第 30 卷 第 2 期
2004 年 3 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 30, No. 2
Mar., 2004

基于统计方法的汉语连续语音中声调模式的研究[1)]

曹 阳 黄泰翼 徐 波

(中国科学院自动化研究所 北京 100080)

(E-mail: {cy, huang, xubo}@nlpr.ia.ac.cn)

摘 要 提出采用决策树的数据驱动方法,结合专家知识,从大规模语料中统计学习出连续语音中声调模式的分布. 在建立决策树的过程中,除了相邻音节的声调外,还考虑了多种可能影响声调模式的因素,如音节声韵母发音特点的分类、音节在词中的位置等. 决策树建立后,共得到 28 种声调模式. 通过对结果的分析发现,除了上下文的声调外,其它因素对连续语音中声调模式的变化也有一定的影响. 声调识别实验的结果证明了该方法的有效性.

关键词 声调分析,决策树,语音识别,声调识别

中图分类号 TP391

A Stochastically-Based Study on Chinese Tone Patterns in Continuous Speech

CAO Yang HUANG Tai-Yi XU Bo

(*Institute of Automation, Chinese Academy of Sciences, Beijing* 100080)

(E-mail: {cy, huang, xubo}@nlpr.ia.ac.cn)

Abstract A decision tree based approach is proposed for obtaining the quantitative result of tone variation patterns in continuous Chinese speech. While constructing the decision tree, besides neighboring tone, many other possible factors are considered such as syllable position in the word and consonant/vowel type of the syllable, which are not studied in conventional analysis. After the tree is established, 28 different tone patterns and their corresponding model parameters are acquired. From the result it is found that many factors in addition to the tone of neighboring syllable affect tone pattern variation in continuous Chinese speech. Tone recognition experiments demonstrate the effectiveness of this approach.

Key words Tone variation patterns, decision tree, speech recognition, tone recognition

1)国家自然科学基金(69835003)和国家"973"项目(G1998030504)资助
Supported by National Natural Science Foundation of P. R. China(69835003) and National Grand Fundamental Research "973" Program of P. R. China(G1998030504)

收稿日期 2002-12-17 收修改稿日期 2003-03-28
Received December 17, 2002; in revised form March 28, 2003

1 引言

众所周知，在汉语连续语音中，受上下文声调等语境的影响，声调的模式和孤立音节时相比发生了很大的变化. 影响声调模式变化的因素不仅包含了上下文的声调，其它的因素如结构信息等也可能有一定的作用. 多年来，在语音学、语音识别和合成领域，许多专家学者在这方面进行了大量的研究分析，取得了很多研究成果.

黄泰翼等在文献[1]中对二字词中可能出现的声调模式变化作了研究分析和定性的描述. 吴宗济[2]曾分别对汉的二字组、三字组、四字组声调模式做过研究，给出一些定性的分析结果. 文献[3,4]总结了应用于语音合成的基频曲线生成规则. 文献[5]采用了定性分析总结出应用于声调识别 23 种声调模式. 此外，在生成音系学领域内也有大量的关于声调变化规律的研究结果.

但上述结果基本上都属于一些定性的描述，很难直接应用到语音识别和合成中，也没有涉及到结构信息对声调变化模式的影响. 在研究方法上，传统研究主要通过手工分析以及定性观察. 这种研究存在着以下一些不足：首先，语音数据变化随机性很大，对少量的语音数据分析处理的结果不能得到比较全面的声调变化规律，而由于工作量等原因，手工分析也不可能处理大量的语料；其次，传统的分析往往很重视特例的分析，如 3-3-3 变调的分析，而对普遍规律的分析不够；最后，传统分析得出的规律是一种很模糊的描述，如“半上”、“降升调”、“变成二声”等.

