第 28 卷第 1 期　　声 学 学 报　　Vol. 28, No. 1
2003 年 1 月　　ACTA ACUSTICA　　Jan., 2003

基于数据挖掘算法的汉语合成韵律参数预测方法*

王 玮　蔡莲红
(清华大学计算机科学与技术系　北京　100084)

2001 年 6 月 13 日收到
2001 年 9 月 13 日定稿

摘要　韵律模块是语音合成系统中的重要组成部分，韵律特征参数的描述正确与否直接影响合成系统的输出，针对目前语音合成系统中缺乏对前后音节的韵律参数之间关系的有效描述，提出一种新的韵律参数预测方法——数据挖掘技术来发现音节韵律参数之间的相互关系，通过其中的关联规则模型对这些关系进行描述，并基于关联发现算法获得汉语韵律参数中基频参数和时长参数的变化规则，研究表明这些规则可以较好地为多样本拼接合成系统的选音提供帮助和指导。
PACS 数：　43.45

Research on predicting prosodic parameters for Chinese synthesis by data mining approach

WANG Wei　CAI Lianhong
(*Department of Computer Science & Technology, Tsinghua University*　Beijing　100084)

Received Jun. 13, 2001
Revised Sept. 13, 2001

Abstract　Prosodic control is an important part of speech synthesis system. Prosodic parameters choice right or wrong influences the quality of synthetic speech directly. At present, text to speech system has less effective describe to reflect data relationships in the corpus. In this paper, we present a new research approach-data mining technology to discover those relationships by association rules modeling. We develop a new algorithm for generating association rules of prosodic parameters including pitch parameters and duration parameters from corpus. The output rules improve the correctness of syllable choice in text to speech system.

引言

随着语音学和计算机技术的发展，文语转换系统 (Text to Speech System) 的研究已经获得了重大进展，并且成功地应用于许多不同的场合，但是文语转换系统合成语音带有浓重的机器味，与人类自然流畅的发音相比还存在一定的差距。其中重要的原因是受制于人们对韵律特征和规律的认识。以往，语音学家利用已有的语音学知识，总结韵律规则；工程师们通过统计的方法，分析归纳韵律规律。但是语音现象复杂多变，人们感知异常灵敏，规则不能穷举，规律只能建在已有知识的基础上。因此我们需要开拓研究语音韵律的新途径。如通过对语音数据的挖掘，突破已有的限制，发现语音韵律的新知识。

数据挖掘技术是发现新颖的、有效的和完全能够被人们理解的数据模式的一种方法，它结合统计和计算技术从大量的数据集中获取有用的模式，进而产生指导性的规则集合，这些规则能够对数据库中数据属性、对象集进行有效描述，为被挖掘处理的数据集产生摘要，提供给决策支持系统。其中关联规则是目前研究得最多的一种数据挖掘问题，关联规则模型能够很好地反映数据集中数据项之间的相互关系，因此发掘汉语中的韵律参数的关联规则发现具有重要的意义。

同以前文语转换系统韵律参数的研究工作相

* 国家自然科学基金资助项目 (69875008)

比，本文工作的优越性主要体现在以下 3 个方面：(1) 提出了一种新的研究方法，采用人工智能领域中数据挖掘技术基于数据驱动方式进行汉语的韵律参数研究。(2) 和先前统计方法不同的是统计模型是对现有数据的分析，不能够产生新的知识，数据挖掘技术则通过对现有数据的分析产生新的指导性知识。(3) 基于关联规则的模型非常适合对大语料进行计算，给出的规则完全是基于数据驱动的方式，克服了传统的语言学家给出的定性描述的弱点，完成了数据从定性到定量的转变。

在文语转换系统韵律模块中，如何选取合适的音节是一个难点问题，在实际的语料库中包含很多某一个音节的音，由于音节具有相应的语境信息，因此同一个音节的韵律参数值在不同语句中表现的值各不相同，什么样的语境下应该选取什么样的音节，就需要根据不同语境下的音节韵律参数进行选取，然而由于缺乏前、后音节之间相互关系的有效描述，因此在确定当前音节的前提下很难对后续的选音工作提供指导，因此对于后续音节的预测就显得十分重要。本文正是基于此种考虑，通过对实际语料库的韵律参数进行数据挖掘产生指导性规则，为系统的合理选音提供依据。

