基于三音子模型的语料自动选择算法*

吴华 徐波 黄泰翼

(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)

E-mail: wh@nlpr.ia.ac.cn

摘要 在语音识别中,如何经济地挑选语音训练语料,使其覆盖尽可能多的语音现象是一个非常重要的问题.传统的语音训练语料采用手工挑选后再进行检验和补充的方法,此方法难以保证所选语料语音现象的覆盖率.该文提出了一种自动地从大规模语料库中挑选语料的搜索算法,此算法不但能使所选语料覆盖几乎所有语音现象,而且能保证训练语料中三音子和类三音子有足够的样本个数,使训练数据不过于稀疏,为训练正确而可靠的语音模型打下了坚实的基础.

关键词 语音识别,模型训练,三音子,类三音子.

中图法分类号 TP391

在语音识别中,不管是语音模型还是语言模型的训练,都存在一个如何选择语料的问题.这些挑选的语料应具有典型性和代表性,而且在语料规模一定的情况下,应既使语料尽可能覆盖所有的语音语言现象,又使用于训练的数据不稀疏.在语音训练中,我们的实验表明,随机选择的训练语料与经过设计的训练语料相比,至少能带来5%的识别率的差别,可见,这个问题是相当重要的.

正是由于这个问题的重要性,国内一些从事语音识别的研究单位对此给予了高度的重视,如中国科学院自动化研究所在1995年设计的cosdic语音库语料[1],采用验证算法进行了词一级的语料选择;清华大学也在1995年构造了一个699个词的词表[2].但这些语料选择方法还存在如下的缺陷:① 语料选择的重点放在词一级的设计,而连续语音识别的对象为句子,势必影响识别效果;② 设计语料时主要考虑音节内和音节间的二元音联现象,而在连续语音中,每个音子同时受其前后音子的影响,存在强烈的协同发音现象.另外,1996年社会科学院语言研究所在"863连续语音数据库"项目的支持下,从人民日报语料中挑选了1 560句,并在此少量语料的基础上由不同人重复发音录制了一个语音数据库.由于该语料选择的部分工作是人工生成的,不但费时、费力,而且无法生成更大规模的语音训练库,因此,这种单一语料的训练不利于模型的鲁棒性.

汉语的音节是由声母和韵母组成的,其中声母由辅音构成,韵母由单元音或复合元音构成.一个汉语句子由许多音节组成,音节之间和音节内部都存在着强烈的协同发音现象,即声母或韵母的发音都受与其相邻的音素的影响.单独以声母或韵母建立语音模型解决不了协同发音的问题,从而影响语音识别的正确率.因此,在连续语音识别中需要建立三音子模型,即考虑声母或韵母左面和右面与之相邻的音素的影响.在挑选语料时,应使所选语料覆盖汉语中所有的三音子,并根据发音特点把三音子分类,以解决数据稀疏的问题(在训练语音模型时,必须保证每个三音子在语料中出现的次数不少于10次,才能基本保证模型的准确性,当出现次数过少时,就称为数据稀疏).

本文提出了一种以基于上下文决策树建模为应用背景的、以某个句子对三音子和类三音子的覆盖贡献的评估函数为依据的自动挑选语料的方法,此方法可以保证将覆盖较多语音现象的句子先挑选出来,并能解决数据

* 本文研究得到国家自然科学基金(No. 69835030)资助.作者吴华,女,1974年生,博士生,主要研究领域为语音识别,自然语言处理.徐波,1966年生,博士,研究员,博士生导师,主要研究领域为语音识别,人机通信.黄泰翼,1934年生,研究员,博士生导师,主要研究领域为语音合成、识别及处理,语言信息处理.

本文通讯联系人:吴华,北京 100080,中国科学院自动化研究所模式识别国家重点实验室

本文1999-01-07收到原稿,1999-03-18收到修改稿

稀疏的难题.

下面首先介绍三音子和类三音子的概念,然后给出挑选语料的标准及算法,最后介绍一些实验结果.