由以上分析可知，要克服传统经验分析的缺点，必须解决：1)对数据随机性的描述；2)从大规模语料中得出结果；3)全面考虑到可能影响声调模式变化的因素. 对于第一个问题我们通过采用随机多项式曲线声调模型来解决[6]，对第二和第三个问题，有效的解决办法是采用数据驱动和专家知识引导的聚类算法. 因此，本文提出采用数据驱动和专家知识结合的聚类的方法，采用随机多项式曲线模型描述声调模式，从大规模语料中统计学习出连续语音中声调模式变化规律的定量结果. 在实现中，我们采用决策树[7]作为基本的聚类工具，它具有以下几种优点：1)数据驱动，所有的结果都是语料聚类的结果，这保证了结果的客观性；2)专家知识引导，可以把有关语音学和语言学的先验知识和决策树的生成过程结合起来，利用专家知识指导决策树的分裂，从而把专家知识和声调模式的聚类过程有机地集成在一起，提高了分类学习的准确性；3)预测能力，可以预测出训练语料中未出现的模式，对于这样的模式可以通过遍历决策树找到与之最匹配的模式.

2 决策树的建立

在我们的研究中，选用二叉树做决策树. 在应用决策树时，必须解决如下几个问题：问题集的设计，评估函数的选择，停止分裂的准则. 通过优化上述参数，以获得更合理的结果.

2.1 问题集的设计

在决策树聚类过程中，有关语音学和语言学的先验知识是以问题的形式来描述的. 因为先验知识指导着决策树的分裂，所以问题集的设计就成为建立决策树时必须重点考虑的问题之一. 为了尽可能准确的分析出连续语音中声调变化的规律，以及影响这些变化的因素，

在问题集中包含了广泛的各种可能影响声调模式变化的因素，它包含以下几类问题.

2.1.1　音节的的声韵母发音特点分类

现有研究表明，不同的音段对基频 *F*0 有一定的影响，因而也就可能对声调模式产生影响，因而我们在问题集的设计中包含了声韵母发音特点(发音方式及部位)及相关问题. 在声母中，浊声母发音时由于声带振动，所以与其它清声母加以区分. 而塞擦音、塞音、擦音由于发音方式不同，对后面韵母的基频可能产生不同的影响，所以归为不同的类，其中塞擦音、塞音又按送气和不送气分开. 因此，声母共分 6 类，具体的分类见表 1.

表 1　声母问题集分类

Table 1　Question set for initials

浊声母	清声母				
	擦音	塞音		塞擦音	
{*m*, *n*, *l*, *r*, "*null*"}	{*f*, *s*, *sh*, *x*, *h*}	不送气 {*b*, *d*, *g*}	送气 {*p*, *t*, *k*}	不送气 {*j*, *zh*, *z*}	送气 {*c*, *ch*, *q*}

在韵母中，单韵母、复韵母、鼻韵母三者发音过程有很大的不同，因此也会对声调产生不同的影响；复韵母中的后响复韵母中的介音都读得相当短，声学特性主要取决于主元音，因此按其主元音与相应的单韵母分成一类；中响复韵母的韵头部分相当短，因此按其后半部分与相应的前响复韵母归为一类；鼻韵母中发音部位或韵腹不相同的韵母，其声学特性不同，对基频的影响也不同，因而按发音部位和韵腹分开. 这样，韵母共分 15 类，见表 2.

表 2　韵母问题集分类

Table 2　Question set for finals

单韵母和后响复韵母			复韵母		鼻韵母	
{*a*, *ia*, *ua*}	{*o*, *uo*}	{*u*} {*v*}	{*ai*, *uai*}	{*ei*, *uei*}	{*an*, *ian*, *uan*, *van*}	{*ang*, *iang*, *uang*}
{*e*, *ie*, *ve*}	{*i*}	{*er*}	{*ao*, *iao*}	{*ou*, *iou*}	{*eng*, *ong*, *ing*, *iong*}	{*en*, *in*, *uen*, *vn*}

2.1.2　该音节前后音节的声调

声调很明显是受上下文声调的影响的. 根据分析，连续语音中声调的变化是以二音节变调为基础的，因此我们不考虑远程因素的影响.

