汉语方言语音数据库建设构想

洪拓夷
湖州师范学院图书馆 湖州 313000

〔摘要〕介绍我国近年来相关研究的情况,论述建设汉语方言语音数据库的重要意义和技术上的可行性,并从汉语方言语音数据库功能、数据库系统构成、语音语料库设计等几个方面进行分析和探讨,构建一款可用于认知和研究等的多功能的汉语方言语音数据库,同时提出需要注意的几个问题。

〔关键词〕汉语方言 方言数据库 语音数据库
〔分类号〕H21 G254

The Design of Chinese Dialect Voice Database

Hong Tuoyi
Huzhou Teachers College Library, Huzhou 313000

〔Abstract〕This paper introduces related research state in recent years, discusses the importance and the technical feasibility of building Chinese dialect voice database. In order to be used for research and awareness of the multi-purpose, it analyses the functions, the softwares and some other aspects of the Chinese dialect voice database. At the same time, it puts forward the points for attention in designing the database.

〔Keywords〕Chinese dialect dialect database voice database

1 引 言

近几年,由于受到普通话和流动人口的影响,使得用纯正方言的人数越来越少,应该说这些方言正处于衰变状态[1]。所以,笔者认为,尽快建立具有多功能的汉语方言语音数据库具有极其重要的意义:①可以通过其存储和学习功能来保护和传播人类非物质文化遗产;②通过其检索和辨识等功能,帮助某些职能部门准确辩别出话语者的乡里籍贯等地域信息,如对公安、安全部门开展刑事侦查等具有重要的应用价值[2];③通过其原生态的语音语料库事实数据有助于深入研究语音现象和文化渊源等。同时,它又可以推动语言现代化处理技术的深入研究和运用。可见,拟建的汉语方言语音数据库具有广阔的应用前景。

多媒体计算机语音处理功能的实现,多媒体软件开发技术的运用,汉语文 - 语转换、自动分词、语音合成、语音检索等处理技术的深入研究等都为多功能汉语方言语音数据库的建设提供了有力的技术保障,如汉语的文语转换系统可以将计算机内的任何文本转换成连续的语音流,再如汉语方言自动辨识技术虽然尚处于起步阶段[3],但它可以通过特征选取、音素匹配等方法对汉语方言进行辨识。

目前,学界相关研究也曾取得了一些成绩,如李永宏和于洪志对“安多藏语语音合成语料库”进行了初步研究,词库以双音和多音节词为主体,句料库却以 7种句型合成为主[4];沈向荣曾提出开发“壮语方言词语在线语料库检索软件”的设想;海柳文曾提出“汉语方言民族语言语音材料处理软件”的开发框架[5];肖双荣和吴道勤曾提出要在建立湖南方言语音特征数据库基础上进行湖南方言语音特征统计和分析;中国社会科学院开发的“北方方言基本词汇数据库”,收录北方话 100余调查点和 2 000余条基本词汇;由丁邦新等开发的“汉藏同源词研究系统”,收录了汉藏语系 122种语言和 12种汉语方言的 1 500余条词汇[6];再如麦耘主持的“汉语方言词汇数据库”,刘丹青主持的“方言语法语料库”,侯精一主持的“现代汉语方言音库”,刘俐李主持的“汉语方言语音词汇库”等[7]。尽管这些相关数据库存在明显不足,如方言偏少,收词量偏少,语音数据缺乏,缺少语音原始情景信息,系统功能单一等,但它们都为汉语方言语音数据库建设打下了良好的基础。

收稿日期:2008 - 08 - 07 修回日期:2008 - 09 - 26 本文起止页码:83 - 86 本文责任编辑:王传清

2 方言语音数据库建构

2.1 数据库功能

·检索功能。数据库应具有多途径检索功能,检索标识可以是文本也可以是语音;可使用多条件进行组合检索,能实现普通话与方言以及方言与方言之间的双向浏览,如由普通话词汇或语音能检索到相应的方言词汇或语音,由方言查找普通话或其它方言等;能以汉语方言语音为检索入口,通过对语音特征进行匹配,如调类、调值、调型、变调等,查出方言的市、县、乡三级地名信息;能根据各种方言实际情况和用户浏览习惯,来确定浏览方式和输出信息。

