第25卷第4期 上海师范大学学报(自然科学版) Vol. 25, No. 4
1996年12月 J. of Shanghai Teachers Univ. (Natural Sciences) Dec., 1996

计算机在汉字自动注音中的应用

潘以锋

提 要 探讨了用计算机进行汉字自动注音的方法,从最基本的思路出发,对其中的难点及多音字注音进行了具体的分析,并提出了解决的方法,同时给出了总的流程图和多音字注音流程图,最后对这方面的应用作了较详细的说明.

关键词 自动注音;多音字;编程;计算机

中图法分类号 TP319

0 引 言

近几年来,英汉字典和汉英字典在计算机上已得到了广泛的应用,输入一个英文单词,马上就可将它的汉语意思显示出来,或输入一个中文单字或词组,立即就可把它的英文意思显示出来,这给许多中国计算机用户带来了很大方便,尤其是对英文不是很通的人来说更是如此.但在计算机应用上,给汉字注上汉语拼音却很少遇到,使许多外国人学中文带来很大不便,因此,笔者对这方面进行了一些探索.

1 总体思路

一般来说,每个汉字都有其固定的发音,有些特殊的汉字有两个或多个发音,也就是说,汉字和拼音之间存在着一定的规律,这就给用计算机给汉字注音带来了一定的基础.事先定义两个集合:汉字集和拼音集,在汉字集中的每个汉字在拼音集中必定有其相对的注音,并据此在两个集合中建立一定的函数关系.其中对于那些只有一个拼音的汉字来说,先在汉字集中找到汉字的位置,然后根据函数关系找到相应的拼音,同时把这拼音反馈出来;而对于多音字来说,就比较复杂,因为有多个拼音与此汉字相关,到底该汉字该注哪个音?如果随便注一个音,那么注音的准确率要大打折扣,至多85%左右,这是解决计算机自动注音问题的难点.经过对汉字的深入研究以及向汉字专家的请教,发现多音字的注音也有一定的规律,给需注音的多音字究竟挑选哪个拼音,可以根据其所在的语境和前后文的关系来挑选相应的拼音,如果该汉字是独立存在的,可根据常用拼音来处理.

收稿日期:1996-05-12

作者潘以锋,男,助教,上海师范大学数学系,上海,200234

2　应解决的问题及方法

2.1　建立两个集合

首先要建立两个集合：汉字集 H 和拼音集 P，汉字集主要以国标规定计算机常用汉字为基础，总共选6763个汉字.

$$H = \{h_n | n = 1,2,\ldots,6763\},$$

拼音集主要以现代汉语字典的音节表为基础，总共有417个音节，而每个音节都有5种音调：阴平、阳平、上声、去声和轻声. 即有发音2085种.

$$P = \{p_n | n = 1,2,\ldots,2085\}.$$

2.2　建立两个集合 H 和 P 之间的关系

用 FOXPRO 建立一个数据库，其结构为：

字段名	类型	宽度	小数位数	含义
HZXH	数值型	4	0	汉字在汉字集 H 中的序号
PYXH1	数值型	4	0	汉字的第1个拼音序号
PYXH2	数值型	4	0	汉字的第2个拼音序号
PYXH3	数值型	4	0	汉字的第3个拼音序号
PYXH4	数值型	4	0	汉字的第4个拼音序号
PYXH5	数值型	4	0	汉字的第5个拼音序号
CYPYH	数值型	1	0	哪一个拼音序号是最常用的

根据区位码的编号顺序建立汉字集中的所有汉字序号，然后通过查汉语字典找出该字所有的汉语拼音(本文设一个汉字的发音最多有5种)，并转换成相应的拼音序号，统计在该汉字的所有发音中哪一个拼音是最常用的，把该拼音的相应序号填入字段 CYPYH 中. 这部分工作量比较大，要通过查阅字典和有关拼音方面的专业书，以及进行必要的统计.