1 韵律参数

语音作为人的发声器官发出来的一种声波具有声音的物理特性，每一种音都具有一定的音色、音调、音强和音长。音色也叫音质，是一种声音区别于其它声音的基本特征。音调是指声音的高低，在汉语语音学中称为音高。声音的强弱叫做音强，它是由声波的振动幅度决定的。声音的长短叫做音长，它体现了发音持续时间的长短。

语音的韵律参数也称为超音段参数，一般是指音节的音高、时长和幅度等参数，通常以声音的基频代表音高，由于汉语是一种声调语言，因此其音高值是声调中最重要的特性。声调的音高值是声带基本振动频率的表现，不同声调的基频值各不相同，从敏感的声学仪器上得出的基频值也是千变万化的，同一个人发同一音节的音，在不同的上下文中，其基频也不一样，如音节“身”，在“身尸”中为 170 Hz，“身躯”中为 150 Hz，“起身”中为 160 Hz，“动身”中为 155 Hz。从以上的例子可以看出，客观存在的各个字音的音高值是千变万化的。不过，在语言听辨中人耳是不可能这样细致地识别基频的，在语言学上，表述一个声调的音高值时，通常采用相对的物理量进行表示，即从一个人或从几个人的声调频率中求出高低曲直的相对的平均关系值来表示。声调的音高值不是一个单一的频率值。由于声调是一个持续性的波段，而且在声调持续的过程中，从起点至终点往往又有频率的变化，通过对基频数据的聚类实验表明，直接对基频矢量进行聚类和将矢量减去均值后进行聚类，所产生的聚类中心一致，只出现少量的坐标平移。从而说明了采用基频中值作为韵律特征参数描述的合理性。

声调是一个持续的音段，每个声调都有一定的时长。在语谱图上，基频曲线随浊音的开始而开始，随浊音终止而终止。语音学家注意到，基频曲线在起始和末尾，出现某些弯头和弯尾的音段，在人们声调感知过程中，并没有成为语言音高信息，那么中部的调型段就成为声调的有效时长，对人们的声调感知所起的作用也最大，因此，基频中值参数和时长参数是描述音节音高特性的最重要属性。

基于以上考虑，本文选取基频中值参数作为处理对象，通过相同音节在大量语句中的音联现象进行数据挖掘，通过关联规则的形式给出前后音节之间的相互关系。

2 关联规则的算法描述

关联规则的挖掘问题的形式化描述如下：假设 $I = \{i_1, i_2, \cdots\cdots, i_m\}$ 是一组 m 个不同项目的集合。一个交易数据集 D，其中每一个交易 T 是 I 中的一组项目的集合，即 $T \subseteq I$。每一个交易都与一个唯一的标识符 TID 相联，其中 $X \subset T, Y \subset T$，且 $X \cap Y = \phi$，如果交易集 D 中有 $c\%$ 的交易既在包含 X 的条件下包含 Y，则称规则 $X \Rightarrow Y$ 的信任度是 c，如果交易集 D 中包含 $X \cup Y$ 的交易占整个的 $s\%$，则称规则 $X \Rightarrow Y$ 的支持度是 s。对于一组交易集 D 来说，一条关联规则的挖掘问题就是发现那些支持度和信任度大于用户指定的最小支持度和信任度的规则，最小支持度反映的是一条规则覆盖的最小数据项目数，最小信任度反映的则是一条规则的预测强度。

关联规则的挖掘问题可以分解为以下两个子问题：(1) 产生满足具有用户指定最小支持度的所有频繁项目集。(2) 根据频繁项目集产生满足最小信任度的关联规则。

在已经提出的许多关联算法中，Agrawal 等人在 1994 年提出的 Apriori 算法是其中核心的算法。Apriori 算法是一种宽度优先算法，通过对数据库 D 的多趟 (pass) 扫描来发现所有的频繁项目集，在每一

