hao[1,2] YANG Zhan-lei[2] LIU Wen-ju[2]
(School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454000, China)[1]
(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)[2]

Abstract A framework was proposed which attempts to incorporate landmarks into segment based Mandarin speech recognition system. In the method, landmarks provide boundary information and phonetic class information, and the information is used to direct the decoding process. To prove the validity of this method, two kinds of landmarks which can be detected reliably were used to direct the decoding process of a segment model(SM) based Mandarin LVCSR system. Experiments conducted on "863-test" set show that decoding time can be saved about 12.92% without obviously decreasing the recognition accuracy. Thus, potential of the method is demonstrated.

Keywords Speech recognition, Stochastic segment modeling, Decoding, Landmark

1 引言

在语音识别领域,针对隐马尔科夫模型(Hidden Markov Model, HMM)帧间独立假设的缺陷,随机段模型(Stochastic Segment Modeling, SSM)作为HMM的一种替代模型被提出[1,2]。相比于HMM,随机段模型不仅拥有更高的声学建模精度,而且拥有更为灵活的模型解码框架,此外其模型结构为直接使用语音段单位上的特征提供了可能性[3]。也正是因为随机段模型的模型结构更为精细,使得声学模型概率计算复杂度高;同时,由于段模型去除了语音帧之间独立的假设,解码过程中路径扩展和合并时不仅要知道当前时间点,还要考虑所有可能的起始点,这样就扩大了搜索范围,使得解码时SSM的计算复杂度和解码时间要比HMM大很多。因此,在保证系统精度的情况下降低解码时的计算复杂度对于随机段模型系统具有很重要的意义。

研究人员已经在这方面开展了深入的研究工作,文献[4]提出的分步段模型概率评估算法和由粗到细一遍解码算法,文献[5]提出的可变步长解码算法,以及文献[6]提出的相邻段并行解码算法,都有效地降低了SSM解码时算法的计算复杂度,使其降到了与HMM相同的数量级,并成功应用于大词汇量连续语音识别(Large Vocabulary Continuous Speech Recognition, LVCSR)任务。尽管如此,随机段模型的计算复杂度仍然高于HMM,所以还需要对其进行进一步的优化。

另一方面,不管是HMM还是SSM,其解码过程都是在统计框架下实现的,采用的是一种盲搜索的方式,即把语音信号的每一部分(或每一语音帧)都看作是同等重要的,解码时都要做同样的处理。而实际上,语音信号中的信息并不是均匀分布于整个句子中,发音时各种发音器官剧烈变化的区域蕴含了更为丰富的语音学信息,反映在语谱上就是各个频带能量的剧烈变化。而人耳的感知实验也已经证明,人们在理解语音信号时并未将注意力平均分配给整个句子,而是将重点放在信号剧烈变化的区域。而这些区域蕴含的语音学信息实际是可以被抽取出来用于语音识别系统的。

语音识别领域利用语音学信息的方式主要有两种:一种

到稿日期:2012-12-11 返修日期:2013-03-26 本文受国家自然科学基金(91120303,90820303,90820011),国家重点基础研究发展计划(973计划)(2004CB318105),国家高技术研究发展计划(863计划)(20060101Z4073,2006AA01Z194)资助。

晁 浩(1981—),男,博士,讲师,主要研究领域为语音识别,E-mail:chaohao@hpu.edu.cn;杨占磊(1984—),男,博士,助理研究员,主要研究领域为语音识别;刘文举(1960—),男,博士,研究员,博士生导师,主要研究领域为语音识别、语音增强、计算听觉场景分析。

方式是在已有的统计语音识别系统如 HMM 系统上融入语音学信息而实现[7,8]。在这种方式中，语音学信息在建模阶段以特征的形式被引入到语音识别系统中，语音学特征既可以单独用于声学建模，也可以和传统的谱特征一起用于声学建模；第二种方式是基于语音学知识的语音识别系统[9]，这种方式基于语音学建立语音识别框架，涉及到更多的语音学知识。

