中 文 信 息 学 报
第 17 卷 第 4 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 17 No. 4

文章编号:1003 - 0077(2003)04 - 0027 - 06

基于综合因素的汉语连续语音库语料自动选取①

康 恒,刘文举

(中国科学院 自动化研究所 模式识别国家重点实验室,北京 100080)

摘要:大词汇量连续语音识别系统的性能很大程度上取决于语音库的质量,而语音库设计的中心环节就是语料选取。但是传统语料选取方法往往考虑因素单一,不利于语音识别系统有效利用语言信息。本语音库的语料选取方法综合考虑了多种因素:三音子覆盖率、三音子覆盖效率、三音子稀疏度、常用词分布等,并完全实现程序自动选取,充分利用了原始语料,使选取结果的信息量更加丰富。程序自动选取结果可以覆盖 94.1 %的三音子,75.4 %的最常用词,覆盖效率和稀疏度也比传统方法有了较大改善。

关键词:计算机应用;中文信息处理;语音库;三音子;高频词;覆盖率

中图分类号:TP391.42 **文献标识码**:A

Automatic Text Selection for Continuous Speech Database of Standard Chinese Based on Comprehensive Factors

KANG Heng, LIU Wen-ju

(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: The performance of continuous speech recognition systems depends much on speech database. Text selection is the key step in designing of the speech database. Conventional text selection methods consider too few factors for the recognition systems to use linguistic information effectually. This paper describes a method which can select text automatically and consider multiple factors: triphone covering rate, triphone covering efficiency, triphone sparse rate and distribution of commonly used words, etc. The set of selected text covers 94.1 % triphones, 75.4 % most commonly used words, and also the covering rate and sparse rate are improved than that of conventional methods.

Key words: computer application; Chinese information processing; speech database; triphone; commonly used words; covering rate

一、引言

目前汉语语音识别已经进入大词汇量连续语音阶段,连续语音库的质量很大程度上影响着语音识别系统的性能[1,2,5]。这个阶段对语音数据库有如下要求:

1. 在语音学和语言学指导下设计,在连续语音层面考虑声学语音学规律;

2. 语料具有典型性和代表性,而且在语料规模一定的条件下,语料应尽可能覆盖所有的语音语言现象;

① **收稿日期**:2003 - 02 - 26
基金项目:国家自然科学基金资助项目(60172055);国家"863"资助项目(2001AA114181);北京市自然科学基金资助项目(4002012)
作者简介:康恒(1978 —),男,硕士生,研究方向为语音合成.

3. 尽量保证每个语音单元在语料中出现的次数不至于太少,从而避免数据稀疏;

4. 综合考虑其它超音段特征,使得语音库包含尽量多的语言信息。

但是,传统语料选取方法实际上考虑的因素往往过于单一,大多只是关注于语音单元的覆盖率,而对于其它超音段特性并未加更多考虑,从而使得语音库的信息量不够大,影响系统的性能。

我们参考了其它同类语音库的设计[3~5],提出一种基于综合因素考虑的汉语连续语音库语料选取方法。该方法以句子对三音子的覆盖贡献为基本标准,同时兼顾三音子的覆盖效率、每个三音子的出现次数等因素。另外本方法还考虑到在语音库中尽量选用最具代表性的常用词组合来覆盖语音单元,使得语音库的适用性更强。总之,我们希望最终的语音库能够给汉语语音识别系统提供更多信息,为提高这些系统的性能提供坚实的基础。

下面首先介绍一下描述汉语连续语音的三音子结构,接下来给出语料选取的目标和算法,然后给出试验结果及对比,最后是总结和展望。

二、描述汉语连续语音的三音子结构

2.1 音子、双音子和三音子

汉语普通话是由音节连接而成,音节又是由更小的语音单元构成。通常将音子定为连续语音的最小单元。汉语普通话共有 37 个基本音子[3]。

虽然音子可以作为描述汉语普通话的最小单位,但在连续语音中,一连串音节紧密连接,发音部位和发音方法不断改变,音节之间互相影响,偏离原来的位置,导致其声学表现和孤立音节有很大区别[6]。这样,在设计连续语音数据库时,就不能仅用音子来描述连续语流中的现象。一般来说,可以使用双音子(diphone)或者三音子(triphone)来描述连续语音的音变和过渡[7,8]。

