基于大学生的汉语说话人识别语音库设计

王 宏 李 鑫 高 阳

(昌吉学院计算机应用研究所 新疆 昌吉 831100)

摘 要:本文设计了一个基于在校大学生的说话人识别语音库 UMSD,其目的主要是用于研究说活人个体特征变迁、文本有关和文本无关的说活人识别。该语音库包含 24名说话人的 12期录音,相邻录音间隔从 1天到 60天不等,在同一间安静的办公室环境下录制完成。录制语料包括:孤立数码,数码串,长度从 1到 10的词句,汉语拼音表,古诗词和短文。为了便于提取感兴趣的音段,本文还基于 Matlab和 Ms - Access设计了相应的语音库管理系统。

关键词:语料库;语音库;说话人识别

中图分类号: TP391 **文献标识码**: A **文章编号**: 1671 - 6469(2008)06 - 0107 - 05

引言

语音是人类通信的自然工具。语音现象非常复杂,无法用一些简单的算法加以描述。因此,目前较好的说话人识别系统大都基于语音信号的概率统计模型,例如 GMM 模型[1]。这些说话人模型需要用相当充分的说话人语料来训练、测试和评估,而且其识别性能也在很大程度上受说话人训练语料的影响。由于说话人自身的语音特征是随时间变迁的,这就需要说话人识别系统用不同时间录制的语音样本来训练说话人模型[2]。尽管可以通过一些模型自适应技术[3]来减少所需的训练样本,但是一个恰当收集和标注的语音库对说话人识别研究来说还是很重要的,它是说话人识别研究的基础和对象。

大学生来自不同地方,口音分布较广,一般要在校学习 4年,因此是获取 18 - 25岁之间说话人语料的有效来源,且语料的采集工作相对容易组织实施,因此本文设计了一种基于在校大学生的说话人识别语音库 UMSD。鉴于 Matlab的高效数值计算功能、快速开发脚本语言、以及其数据库工具箱提供对诸如 MS ACCESS、MS SQL Server、Oracle、IBM DB2等数据库系统的支持[4],我们选择 Matlab做为信号处理平台设计了 UMSD语音库管理系统。初步测试表明该语音库能够满足我们目前说话人识别研究的需要。

1 说话人识别语音库现状

国际上长期从事语料库大规模开发的机构主要有:美国的 LDC(Linguistic Data Consortium)[5]、OGI(Oregon Graduate Institute)[6],欧洲的 ELRA(European Language Resources Association)[7]等。除少数说话人识别语音库之外,这些机构开发的语音库大多是面向语音识别和语音合成等应用的,其中只有一小部分可以间接用于说话人识别研究。

LDC开发的 TIMIT语音库包含有 630(男 438/女 192)名美式英语说话人,每人 10句话,是最早应用于说话人识别研究的语音库之一。TIMIT同时包含有一系列经过二次处理的派生语音库,例如 FRMTIMIT(较远距离麦克风录制)、NTIMIT(通过长途电话网传输后录制)、CTIMIT(通过移动通信网传输后录制)、HTIMIT(直接通过固定电话录制)等。LDC开发的 Switchboard I - II语音库分别包含了 543人和 657人的电话对话,可用于说话人识别辨认和确认研究。此外,LDC还发布了专门用于政府门禁安全控制应用的 YOHO语音库,该语音库包含 138(男 106/女 32)名

收稿日期: 2008 - 10 - 20

基金项目:新疆自治区青年教师启动基金项目(XJEDU2006S34)

第一作者简介:王宏(1972—),男,山西长治市人,昌吉学院计算机应用研究所,副教授,研究方向:信息与信号处理。

说话人,在安静的办公室录制,适用于文本有关的说话人确认研究。

OGI开发的 CSLU说话人识别语音库,在两年内采集了 500名说话人的至少 12次电话语音。该语音库适用于多种大容量的文本有关及文本无关的说话人鉴别和说话人确认研究。

ELRA开发的 SIVA电话语音库是专门针对意大利语说话人识别的,它通过公用电话交换网录制了 671(男 335/女 336)人的超过 2000次的电话语音。ELRA开发的 Polycost说话人识别语音库包含了 134(男 74/女 59)名欧洲人的英语电话录音。

