第32卷 第21期 计 算 机 工 程 2006年11月
Vol.32 No.21 Computer Engineering November 2006

·博士论文· 文章编号：1000—3428(2006)21—0003—03 文献标识码：A 中图分类号：TP391

基于决策树的语料库分析

崔丹丹，蔡莲红

(清华大学计算机系普适计算教育部重点实验室，北京 100084)

摘 要：利用CART决策树算法，对TH-CoSS语料库音节的韵律参数进行聚类，分析语境特征的分布：出现率、平均层级。为了评价语境特征对语音韵律表现的影响程度，设计了一个影响权重的重要性函数，对语料库文本设计和TTS系统选音参数的权重设定具有较高的参考价值。

关键词：语境特征；韵律参数；TH-CoSS；CART；影响权重

Corpus Analysis Based on Decision Tree

CUI Dandan, CAI Lianhong

(Key Lab of Pervasive Computing, Ministry of Education, Dept. of Computer, Tsinghua Univ., Beijing 100084)

【Abstract】This paper uses a CART to cluster the syllables in the TH-CoSS corpus by their prosodic parameters, analyses the distribution of context features (appearance rates and average levels), and proposes an importance function of context features to evaluate their weights of influence on the prosody of speech, which shows to be a valuable reference to both text script design of speech corpus and weight setting in TTS unit selection.

【Key words】Context features; Prosodic parameters; TH-CoSS; Classification and regression trees(CART); Weight of influence

1 概述

近年来，随着语料库建设向大规模高层次发展，基于大语料库的TTS系统的合成效果也得到了大幅度的提高。然而面对大规模的语音数据，如何分析语音特征及其分布、评价语料的科学性是人们面临的新问题。

对语音特征的分析已有很多的研究成果：王蓓通过音节相似度听辨实验考察音节间协同发音现象所引起的音节知觉差异，并进行韵律边界的知觉标定实验，统计韵律边界声学特征的平均值来研究韵律边界的声学表现；王蕴佳通过重音感知实验，归纳双音节词重音感知与韵律边界和音节调型的关系；刘涛则利用k中心点算法，基于基频包络参数对有调音节样本进行无监督聚类。但要综合评估语境特征对语音自然度的影响，简单的统计方法难以胜任。机器学习、决策树则能揭示出语音内在规律和相关信息。

CART(Classification and Regression Trees)树是一种二叉决策树模型。通过自上而下的逐层分裂，生成由最相似样本组成的叶节点(类)，从而建立起特征与样本相似性之间的联系。其中，每个分枝节点对应一个问题(测试属性)，根据回答的“是”或“否”来决定样本属于左边或右边的子节点，直到分裂为叶节点；问题选取的标准是能够使样本集的纯度增量(信息增益)最大。

本文将决策树方法用于语料库分析。利用CART决策树，对标注好的大规模语料库TH-CoSS中的音节，按照韵律参数向量的距离进行聚类。分析了常用的9种不同语境特征(问题集)对聚类结果的影响，统计出每个特征在决策树中的出现率以及平均层级，并设计了一个权重函数，计算出每个语境特征的权重值，综合衡量语境特征对韵律参数的相对影响。分析结果表明，音节在低层韵律结构中的位置特征对语音的韵律表现具有较大的影响，声调音联后音节比前音节影响略大，按首尾音分类的音段音联特征则表现出比按声韵母分类更大的影响。本研究成果可为语料库语料选取和TTS系统选音等工作提供参考。

2 分析用语料库简介

本文分析的语料库是面向语音合成的语音数据库TH-CoSS。这是由清华大学创建，适用于语音分析、语音合成系统开发和评测的朗读语音数据库。

TH-CoSS包含多个不同规模的子库，其中03MR00和03FR00分别录制了男/女声播音员的朗读语音。包括主体语句、系统测试语句、特殊音节词、其它语调语句等，音节个数为85 000多个。数据的采样率为16kHz，量化精度为16位。语料库中的文本信息包括汉字、拼音、声调和韵律边界(韵律词、韵律短语、句子)标记；语音波形数据中标注了音节切分和基频信息。所有标注信息均经过人工校对。

本文将重点对主体数据进行分析。主体数据以陈述句为主，长度为5~25个音节，语速约为每秒2.7个音节。涵盖了汉语普通话中的全部有调音节，以及丰富的音联现象和韵律表现，且分布率与自然语音相似。表1给出了女声主体数据(03FR00)所包含韵律单元的个数。