本文主要的研究工作就是将语音学信息应用于随机段模型语音识别系统中。与将语音学信息以特征的形式用于声学建模不同，声学信息用于指导随机段模型的解码过程：首先检查出语音信号中的能量突变时间点——声学界标点，然后获取界标点蕴含的发音单元边界信息和发音单元的类别信息，并在界标点的基础上检测出元音稳定段，最终将发音单元的边界信息、类别信息以及元音稳定段用于随机段模型系统，缩小了解码时的搜索范围，从而达到了降低计算复杂度的目的。

2 随机段模型

随机段模型的定义如下：假定语音段对应的观测矢量序列 $X=\{x_1,x_2,\cdots,x_l\}$ 是由随机段模型 α 产生，α 用一个定长的点序列 $R=\{r_1,r_2,\cdots,r_L\}$ 来拟合 X 的均值轨迹。观测矢量序列 X 一般是变长的，为了对齐 X 和 R，必须对 X 进行重采样 T，得到与 R 时长相同的新的观测矢量序列 $Y=\{y_1,y_2,\cdots,y_L\}$：

$$Y=XT \tag{1}$$

重采样 T 有多种方法，比较常用的方法是近似采样和插值采样。

经过重采样后，Y 中观测矢量 y_i 与随机段模型 α 中的域模型 r_i 根据在序列中的位置一一对应。因此，随机段模型 α 产生语音段 X 的概率可以转换为由域模型对 Y 中的特征矢量进行评估：

$$p(x_1^l|\alpha)=\prod_{i=1}^{L}p(y_i|\alpha,r_i,l) \tag{2}$$

若随机段模型建模时加入了段特征(如时长特征)，则式(2)变为：

$$P(x_1^l|\alpha)=\prod_{i=1}^{L}p(y_i|\alpha,r_i,l)\,p_s(x_1^l|\alpha) \tag{3}$$

式中，$p_s(x_1^l|\alpha)$ 是段模型 α 对特征序列 x_1^l 的段特征得分。

段模型的解码是一个双层的搜索过程。第一层是寻找最优的声学模型——对起点为 τ、终点为 m 的语音段 x_τ^m，找到有最大似然概率的段模型 α：

$$D_m(\tau)=\max\{\ln[p(x_\tau^m|\alpha)](m-\tau)+\ln[p(\alpha)]+\ln[P_s(x_\tau^m|\alpha)]\} \tag{4}$$

$$0\leqslant\tau<m<T$$

式中，$D_m(\tau)$ 是该段的最大似然得分，$p(\alpha)$ 是语言模型得分，$P_s(x_\tau^m|\alpha)$ 是语音段特征得分(时长等)。第二层是寻找最优切分的过程，具体到当前帧 m，是寻找以其为结束点的语音段所对应的最佳起始点 τ。这样，对 $[0,T]$ 范围内的所有语音帧，依次找到以其为结束点的语音段所对应的最佳起始点：

$$J^*(m)=\max_{\tau}\{J^*(\tau)+D_m(\tau)+C\} \tag{5}$$

$$J^*(0)=0$$

当前帧 m 和起始帧 τ 的取值范围是：

$$0<m\leqslant T \tag{6}$$

$$\max\{m-L_{ext},0\}\leqslant\tau<m \tag{7}$$

式中，$J^*(m)$ 是到 m 点为止的语音段序列的累积得分，C 是插入因子，L_{ext} 是允许最大段长。而解码器将对所有起点在允许范围内、终点为 m 的语音段进行段模型的扩展和解码。

3 声学界标点

3.1 声学界标点定义

根据文献[9]的定义，界标点可以分为声门界标点(Glottis Landmark)、响音界标点(Sonorant Landmark)、突发界标点(Burst Landmark)、滑音界标点(Glide Landmark)、元音界标点(Vowel Landmark)等几类，其中只有声门界标点和语音界标点能够被较准确地检测出来。从界标点得到的语音学信息用于语音识别系统时，界标点检测的准确率对语音识别系统的精度有很大的影响，所以这里只选用声门界标点和元音界标点用于后续的处理。

声门界标点反映了声带动作变化的时间点，其具体分为两类：+g landmark，声带开始振动的时间点；-g landmark，声带停止振动的时间点。元音界标点是指元音中声学信号最显著的时刻。