1 用于语音识别的三音子和类三音子

汉语的单音节由声母和韵母构成,声母由21个辅音组成,韵母由9个单韵母、13个复韵母和15个鼻韵母组成[3].汉语是单音节结构语言,在语音识别中考虑上下文相关模型(例如,三音子模型)时,一般把音节中的声母和韵母作为中心建模单位.而在考虑左右上下文交体时,只考虑其左面的声母或韵尾与其右面声母或韵头的影响.这种三音子模型可以写成X-Y-Z的形式,其中X代表左面与其相邻的声母或韵尾,Y代表声母或韵母,Z代表右面与其相邻的声母或韵头.考虑到语音训练中的静音模型,我们得到了8种类型的三音子模型:(1) 韵尾+声母+韵头;(2) 声母+韵母—声母;(3) 声母+韵母+韵头;(4) 韵尾+韵母+声母;(5) 韵尾+韵母+韵头;(6) 静音+声母+韵头;(7) 声母+韵母+静音;(8) 韵尾+韵母+静音.整个三音子的组成见表1.

Table 1 Constituents of triphones

表1 三音子的可能组成

X	Y		Z
	Initials①	Finals②	
a,o,e,er,i, i1,i2,u,v,n,ng, silence,21 (initials)	b,d,g,p,t,k, z,zh,j,c,ch,q,f, s,sh,x,h,m,n,l,r	a,o,e,i,iI,i2,u,v,er, ai,ei,ao,ou,ia,ie,ua,uo,ve, iao,iou,uai,uei,an,ian,uan,van,en,in, un,vn,ang,iang,uang,eng,ing,ong,iong	a,o,e,er,i, i1,i2,u,v,silence, 21 (initials)

说明:(1) X中的韵母尾"n"与声母中的"n"形相同,但前者不能处于Y的位置,在此分开讨论,

(2) i1是zi,ci.si中的i,i2是zhi,chi,shi中的i,

(3) 由于er的发音比较特殊,我们不把er拆成e为韵头和r为韵尾的形式,

(4) silence代表静音,

(5) 表中①为声母,②为韵母.

在普通话中,并不是所有的声母和韵母都可以组合,按照21个声母加上一个零声母和37个韵母组合,应该有814个音节,但实际上只有410个左右.同样地,在二音子的组合中,由于声、韵母的组合特点及韵母的构成特点,在不包括静音的情况下为22 295个,包括静音时为23 745个.

在挑选语料时,考虑到三音子的组合比较多,容易造成数据稀疏.而且,在语音模型的训练框架中,一般采用自上而下或自下而上的聚类方法来解决数据稀疏问题,例如,采用基于决策树的方法[4].所以,在挑选语料时,充分考虑这种在未能满足三音子的覆盖率的情况下,退而求其次而采用语音学意义上的类三音子来满足数据的覆盖率的方法就显得尤为重要.鉴于此,我们把X和Z中的声母按照其发音部位和发音特点分为8类[3].

(1) 不送气塞音: b,d,g;

(2) 送气塞音: p,t,k;

(3) 不送气塞擦音: z,zh,j;

(4) 送气塞擦音: c,ch,q;

(5) 擦音: f,s,sh;

(6) 鼻音: m,n;

(7) 边音: l;

(8) 通音: r.

分类后的三音子在不包括静音的情况下为10 374个,包括静音时为11 378个.可见,分类后的三音子个数减少了大约一半.

2 训练语料自动搜索的算法

我们的目的是从大规模语料中挑选一定数量的句子作为语音训练语料,结合汉语的特点提出了如下的设计

思想.

(1) 考虑到汉语句子中音节内和音节间的强烈协同发音现象,为了真实地反映协同发音现象,我们采用了上面提到的三音子模型.

(2) 根据在训练语音模型时存在的数据稀疏问题,把上节提到的 X 和 Z 中的声母按其发音部位和发音特点进行分类,这和声学训练中的聚类思想非常吻合.

(3) 我们考虑到在连续语音识别中,句子是一个基本单位,因此,我们挑选的对象是真实语料中的句子.