2.3　多音字发音字典

对于多音字来说，在几个都可能的拼音中挑选哪个拼音作为在这个环境下汉字的拼音，是注音中关键的问题. 经过观察和探讨，发现绝大多数多音字在一定的词组中，注音一般来说是一个音. 例如："没"的发音有两个：méi 和 mò. 在组字："没精打采"、"没门儿"、"没趣"、"没有"、"没辙"、"没治"中发音为"méi"；在组字"没齿不忘 "、"没落"、" 没奈何"、" 没世"、"没收"、"没药"中发音为"mò". 用数据库来表示多音字发音字典的格式为：

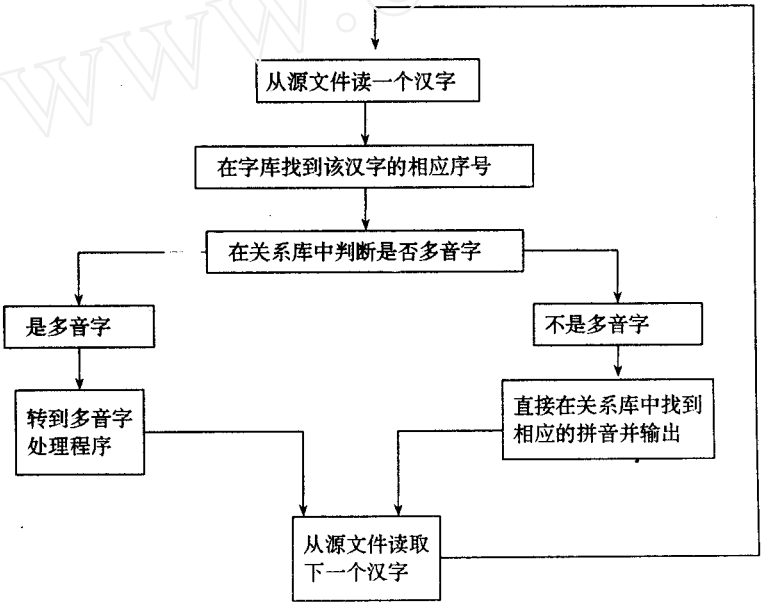
HZ(汉字)	FY1(发音1)	CZ1(词组1)	FY1(发音2)	CZ1(词组2)	…
没	méi	没精打采 没门儿 没趣 …	mò	没齿不忘 没落 没奈何 没世 …	…
糜	méi	糜子 糜黍	mí	肉糜 糜烂 糜费	…
…	…	…	…	…	…

多音字字典数据库建立起来确实比较麻烦,工作量特别大,但注音中关键的一部分工作是不能有差错的.