3.2 界标点蕴含的信息

界标点作为语音中的声学线索，其附近蕴含了比其它区域更丰富的信息。这里提出根据声门界标点和元音界标点获取界标点附近声学单元的边界信息、类别信息以及元音稳定段信息。

在判断界标点附近发音单元的类别前，首先要介绍下发音单元的类别信息。本文所用的声学模型是基于声韵母的声学单元，声母和韵母的类别信息如下：按照发音方式，声母分为辅音和滑音(半元音)两类，而辅音更可细分为塞音、擦音、赛擦音、边音和鼻音 5 类，其中滑音、边音、鼻音以及擦音中的'r'是浊音，其余的声母都为清音，汉语声母的分类见表 1。韵母按照结构和发音方式分为单元音、复元音和复鼻尾音 3 类，复元音是指由两个以上的元音连接而成，复鼻尾音是由单元音或复元音后跟鼻韵尾(n,ng)形成的，汉语韵母的分类见表 2。

表 1 声母的类别

		清音	浊音
辅音	塞音	b,p,d,t,g,k	
	擦音	f,s,sh,x,h	r
	塞擦音	z,zh,j,c,ch,q	
	鼻音		m,n
	边音		l
滑音			w,y

表 2 韵母的类别

类别	韵母
元音	a,o,e,i,u,ü,er
复元音	ai,ao,ei,ia,iao,ie,iu,ua,ui,ue,uo,uai,ou
复鼻尾音	an,ian,uan,üan,en,in,uen,ün,ang,iang,uang,eng,ing,ueng,ong,iong

根据声门界标点的定义可知，其在一定程度上反映了语音单元的边界信息[10]；同时，根据发音动作的不同可以判断界标点前后语音单元的类别信息，具体如下：

+g landmark 反映了声带由不振动转向振动的时间点，根据这个性质、声韵母的发音特点以及汉语音节声韵母的结

构，可以判断+g landmark要么位于清辅音的声母和韵母之间(图1中的'z'和'engl')，要么位于静音段和浊音类的声母之间，这样才能使+g landmark之前的语音段发音时声带不振动，之后的语音段声带振动。

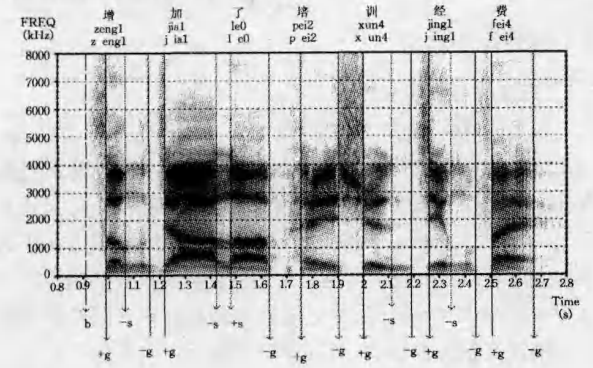

图1 声母界标点

−g landmark反映了声带由振动转向不振动的时间点，因此−g landmark要么位于韵母和清音类声母之间(图1中的'engl'和'j')，要么位于韵母和静音段之间。声门界标点前后声韵母的类型具体见图2。

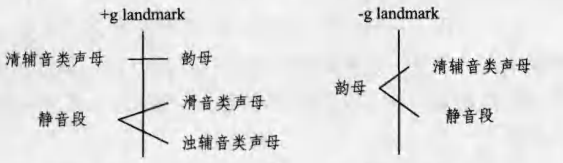

图2 声门界标点提供的声韵母类别信息

元音界标点虽然不能用来确定附近声学单元的边界和类别，但从对立的角度考虑可以引出语音信号中的非边界区域，即元音的发音稳定段(Vowel Steady Segment，VSS)。语音的产生是由发音器官的连续运动引起的，而不同的语音间之所以存在听觉差异，其根源在于发声时发音器官处于不同的位置。当语音由一个音素变换到另一个音素时，发音器官产生了快速而明显的运动；这种运动反映到语谱图中，表现为语音单元边界附近的能量跳变。而当处于音素快速变化之间的平稳段时，发音器官处于暂时的平稳状态或平缓运动状态；反映到语谱图中，对应于语音单元边界之外的能量平稳变化，这就是发音稳定段。因此，发音稳定段内是不包括语音的转换边界的。元音稳定段具体是指以元音界标点为起点，分别向前后两个方向扩展形成语音段，该语音段内信号频谱能量变化平稳，对应着发音动作的平缓运动过程，见图3。