上述语音库都是基于英语等西方语言的。近年来,国内建成的语音语料库也很多。例如中国科技大学、中国科学院声学研究所、中国社会科学院语言研究所联合开发的汉语语音识别资料库;中国社会科学院语言所开发的现代汉语自然口语语料库、自然对话语料库、现代汉语方言自然口语语料库;中国科学院自动化所开发的旅游咨询口语对话语料库和旅馆预定口语对话语料库;香港大学和香港理工大学联合开发的香港广州话语音资料库;在汉语说话人识别领域应用较为广泛的 863中文语音数据库[8],等等。

2 语音库设计

语音库包括语音语料库、语音库管理系统两部分。前者通常是按目录分类存放的一大批录音文件,以及与每个录音文件对应的附加标注文件;由于语料库中的录音文件很多,为了便于快速提取出感兴趣的音段,往往需要利用后者来完成对语料库的查询、检索和管理等工作。

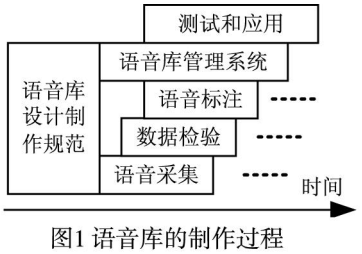

图1 语音库的制作过程

Fig.1 The process of developing a speech database

语音库的制作过程如图 1所示。其中,设计阶段的任务包括:根据设计要求确定各种技术规范,设计语料,选择说话人,搭建录音环境,检验录音系统,制定语料采集计划。然后按计划进行语音采集,并逐步完成数据检验、语音标注、语音库管理系统开发、语音库测试等工作。

2.1 语料设计

语音库按照说话的自然程度大致分为朗读语、口语和对话三种。第一种,说话人按指定文本朗读,语音的自然度不高,录音后期人工校对的工作量较小;第二种和第三种,一般不指定文本,有时会指定一个话题,语音的自然度高,但是录音后期校对的工作量较大。而且,若不对语料(即录音内容)进行设计,则很难保证语音库的完备性,例如很可能造成有些词汇反复出现,而有些则不出现。用这样的语音库来训练说话人模型,系统的识别性能会受到一定影响。

语音库的完备性要求[9]是指,语音库要符合语言的概率模型,在保证文本真实性和口话自然度的前提下,用尽可能少的语句来覆盖所有的汉语发音现象,即包含所有合理的音联关系,包含各种音节内和音节间的元辅音搭配关系,能体现协同发音现象及发音的韵律特征,能体现汉语语音学、声学的各种特征[10]。

汉语有 6800多个常用汉字,400多个音节,要用有限的文本设计出绝对完备的语料是非常困难的。为此,我们按照汉语的口语习惯设计了 UMSD的语料文本,其内容包括:孤立数码(0～9),电话号码(5个),邮编(2个),汉语拼音表(21个声母和 35个韵母),长短不等的词句(1个字～10个字)各 10条,古诗词各 2首,200字的短文 1篇。全部语料按正常语速计约为 8分钟。这些语料按顺序分别编号为 1～17,录音时说话人照上述语料文本自然重复即可。

2.2 录音条件

UMSD的 24(男 12/女 12)名说话人全部来自在校大学生,年龄介于 19～22岁之间。录音要求为:在自然放松的条件下,用正常语速和语调的普通话说出指定的语料。录音时,先请说话人熟悉语料文本和录音系统,然后说话人自行开始录音。所有录音均在同一间安静的办公室(4.5m * 5m)进行。录音设备为:普通 PC机 +声卡 + Cool Edit录音软件 +头戴式话筒。录音数据是单声道的,其采样频率为 16KHz,量化精度为 16比特,保存格式为 PCM方式的 WAV文件。语音采集工作分 12期完成,每期同样的语料录 2次音。期次