这样，上下文声调的问题分为：针对前音节提出的 5 个问题，包含前音节的声调是否是一声、二声、三声、四声、轻声；针对后音节提出的 5 个问题，与前音节提出的 5 个问题类似. 另外在现有的结论中通常认为一声、二声(三声、四声)对后面音节声调的影响是相似的，一声、四声(二声、三声)对前面音节声调的影响是类似的. 因此，为了提高问题集的广泛性，在问题集中包含了 4 个组合问题：前音节的声调是一声或二声，前音节的声调是三声或四声；后音节的声调是一声或四声，后音节的声调是二声或三声. 这样共有 14 个关于上下文声调的问题.

2.1.3　该音节在它所在的词中的位置

通过定性的观察分析，发现位于词边界上的音节声调受协同发音的影响和词中间音节的声调所受影响并不相同(其原因可能是词边界处由于不连续大大减弱了受其它词音节影响的可能)，因而在问题集中包含了关于词位置的 4 个问题：是否是单字词，是否在词首，词尾，还是词中. 另外，在句子中静音段前后的音节的声调变化和其它位置上音节的声调明显

不同,因此问题集中还包含了 2 个静音问题:音节左边是静音,音节右边是静音.句首音节归入左边是静音,句尾音节归入右边是静音.

还有其它的一些因素也可能会影响声调模式的变化,如语速、轻重音、语句的韵律结构等.但由于现有的语音库中并不包含轻重音的标注,手工标注的工作量太大,对于语速也有类似的情况,因此在本文的研究中未包括轻重音和语速的影响.对于语句的韵律结构,如韵律词、词组、短语等,由于本身的定义和标注方法还不是非常清晰,因此本文也未考虑它们的影响.

2.2 评价函数和停止准则

决策树从根节点开始分裂,一直到满足停止条件而分裂结束.在分裂的过程中,需要一个评价函数来确定分裂后是否"值得".评价函数是为了度量样本之间的相似度,在本文中采用随机多项式声调模型描述声调模式,因此评价函数为样本对模型的似然概率.决策树某一节点的似然概率计算公式如下:

$$L(S)=\sum_{x\in S}\log P(X)=\sum_{x\in S}\log(P_f(X)P_l(X)^{\eta}) \tag{1}$$

上式中 S 表示由属于节点 n 的样本组成的样本集合,$P_f(X)$表示样本 X 对描述本节点的随机多项式曲线声调模型的概率,$P_l(X)$表示时长模型的概率,η是权系数.

当决策树的某一个节点(称之为父节点)分裂成两个节点(称之为子节点)时,似然度会增加.假设 P 表示父节点 n 的样本集合,P_{left}和 P_{right}分别表示按照问题 q 分裂后的两个子节点的样本集合,并且它们满足关系 $P=P_{\text{left}}\cup P_{\text{right}}$,$P_{\text{left}}\cap P_{\text{right}}=\varnothing$.那么当父节点分裂成两个子节点时似然度的增加为

$$m(n,q)=L(P_{\text{left}})+L(P_{\text{right}})-L(P) \tag{2}$$

停止准则决定了决策树何时停止.为了保证每一步结果的可靠性,必须确保每个节点有充分的样本,这样我们的停止准则是,当一个节点的样本少于一定的门限或者 $m(n,q^*)=\max_q m(n,q)$小于一定的门限时,本节点停止分裂,标记为叶节点.

2.3 决策树生成

解决了问题集的设计、评价函数、停止准则后,我们可以在预处理生成的数据上构造决策树.通常的二叉决策树的生成算法中,每一步都选当前最优的问题分裂,这种生成算法是一种局部最优的算法,很容易陷入局部最优.但是要直接寻找全局最优的计算代价太大,在这里我们采用了一种改进的算法,其思想是向前预测一步,以得到更合理的估计.定义

$$p(n,q)=m(n,q)+\max_{q1}m(n_{\text{left}},q1)+\max_{q2}m(n_{\text{right}},q2) \tag{3}$$

上式中 n_{left},n_{right}表示按照问题 q 分裂后的两个子节点.$p(n,q)$表示了按问题 q 分裂后,向前预测一步后的相似度增加.显然 $p(n,q)$比 $m(n,q)$描述了更多的全局信息,以 $p(n,q)$作为选择节点最优分裂函数要比 $m(n,q)$更为合理.这样对于节点 n,最优的分裂问题为

$$q^*=\arg\max_q p(n,q) \tag{4}$$

但是,直接计算 $m(n,q)$的代价非常高,为了减少计算时间,必须对搜索空间作压缩.一般而言,使 $p(n,q)$最大的问题 q^*,对应的 $m(n,q^*)$应该是 $m(n,q)$的某个极值,因此,可以把搜索空间减小到的 $m(n,q)$的前 N 个极值点.