表1 03FR00中主体语料的韵律单元统计结果

韵律单元	句子	韵律短语	韵律词
个数	5 406	16 769	44,658
平均音节数	18.3	5.9	2.2
最多音节数	24	9	4

可以看出，该语料的韵律成分数量充足，且平均分布较

基金项目：国家自然科学基金资助项目(60433030，60418012)

作者简介：崔丹丹(1981－)，女，博士生，主研方向：语音合成和语料库；蔡莲红，教授

收稿日期：2005-12-09 **E-mail：**cuidd02@mails.fsinghua.edu.cn

为合理。本文对该语料进行聚类分析，以期反映出中性风格下语境信息对自然语音韵律表现的影响。

3 决策树聚类

采用决策树来分析语料库中语境特征对韵律表现的影响，首先要解决如何选取合适的特征构成问题集以及如何度量样本距离两个问题。本文从音节位置、前后音联角度选取了 9 个常用语境特征构成 CART 树的问题集，用时长、能量和基频向量所构成的韵律参数向量来度量样本间的距离。节点分裂时，按照样本方差减少量最大的标准来选择问题。

3.1 语境特征的选取

通常韵律结构可以表示为语句、韵律短语和韵律词 3 层。音节在韵律结构中的位置是影响语音韵律表现的重要参数。在连续语流中，音联导致音节声学特征改变，如连读变调和变音。本文采用声调音联和音段音联来刻画之。分析中从音节位置和前后音联两个角度选取了语料库设计和 TTS 系统选音中常用的 9 个语境特征构成决策树的问题集。语境特征集简述如下：

(1) 音节位置：本文选用音节在韵律词中的位置(PinW)、在韵律短语中的位置(PinP)、在句子中的位置(PinS) 3 种位置信息。

(2)音联特征：

声调音联：前音节声调类型(LT，5 种)，后音节声调类型(RT，5 种)。

音段音联：

后音节首音类别(RIC，8 类)，前音节尾音类别(LFC，3 类)。

后音节声母类别(RTp，8 类)，前音节韵母类别(LTp，4 类)。

3.2 韵律参数向量的构成

时长、基频、能量都是与语音感知密切相关的重要韵律参数。本研究中，选取时长、均方根能量和基频向量构成衡量样本距离的韵律参数向量。

(1)时长 D，音节的时长，以采样点为单位。

(2)能量 E，音节的均方根能量。

(3)基频向量 $P=\{p_1,p_2,p_3\}$，其中 p_1，p_2，p_3 分别是该音节长度的 0.15、0.5、0.85 处的基频平滑值。

4 聚类结果统计

本文采用上述语境特征作为问题集，将韵律参数用于度量样本距离，对语料库的有调音节进行决策树聚类。男女声语料各生成超过 8 000 个叶节点，叶节点内的样本个数为 3~39。

决策树生成过程中，问题的出现率及其在决策树中的层级，反映了相应的语境特征对韵律表现的影响。为此作者统计了聚类结果中各决策树的规模。计算语境特征在决策过程中出现的次数及其所处的层级，以反映不同语境特征的信息增益。

4.1 聚类结果概述

聚类生成了大量的音节类(叶节点)，而每个类包含一定数量的韵律表现相近的音节样本。决策树则代表了语音中丰富的韵律表现。表 2 对比了语料库中各声调的音节数与决策树中出现的语境特征数(分枝节点数)。

考虑到音节声调不同将导致分类的差异，表 2 按声调列出统计结果。实际上，声韵母、元辅音等对语境的韵律反应也存在差异，但类别较多，本文将不再列出。

从表 2 可以看出，音节数/特征数之比约为 14，类规模比较合理。男女声两部分语料的样本数和类规模相近，表明语料的平衡性。分析还表明两部分语料特征分布基本相同，研究结果具有普遍意义。

表 2 音节/特征数统计结果

语料		全部音节	阴平	阳平	上声	去声	轻声
03MR00	音节数	85 679	18 040	21 006	13 306	26 819	6 508
	特征数	6 224	1 297	1 564	919	2 007	437
03FR00	音节数	98 681	20 767	23 839	15 486	31 284	7 305
	特征数	6 985	1 463	1 685	1 021	2 301	415

4.2 特征出现率

本文为排除音段不同对分类的影响，采用每个音节单独聚类的方法，集中分析语境特征对韵律的影响。首先统计了两部分语料音节的特征出现率(如图 1)。

特征出现率反映特征用来分类的频度。图 1 表明男/女声语料的特征出现率分布基本一致，而不同特征出现率的差别较大。声调音联特征的出现率最高，且后音节声调出现率比前音节的略高。位置信息中，韵律词和韵律短语内位置的出现率也很高，但句内位置的出现率很小。音段音联的两种分类方法出现率基本相当。与声调音联不同，音段音联的两种分类方式都表现出前音节特征的出现率大于后音节的特点。

