粤拼序列自动切分算法的研究

肖镜辉，刘秉权

（哈尔滨工业大学计算机科学与技术学院，哈尔滨 150001）

摘　要： 语句级拼音输入法是当前汉字键盘输入的主流方法，拼音序列的自动切分是语句级汉字输入的前提。本文面向语句级粤语拼音输入法，针对粤语拼音序列的自动切分问题首次提出有效的解决方法，并提出相应的评价标准。目前本文成果已成功应用于产品中。

关键词： 粤语拼音，自动切分，评价

The Study of the Segmentation Algorithm for Jyutping Sequence

Jinghui Xiao, Bingquan Liu

(School of Computer Science and Techniques, Harbin Institute of Technology, Harbin, 150001, China)

Abstract: Nowadays, the intelligent pinyin input methods of sentence-level dominate the market of Chinese keyboard input method. The segmentation for the pinyin sequence is the preprocess step of the pinyin input methods. For the Jyutping input method, this paper proposes an effective algorithm to segment the Jyutping sequence automatically. Principles of evaluation and the experimental results are presented. The techniques of this paper have been successfully applied into the commercial product.

key words: keywords: Jyutping, Segmentation, Evaluation

1 引言

汉字键盘输入是中文信息处理的一个重要的应用领域。当前流行的中文输入法主要可以分为两类：基于笔画的输入法和基于拼音的输入法。笔画输入法的特点是输入速度快，识别精度高，但需要用户记忆相当多的输入规则，通常用户需要经过一段时间的训练才能够熟练掌握，一般受到专业用户的喜爱。拼音输入法的特点是易学易用，用户无需记忆繁复的规则，只要懂得拼音就可以顺利输入中文；缺点是用户需要有选字的过程，输入速度相对较慢。目前，拼音的输入法被广大的电脑用户所喜爱，当前比较流行的有：微软拼音输入法[1]，紫光拼音输入法[2]、智能狂拼[3]、拼音加加[4]等等。这些中文输入方法均使用简体拼音作为输入手段。然而，在中国广大南方地区和香港、澳门地区，粤语是主要使用的语言。同简体拼音不同，粤语发音要求使用粤语拼音[5]。上述

基金资助：本课题得到国家自然科学基金重点项目“问答式信息检索的理论与方法”资助（项目编号60435020）和教育部微软语言语音重点实验室基金项目“面向特定领域的词典获取和统计语言模型的建立”的支持（项目编号01307620）

作者简介：肖镜辉（1978-），男，黑龙江省哈尔滨市，博士生，xiaojinghui@insun.hit.edu.cn

拼音输入法无法兼容粤语拼音作为输入方式，因此无法在上述地区使用。本文作者以粤语拼音作为输入手段，将语句级拼音输入技术同粤语拼音体系相结合，设计并实现了智能粤语拼音输入法，并在此基础上进行了粤拼序列自动切分算法的研究。目前，该系统已形成产品，并成功推向市场[6] 。

本文内容按照如下方式进行组织。本文在下一部分将介绍粤语拼音的拼音体系；在第三部分将详细介绍粤拼序列自动切分算法；在第四部分，本文给出粤拼序列切分算法的评价原则和本文的实验结果；最后，本文在第五部分给出结论。

2 粤音体系

香港语言学学会（LSHK）在一九九三年十二月公布了香港语言学学会粤语拼音方案， 简称粤拼[5]，成为当前粤语拼音的标准。本文所介绍的粤语拼音自动切分算法就是在这个方案的基础上建立和实现的。本文首先对这一套方案进行简要地介绍，并同简体拼音体系作简单的比较。

粤语拼音由声母（On sets）、韵腹（Nuclei）、韵尾（Codas ）三部分构成。声母作为拼音的首部，接下来是韵腹，韵尾作为拼音的尾部。下面分别给出粤拼中的所有声母、韵腹和韵尾，以及粤拼的字调信息。