的间隔不等,如表 1所示。

表 1 采集说话人语料的期次和间隔

期次	1	2	3	4	5	6	7	8	9	10	11	12
间隔天数	0	1	1	7	7	7	30	30	30	60	60	60

2.3 语音库的组织

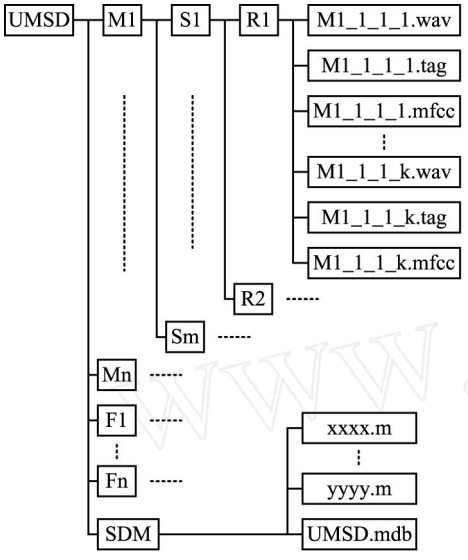

图2 语音库的目录结构
Fig.2 The Directory structure of UMDB

语音文件的命名规则是“说话人编号 _期号 _次号 _语料编号 .wav”。例如,M1_S3_R2_12.wav表示说话人 M1在第 3期录音时第二次读 12号语料的录音文件。录制好的语音文件按目录存放,先建立语料库的根目录 UMSD,然后建立 24个说话人子目录 M*或 F*,在每个 M*下建立 12个期目录 S*,在每个 S*下建立两个子目录 R1和 R2,每期的第一次录音文件 (17个)和第二次录音文件 (17个)分别保存在这两个子目录下,如图 2所示。图中,扩展名 tag表示同名语音文件的标注文件,是纯文本文件;扩展名 mfcc表示同名语音文件的 MFCC特征参数文件,是 Matlab的 mat文件; SDM子目录用来保存语音数据库文件及其它 Matlab程序。语音库录制完后,每人有 408个语音文件,数据量约为 168M,时长 192分钟;语音库共有 9792个语音文件,数据量约为 4032M,时长总计 4608分钟。

2.4 数据检验

每期录音后,对语音文件进行人工检验,以排除录音过程中可能出现的错误。例如,查看并剔除语音中的信号过载音段、不规则噪声 (例如咳嗽等)和非正常停顿造成的长时静音等。对于错误严重的录音文件,必要时可以请求说话人重新补录。

3 语音标注

简单地说,语音标注就是利用某种可书写文本符号来说明或描述语音波形中的各种语言现象。最基本的标注符号是相应语种的语言文字和发音字母表[11]。标注通常是由“标注符号 +标注起点 +标注终点”组成的三元组,以指示语音波形中某特定语音基元 (例如音节、字、词、句、段落等)对应的语段。依靠这些标注文本,可以用字符串在语音库中检索出感兴趣的语音段。

语音标注通常是分层次进行的。汉语语音标注大致分为如下六个层次:第 1层:汉字层,标注汉字;第 2层:拼音层,标注拼音,含声调;第 3层:音节层,标注声母和韵母。这一层标注较困难,因为声母和韵母之间存在过渡,有时很难精确划分它们之间的音节边界;第 4层:声韵层,标注发音人语流的音变情况。包括韵母丢失,声母丢失,声母增加,浊化,鼻化,清化,儿化,口音错误等等。这些现象在连续话语中有时非常明显;第 5层:音段层,标注音节在连续语流中的变化情况。因为受协同发音、语义、韵律等因素的影响,连续话语中的音节与单独发音的音节相比变化较大。音段标注常用 IPA(国际音标)或 SAMPA-C符号系统[12];第 6层:韵律层,标注语句中具有语言学功能的声调变化、语调模式、音质、重音模式和韵律结构。该层标注可使用 C-ToBI韵律标注系统[13]。

本文采用第 1、2层合并的单层标注,标注基元为单个汉字。其中,四个音调和轻声分别用“12340”标记,韵母“ü”用“v”标记,韵母“üe”用“ue”标记。例如,“南京和上海”语音的标注如图 3所示。此外,我们还定义了主要用于语音库检