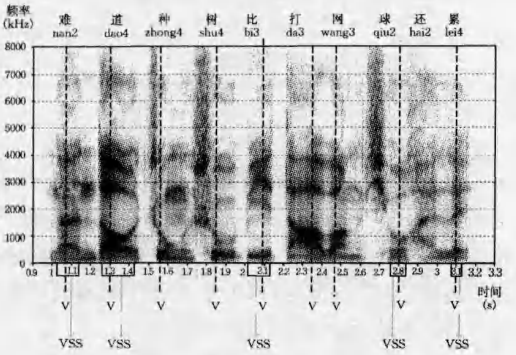

图3 元音界标点和元音稳定段

4 声学界标点引导的解码算法

在检测出声门界标点和元音稳定段后，元音稳定段、声门界标点提供的边界信息和声韵母类别信息被用于指导随机段模型的解码。

声门界标点提供的边界信息首先被用于解码，用于缩小随机段模型解码第二层(见式(5))的搜索范围。具体操作如下：

对当前帧 m 解码时，其允许的待解码语音段的起点帧 τ 的取值范围见式(7)。如果有声门界标点 g 位于 $m-L_{ext}$ 和 m 之间，即：

$$\max\{m-L_{ext},0\}\leqslant g<m \tag{8}$$

就认为 g 为实际的边界，也就是语音段实际的起始点。式(7)变为：

$$g\leqslant\tau<m \tag{9}$$

这样 g 之前的语音帧就不用参与搜索。

但是需要注意的是边界信息并没有精确到语音帧，而是大致的边界位置，所以加入了调节参数 q 确保实际的边界位于搜索范围内，即：

$$g-q\leqslant\tau<m \tag{10}$$

接下来声门界标点提供的声韵类别信息被用于解码，用于缩小随机段模型解码第一层(见式(4))的搜索范围。上文提到界标点 g 不是精确的声韵母边界，这里加入调节参数 d 和 b，使得实际的边界位于$[g-d,g+b]$中。解码时当前帧 m 如果也位于$[g-d,g+b]$范围内：

$$g-d\leqslant m\leqslant g+b \tag{11}$$

待解码语音段对应的声学模型要么属于图2中界标点之前的类别，要么属于界标点之后的类别。例如当界标点的类型为$-g$界标点时，当前语音段对应的声学模型可能是韵母、清辅音类声母或静音。由于基线系统解码时路径是按照声母-韵母-声母-韵母的方式进行扩展的，路径中前一个声学模型的类型实际上已经限定了当前所求的声学模型的范围：若前一个声学模型对应韵母，则当前所求的声学模型必定为声母，反之亦然，这就使得寻找最优的声学模型时搜索范围更小。上面的例子中若前一个声学模型对应韵母，那么当前语音段对应的声学模型只能是清辅音类声母或静音。

最后元音稳定段被用于指导随机段模型的解码，用于缩小随机段模型解码第二层(式(5))的搜索范围。上文已经介绍了发音稳定段内不存在语音单元的边界。随机段模型在解码时，语音信号中所有长度从1帧到 L_{ext} 帧的可能语音段都要进行段模型的搜索。这其中有许多语音帧都位于已经检测出来的元音稳定段内，而将这些语音帧作为解码时语音段的起始点或者结束点，在扩大了搜索的范围的同时却不能提高系统识别的准确率。因此，对所有开始帧或结束帧位于发音稳定段上的语音段，都不再进行段模型的扩展和解码，从而缩小了搜索范围，降低了计算复杂度。

假设元音稳定段的集合为 Θ，则段模型解码公式中的语音段起始帧 τ 和终止帧 m 的取值范围由式(6)和式(7)改变为：

$$0<m\leqslant T,m\notin\Theta \tag{12}$$

$$\max\{m-L_{ext},0\}\leqslant\tau<m,\tau\notin\Theta \tag{13}$$

图4和图5分别显示了基线系统的解码方式和元音稳定

段引导的解码方式。图中，语音信号表示为矩形序列，每个矩形代表一帧语音的声学特征。图 5 中深色的矩形序列表示元音稳定段。两图中演示了对连续 5 个终止帧的解码过程，带箭头的竖线代表终止帧，与其平行的竖线代表待扩展段的起始帧。这里设最大允许段长为 12 帧。由图中可以直观地看出，在原解码方法中，对 5 个终止帧的解码需要扩展 60 个语音段；而在元音温度段引导的解码中，只需要扩展 23 个语音段，这就大大降低了解码时的计算复杂度。