趟 k 中只考虑具有同一长度的 k(即项目集中所含有项目的个数) 的所有项目集, 在第 1 趟扫描中, Apriori 算法计算 I 中所有单个项目的支持度, 生成所有长度为 1 的频繁项目集。在后续的每一趟 k 中, 首先以前一趟中所发现的所有频繁项目集为基础, 生成所有新的候选项目集 (candidate itemsets), 即潜在的频繁项目集, 然后扫描数据库 D, 计算这些候选项目集的支持度, 最后确定候选项目集中哪些真正成为频繁项目集, 重复上述过程, 直到不再产生新的频繁项目集, 但是 Apriori 算法是一种处理无序数据集的算法, 然而语音数据是具有一定的时间顺序的, 即语句中的各个音节间先后顺序对于数据挖掘的结果是有直接影响的, 因此必须对该算法进行改进, 基于这种思想, 本文给出一种适应语音数据的关联算法 SPApriori 算法。

SPApriori 算法过程描述如下:

```
(1)  C1 = count(c, D)              // 统计各个韵律参数数据在整个数据库中出现的频率
(2)  L1 = {c ∈ C1|c.count ≥ minsupport};      // 产生大于用户设定最小支持度的韵律参数数据
(3)  for (k = 2; Lk−1 = φ; k++) do            // 产生两字词及多字词的频繁项目集
(4)      for (k = 0; k < C1.size; k++) do // 对于产生的频繁项目集进行排序操作
(5)          f = C1[k];
(6)          if f.count ≥ minsupport(f) then
(7)              insert {f, f} into C2
(8)              for (j = k + 1; j < C1.size; j++) do
(9)                  g = C1[j];
(10)                 if g.count ≥ minsupport(f) then
(11)                     insert {f, g} {g, f}into C2
(12)                 end
(13)             end
(14)     end
(15)   end
(16)   for all transactions ∈ D
(17)       Ct = count_support(Ck, t);     // 包含在事务 t 中的候选项目集
(18)       for all transaction c ∈ Ct
(19)           c.count++;
(20)       end
(21)   Lk = {c ∈ Ck|c.count ≥ minsupport};
(22)   end
(23) end
(24) resultset = resultset ∪ Lk            // 产生所需要的规则
```

事实上, 挖掘关联规则的整个执行过程中第一个子问题是核心问题, 当找到所有的最大项目集后, 相应的关联规则就很容易生成。

寻找最大项目集的基本思想是: 算法需要对数据集进行多步处理, 第一步是统计所有含有一个元素项目集出现的频率, 并找出那些不小于最小支持度的项目集, 即一维最大项目集。从第二步开始循环处理直到再没有最大项目集生成, 循环的过程是: 第 k 步中, 根据第 $k-1$ 步生成的 $(k-1)$ 维最大项目集产生 K 维候选项目集, 然后对数据库进行搜索, 得到候选项目集的项集支持度, 与最小支持度比较, 从而确定 K 维最大项目集, 这里, k-itemset 是 K 维项目集, L_k 表示具有最小支持度的最大 k-itemset, C_k 表示候选的 k-itemset(潜在的最大项目集) 。

在 SPAprior 算法中, 由于数据元素众多, 采用穷举的方法寻找符合要求的关联规则, 无论空间存储和时间消费上都是不可行的, 同时语音数据是一个时间序列数据, 前后音节的不同位置对于音联有重大影响, 因此必须采用序列方式记录数据形式, 这里考虑的是相邻音节的关联关系, 因此将长度定为 2, 分成两步: (1) 搜索当前的最大序列, (2) 用最大序列产生候选序列, 所有的候选序列构成一个集合, 经过对数据库的遍历去除非最大序列, 留下最大序列, 如此反复可以得到所需要的关联规则。