索的段落起止标注符“[ps]”、“[pe]”和语句起止标注符“[us]”、“[ue]”。

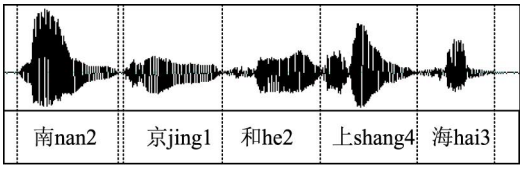

图3 语音文件的标注
Fig. 3 Annotation of the speech file

4 语音库管理系统

4.1 数据库设计

语音管理系统的核心功能是实现基于语音库存储结构的快速音段检索和分析处理。为此,我们基于 Matlab + MS ACCESS设计了 UMSD语音库管理系统,其中,语音数据库的基本表设计如下:

1)说话人信息表 (编号,姓名,性别,出生年月,民族,口音);

2)录制环境信息表 (编号,场所,录音系统,语种,录制方式);

3)数据格式表 (编号,语音文件格式,采样频率,量化比特数);

表 2标注表中的记录内容示例

序号	字	拼音	文件名	起点	止点
#	…	…	…	…	…
# + 1	[us]	[us]	-	-	-
# + 2	南	nan2	M1_S3_R2_12	1400	10000
# + 3	京	jing1	M1_S3_R2_12	10300	18400
# + 4	和	he0	M1_S3_R2_12	18400	26360
# + 5	上	shang4	M1_S3_R2_12	26360	34300
# + 6	海	hai3	M1_S3_R2_12	34300	41800
# + 7	[ue]	[ue]	-	-	-
# + 8	…	…	…	…	…

4)系统表 (单次录音标号,文本标注,拼音标注,录音日期,录音环境编号,数据格式编号)。其中,“单次录音标号”的格式为“说话人编号_期号_次号”;以“单次录音标号”为 M1_S3_R2的记录为例,“文本标注”字段中的内容是路径 UMSD\M1\S3\R2下 17个录音文件的标注文件中的汉字严格按语料编号顺序级联而成的字符串,且插入了段落标志和语句标注;“拼音标注”字段与“文本标注”雷同;其它字段记录了该次录音的环境信息。此外,M1_S3_R2也是该次录音对应的标注表的名字;

5)标注表 (序号,字,拼音,文件名,起点,止点)。每次录音对应一个标注表,数据库中共有 24 * 24 = 576个标注表。与系统表类似,每个标注表中的记录也是根据标注文件生成的,只是在这里每个标注三元组都对应一条记录。其中,“序号”字段是严格按顺序递增的长整数,它维系着“字”或“拼音”字段中标注文本或拼音的前后位置关系,是语音库检索的关键信息。表 2示意了标注表 M1_S3_R2中与句子“南京和上海”对应的部分记录内容。

4.2 语音库的查询和使用

下面以“找出所有新疆口音的男性发音词[我们]”为例说明在 Matlab环境下使用 UMSD语音库的基本步骤:1)新建一张包含字段 (文件名,起点,止点)的空表,将其命名为 RT,用于保存检索结果;2)从说话人信息表中查询出符合口音要求的说话人编号;3)根据说话人编号和发音词“我们”在系统表中查询出所有满足要求的记录集 (单次录音标号,文本标注);4)对于记录集中的第一条记录,执行如下操作:将“文本标注”读入 Matlab,记算字符串中所有与“我们”匹配的位置标号,打开与“单次录音标号”同名的标注表,查询出“序号”与位置标号相同的记录集 (文件名,起点,止点) (注意:这里起点是“我”的起点,止点是“们”的止点),把该记录集追加到表 RT中;5)按第 4步操作轮询第 3步返回的所有记录;6)查询结束,返回表 RT;7)遍历表 RT中的记录,用 waveread函数读取检索到的每个语音段中的语音数据,然后完成需要的各种分析处理工作。

5 基本测试

我们首先利用 UMSD语音库进行了基于长时平均频谱的说话人辨认 (SI)实验。任意指定 5个男性说话人和 5个女性说话人,在语料中选择连续的 60个字作为研究对象,取出与这 60个字对应的 2次不同期录制的音段分别作为说话人的参考样本和待识别样本,分别计算男性组和女性组的非对称长时平均频谱距离[14],实验结果如表 3和表 4所示,可见在小人群范围内采用该方法的说话人识别率为 100%。