双音子代表介乎两个相邻音之间的声学音段,通常由一个语音单元的末尾部分跟下一个单元的开头部分组成。它包括两个音段各自稳定段的一部分和它们之间的过渡段。

三音子考虑了一个音子和其左右语言环境对其造成的影响,它包括音子本身(中心音子)以及和它左右相邻的音子之间的过渡段。采用三音子能够较好地描述语音的变化和过渡状态。

许多语音识别系统都是用三音子作为识别单元。也有一些语音合成系统采用部分三音子加双音子。在连续语音中使用三音子作为基本描述单元,然后经过合理的分类和精简,就可以很好地描述连续语音中的声学特性。在本语音库的设计中,也是主要采用三音子为基本单位来描述连续语音的。

2.2 音节间三音子的分类和精简

由 37 个基本音子组合而成的汉语三音子理论上有 8000 多个,数目相当大,然而在普通话中,并不是所有的音子组合都出现。在挑选语料时,考虑到三音子组合的数量比较多,容易造成每个语音单元在语料中出现的次数太少,即数据稀疏。所以,可以对这些三音子根据发音部位和发音特点进行合理的归类和精简。为了便于语料选取方法的比较,在此我们参考了 863 语音数据库的设计,沿用了他们的三音子分类和精简的方法[3]:

(1) 单音子音节构成的三音子,如 a(?*, ?*) (符号 *代表音节边界)

可将这些三音子看成两个双音子,如三音子 a(i*, n*) 可以看作是两个双音子 i-a 和 a-n。

(2) 中心音子为韵尾的三音子(右边是音节边界: ? (?, ?*))

将第二音节的声母按照发音部位归类,前一音节的韵尾向相同部位的声母过渡时,变化基本相同(鼻音除外)。所以前一音节的韵尾向后一音节的三十多个音节过渡可以简化为向六个部位过渡:

韵尾 → 双唇{b, p, m, f}
舌尖{d, t, n, l}
舌根{g, k, h}
齿间{z, c, s}
齿龈{zh, ch, sh, r}
舌面{j, q, x}

鼻尾(n, ng) → {b, p} + m, f
{d, t} + n, l
{g, k} + h
{z, c} + s
{zh, ch} + sh, r
{j, q} + x

(3)中心音子为声母的三音子(左边为音节边界:? (?*, ?))

所有的塞音和塞擦音前都有一闭塞段,可看作与孤立字发音情况相同。

(4)生僻搭配不计

与 ei、yo、nou 等生僻音节形成的三音子不予考虑,它们多为自由口语中使用,在朗读口语中基本不出现。

经过上述分类和精简,音节间三音字数量为 3035 个。

三、汉语连续语句语料的自动选取

3.1 连续语句选取的目标准则

根据大词汇量连续语音识别系统对语音库的要求,并考虑汉语连续语音的特点,我们对连续语句选取提出下面的目标准则[9]:选用三音子作为描述连续语音的基本单位,对其进行分类和精简,要求选取的句子包含尽量多的语言现象(三音子等);用尽量少的语料覆盖尽量多的语言现象,即保证一定的覆盖效率;考虑到连续语音识别中,高词频词的出现几率更多,所以要求选取的句子中包含尽量多的高词频词;考虑到避免数据稀疏,要求每个三音子都应该出现一定的次数;语料选取过程不需人工干预由计算机自动进行。

3.2 连续语句自动选取流程

针对上述的目标准则,我们设计了一个从原始语料库中选取语料的全自动算法。整个过程以三音子为中心,兼顾常用词分布和稀疏度的要求,以评估函数为手段。整个自动选取的过程如下:

(1)将原始语料库根据标点分割成句子形式,丢弃太长或者太短的句子,保留适当长度的句子(本实验选取句子的长度为 4 - 20 个汉字),丢弃含有特殊符号的句子;

(2)给每个句子分词,并根据句子中每个词在词频表中的分布调整句子库中句子的顺序,使得含有高词频词的句子被优先选取;

(3)根据拼音库对每个句子注音;

(4)根据每个句子中包含的语音现象的多少,用一个评估函数给每个句子打分,选取得分最多的若干句子;

(5)手工对选取得到的结果中的个别句子进行修整,作为最后语料选取的结果。

3.3 根据词频调整句子库

在一般的语料自动选取算法中,大多只关注于语音单元覆盖度。而由于生僻词组合和人名、地名等词语比普通词的语音单元覆盖效率更高,所以含有生僻词的句子被选中的可能性要大得多,这样语料选取结果不能代表真实语境,适用性不强。

实际上在连续语音识别和合成中,正常语境下,常用词出现的机率更大。所以我们考虑语料库应该使用较多的具有代表性的常用词组合覆盖尽量多的三音子,提高语音库的适用性。我们在语料选取之前首先对原始语料库中的句子按照词频进行调整,这样如果在句子包含语音现象相同的情况下,包含高词频词较多的句子有优先权。这就保证了在后续的语料选取过程中,包含常用词的句子首先被考虑。