(4) 采用全自动的无需人工干预的挑选方法,而且,使用者可以按照自己的意愿挑选任意多的语料,这就解决了挑选语料少的问题,提高了语料的鲁棒性.

(5) 采用优先原则,包含语音现象最多的句子首先被挑选出来.

(6) 采用全面覆盖原则,保证选择包含有已选语料中未出现过的三音子的句子.

(7) 如果某个句子包含汉语中出现频率低的三音子,则此句子可被重复选择.

针对现有挑选语料方法的缺点和上述设计思想,我们设计了一种以三元音子模型为基础的、从真实语料中挑选句子的全自动语料选择算法.作为语料的自动搜索算法,我们必须有一种方法,能够评价一个句子所反映的语音现象的多少.因此,我们设计了一个评估函数,此函数能够保证上面提到的优先原则和全面覆盖原则等.

2.1 评估函数

评估函数计算的是实际语料中每个句子的得分,其目标有两个:① 满足优先原则和全面覆盖原则;② 解决数据稀疏问题,并保证在选中语料中将三音子样本次数的方差限制在一定的范围内.在挑选过程中,我们设计了两个表,其中一个为三音子表,存放所有的三音子及其在已选语料中出现的次数.另一个为类三音子表,存放所有的类三音子及其在已选语料中出现的次数.评估函数的实现形式如下:

```
如果句子中某三音子在相应表中计数为零,则
    如果其所属的类在类三音子表中的计数也为零,则
        score+=W3        //式中的"+="与C++语言中的赋值符号的意义相同
    否则,
        score+=W2
如果句子中某三音子在表中的计数大于零,则
    如果其所属类对应计数小于某一门限 δ1,则
        score+=W1+W4/所属类对应计数;
    如果其所属类对应计数大于门限 δ1 且小于某一门限 δ2,则
        score+=W1+W5/所属类对应计数;
    否则,
        score+=W1;
    计算整个句子的得分:
        score+=score/num;
```

其中 $score$ 是一个变量,评估每个句子对三音子和类三音子的覆盖贡献.贡献越大,则被选中的可能性越大.W_i,$i=1,2,3,4,5$ 是根据实验确定的权值,且 $W_3>W_2>W_1$,$W_4>W_5$,$\delta_2>\delta_1$,num 为句子所包含的三音子数目.

以上参数是根据实验确定的,其中 $W_3>W_2>W_1$,保证能够覆盖尽可能多的三音子;而 $\delta_2>\delta_1$ 的设置使找到的三音子不过于稀疏,而且使三音子的方差限制在一定范围内.

2.2 语料选择算法

整个算法是以三音子和类三音子为中心、以评估函数为评价手段的算法,其实现形式如下:

(1) 对实际语料进行预处理,将太长和太短的句子滤掉,并将包含字母书写、符号的句子过滤掉.

(2) 对处理后的语料注音.

(3) 初始化.置三音子和类三音子两个表为零,并置存放已选句子数的寄存器 n 为零.

(4) 输入供选择的语料.

(5) 根据三音子和类三音子由评估函数对每个句子计分.

(6) 对每个句子按照得分高低降序排列.

(7) 选择得分最高的句子,将寄存器 n 的计数加 1.

(8) 根据已选句子所包含的三音子和类三音子,更新两个相应的表.

(9) 检验 n 是否达到预置值,若是,则转(10),否则,转(5),继续循环.

(10) 结束.

3 算法的评测标准和实验结果

一个算法是否可靠,需要有一定的评测手段和标准.我们挑选语料的算法是否可靠以及在 *score* 函数中 W_3, W_2,W_1 和 δ_1,δ_2 的设置是否恰当,是以已被选中的语料所覆盖的三音子数目和类三音子数目以及相对应的平均次数和数据稀疏状况等作为指标来衡量的.显然,找到的三音子和类三音子数目越多,出现的平均次数越大,算法就越好.