3 实验研究

3.1 实验数据

目标测试集是基于本实验室中实际语料库的2259个句子进行的测试，该实验采用的数据库是一个女声数据库，采样率为16 kHz、量化精度为16位。包含近两万个音节。覆盖汉语的417种有调音节以及多种声学特征的搭配关系。该数据中的所有音节利用语音分析软件Speech进行了音节边界和基频的标注。此外，Speech软件可以方便地显示语音数据的波形和语谱图、测听选中的任何语句，因此可以对以上分析结果进行人工修正，基本杜绝了音节切分和标注的错误。

部分测试数据集的格式如表1所示。

表1 ‘部分实际语料数据库

语句序号1	一	一	奉	陪	吃	喝				
基频中值	291	300	308	161	249	264				
语句序号2	1	9	9	2	年	体	育	建	筑	奖
基频中值	334	263	178	249	172	171	232	329	175	108

经过合成系统分词后的结果语料是：一一|奉陪|吃喝，1992年|体育|建筑奖，这里只举例说明。上述表格是实际语料库中的数据格式，我们将其转换成数据挖掘方法的需求对象数据形式。

3.2 基频数据量化研究

语料库中基频中值的变化范围比较大，这主要是由于录音人说话时的语速和语调造成的，因此需要对数据段进行区间离散化处理。这里我们采用两种方法进行离散处理，均值离散和正态离散，均匀离散为50个区间的结果如表2所示。

表中给出了基频中值数据取值在不同区间段内对应的数据出现次数的统计结果，从表2的数据统计次数不难看出，整个数据库中的基频中值数据分布是有一定规律的，处于区间段9-42内的数据出现次数的值都比较多，而在区间1-8和43-48内的出现次数就比较少，因此近乎满足正态分布的趋势规律。由此可见采用均匀量化的方法不能很好地反映数据项的实际数据分布规律，因此考虑采用正态分布的量化措施。将上述数据集采用正态分布量化的结果如表3所示。

从表3中可以看出在每个区间段数据出现的次数差不多，这主要是由于在数据稀疏的地方我们取的区间范围较大而对于数据相对集中的部分分割的区间较小，这样可以更好地反映出实际数据的分布规律。下面我们就用正态分布的数据进行规则分析。

表2 均匀量化及均匀量化的频繁项目集1对应表

基频数据	区间段	出现次数
84-90	0	/
91-97	1	9
98-104	2	12
105-111	3	25
112-117	4	33
118-124	5	81
125-131	6	115
132-138	7	146
139-145	8	205
146-151	9	220
152-158	10	288
...	...	...
322-328	35	501
329-334	36	458
335-341	37	463
342-348	38	477
349-355	39	386
356-361	40	289
362-368	41	263
369-375	42	200
376-382	43	137
383-389	44	91
390-395	45	43
396-402	46	52
403-409	47	27
410-416	48	9
417-422	49	/
423	50	/

注：最小支持度设置为0.5%。

表3 正态量化及正态量化的频繁项目集1对应表

基频数据	量化段	出现次数
84-135	0	354
136-148	1	362
149-157	2	371
158-164	3	371
165-170	4	376
171-175	5	298
176-181	6	385
182-187	7	418
188-192	8	361

(表 3 续)

193-197	9	367
198-202	10	372
...	...	...
318-322	40	384
323-327	41	338
328-332	42	408
333-337	43	367
338-342	44	321
343-348	45	410
349-355	46	386
356-363	47	371
364-374	48	351
375-422	49	388
423	50	354

注：最小支持度设置仍为 0.5%

4　规则的发现

4.1　规则的产生

首先设置产生规则的最小支持度，最小支持度的值是根据用户需求进行设定的。然后计算不同区间段的数据出现频率，产生出大于或等于用户指定最小支持度的韵律参数数据，作为候选项目集。这里只列出了少量例子进行方法的说明。如区间段 48 和 12，基于SPAprioir算法，产生如 $48 \rightarrow 12, 48 \rightarrow 20, 48 \rightarrow 21$ 等出现频率满足最小支持度的设定的规则，产生的规则数目取决于最小支持度的阈值设置。其中得到规则 $48 \rightarrow 12$ 的含义是说如果当前音节基频数据值在量化区间 48 范围内时，则后续的音节可能值为量化区间 12 范围内的值。规则的支持度为 15 。这里的支持度表明满足这种规则的数据频率。