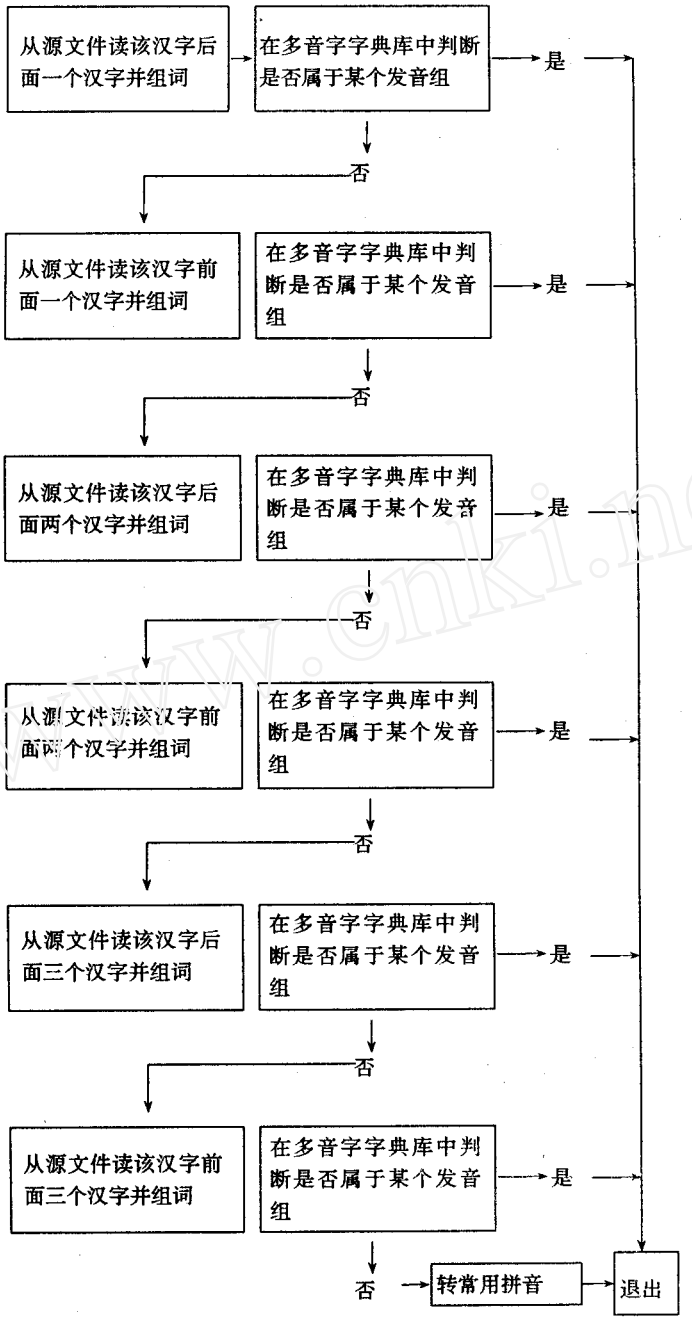
3 编程及流程图

有了一定的思路,把自动注音思想搞清楚,就可以编程,编程的语言可以用C,DELPHA,PASCAL 以及 FOXPRO 来编,其中用 C 来编速度比较快,但其中必须要了解 C 操作数据库的方法和技巧;DELPHA 是在 PASCAL 的基础上最近发展起来的,具有可视性及面向对象等优点,操作数据库比较方便;用 FOXPRO 来操作数据库是最方便的,只是其速度比前者要稍差一些.

编程的总流程图如下(具体程序略):



其中最难编的是关于多音字的自动注音,现把多音字的注音这一段的编程流程图介绍如下(具体程序略):



4 应 用

4.1 为学汉语的人提供方便

随着中国的改革开放、经济腾飞，全球学汉语的人越来越多，对于许多外国人来说，发音是比较大的难关，鼠标在屏幕上指到什么汉字，在其旁边就能够自动出现拼音，同时让电脑发出这汉字的读音，给学汉语的人带来很大方便.

4.2 为出版社或印刷行业服务

在小学课本和外国人学汉语等书籍中，有大量的汉字需要注音或直接出现汉语拼音，这给录入人员带来很大的困难，而且容易出差错，又给校对人员带来困难. 用计算机自动注音不但可节约大量的工作量，而且具有极大的准确率. 例如：给“没”注音，需输入如下字符：“$\overset{\text{méi}}{\text{没}}$”；而用自动注音只需录入“没”一个字，然后用程序转换就可以了，工作量可以节约80%左右. 我们在这方面做过尝试，使用方便，速度快捷，准确率高，效果十分理想，深受录入人员和校对人员的欢迎.

4.3 在多媒体以及电化教学上的应用

在多媒体以及电化教学中，电脑发出声音是很重要的标志，同时电脑发声也是许多电脑爱好者以及研究者所关注的，虽然市面上已有类似的产品出现，但其技术还不为广大电脑爱好者所熟悉.

参 考 文 献

1 中国社会科学院语言研究所词典编辑室编. 现代汉语小词典. 北京：商务印书馆，1984

2 章立民. FOXPRO2.5 FOR DOS 程序设计——提高篇. 北京：人民邮电出版社，1994

Applications of the Chinese Character Automatic Phonetic Notation to the Computer Science

Pan Yifeng

(Department of Mathematics)

Abstract We discuss the Chinese character automatic phonetic method with the computer. From the initial thought, we analyse the difficulties and polyphonic words, and put forward a solution. Meanwhile we give the general flow chart and the polyphonic words flow chart. Finally, we explain applications in this field in detail.

Key words automatic phoneticism; polyphonic word; program; computer