1. 声母表

b	p	m	f
d	t	n	l
g	k	ng	h
gw	kw	j	w
z	c	s	

表中列出了粤拼体系的所有声母。同简体拼音体系中的 23 个声母相比，粤拼的声母个数（19 个）明显要少。很多简体拼音体系中的声母发音，如：“zh”、“ch”、“sh”，在粤拼体系中没有相应的声母对应；而且，粤拼体系中还有很多简体拼音系统中没有的声母发音，如：“ng”、“gw”等。简体拼音对应的是大陆普通话，而粤拼对应的主要是广东、港台地区的粤语，语言发音的差异和使用的习惯造成了简拼和粤拼之间的差异。

值得注意的是，粤语拼音中存在零声母的情况，即一个拼音仅仅由韵腹和韵尾组成。在粤拼体系中，零声母没有特殊的字母标记，例如：“呀”只拼做“aa”。

2. 韵腹表

粤语通常在讲的时候尾音拖得比较长，体现在粤拼体系上的特点就是，粤拼将简拼中的韵母又细分成了韵腹和韵尾两部分，从而可以更详细地刻画粤语的声音信息。这是粤拼同简拼的一个显著区别。韵腹是在声母之后、韵尾之前的部分。粤拼体系共有 9 个韵腹，如下表所示：

aa	i	u	e	o
a	yu	oe	eo	

3. 韵尾表

韵尾紧接在韵腹之后，作为一个拼音的结束。粤拼体系中共有 8 个韵尾，如下表所示：

p	t	k
m	n	ng
i	u	

声母、韵腹和韵尾相组合，共组成了 625 个粤语拼音，构成了粤拼体系的主要部分。

4. 鼻音

此外，粤拼中还有鼻音，鼻音单独成韵。

m	ng

5. 粤语字调

同简体拼音不同，粤拼共有 6 个声调。下表列出每个声调及其对应的拼音和汉字的例子：

1（fu1/夫）	2（fu2/虎）	3（fu3/副）
4（fu4/妇）	5（fu5/斧）	6（fu6/父）

3 粤拼序列切分算法

3.1 粤拼序列切分算法的总体思路

粤语拼音切分算法利用系统缓冲区接受用户的输入字符，根据一定的切分规则对缓冲区中的字符进行切分。

系统缓冲区存储的是用户输入当前拼音的前部，或者是前一个完整拼音和当前拼音的前部的字符串连接。算法首先对缓冲区中的内容进行判断，如果是第一种情况，则继续接受用户的输入；如果是第二种情况，系统需要根据一定的规则来对系统缓冲区中的字符序列进行切分，得到前一个完整拼音和当前拼音的前部。系统何时对缓冲区中的字符序列进行切分，以及如何切分是切分算法的关键。

本文首先根据上文介绍的粤拼体系对组成粤语拼音的字符进一步进行详细分类，根据分类结果抽象出一些切分规则，利用规则的方法对上述情况进行判断和处理。

3.2 对组成粤语拼音的字符进行分类

虽然香港语言学学会已经把粤语拼音分解成了声母、韵腹和韵尾，但是经过观察并不是所有的组成声母的字符都只能出现在粤语拼音的首部，而是既可以出现在拼音首部做声母又能够出现在拼音尾部作韵尾。况且并不是所有的拼音都有声母（有零声母现象），也并不是所有拼音都是由声母、韵腹和韵尾三部分完整的组成。因此有必要对组成粤语的字符进一步分类，一方面便于了解粤语拼音的规律，另一方面也便于算法的设计实现。