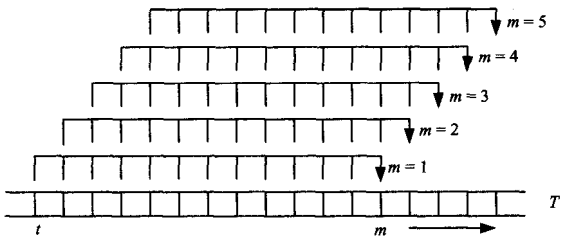

图 4　基线系统解码

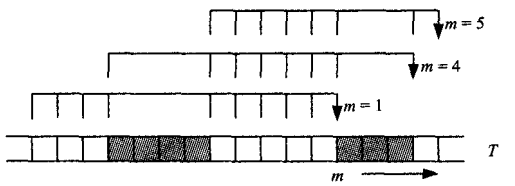

图 5　元音稳定段引导的解码

5　实验及分析

5.1　实验设定

实验所用的数据库为国家 863 项目汉语广播语音库。使用全部的 83 位男性说话人的数据训练声学模型，共 48373 句话，约 55.6 小时。使用 6 个男说话人数据做测试，共 240 句话，约 17.1 分钟。说话人来自北京等 6 省 2 市，没有明显口音。文本取自于《人民日报》，考虑了语料的声学平衡和覆盖性。所有语料均为安静实验室环境录制，信噪比小于 50dB，采样率为 16kHz，16bit 量化。声学特征包括 12 维梅尔频率倒谱系数(MFCC)及 1 维标准化能量，以及它们的一阶及二阶差分。实验中采用的汉语普通话音素集包含 24 个声母及 37 个韵母，每一个韵母含有 5 个声调。去除训练库中没有出现的声韵母，音素集中共包含 191 个基本音素。

本文采用文献[11]中的随机段模型系统作为基线系统。建模单元为声韵母，采用背景相关的三音子结构，每个段模型包含 15 个域模型和一个基于伽马分布的时长模型。每个域模型由 12 混合数的高斯混合模型模拟。域模型采用基于音素的决策树进行参数合并。训练阶段完成后，三音子模型一共有 202,984 个，域模型的个数为 24180 个。

5.2　声门界标点以及元音稳定段检测结果

声门界标点检测方法主要参照文献[8]提出的语音信号界标点检测系统中的相关算法并做了适当的修改，使得检测出的界标点尽可能的准确，而对检测出的声门界标点的数量则不作太高的要求。

检测出声门界标点后，将每一对声门界标点：$+g$，$-g$ 之间的语音段作为候选元音稳定段，然后检测出各个频带位于候选元音稳定段中的能量跳变点 t 后，将以 t 为中心的一段语音$[t-r,t+r]$作为发音非稳定段从候选元音稳定段中剔除，同时将频带 1 中能量小于设定阈值的语音段也从候选元音稳定段中剔除，目的是去掉非元音段，余下的连续语音段被认为是元音稳定段。

实验的目的是检测出测试集中 240 句话中的声门界标点。界标点检测结果可以用插入错误、删除错误以及与语音边界的距离来表示，这里所述的语音边界是指利用基线系统通过强制对齐方式来得到语音边界。界标点的插入错误是指在语音信号中声带一直振动或者一直不振动的语音段内检测出了声门界标点；界标点的删除错误指语音信号中声带停止振动或者开始振动的时间点附近区域没有检测出声门界标点。就本文所提的引导算法来说，界标点的插入错误会提供错误的声韵母边界信息和类别信息，同时对后续的元音稳定段检测也会造成影响，从而造成语音识别系统准确率的下降；而界标点的删除错误只是降低本章所提算法的效率，对系统的准确率没有影响。表 3 给出声门界标点的检测结果。