3.1 实验用的原始语料

所有原始语料是从人民日报上挑选出来的,共有 5 个文件,每个文件经过预处理后的大小大约为 5M 字节.我们评价了挑选的原始语料对整个语音现象的覆盖率,表 2 是其结果,表中第 1 列中的"+"号说明右边的所有统计结果是在其前面文件统计结果的基础上得出的.5 个文件包含的三音子和类三音子总数分别为 20 377 和 9 505,对整个语音现象的覆盖率分别为 85.7%和 83.5%.针对上述统计数字,我们分析了语料中尚未包含的三音子的情况,它们是汉语中出现概率及其微小的情况.主要有两种类型,分别是"声母+韵母+韵头"和"韵尾+韵母+韵头"的情况,分别占 1/3 和 2/3.前者在汉语中表现为有声母音节和零声母音节相邻的情况,而且此韵头同时又可单独作为一个音节,如"o";后者主要是 3 个零声母音节相邻的情况,如"a-a+e"之类的三音子,在实际中这类情况几乎不出现.从上面的分析可以判断,原始语料是具有代表性和普遍性的.

Table 2 Initial information of corpus

表 2 语料的原始信息

	total number of included triphones①	total number of included class-triphones②
+File 1	17 929	8 447
+File 2	19 304	9 034
+File 3	19 862	9 265
+File 4	20 167	9 398
+File 5	20 377	9 505

①包含的三音子总数,②包含的类三音子总数.

3.2 实验结果

下面是我们所做的两个实验.实验 1 是对采用上节介绍的算法来解决数据稀疏和数据鲁棒性问题的评价;实验 2 是用我们的算法挑选出来的语料与语言所的 1 560 个句子以及等距抽样选择语料的方法的比较.实验 1 的结果见表 3,实验 2 的结果见表 4.

Table 3 Information of selected corpus

表 3 挑选后语料的信息

	+File 1	+File 2	+File 3	+File 4	+File 5
total number of selected triphones①	17 929	19 304	19 862	20 167	20 377
total number of selected class-triphones②	8 447	9 034	9 265	9 398	9 505
average frequency of each triphone③	21.9	40.4	63.5	82.7	98.5
average frequency of each class-triphone④	43.7	62.7	81.5	101.3	119.0
the number of triphones appearing more than ten times⑤	6 401	8 898	10 085	11 102	12 098
the number of class-triphones appearing more than ten times⑥	5 125	5 958	7 031	7 460	7 790

①被选三音子总数, ②被选类三音子总数,
③三音子出现的平均次数, ④类三音子出现的平均次数,
⑤在选中语料中出现次数超过 10 次的三音子, ⑥在选中语料中出现次数超过 10 次的类三音子.

在上述5个文件中,从每个文件中挑20 000句,一共100 000句,大约为3M字节.即每个文件只挑20 000句,并把前面文件所得到的三音子和类三音子两个表的信息都保留下来作为后面语料挑选的依据.实验时,W_1,W_2,W_3分别为20.0,18.0,2.0;δ_1,δ_2分别为20.0,1.0.从表2和表3中可以看出,此算法可以把原始语料中包含的所有三音子和类三音子都挑选出来,即覆盖了语料中所有的三音子和类三音子信息.而且,随着被选语料的增加,出现次数超过10次的三音子和类三音子数目都随之增加,但三音子所占总数的比例比类三音子所占总数的比例低大约30%,这说明,类三音子能大大降低数据稀疏的程度,这也正是我们把三音子分类的主要原因.因为现有的解决数据稀疏的方法是用少量的语料由不同的人录同样的语料,这样会带来我们在前面所提到的鲁棒性问题,而我们的算法能够全自动地挑选大量语料由不同的人念不同的语料,避免了语料单一性所带来的偏差.

表4是用我们的算法挑选出来的语料和语言所的1 560个句子及等距抽样方法的比较.为了保证比较的可行性,假设语言所的训练语料由64个不同的人来录制而成,共形成99 840个句子,与我们挑选的100 000个句子在数量上相当.很明显,前者的语料不管由多少人来录制,它所覆盖的三音子和类三音子的总数是不变的,分别只覆盖总数的30%和38%左右,等距抽样方法分别覆盖三音子和类三音子总数的71.8%和71.0%,而我们的语料可覆盖85.7%和83.5%.同时,前者的方差随着人数的增加而递增,它的数据离散度也会随之增大.同样,等距抽样方法的方差相对也比较大,而我们所得出的方差相对比较小,每个三音子在语料中出现的次数相对比较平均.可见,我们的算法在覆盖率和方差两方面都具有不可比拟的优越性.