按照字符在拼音中出现的位置和功能可以作如下的归纳：

1. 仅可以在拼音首部出现的字符（**Head character**）：这部分字符完全由声母构成，包括b、d、l、h、w、z、c、s、j、f。
2. 既可以在拼音首部又可以在拼音尾部出现的字母（**Head-end character**）：这部分同样完全由声母构成，包括p、m、t、n、g、k。其中'g'在拼音尾部出现时，总是同'n'以'ng'的形式出现在拼音尾部。
3. 复合声母，是指有两个字符联合组成、履行声母职责拼音组成单位。在粤语拼音中有三个复合声母，是ng、gw、kw。
4. 可以单独成音的韵母（零声母现象），有以下几类：aa+x ('x'表示任意字符。例如：aa、aai、aak)，e+x(例如：e、ei)，m，ng，o+x(例如：oi)，nk，ung。
5. 既可以在拼音尾部又可以在拼音中部出现的字母，包括i、u。
6. 在粤语拼音中没有出现的字母，包括q、r、v、x。

以上几点既是对字符在拼音中出现的位置和功能的分类，也是对粤语拼音的组成规律的一种归纳，粤语拼音的切分算法就是根据这些规律设计和实现的。

3.3 粤拼序列自动切分算法

本文在这部分详细介绍粤拼序列自动切分算法，给出算法流程图和切分规则。首先，为了讨论方便，本文作如下定义：定义当前系统缓冲区的尾部为零位置。定义当前系统缓冲区中倒数第i 个字符前的位置为负i 位置，如果切分算法在此位置处对缓冲区进行切分，称为系统在负i 处切分。

算法流程图如下：

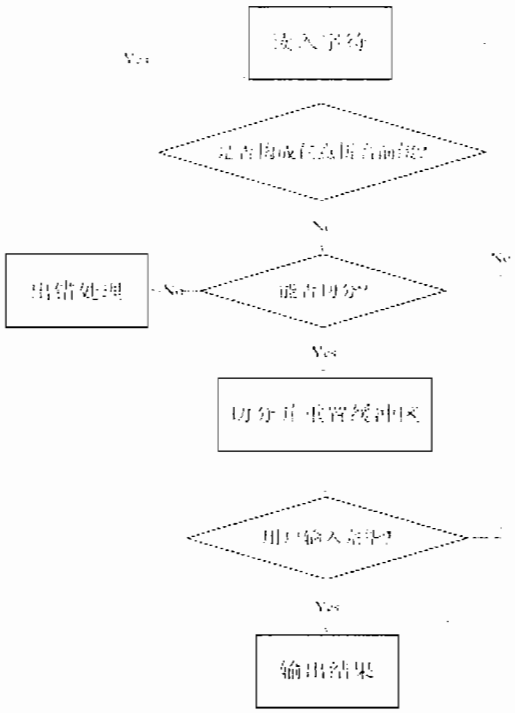

图1 算法流程图

Fig.1 The Flowchart of the Segmentation Algorithm

从流程图中可以看出，算法的关键在于怎样判断何时对当前缓冲区的字符串进行切分以及如何进行切分。本文设计了如下规则对缓冲区中的字符串进行切分：

1. 新读入的字符是Head Character ，这种字符只能作为拼音的首部，是切分的标志。在这之前是用户输入的完整的拼音，而这个字符作为新的拼音的首部。此时算法在-1 处切分。

2. 如果新读入的字符不是Head Character ，继续判断倒数第二个字符时候是否是Head-end character 。Head-end character 既可能是前一个拼音的尾部，也可能是当前用户输入的拼音的首部，因此有可能是在-2 处切分，也有可能是在-1 处切分，还有可能是既可以在-1 处切分也可以在-2 处切分。对于前两种可能，只需要稍作判断，就可以确定如何切分。对于第三种情况，如何切分，就取决于算法的策略。本文介绍的算法采用“短拼音优先策略”优先在-2 处切分。