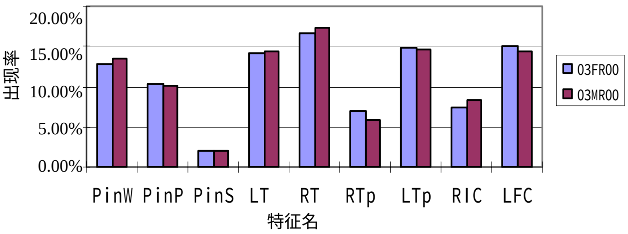

图 1 男/女声语料音节的特征出现率对比

分声调的统计结果还表明：正如前面提到过的，不同声调音节样本的聚类结果存在差异。最为明显的是轻声音节的结果与其它 4 声差别较大，后音节声调的出现率非常高，而韵律词内位置的出现率则非常小，后者与轻声音节多位于韵律词尾可能有一定关系。其它 4 声音节的特征分布率则比较相似，部分特征略有区别。其中，区别比较显著的是上声，韵律短语内位置和后音节声调的出现率更高了；阴平和阳平的分布规律基本一致，左音节韵母类型和右音节首音类型两个特征的出现率区别稍大；与上声相反，去声的韵律短语内位置的出现率偏低。阴阳上去 4 声在韵律短语内位置和前音节韵母类型两个特征的出现率上表现出的差别较其它特征略大。

4.3 特征平均层级统计

聚类结果中决策树的层数由 2 层到 12 层不等，不同规模的决策树里，以及同一决策树不同层级上的特征，引起的韵律区分程度都是不同的。03FR00 的聚类结果中特征出现平均层级的统计结果如图 2。

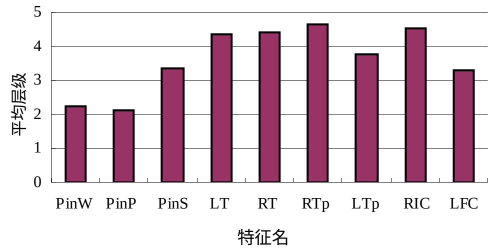

图 2 03FR00 语境特征的平均层级

与分布率的统计结果不同，韵律词内位置和韵律短语内

位置的平均层级明显低于其它特征，反映出位置对语音韵律表现的高区分度。同样地，分声调的统计结果表明轻声分布较为特殊，而其它 4 声则相对接近，上声差别略大。

5 影响权重分析

在第 4 节中，首先统计了聚类结果中各特征的相关数据。特征出现的频度和所处的层级从影响有无和区分大小两个不同侧面反映了语境特征对韵律参数的影响程度，但就其中任何一个参数得出的重要性结论都是不够全面的。对比图 1 和图 2 的结果可以看出，由特征出现频度和平均出现层级得到的特征影响程度排序结果有相当的不同。

为更好地体现特征在决策中的重要性，即特征对韵律影响的相对程度，需要综合考虑特征出现频度及其在决策树中的位置等因素。本文找到了满足如下条件和假设的重要性权重函数作为综合衡量的指标。

5.1 重要性权重函数

首先，在每棵决策树I（1≤i≤S，S为音节总数）中，特征x（x∈X，X为语境特征全集）的重要性权重函数$If_i(x)$应满足以下条件：

(1) $IF_i(x)$的值随属性为x的分枝节点数的增加而增大。

(2) $IF_i(x)$的值随属性为x的分枝节点所在层级的降低而增大，且越低的层对函数值影响越大。

(3)$IF_i(x)$的值与决策树的规模有关，条件 1 和 2 都相同的特征x，在规模大的树中$IF_i(x)$的值较小。

(4)对于全部特征 $x\in X$，有 $\sum_x IF_i(x)=1$。

假设：

(1)各类别(即决策树叶子节点)的重要性相等。将产生每一个叶子节点所经过的所有特征的重要性和设为 1。

(2)一个叶子节点产生过程中出现的所有分枝节点对产生该叶子的贡献相同。产生该叶子节点的路径上每一个分枝节点对于该叶子节点的重要性均为该叶子节点的层级的倒数(根节点的层级为 0)。