在词频调整算法中,程序自动从事先统计好的词频表中查找每个词的词频,并根据词频对每个句子记分。最后根据每个句子的得分多少调整语料,让得分高的句子出现在最前面。在根据词频给每个句子记分的时候,我们对句中重复出现的词只记一次。

经过词频调整后的语料作为下一步语料选取的输入。

3.4 语料选取

语料选取过程是用一个评价函数根据三音子对覆盖率的贡献对每个句子打分,然后从原始语料中摘取出得分最高的若干句子。这个过程如下:

(1)初始化,置已经处理的句子数 n = 0,清空句子分数表 ssTable;

(2)从拼音形式的句子库中取一个句子,n = n + 1;

(3)根据三音子分割表,对这个句子分割三音子;

(4)使用评价函数给该句中出现的三音子打分 ss(n);

(5)把该句的编号 n 和该句的三音子得分 ss(n)保存到句子三音子得分表 ssTable 中;

(6)如果 n = = N,转到(7),否则转到(2);

(7)句子三音子得分表 ssTable 按三音子得分 ss(n)降序排序;

(8)根据排序后的 ssTable,从句子库中摘取得分最高的若干句子作为选取结果。

其中,第(4)步中,句子得分的评价函数的设计要求选取的结果要有一定的覆盖效率,又要求数据不过分稀疏,即每个三音子出现的次数不应太少。我们设计了一个类似 Hash 表的结构 TriphoneMap,该结构保存选取过程中已经出现过的三音子和对应的该三音子已经出现的次数。评价函数的算法如下所示:

(1)初始化句子得分 ss(n) = 0;

(2)取句子中的一个三音子:

如果该三音子在 TriphoneMap 中出现的次数 ts ≤= t1,则 ss(n) = ss(n) + s1;

如果该三音子在 TriphoneMap 中出现的次数 t1 < ts ≤= t2,则 ss(n) = ss(n) + s2;

如果该三音子在 TriphoneMap 中出现的次数 ts > t2,则 ss(n)不变;

(3)该三音子在 TriphoneMap 中出现的次数 ts = ts + 1;

(4)如果该句中的三音子全部处理完毕,转到(5),否则转到(2);

(5)ss(n) = ss(n) / trNumber,其中 trNumber 是该句中包含的三音子的总数。

其中第(5)步中,把三音子得分除以该句中包含的三音子总数 trNumber 是一个归一化处理,如果不进行这个处理则含有较多音节的长句子更容易得到较高分数。我们在实验中也考虑了其它归一化方法,比如句子得分除以句子音节数,但是这个方法则使得含有较多复杂音节(比如 zhang、qiong 等)的句子占有一定的优势,也不符合语音库的选取原则。

这个评价函数中的四个参数 t1,t2,s1,s2,要求 t1 < t2,s1 > s2。它们的值要根据实验来确定,保证即能达到一定的覆盖效率,又使得每个三音子出现的次数不至于太少。

另外,在实验中,我们发现如果对所有的语料用该算法进行一次性选取,效果并不理想。这是因为在原始语料规模较大的情况下,较后面出现的数据基本对选取结果没有影响。实验

中的数据表明,在算法处理了约 10000 句左右的时候,再后面出现的句子的得分绝大部分是 0 分。这样使得我们的数据有了很大的冗余,选取的结果还不够优化。所以我们采用了这样的策略:把原始语料的数据分割成几个单独的文件,然后对这几个文件分别采用这样的算法,每次都选出若干句子(比如 20000 句),然后把这些结果合并,再对合并后的数据使用这个算法,选出最后的结果。实验证明这个策略是行之有效的。

四、语料选取结果和分析对比

4.1 原始语料库

实验所使用的原始语料主要是《人民日报》2000、2001、2002 年的内容。另外由于《人民日报》的内容以政论性质的文章为主,句型过于单一,短语组合往往非常固定,包含的语音现象不够丰富,因此我们另外从《中国大百科全书》语言文字、教育、历史、政治等卷挑选出来部分内容。这样原始语料内容更加广泛,包含了更多的语言现象。

我们对这些原始语料进行分句并丢弃过长或过短的句子,形成了表 1 的原始语料库。

表 1 原始语料基本信息

	大小	句子数	音节数
人民日报 2000	24.5M	1,014,041	约 1280 万
人民日报 2001	23.7M	976,124	约 1240 万
人民日报 2002	20.0M	822,356	约 1000 万
大百科全书	2.54M	100,000	约 120 万