Table 4 Comparison of training corpus

表4 训练语料比较

	100 000 sentences selected automatically①	1560 * 64 sentences selected by the Institute of Linguistics②	100 000 sentences selected through equidistant sampling③
total number of selected triphones④	20 377	7 307	17 049
total number of selected class-triphones⑤	9 505	4 374	8 078
average frequency of each triphone⑥	98.5	5.1 * 64	120.0
average frequency of each class-triphone⑦	119.0	8.6 * 64	269.8
variance of selected triphones⑧	450.1	11.5 * 64	471.0
variance of selected class-triphones⑨	673.5	16.4 * 64	764.6

①自动挑选100 000句, ②语言研究所1560 * 64, ③等距抽样选100 000句,
④被选三音子总数, ⑤被选类三音子总数, ⑥三音子出现的平均次数,
⑦类三音子出现的平均次数, ⑧被选三音子的方差, ⑨被选类三音子的方差.

4 结 论

本文提出了一种自动选取语音训练语料的行之有效的方法,它既能覆盖所选语料中包含的所有语音现象,又能解决数据稀疏的问题.从语音识别的角度看,训练语料越多、越全面,识别效果越好,而我们的算法在挑选保证适度语料的情况下,覆盖的语音现象也越多,数据稀疏的问题也能得到更好的解决.

参考文献

1 Qu Fei, Huang Tai-yi, Zhang Xi-jun. Acoustic library design for mandarin Chinese. In: Proceedings of the 4th National Conference on Man-Machine Speech Communication. Beijing, 1996. 337~341
(曲菲,黄泰翼,张希军.汉语综合语音库语料设计.见:第4届全国人机语音通讯学术会议论文集.北京,1996.337~341)

2 Sun Jia-song, Wang Zuo-ying, Wang Xia *et al*. Construction of the lexicon for continuous acoustic model training. In: Wu Quan-yuan, Gao Wen eds. Proceeding of the Improvement of Intelligence Computer Interface and Application'95. Beijing: Beijing University Publisher, 1995. 116~121
(孙甲松,王作英,王侠等.连续语音训练词表的构造.见:吴泉源,高文编.'95智能计算机接口与应用进展.北京:清华大学出版社,1995.116~121)

3 Lin Tao, Wang Li-jia. A Course Book of Phonetics. Beijing: Beijing University Press, 1991
(林焘,王理嘉.语音学教程.北京:北京大学出版社,1991)

4 Gao Sheng, Xu Bo, Huang Tai-yi. Class-triphone acoustic modeling based on decision tree for mandarin continuous speech recognition. In: Li Hai-zhou, Guan Cun-tai eds. Proceedings of the '98 International Symposium on Chinese Spoken Language Processing Symposium Processing. Singapore, The Chinese and Oriental Language Information Processing Society, 1998. 44～48

Automatic Corpus Selecting Algorithm Based on Triphone Models

WU Hua　XU Bo　HUANG Tai-yi

(*National Laboratory of Pattern Recognition　Institute of Automation. The Chinese Academy of Sciences　Beijing* 100080)

Abstract　In speech recognition, the selection of training corpus for robust acoustic modeling which can cover almost all phone phenomena is very important. Traditionally, corpus is selected manually first, and then tested and supplemented, which can't provide sufficient coverage of samples for various statistical modeling methods. An algorithm for automatically selecting the training samples from large-scale text corpus is developed in this paper. This algorithm can not only cover almost all phone phenomena but also ensure to include ideal samples of triphones or class-triphones and ensure enough data for training, which makes it possible to train acoustic model reliably.

Key words　Speech recognition, model training, triphone, class-triphone.