3. 在第二种情况里还有一种特殊的情况，现在单独拿出来讲。在粤语拼音中，‘n’和‘g’都是Head-end character ，它们可以组合成‘ng’作为声母出现在某个拼音的首部，也可以作为韵尾出现在拼音的尾部，还可以作为鼻音单独成韵；‘n’和‘g’还可以拆开，‘n’作为上一个拼音的尾部，而‘g’作为下一个拼音的首部。因此对于‘ng’就有三种切分的可能：-1 处切分，-2 处切分和-3 处切分。在实际切分过程中，如果实际情况仅满足上面的一种可能，那么就可以按照相应的方法进行切分。但是，通常实际情况可以满足上面的多种可能，也就是存在切分歧义的情况，如何进行切分，就在于算法的策略。本算法介绍的算法采用“短拼音优先策略”，优先在-3 处和-2 处切分。

4. 如果不是以上的任意一种情况，那么是遇到了零声母情况。此时可以在-1 处切分。此外，当用户输入q、r、v 等不能构成粤拼的字母，则自动转向出错程序。

4 算法评测

作为一种拼音序列切分算法，首先必须要求它是“完备的”。完备性有两方面含义：一方面，要求算法对拼音表中的任意多个拼音组成的字符串都能够进行切分，并且切分的结果是一组在拼音表中的拼音；另一方面，

对于拼音表中的某个拼音串，经过切分算法之后，得到的结果应该是它本身，而不是多个拼音组成的一组拼音串——该性质能够保证在拼音输入的过程中任意一个拼音都能够完整输入。对于第二方面，很容易验证，将原始拼音表作为算法的输入，再检查输出就可以了。经过检查，本算法符合该方面的要求。对于第一方面，如果算法对任意相邻的两个拼音能够进行切分，并且切分结果是两个在拼音表中的拼音，那么对于多个拼音组成的字符串也能够进行切分，且满足上述要求。粤语一共有625 个拼音，将任意两个拼音两两连接组成约39 万个字符串，作为算法的输入，然后检查算法输出结果。实验表明，本文算法的切分结果满足以上的要求。这说明本文的切分算法满足"完备性"。

满足"完备性"表示拼音切分算法可以应用到实际的拼音输入中去，是对拼音切分算法的最基本的要求。但是一个好的拼音切分算法仅仅满足完备性是不够的。在任意两个拼音两两连接的39 万个字符串中，有一些重复的字符串，这些字符串是由两组或多组拼音连接组成的，从而也就对应了两种或多种的切分方法，也就是说存在着切分歧义。在上述39 万个字符串中，这样的字符串有12298 个，占总数的3.15% 。切分歧义的存在，使得拼音序列自动切分的结果具有"多样性"，即对于同一个待切分字符串，可以有多种合理的切分方式。然而在实际应用中，仅有一种切分方式是用户期望得到的。切分歧义的存在会降低切分算法的准确性。不同的切分算法处理歧义的方法不同，不同的处理方法直接影响了切分算法的切分准确率。定义拼音切分算法的切分准确率P为：

$$P=\frac{\text{算法切分的正确的拼音数目}}{\text{语料中总的拼音数目}}$$

本文提出的算法根据粤语拼音的规律，设计一定的规则来处理切分歧义，能够有效地解决切分歧义问题，取得较好的切分准确率。本算法用50 万字的语料进行测试，算法的切分准确率为95.47% 。据我们所知，本文是首次根据粤语拼音规律对粤语拼音序列作自动切分的工作，因此，无法将实验结果同其他人的结果进行对比。

5 结论

本文分析了粤语拼音的规律和特点，在此基础上提出了一种粤拼序列的自动切分算法，并给出拼音序列切分算法的评测的标准和实验结果。本文提出的粤语拼音切分算法已经成用应用于粤语拼音智能输入法，实现产品化。

参考文献：

[1] 微软公司，http://www.microsoft.com/

[2] 北京紫光华宇软件公司，http://www.thunisoft.com/unispim/overview.shtml

[3] 中文之星数码科技有限公司，http://www.cstar.com.cn/index1.html

[4] 北京六合源软件技术有限公司

[5] 香港语言学学会，http://www.hku.hk/linguist/lshk/

[6] 香港语言技术公司《智能粤语拼音语句输入法》，http://www.langcomp.com.hk/