表 3　声门界标点检测结果

界标点类型	删除率(%)	插入率(%)	距离(ms)
$+g$	27.3	0.3	+42.7
$-g$	26.9	0.6	+18.5

表中距描述界标点与边界距离中的'+'是指界标点位于边界点之后。可以看出$+g$界标点与边界的平均距离为42.7 ms，远大于$-g$。由于系统所用 MFCC 特征的帧移为 10ms，为了使确保界标点提供的边界信息和类别信息尽量正确，式(10)中的调节参数 q 设为 6 帧，式(11)中的调节参数 d 设为 6 帧，b 设为 2 帧。

在元音稳定段的检测中非稳定区域范围划定参数 r 决定了检测出的元音稳定段的长度。当 r 取值越小时，检测出的元音稳定段越多，但检测错误即单稳定段内的语音边界也越多，识别中引入错误的风险越大。经过多次实验，r 的值设定为 4，表 4 给出了元音稳定段的检测结果。

表 4　元音稳定段的检测结果

	准确率(%)	帧数	时间(min)	VSS/总时长
VSS	98.2	22998	3.8	22.3%

5.3　LVCSR 系统实验结果

首先将声韵母边界信息、类别信息以及元音稳定段分别应用于随机段模型系统，结果见表 5。系统 1 将声门界标点提供的声韵母边界信息用于随机段模型，系统 2 将声门界标点提供的声韵母类别信息用于随机段模型，系统 3 将元音稳定段用于随机段模型。与基线系统相比，系统 1 的解码时间减少了 1.6min。同时，尽管式(10)中的调节参数 q 设为 6 帧，要大于界标点与边界的平均距离，但是还是有少数的边界与界标点的距离大于 60ms，处于有效的搜索范围之外，造成系统汉语字的误识率轻微上升。

表 5　不同信息引导的解码算法识别结果

系统类型	WER(%)	解码时间(min)
基线系统	13.67	24.0
系统 1	13.98	22.4
系统 2	14.12	23.6
系统 3	13.32	21.9

由于随机段基线系统采用染色算法对语音段的候选声学模型集进行了限制，因此声门界标点提供的声韵母类别信息的作用大大降低了，系统 2 的解码时间只减少了 0.4min。同时依旧存在少数的边界位于$[g-d,g+b]$范围之外，使得错

误的类别信息被用于指导解码，造成系统误识率轻微上升。

系统3用元音稳定段排除了解码时部分不符合声学实质的语音候选段，避免了由此产生的解码错误，因此在提高运行时间的同时，识别系统的错误率有所降低。

最后将界标点提供的所有信息同时应用于随机段模型系统，结果如表6所列。

表6 所有信息引导的解码算法识别结果

系统类型	WER(%)	解码时间(min)
基线系统	13.67	24.0
系统4	14.01	20.9

系统4将声韵母边界信息、类别信息以及元音稳定段同时用于指导随机段模型的解码。与基线系统相比，系统4的系统误识率轻微上升，解码时间下降了12.9%。

结束语 根据声门界标点分析语音单元的边界信息、类别信息以及元音稳定段，并将上述语音学信息用于指导随机段模型的解码。实验结果表明，本文提出的方法能够有效地提高段模型大词汇量汉语连续语音识别系统的解码速度。接下来的研究工作将着重提高其它类型声学界标点的检测精度，并分析不同类型声学界标点相结合时蕴含的语音学信息，将其用于改进语音识别的解码。

参 考 文 献

[1] Kimball O, Ostendorf M, Bechwati I. Context Modeling with the Stochastic Segment model[J]. IEEE Trans. on Signal Processing, 1992, 40(6): 1584-1587

[2] 唐赟，刘文举，徐波. 基于后验概率解码段模型的汉语语音数字串识别[J]. 计算机学报，2006，29(4)：635-642

[3] Chao Hao, Yang Zhan-lei, Liu Wen-ju. Improved Tone Modeling by Exploiting Articulatory Features for Mandarin Speech Recognition[C]// Proceedings of ICASSP. 2012: 4741-4744

[4] Tang Yun, Liu Wen-ju, Zhang Hua. One-pass coarse-to-fine segmental speech decoding algorithm[C]// Proceedings of ICASSP. 2006: 441-444

[5] Zhang Hua, Liu Wen-ju, Xu Bo. Research on Adaptive Step Decoding in Segment-Based LVCSR[C]// Proceedings of IEEE NLP-KE'07. 2007: 463-467