将该分枝对所有由它直接或间接分裂而成的叶子节点的重要性相加，再把该特征的全部分枝相加，可以得到该树的满足条件 1、2、3 的特征重要性函数。为满足归一化条件 4，将计算的结果再除以该树的叶子节点总数，即得到一个满足以上 4 个条件的简单的重要性权重函数(证明从略)。

$$IF_i(x)=\frac{\sum_{branch_j\in x}\sum_{leaf_k\subset branch_j}\frac{1}{level(k)}}{count(leaf\in i)}$$

其中 $branch_j\in x$ 表示分枝节点 j 属性为 x，$leaf_k\subset branch_j$ 表示叶子节点 k 由分枝节点 j 直接或间接分裂而成。$count(leaf\in i)$ 表示树 i 的叶子节点数。

同样，特征x在所有决策树即全部音节分类中的全局重要性权重函数IF(x)，仍应满足同$IF_i(x)$一样的单调性条件 1、2、3 和归一化条件 4 以及假设 1 和 2，从而同时保证了与决策树规模的对应性。因而有

$$IF(x)=\frac{1}{\sum_i count(leaf\in i)}\sum_i (IF_i(x)\times count(leaf\in i))$$

其中 $count(leaf\in i)$ 表示树 i 中的叶子节点数，它代表了树的规模。

5.2 统计结果分析

按照此函数得到 03FR00 聚类结果的分声调重要性权重统计结果如表 3。

表 3 03FR00 声调权重的统计结果(%)

特征	全部音节	阴平	阳平	上声	去声	轻声
PinW	21.67	27.27	25.10	16.39	21.04	2.23
PinP	17.84	16.58	19.28	25.38	15.15	10.86
PinS	1.73	1.79	1.60	1.84	1.86	1.09
LT	9.73	7.47	8.93	9.73	9.71	22.69
RT	11.42	10.40	9.40	15.05	11.03	16.52
RTp	4.34	4.44	3.98	3.40	4.63	6.60
LTp	12.45	11.80	11.49	9.48	14.85	13.51
RIC	4.92	4.82	4.98	5.58	4.28	7.15
LFC	15.90	15.42	15.25	13.14	17.46	19.36

同样，轻声的差别较大；其它 4 声在分布大致接近的同时，差异也比前面的统计略有扩大，尤其是韵律词和韵律短语内位置的权重分布，另外后音节声调对上声音节的影响也更加显著。总体来看，韵律词和韵律短语内位置表现出了明显高于其它特征的权重，前音节音段特征与后音节音段特征权重差别扩大，而按尾音进行的前音节分类的权重值比按韵母进行的前音节分类的更高。

与前面以出现率和平均层级为依据的重要性排序对比如表 4。

表 4 03FR00 语境特征重要性排序

标准 / 排序	出现率		平均层级		IF(x)(%)	
	特征名	统计值	特征名	统计值	特征名	统计值
1	RT	16.56	PinP	2.111 1	PinW	21.67
2	LFC	14.97	PinW	2.255	PinP	17.84
3	LTp	14.79	LFC	3.292 4	LFC	15.90
4	LT	14.12	PinS	3.351 1	LTp	12.45
5	PinW	12.78	LTp	3.765 7	RT	11.42
6	PinP	10.35	LT	4.376 7	LT	9.73
7	RIC	7.47	RT	4.437 8	RIC	4.92
8	RTp	6.87	RIC	4.667 4	RTp	4.34
9	PinS	2.09	RTp	5.1404	PinS	1.73

从该结果可以看出，重要性权重函数综合了特征在聚类决策过程中的出现率和位置信息，更好地反映出语境特征对语音韵律表现的影响方式和相对程度。其中，音节在韵律词内的位置被统计出是对韵律表现影响最大的语境特征，韵律短语内位置则位于第 2——事实上，某特定音节在韵律结构内的位置很大程度上也影响着前后音联对该音节韵律表现的影响程度和方式。其次是前音节的音段特征，且首尾音分类的权重要高于声韵母分类。然后是声调音联，与音段音联相反，后音节声调的影响略大。影响较小的是后音节的音段特征。而句内位置在聚类决策过程中的相对权重则非常小。

虽然前后音段和语调音联特征权重的结果是否受到了与位置特征耦合作用的影响还有待商榷，但以上的结论还是基本符合语音学理论和 TTS 系统选音的实践经验的，具有一定的合理性。与听辨实验指导的选音代价函数权重优化的结果相对比，特征的权重排序也有一定一致性。

6 总结与讨论

本文对 TH-CoSS 的大量音节按照韵律距离进行了 CART 树聚类。统计了聚类结果中语境特征的出现率、平均层级，设计了一个综合衡量聚类特征影响程度的权重函数，以此计

（下转第 11 页）