表 4 列出基于本试验数据库发现的基频量化数据段48和49的相邻后音节基频量化段的有关规则。

从表中看出，基频量化数据段 48 和 49 的相邻后音节基频中值量化段的有关规则。其中 $48 \rightarrow 12$ 和 $49 \rightarrow 27$ 是支持度最大的两条规则。

这些规则意义在于，如果已知当前音节的韵律特征如基频中值，可以有效地预测后音节的韵律特征，通过这样的规则描述可以在相同后续音节进行合理的选音，后续音节由于受到不同音联关系或自身是重音等声学特征的影响而表现出不同的韵律特征候选音节中进行筛选，具体体现为不同韵律参数数据值，为合成选音提供指导。在当前相同的音节前提下，根据后续音节的不同取值范围进行规则查找，示例规则给出了音高特征候选音节的音联的可能性，其余的规则信息可以类似进行分析，这样的规则描述给出了语句中不同语音数据项之间的音联关系，为合成系统的正确选音提供指导。

表4　基频量化数据段48和49的相邻后音节基频量化段的有关规则

当前音节基频中值	相邻后音节的基频中值	规则的支持度
48	12	15
48	20	10
48	21	12
49	6	10
49	18	10
49	13	10
49	28	10
49	20	11
49	30	12
49	27	16
49	29	10
49	24	11
49	32	10
49	12	11

4.2　规则的分析

当前音节量化段为 48 时，对应基频中值正态分布的数据段为 364-374，它在数据库中出现次数是 351，大于我们设置的最小支持度的阈值，即数据库中的总的音节数乘上 0.5%(最小支持度) 后得到的值。其中在量化段为 48 范围内的 “物” 这个音节对应的基频中值为 364，它在数据库中出现次数是 33 。后续音节量化段出现情况如表 5 。

表 5　后续音节的出现频率

后续音节量化段	基频中值数据	出现次数	“质” 出现的次数
12	207-211	388	50
20	239-242	384	31
21	243-246	382	36

说明如果当前音节 “物” 的后续音节为 “质” 的在不同的数据段出现的可能性分别是 50 、 31 和 36 三种情况。选取其中出现次数最大的规则并返回数据库进行验证，如果在数据库中确实存在这种情况，则说明这条规则给出合适的后选音节的最佳可能性，其余的规则信息可以类似进行分析，这样的规则描述描述了语句中不同语音数据项之间的音联关系。