[6] 彭守业，刘文举，张华. 基于相邻段的随机分段模型解码算法及其在LVCSR中的应用[C]// 2008年全国模式识别学术会议. 2008: 432-436

[7] 张晴晴，潘接林，颜永红. 基于发音特征的汉语普通话语音声学建模[J]. 声学学报，2010，35(2)：261-266

[8] Yang Zhan-lei, Liu Wen-ju. A Novel Path Extension Framework Using Steady Segment Detection for Mandarin Speech Recognition[C]// Proceedings of InterSpeech. 2010: 226-229

[9] Liu S A. Landmark Detection for Distinctive Feature-based Speech Recognition[J]. Journal of the Acoustical Society of America, 1996, 100(5): 3417-3430

[10] Park C. Consonant Landmark Detection for Speech Recognition [D]. Massachusetts, Cambriage: Massachusetts Institute of Technology, 2008

[11] 唐赟. 基于随机段模型的汉语语音识别算法研究[D]. 北京：中国科学院自动化研究，2006

(上接第193页)

同Hadoop集群规模运行算法。本算法利用Hadoop加载不同的Map和Reduce算法过程，对图结构数据进行处理，即完成图数据的稀疏化。本文提出的算法加速比性能测试结果如图7所示。

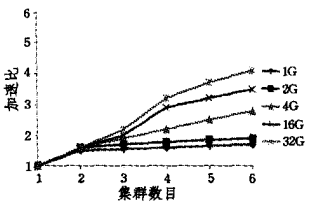

图7 算法性能测试结果

从图7中可以看出，尤其是面对大数据，利用Hadoop集群可以高效地减少时间的消耗，即加速比 $S_{speedup}$ 明显增大。根据Hadoop框架运行的特点，图数据量较大时，对图的稀疏化速率明显加快，有线性趋势；但是由于集群结点之间的数据通讯会有一定的数据开销，当图数据量较小时，图的稀疏化的优势会有所减弱，甚至需要的速率会略有下滑。同时加速比性能和集群的节点数成正比，即随着集群中节点数的增多，算法的加速比 $S_{speedup}$ 也逐渐渐增大。

上述结果表明，本文提出的基于MapReduce的图稀疏化算法更加适用于大规模图数据。即图数据量越大，算法的性能越好。其原因是算法设计中，增加了排序，合并了一些额外操作，使主节点和从节点之间的通讯代价大幅度减小，并且数据集规模越大，通讯量减少的比例越高。因此，当数据集规模越大时，算法的加速比性能越好。

结束语 本文深入研究了一种基于MapReduce的大图稀疏化算法。本文在介绍Minhash算法在图稀疏化处理中应用的基础上，在MapReduce计算框架上实现了Minhash算法的并行化改造，设计并实现了面向大规模图数据的分布式稀疏化算法，给出了具体的算法处理流程。最后，通过在多组不同大小数据集上的实验表明，本文提出的MR-GSpar算法适合运行于大规模并行平台下，提高了图稀疏化效率。

参 考 文 献

[1] Satuluri V, Parthasarathy S. Scalable graph clustering using stochastic flows: applications to community discovery[C]// ACM SIGKDD. 2009: 737-746

[2] Kulis B, Basu S, Dhillon I, et al. Semi-supervised Graph Clustering: A Kernel Approach[J]. Machine Learning, 2009, 74(1): 1-22

[3] Satuluri V, Parthasarathy S, Ruan Y. Local graph Sparsification for Scalable Clustering[C]// SIGMOD. 2011: 737-746

[4] 李建江，崔健，王聃，等. MapReduce并行编程模型研究综述[J]. 电子学报，2011，39(11)：2635-2642

[5] Lin J, Schatz M. Design patterns for efficient graph algorithms in mapreduce[C]// MLG. 2010: 78-85

[6] 尹丹，高宏，邹兆年，等. 一种新的高效图聚集算法[J]. 计算机研究与发展，2011，48(10)：1831-1841

[7] Lv Qin, Josephson W, Wang Zhe, et al. Multi-probe LSH: efficient indexing for high-dimensional similarity search[C]// Proc of the 33rd Int Conf on Very Large Data Bases(VLDB'07). Vienna Austria: VLDB Endowment, 2007: 950-961