·学习功能。该数据库的建成将成为人们了解和学习各种方言不可或缺的工具。可以通过直接点击数据库中的词语或句子并选择方言类型,便能听到该方言的发音和相关例证等信息。可以通过输入词、句、段等文本,输出相对应的语音和相关例证等信息,这些语音文件有的是语料库中的原生态发音,有的是通过自动切分技术,采用音节及词汇的语音合成技术模拟而成的仿自然语句。

·分析功能。系统可以实现各种知识库间的有机联系,对各种方言的语法、语义和语用等资源描述信息进行比较分析等,如能以方言语音语料库、方言词语词汇库、方言语音语法知识库等为基本依据进行各种特征相似度比较、匹配和分析等,从而确定检索结果与要输出的内容。

·下载功能。它用来实现用户对检索、分析和比较结果的下载和打印;提供给用户对语言数据的统计(表格)等的输出;可以输出用户使用情况的统计数据与分析等。可以选择语音文件的某种格式进行下载,也可以输出带方言注音的汉语学习文本。

·维护功能。它主要包括三方面的维护功能:一是数据编辑功能,系统对载入的语音数据可以进行复制、剪切、替换、插入等;二是系统维护功能,系统管理员可进行数据管理、用户管理、日志管理以及系统升级等;三是拓展功能,根据发展需要对数据库再设计或添加新模块,以加强或拓展数据库功能。

·用户验证。系统对使用者身份进行确认从而分配不同的权限,主要分为系统管理员、数据管理员和普通用户。

·辅助功能。为用户提供每种方言的语音系统介绍,为用户提供输入输出音标系统说明以及提供汉语方言调查表和相关语音对照表等。

2.2 数据库系统构成

2.2.1 系统基本模式 建议采用 C/S模式即服务器/客户端模式。服务器端主要用于存放与管理数据,可使用具有强大伸缩性和可靠性的网络后台数据库软件,如 SQL SERVER等。客户端软件可采用相关开发工具自主开发,如借助 ASP并结合相关语音录入、合成、辨识和输出等开发软件共同制作而成,主要用来输入、输出及互传信息等。

2.2.2 数据库基本结构

·语音数据库。语音数据库用于存放汉语方言字、词、句等各语音数据及其属性、特征、标注、链接等相关信息。其中方言语音特征信息是方言相关度计算的前提,需要把纯粹音系特征和字音特征相结合来确定方言语音特征,这样即能体现出不同方言在音系特征方面的异同,也能体现出字音特征方面的异同[8]。语音数据既包括每个字、词、句可能的正常发音,又包括其变调后的发音,特别是变调后那些“半阴”、“半阳”等模糊声调音,由于变调都遵循规则,因此尽量录制存储音节单元的变调,这样不会使语音库无限扩大[9]。语音数据库也可细分为词音库、句音库、段音库等,也可分为方言音库和普通话音库等。

·文本数据库。文本数据库用于存放汉语方言字、词、句、地名等各文本数据及其属性、标注、链接等相关信息。文本数据库具体可包括词汇库、地名库、语料库等。

·知识库。该库用于存放各种词典、语法和语义等关系数据及各种规则等,它是集各种知识文档和关系文档于一体的大型集成系统。它可存放检索标识、特征和关系信息等,这既是实现具体检索方法的基础,又是对检索标识属性的描述。这些关系离不开各种规则,即事实性规则、关联规则、推理规则、认知规则和模糊规则等。知识库具体可分若干子库,如方言词汇对应规则知识库、方言语音对应规则知识库、方言属性对应规则知识库、语根知识库等[10]。

·索引库。它用于存放各种索引,包括分类索引、主题索引、语音特征索引、语音代码索引等。

·辅助库。它用于存放在检索或维护过程中调用或形成的各种临时数据或辅助数据等,如可根据需要建立一个临时用户代码库等,方便高级用户在检索时使用。也存放各级用户相关的背景资料信息,即所谓的用户库;或存放系统维护的相关控制信息等,即所谓的控制库;或存放用户自定义的数据资料和输出结

果,即所谓的自定义库等[11]。

2.2.3 数据基本结构 数据项涉及多种数据属性,包括方言域、方言类型、语音词、音节、调类、音频、释义等。每个数据包括许多匹配与辨识所用的关联与指示,如标识域、描述域、分类域、关系域等,以及其它相关属性等。如某一类数据基本结构为:

Key	T	C	N	Pc	Pi	Pn	……

Key:检索键值

T:类型

C:族性类别代码

N:出现频次

P:地址指针 (其中 Pc为域指针, Pi为信息指针, Pn为其它指针)

2.2.4 检索机制 汉语方言语音检索主要是通过语音、语法、词汇等关键特征的匹配来完成的,可以通过方言语音典型特征及相关控制等因素来判断,也可通过对其综合特征进行分析等来准确判断,或利用方言亲疏关系聚类分析等来判断。不管哪种方法都是要利用语音处理软件把语音特征及相关数据转换成与知识库规则相一致的可比数据,再通过辨识系统进行对比分析,最后输出检索结果。

特征信息量越大,排他性越强,越利于检索匹配。所以,语音辨识,首先进行方言语音声调和音长典型特征的匹配,声调特征涉及面广,具有强烈的排他性,各种方言的声调系统间极少有在调类数、调型、调值、声调来源、变调规律各方面都完全重合的;其次,可根据需要进一步进行方言其它特征的匹配,如声波频变、叠加、滑变等。

3 语音语料库设计

3.1 语料库

虽然语音语料库搜集哪些语料、搜集多少,并无统一标准,但要建立具有一定数量规模和特征的词汇库、句子库、语段库等,就要搜集方言地域人们所经常使用的语言文字材料,如文化与生活、历史与宗教、教育与科技等,越土、越俗的越要选用。

对于词汇,美国普林斯顿大学 1972年出版的 *Handbook of Chinese Dialect Vocabulary* (汉语方言词汇调查手册)将词目分为 33个义类,共 5 000余条目;我国 2003年修订的《汉语方言词语调查条目表》版将词汇分为 29大类,词目 4 000余条。方言的核心词主要包括名词、动词、形容词等,而方言中对同一事物的不同表述 (或说法)的词语,要尽量搜集全面,对于那些有本地方言特征或掺杂本地音调的外来词,也要适当搜集,增加例词、例句、释义等,力求能够全面反映某地方言的语音特点。

对于句子,结合方言自身生活习俗、语言习惯等实际情况采集语料 (包括长篇的话语材料),按其语法特点,提炼含有各种句型的句子样本,它们包括叙述句、判断句、疑问句、否定句、祈使句等。这些样本是在一定的情景下以日常生活为题材的自然话语。

3.2 音源选择

在语音数据库建设中,对某种方言的典型地域及发音人的选择至关重要。中国语言状况极为复杂,每个地方都有自己的"语言",这既是方言魅力所在,又给音源选择带来极大难度,所以,要对某种方言状况进行较全面的调查了解,才能科学地选择具有代表性的地域、方言及发音人。

为确保采集方言语料的质量和代表性,所选择的方言发音人必须土生土长、口齿清晰、操音熟练、用语传统、语速适中,是当地公认发音准确的。应选择那些文化程度不高、生活范围狭小、善于交际聊天、但很少受普通话影响的年龄在五、六十岁的发音人,这个年龄段的人讲话相对较"土"一些[12]。

至少要选择三组平行音源,进行平行录音和重复录音,以便采集准确语音发音样本。

3.3 语音录制

应选配专用录音房、专业录音麦克风、电平监视器等设备,采用先进的录音合成软件,音频控制要在 16 000Hz采样率和 16位精度以上,设置为清晰的单声道音频信号,存储为相应的文件格式。

麦克风是录音中重要的设备,既要保证在专业环境下的高灵敏性,也要保证其能在非专业录音环境中正常使用。对于具体环境下的情感语音录音可随机应变,尽量选用不会影响发音人情绪的录音设备,如录音笔等。如果朗读情感式录音文本,建议配戴袖珍麦克或头戴式麦克[13]。

整个录音过程应在专业技术人员指导下进行,有些录音可在正式录音前安排模拟录音实验,但有些实时录音必须一次成功。所有方言发音文件应配备对应的普通话读音文件,便于理解与学习。

4 需要注意的几个方面

4.1 数据库功能

数据库的设计既要保证数据的可靠性和完整性,又要保证系统的兼容性和共享性;既要成为通用的数据库检索系统,又要成为语言学习与研究的共享软件。对方言文本发音的标音不仅要易标,关键是要易读、易懂,能够保证它的准确性和连贯性。