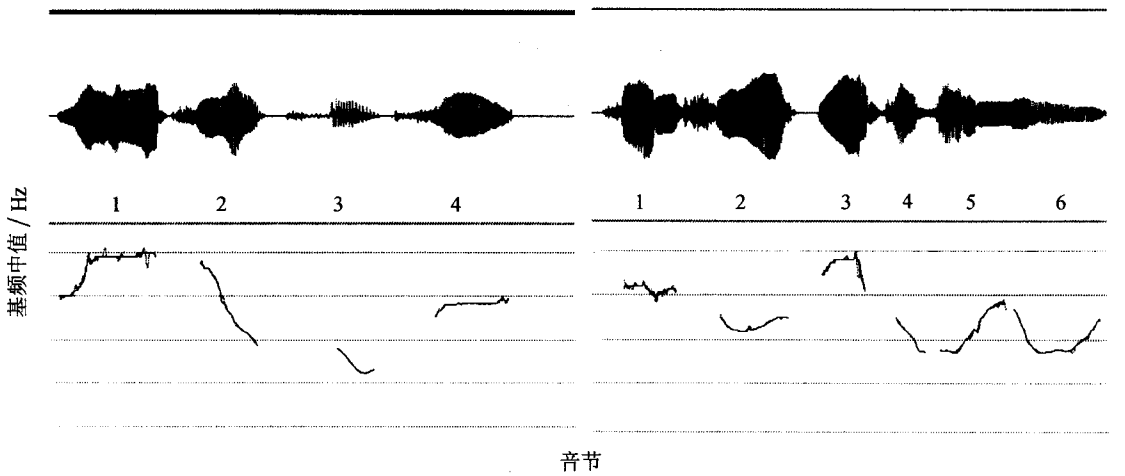

图 1 “优质苦咖” 和 “坚持物质文明” 的波形和基频中值曲线

4.3 规则的实施示例

下面以本实验室合成系统中的实例进行说明。如短语 “坚持物质文明” 和 “优质苦咖” 经过文本分词处理得到如下形式: | 优质 | 苦咖 | 和 | 坚持 | 物质 | 文明 |, 将分词后的结果 “物质” 和 “优质” 的基频中值作为关联算法的输入, 采用正态分布离散基频中值数据段, 有关的规则如表 4 所示。 “物” 的离散基频中值数据段含有 48 这个段, 其中 $48 \rightarrow 12$ 是支持度最大的一条规则。对照语音数据库, 与 12 对应的数据中包括 “质” 的音节。说明对于前音节是 “物” 的情况, 后音节选择的 “质” 应该是对应于数据段 12 内的 “质” 音, 而非数据段 32 内的 “质” 音。表中 $49 \rightarrow 27$ 是支持度最大的另一条规则。 “优” 的离散基频中值数据段含有 49 这个段, 但是 32 离散基频中值数据段中并没有 “质” 的数据。这是我们应降低支持度, 直到找到含有 “质” 的所对应的规则。以此, 找到 $49 \rightarrow 32$ 的规则。由于 “质” 这个音节具有不同的基频中值段, 这主要是由于不同的语境下 “质” 体现出不同的韵律特征。

通过挖掘, 得到相邻音节基频中值的关联规则。基于上述规则可以有效地系统选音提供指导信息, 提高选取音节的准确度。

图 1 显示原始录音人发音的两个短语的波形和基频中值曲线。 “物质” 在语料中对应值为 364 和 207, 量化后的基频中值数据段是 48 和 12 。 “优质” 在语料中对应值为 375 和 288, 量化后的基频中值数据段是 49 和 32 。图 1 从直观上说明对于前音节是 “物” 的情况, 后音节选择的 “质” 应该是对应于数据段 12 内的 “质” 音, 而非数据段 32 内的 “质” 音。

最后需要指出表 4 的实验结果由于实验语料的限制, 可能不能够完全覆盖某种数据在所有数据库中出现的各种情况, 然而能够在特定的数据库中得出相应的结论仍然说明了这种新方法的意义。我们有理由相信只要语料库覆盖面足够的宽, 会得到更好的结论。

5 结束语

在合成系统中, 韵律模块中韵律参数之间的相互关系直接影响合成系统的输出质量。本文提出了将数据挖掘方法应用于汉语韵律特征的基频中值参数和时长参数数据进行知识发现, 并提出一种适合语音信号处理的关联规则的发现算法, 能够得到统计方法所无法获得的潜在知识, 这样得到的规则可以有效地对系统中选音提供指导, 提高选音的正确度。但是最小支持度的选取对挖掘结果有直接的影响, 因此如何合理的设置最小支持度或者寻找其它确定最小支持度的方法是本文下一步的研究方向。

参 考 文 献

1 郭锦桴. 汉语声调语调阐要与探索. 北京: 北京语言学院出版社, 1993

2 姚天任. 数字语音处理. 武汉: 华中理工大学出版社, 1992

3 Rakesh Agrawal, Ramakrishnan Srikant. Fast algorithm for mining association rules. In Proceedings of the 20th VLDB Conference, Santigo, Chile, 1994: 487—499

4 Ramakrishnan Srikant, Rakesh Agrawal. Mining generalized association rules. In Proceedings of the 21st VLDB Conference, Zurich, Switzerland, 1995: 250—259

5 Rakesh Agrawal, Tomasz Imielinski, Arun Swami. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD Conference, Washington DC, USA, 1993: 207—216