这样,对于任意声调,其决策树的建立过程如下.

1)选择预处理后的样本集 X.定义开始节点为根节点,它包含样本集中的所有样本,估

计出该节点的随机多项式曲线声调模型参数，从而计算出节点的相似度，标记根节点为“没有处理过”.

2)从所有节点中选择一个“没有处理过”的节点 n，如果当前节点所包含的样本数小于停止门限，则记该节点为叶节点；否则，对问题集中该节点未处理的每一问题 q 都计算该节点. 如果这个问题分裂为两个子节点时(一个子节点对该问题回答“Yes”，一个子节点回答“No”)，则计算两个子节点的随机多项式模型，以及相似度和相似度的增加值 $m(n,q)$. 由此求出 $m(n,q)$ 的 N 个极值以及对应的问题，并由此求出 q^*. 如果 $p(n,q^*)$ 小于设定的门限，该节点就停止分裂，标记为叶节点；否则，按问题 $m(n,q)$ 对该节点进行分裂，并记录该问题，标记该节点“已经处理过”，同时标记新产生的节点为“没有处理过”，估计新产生节点的模型参数，计算出节点的相似度.

3) 如果所有的节点都已经处理过，则决策树已经形成；否则，转入 2).

3　实验结果

3.1　数据库

研究用的数据库为 863 连续语音库，在语音库的设计过程中，已经尽量覆盖了音段和超音段的音联现象. 我们选取了其中 8 个男声数据和 8 个女声数据，共 9740 句，110696 个音节，其中包含一声数据 23048 个，二声数据 27128 个，三声数据 16372 个，四声数据 37712 个，以及轻声数据 6472 个. 库中不同的句子共有 2435 句，也就是说每一句共有 4 个不同人的样本. 这样大规模的全面均衡的语料保证了统计学习结果的可靠性.

3.2　预处理

预处理完成将数据库中的数据转换为决策树的输入数据的过程，它包括基频提取和归一化、声韵母切分和标注、训练语料的文本分析、语料综合以及特殊变调纠正等步骤，最终生成带属性标注的样本序列.

基频提取采用的是基于动态规划的集成算法，然后统计每一发音人的基频上下限，最后按说话人的基频上下限对每一句的基频进行归一化处理以消除不同说话人的影响. 我们采用语音识别系统对语料进行了自动的声韵母切分，为了保证切分的准确性，识别系统的声学模型采用 32 个混合高斯密度的三音子(triphone)模型. 文本分析的目的是为了获得训练语料中每一个音节的语境参数，其主要步骤为文本预处理、分词、分词后处理、声韵母分类. 文本预处理是将训练语料中的数字、字母以及其他符号转换成可处理的汉字. 分词采用了双向的机械分词算法，分词用词典共包含词 39926 个，词长从 1 字到 7 字词，分词结束后手工对由词典中不存在的人名、地名等分成的单字等进行合并. 特殊变调的纠正是针对一些特殊字(主要指一、不)的声调在组合语音环境中发生规律性的变调进行纠正. 预处理的最后步骤是对训练样本中所有的文本扫描，用属于同一音节的所有样本标注形成本声调的样本库. 这样，每一声调的样本库由以下的每一样本音节的结构组成：本音节的韵母部分时长、本音节的韵母部分的归一化基频序列、本音节在所在词中的位置以及本音节左右是不是静音、本音节的声母分类、本音节的韵母分类、本音节前一音节的声调分类、本音节后一音节的声调分类.