表 3男性组说话人辨认实验结果

	M1	M2	M3	M4	M5
M1	0. 2184	1. 1832	1. 0480	1. 1385	0. 7149
M2	0. 9511	0. 1415	1. 1493	0. 3582	0. 8779
M3	0. 9105	1. 2502	0. 1827	1. 2475	0. 8160
M4	0. 7172	0. 3535	1. 2345	0. 1335	0. 6213
M5	0. 4766	0. 7385	0. 8839	0. 5582	0. 1775

表 4女性组说话人辨认实验结果

	F1	F2	F3	F4	F5
F1	0. 1548	0. 5516	0. 8402	0. 4605	0. 5426
F2	0. 5321	0. 1394	1. 1734	0. 6616	0. 8222
F3	0. 7416	1. 3916	0. 2055	0. 5274	0. 7730
F4	0. 4310	0. 8223	0. 5375	0. 1927	0. 6525
F5	0. 7810	1. 1112	0. 9802	0. 6977	0. 2360

6 结论和进一步工作

本文设计了一种基于在校大学生的说话人识别语音库 UMSD,基本测试表明该语音库能够满足我们目前对说话人识别研究的需要。按照本文提出的设计方案,语音库中的大学生说话人可以逐步增加。目前 UMSD的语音采集工作还在继续,我们的目标是达到 100名说话人。进一步的工作还包括分析语音的各种特征参数,例如基频、共振峰、LPCC和 MFCC等,建立与标注表结构类似的特征参数表,以方便今后的说话人识别研究。

参考文献:

[1]赵力. 语音信号处理[M]. 北京:机械工业出版社, 2003.

[2]Reynolds, D. A. An Overview of Automatic Speaker Recognition Technology[A]. Proc of ICASSP[C]. Orlando, FL, USA: IEEE, 2002: 300 - 304.

[3]Reynolds, D. A., Quatieri, T. F., Dunn, R. B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, 10: 19 - 41.

[4] The Mathworks Inc. Using MATLAB [EB/OL]. http://www.mathworks.com/access/helpdesk/help/techdoc/matlab.shtml, 2006 - 09 - 10/2006 - 10 - 12.

[5]LDC. TIMIT acoustic - phonetic continuous speech corpus[EB/OL]. http://wave.ldc.upenn.edu/, 2004 - 11 - 10/2006 - 1 - 8.

[6]OGI. Numbers Corpus V1.2 [EB/OL]. http://cslu.cse.ogi.edu/corpora/numbers, 2001 - 05 - 02/2006 - 1 - 10.

[7]ELRA. ELRA Catalogue of Language Resources[EB/OL]. http://catalog.elra.info/product_info.php, 2003 - 04 - 29/2006 - 1 - 12.

[8]刘岩. 关于中国少数民族濒危语言语音语料库的设计[J]. 中央民族大学学报(哲学社会科学版), 2006, 4: 133 - 136.

[9]Henk van den Heuvel, Dorota Iskra, Eric Sanders, Folkert de Vriend. SLR Validation: Current Trends and Developments[A]. LREC 2004 Workshop: Speech Corpus Production and Validation[C]. Lisbon, Portugal: ELRA, 2004: 107 - 111.

[10]祖漪清. 汉语连续语音数据库的语料设计[J]. 声学学报, 1999, 3: 236 - 247.

[11]Bird, S. Liberman, M. A Formal Framework for Linguistic Annotation[J]. Speech Communication, 2001, 33: 23 - 60.

[12]Hua Wu, Yin Zhigang. An Application of SAMPA - C in Standard Chinese[A]. Proc of ICSLP[C]. Beijing, China: ICSLP, 2000: 17 - 20.

[13]Li Aijun. Chinese Prosody and Prosodic Labeling of Spontaneous Speech[A]. Proc of Speech Prosody [C]. Aix - en - Provence, France: ISCA, 2002: 11 - 13.

[14]王宏,向大威. 基于长时平均频谱的文本无关话者识别[J]. 声学技术, 2002, 1 - 2: 89 - 92.

(责任编辑:香丽芸)