4.2 方言数据采集

数据是数据库各种功能得以实现的最基本保证,而方言所涉及种类多、范围广,所以,要在数据采集上加大投入,建立数据搜集的各级组织,以便把那些很土的方言采集齐全。应减少朗读普通话提示文本录制语音数据,尽量采集原生态语法现象与发音习惯,以保证某种方言的客观性和特殊性。

4.3 检索预处理

一般检索系统往往采用禁用词表、运算规则等进行初步检索规范,但在方言数据库检索过程中无法使用这些规则,因为每个词都有其发音,都有检索意义,特别是在进行语音检索时,某些超失范语句必须进行预处理,否则容易匹配失误,导致检索错误,而人工预处理需要有一定的检索知识和语言知识,所以设计智能预处理系统非常必要。

4.4 软件开发

尽管已经开发出一些语音处理与识别软件,但尚需进一步研究和开发具有"自然语言认知和情感理解能力"的语音特征自动识别与提取软件、语义自动分析软件、语-文自动转换软件以及汉语方言智能预处理软件、自动辨识软件、自动合成软件等。

4.5 数据库标准化

我国数据库系统的研制与开发需要统一的数据标准和建库规范。所以,有必要对语音数据库的有关数据和功能制定一个统一的标准,而方言语音数据库的建设也亟需采取统一的规划措施。

5 结 语

中国地域辽阔,语言状况相当复杂,每个地方都有自己的方言,方言研究涉及很多问题,是一个很大的课题,为此,有必要对方言语音数据库进行规模研究,以促进我国汉语语言文化的留存与传承,并推动学界对我国汉语方言语言深入而持久的研究。

参考文献:

[1] 田珍都,张振礼. 基于数据库的新词语中的方言词语研究. 烟台教育学院学报, 2003, 9(1): 35-40.

[2] 陈海伦. 方言机器识别技术研究. 公安大学学报(自然科学版), 2000(1): 33-38.

[3] 顾明亮,沈兆勇. 基于语音配列的汉语方言自动辨识. 中文信息学报, 2006(5): 77-82.

[4] 李永宏,于洪志. 安多藏语语音合成语料库的设计. 西北民族大学学报(自然科学版), 2006, 27(1): 36-39.

[5] 海柳文. 汉语方言民族语言语音材料处理软件设计. 广西民族学院学报(自然科学版), 2005, 11(3): 60-64.

[6] 范俊军. 汉语方言词汇数据库研制的思路. 广东教育学院学报, 2006(1): 62-64.

[7] 周杨. 计算机技术与汉语方言研究. 现代语文, 2007(2): 17-18.

[8] 肖双荣,吴道勤. 湖南方言语音相关度计算与亲疏关系聚类分析. 湖南社会科学, 2004(1): 138-140.

[9] 陈蓁,张明. 文语转换系统中语音库录制方法的研究. 南京师大学报(自然科学版), 2000(1): 217-23.

[10] 陈海波. 关于数据库在古汉语研究中的应用. 古汉语研究, 2000(3): 61-65.

[11] 沈向荣. 壮语方言词在线语料库检索系统设计. 科技信息. 2007(13): 65.

[12] 王小龙. 基于语料库的东台方言特色词释义[学位论文]. 南京:南京师范大学, 2007.

[13] 谢波. 普通话语音情感识别关键技术研究[学位论文]. 浙江:浙江大学, 2006.

作者简介〕洪拓夷,男,1963年生,研究馆员,副馆长,硕士,发表论文30余篇。

(上接第46页)

[8] 周宁. 信息组织. 武汉:武汉大学出版社, 2004.

[9] 周宁,张驰,张会平. 信息可视化与知识检索系统设计. 情报科学, 2006(4): 571-574.

[10] Noy N F, Sintek M, Decker S, et al. Creating semantic web contents with protégé-2000. IEEE Intelligent Systems, 2001, 16(2): 60-71.

[11] 王应华.《中图法》电子版实现了类目的多维检索和多维显示. 图书馆论坛, 2003(6): 64-66.

[12] OCLC Online Computer Library Center, Inc. Using a Classification-Based Information Space. [2007-12-08]. http://www.oclc.org

作者简介〕欧阳宁,女,1980年生,助理馆员,发表论文2篇。
胡飞燕,女,1986年生。