3.3　决策树学习结果和分析

在决策树学习过程中，我们为每一种声调建立了一颗决策树，决策树建立后的每一个叶

节点对应了一种声调模式,叶节点上的问题则对应了这种声调模式的语音环境.决策树的参数设置如下:随机多项式曲线的阶为 3;停止准则为节点样本数目小于 300 或分裂的最大评估函数增加小于样本数乘以 0.1;最佳问题的搜索空间为 2.

通过决策树学习,最终达到 28 种声调变化模式,其中一声 3 种模式,二声 8 种模式,三声 8 种模式,四声 5 种模式,轻声 4 种模式.我们对最终模式的问题进行了分析,得出了声调模式受哪些因素的影响,以及影响程度的结果:

1)关于音节声韵母发音方式分类问题,占总因素的比例为 19%;

2)声韵母分类问题中以声母是否是浊声母为主;

3)关于音节在词中位置的问题,占总因素的比例为 19%;

4)关于上下文声调(前后音节声调)的问题,占总因素的比例为 62%;前、后音节声调的影响大致相当.

3.4 与传统研究结果的比较与新的发现

对上述结果与传统方法研究所取得的结果进行比较分析,可以得到以下几点结论:

1)本方法所生成的声调变化模式比传统模式更全面和细致地描述了连续语流中的变调现象,传统研究结果只是我们研究结果中的一个子集;如对于三声在连续语音中的声调模式的变化,我们的实验结果不仅包含了传统研究得出的的三三变调规律(对应声调模式 12),而且得到了大量传统研究中未能得出规律的变化模式;

2)特别要提到的是研究结果揭示了影响连续语音中声调变化模式的多种因素,如所在音节中声韵母的发音方式、音节在词中的位置等,这些因素都是不可忽略的;而传统研究只考虑了前后音节声调的影响;

3)研究结果纠正了传统研究的一些错误,例如传统分析中认为一声和二声(三声及四声)对随后的音节声调、一声和四声(二声和三声)对前一音节声调影响相似,而从我们研究的结果可以看出它们的影响并不相似.

3.5 在声调识别中的应用

我们将数据驱动所获得的声调模式变化学习的结果应用于连续语音中的声调识别,分别对 HMM 模型及随机多项式声调模型进行了识别实验.测试集来自 863 的连续语音库,共包括 4 个男声和 4 个女声数据,共 4370 句,55348 个音节.

在 HMM 模型的识别实验中,我们共进行了三类模型的比较,第一种是不考虑语境影响的声调模型,第二种是考虑了前后音节声调影响定性总结出的 23 种前后音节相关 HMM 声调模型,第三种是由决策树生成的 28 种声调模式对应的 HMM 声调模型,结果如表 3 所示.实验中所采用的 HMM 模型状态数为 5,输出分布为 3,每一个输出用 4 个混合的高斯分布来描述.表 3 中最后一行给出采用决策树生成模型后误识率下降程度.

表 3 HMM 声调模型声调识别结果(识别率%)

Table 3 Tone recognition based on HMM tone models(accuracy %)

	一声	二声	三声	四声	轻声	平均
语境无关模型	71.1	54.2	40.0	62.3	30.9	57.1
前后音节相关模型	72.5	61.3	55.4	67.1	46.3	63.9
决策树生成模型	73.4	64.2	59.1	68.8	54.1	66.3
误识率下降(%)	7.96	21.8	31.8	17.2	33.6	21.4

从实验结果可见，当采用考虑了前后音节声调影响的模型后，除一声的改善很不明显外，三声、二声和轻声的识别率得到明显的改善，这反映了三声、二声和轻声受音联影响严重，而一声相对几乎不受上下文的影响. 当采用了决策树聚类的模型后，识别率又得到进一步提高，结果说明定量分析的结果要优于定性观察的结果，同时也表明影响连续语音中汉语声调模型变化的不仅仅是上下文的声调，还应该包括音节声韵母分类和音节的位置等因素.

同时，我们还采用了作者在文献[6]中提出的随机多项式声调模型进行了声调识别测试. 测试模型共分两种，第一种是语境无关的随机多项式模型，第二种是决策树生成的语境相关随机多项式模型，其结果如表 4 所示. 实验中，随机多项式声调模型的阶数为 3，方差数为 4，实验中未包括时长模型.