4.2 选取结果

表 2 是自动选取出来的三个库的统计结果,其中,“A”和“B”是从人民日报三年的数据中选取出来的,“C”是从《中国大百科全书》的数据中选取出来的。其中“B”主要是为了提高常用词出现几率和降低稀疏度,对三音子覆盖没有影响。使用时可以根据实际情况考虑是否选用“B”部分的数据。

表 2 语料库选取结果统计

文件	来源	句子数	三音子(累计)	覆盖率(累计)
A	人民日报 2000/ 2001/ 2002	1,296 句	2717	89.5 %
C	中国大百科全书	737 句	2856	94.1 %
B	人民日报 2000/ 2001/ 2002	1,506 句	2856	94.1 %

注: (累计)的统计项指的是该项目中后一个文件的统计结果包括了前一个文件的结果

对比 863 语音数据库的自动选取结果[3]:

863a	人民日报 1993/ 1994	1,560 句	2128	70 %
863b	百家报刊精选/ 电视剧本	625 句	2644	87 %

对比可以看出,本文的无需人工干预的自动选取方法实现了很高的三音子覆盖率,达到 94.1 %,比 863 数据库的自动选取结果高 7.1 %。

另外,我们还从以下几个角度与 863 语音库的数据进行了对比:

(1) 常用词分布

最常用的 4000 个词在库中出现的个数如表 3 所示。

从统计数据可以看出,本语音库覆盖了绝大部分最常用词,高于 863 语音库。这使得语音库的适用性更强,更能代表真实语境中的语言现象,有利于提高语音识别、合成系统的性能。

(2) 覆盖效率

我们定义:覆盖效率 = 覆盖的三音子数/ 语料总音节数。由于语音库要求尽量用较少的语料覆盖尽量多的语音单元,所以,该参数越高,说明语音库的效率越高。

从表 4 可以看出,本文方法用较少的语料实现了较大的覆盖率,覆盖效率比 863 库有了很大的改善。

表 3　最常用 4000 个词统计

文件	次数	百分比
863	2699 个	67.4 %
A + B + C	3016 个	75.4 %

表 4　覆盖效率统计

文件	三音子数	总音节数	覆盖效率
863	2947	33,121	0.08979
A + B	2856	21,322	0.13395

(3) 稀疏度

表 5 是两个数据库中出现次数大于 4 次的三音子个数的统计数据。

表 5　出现次数大于 4 次的三音子个数统计

文件	个数	百分比
863	1499 个	49.4 %
A + B + C	2073 个	68.3 %

一般来说,如果只考虑覆盖效率,就要求每个三音子出现的次数尽可能的少,所以覆盖效率和稀疏度是两个矛盾的因素。从表 5 可以看出,本文方法较好的解决了这个矛盾,使得选取的语料即能达到较高的覆盖效率,又解决了三音子稀疏问题。

五、总结和展望

本文描述了基于综合因素的连续语音数据库的语料选取过程,介绍了文本选取满足语音单元覆盖和词频调整的算法。

但是在语料选取算法的设计中,我们只考虑了音段方面和词一级的问题,对于连续语句的韵律结构并未作过多地考虑。而现阶段的连续语音合成技术对语音库的要求越来越高,要求语音库能较多地考虑韵律结构的覆盖和平衡。不过由于现阶段中文信息处理技术在这一层次还并未发展成熟,完全靠机器自动实现还比较困难。所以,考虑连续语音韵律结构的语音库设计是今后工作的重点。

参 考 文 献:

[1] Jan P. H. van Santen, Adam L. Buchsbaum. Methods for Optimal Text Selection [C]. Proceedings of Eurospeech '97. 1997, (2) :557 - 561.

[2] Helene Francois, Olivier Boeffard. The Greedy Algorithm and its Applicationto the Construction of a Continuous Speech Database [C]. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002). (5) :1420 - 1426.

[3] 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.

[4] 祖漪清. 连续语音数据库设计的科学性问题[Z]. 语音研究报告 http://www.cass.net.cn/s18-yys/yuyin/rpt-il/zuyq-98.htm.

[5] 吴华,徐波,黄泰翼. 基于三音子模型的语料自动选取算法[J]. 软件学报. 2000, 11 (2) :271 - 276.

[6] 林焘,王理嘉. "语音学教程"[M]. 北京:北京大学出版社, 1999.

[7] 曹剑芬. 普通话语音的环境音变与双音子和三音子结构[J]. 语言文字应用. 1996, (2) :58 - 63.

[8] 曹剑芬. 普通话双音子和三音子结构系统代表语料集[J]. 语言文字应用. 1997, (1) :60 - 68.

[9] 祖漪清. 实现语音数据库科学性的重要环节[J]. 语言文字应用. 1998, (1) :93 - 97.