表 4　随机多项式声调模型声调识别结果(识别率%)

Table 4　Tone recognition based on stochastic polonomial tone models(accuracy %)

	一声	二声	三声	四声	轻声	平均
语境无关模型	74.9	54.1	55.0	62.8	38.3	60.6
决策树生成模型	76.6	69.5	66.1	68.4	52.9	69.1
误识率下降(%)	6.8	33.6	24.7	15.0	23.7	21.6

由结果可见，当采用了决策树聚类的模型后，识别率获得了很大的提高，这表明了决策树分析结果的有效性. 与表 3 相比较，还表明随机多项式模型可以更好地描述声调变化的模式，性能优于 HMM 模型.

4　结论

本章讨论了如何应用决策树的聚类算法通过大规模语料学习连续语音中声调模式的变化. 在决策树的问题集设计中，我们广泛地考虑了各种可能影响声调模式的因素. 通过决策树学习的结果及相应的实验得出以下的结论：

1)前后音节的声调是影响连续语音中声调模式变化的最重要的因素；音节的声韵母发音特点(发音方式和部位)的分类、音节在词中的位置对声调模式的变化也是非常重要的；

2) 根据决策树学习的结果建立的声调模型能显著提高声调识别的准确性.

References

1　Huang Tai-Yi, Wang Cai-Fei, Yoh-Han Pao. Speech analysis for Chinese putonghua(Mandarin). In: Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing. Atlanta: IEEE Press, 1981. **1**:370～373

2　Wu Zong-Ji. Tone variation in Chinese language. *Chinese Language*, 1982, **28**(6): 439～449(in Chinese)

3　Lee Lin-Shan, Tseng Chiu-Yu, Ming Ouh-Young. The synthesis rules in Chinese text-to-speech system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1989, **37**(9):1309～1320

4　Lee Lin-Shan, Tseng Chiu-Yu, Hsieh Ching-Jiang. Improved tone concatenation rules in a formant based Chinese text-to-speech system. *IEEE Transactions on Speech and Signal Processing*, 1993, **1**(3):287～294

5　Wang Hsin-Min, Ho Tai-Hsuan, Yang Rung-Chiung *et al*. Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data. *IEEE Transactions on Speech and Audio Processing*, 1997, **5**(2):195～200

6　Cao Yang, Huang Tai Yi, Xu Bo, Li Cheng-Rong. A stochastic polynomial tone model for continuous Mandarin

speech. In: Proceedings of International Conference on Spoken Language Processing. Beijing: 2000. **3**:674～677

7 Bahl L R, de Souza P V, Gopalakrishnan P S, Nahamoo D, Picheny M A. Decision tree for phonological rules in continuous speech. In: Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing. Glasgow, Scotland: 1989. **1**: 185～188

曹 阳 1994 年于上海交通大学电机工程系获学士学位,2001 年于中国科学院自动化研究所获博士学位,现工作在北京 Nokia 研究中心. 主要研究领域为语音识别和语音合成.

(**CAO Yao** Received his bachelor degree from Shanghai Jiaotong University in 1994 and the Ph. D. degree from Institute of Automation, Chinese Academy of Sciences in 2001. His research interests include speech recognititon and speech syntthesis.)

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研究员. 主要研究领域为口语处理、语音识别及理解、语音翻译以及基于统计方法的语言处理.

(**HUANG Tai-Yi** Professor in National Laboratory of Pattern Recognition at Institute of Automation, Chinese Academy of Sciences. His research interests include spoken language processing, speech recognition and understanding, and speech translation and statistical natural language processing.)

徐 波 中国科学院自动化研究所模式识别国家重点实验室研究员. 主要研究领域为口语处理、语音识别及理解、分布式语音处理、语音翻译和基于统计方法的自然语言处理.

(**XU Bo** Professor in National Laboratory of Pattern Recognition at Institute of Automation, Chinese Academy of Sciences. His research interests include spoken language processing, speech recognition and under standing, distributed speech recognition, and speech translation and statistical natural language processing.)