中 文 信 息 学 报
第 19 卷 第 1 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 19 No. 1

文章编号:1003 - 0077(2005)01 - 0084 - 07

多项式回归的汉语时长预测模型①

孙 璐,胡 郁,王仁华

(讯飞语音实验室 中国科学技术大学电子工程与信息科学系,安徽 合肥 230027)

摘要:时长信息是韵律的重要组成部分,对于语音合成的自然度和可懂度都有不可忽视的作用。时长预测是建立对时长有影响的韵律环境与自然语流中音段时长的对应关系。本文引入了统计学中 eta squared 的概念研究汉语中韵律环境因素对时长的影响,设计了残差算法定量分析属性之间的交互作用,由此建立了多项式回归的汉语时长预测模型。实验结果表明,使用 5～6 个韵律属性基本上就能够建立比较相关的对应关系,和使用同样韵律属性的 Wagon 回归树的效果相比有明显的优势。

关键词:计算机应用;中文信息处理;时长建模;多项式;交互作用

中图分类号:TP391.4 **文献标识码**:A

Polynomial Regression Model for Duration Prediction in Mandarin

SUN Lu, HU Yu, WANG Ren-hua

(Iflytek Speech Laboratory, Department of Electronic Engineering Information Science,
University of Science and Technology of China, Hefei, Anhui 230027, China)

Abstract: Duration information is an essential part of speech prosody, and plays a critical role in improving the naturalness and understandability of synthesized speech. Duration modeling is to establish a mapping relationship between the prosodic environment and the final duration engendered in natural speech. In this paper, we first study the effect of prosodic features on segmental duration by introducing a statistical concept —eta squared, then choose more forceful prosodic features and design an algorithm to quantify the interaction among them, and finally bring forward the method of determining the duration model using a polynomial equation and obtain the coefficients through non - linear regression. Our research work indicates that 5 or 6 prosodic features might by and large assist a close and accurate mapping between prosodic environment and perceived duration. Compared to Wagon tree method, this method has undeniable merits.

Key words: computer application; Chinese information processing; duration modeling; polynomial; interaction

1 引言

时长信息是韵律的重要组成部分,时长的准确预测对于语音合成(TTS)的自然度和可懂度有着不可忽视的作用。

时长预测有两种基本方法,一种是规则驱动的方法(rule-driven),一种是数据驱动的方法(data-driven)。前者是用语法规则对时长赋值并稍作调整的方法,后者完全建立在数据处理的基础上,并以统计方法和数值计算为依托。随着计算机技术的发展,数据存储和数据处理的能力大大提高,数据驱动方法的优势日益明显。

① **收稿日期**:2004 - 04 - 20
作者简介:孙璐(1982 —),女,硕士生,主要研究领域为汉语韵律模型.

目前,数据驱动时长预测主要以对时长建模为主。时长模型是体现对时长有影响的韵律环境即影响因子和最终在自然话语中表现出的音段时长之间的对应关系。捕捉这种对应关系的方法有很多种,用韵律属性对时长进行直接训练的有神经网络的方法[1,2]和决策树[3]的方法,进行韵律属性因子分析后构造多项式,然后对多项式系数进行训练的有非线性回归[4]、MARS 算法[5]和 EM 算法[6],也可以根据因子分析的结果通过构造贝叶斯网络实现。

本文中,我们采用分析影响因子,多项式建模并用非线性回归进行模型(polynomial regression,后面称 PR 模型)训练。本文在 10000 句实现数据下完成了时长预测器,并对预测性能进行评测。结果显示,和其他方法相比,该方法有不可忽视的优势。

2 影响因子分析

研究影响因子对时长作用的基本方法是统计分析,我们使用统计分析软件 SPSS 完成该过程。首先,对于后面诸多假设检验,要求时长数据符合正态分布,我们首先分析时长的分布情况;然后,由于韵律环境属性为数较多,中间又有冗余的现象,于是引入统计学中 eta squared 的概念进行属性筛选。最后设计了残差计算的算法进行因子之间交互作用的分析,即两个因素对时长的联合影响。

2.1 时长分布

1. 声母的时长分布

我们用偏态系数和峰度描述各个声母时长分布情况,其正态分布特性很容易观察到。偏态系数是分布偏斜方向和程度的测度,峰度描述分布的集中趋势。标准正态分布的偏态系数和峰度皆为 0,峰度大于 0,则分布高出标准正态分布曲线,反之则较为平坦。

表 1 声母时长分布参数表

声母	b	p	m	F	d	t	n
偏态系数	1.68	-0.12	0.06	0.45	3.20	-0.19	0.67
峰度	8.61	0.90	1.14	0.55	30.70	0.57	6.82
声母	l	g	k	h	j	q	x
偏态系数	0.35	1.06	-0.37	-0.11	0.20	0.20	0.57
峰度	0.64	2.45	1.14	1.56	0.50	0.36	2.88
声母	z	c	s	zh	ch	sh	r
偏态系数	0.39	-0.05	0.40	1.18	0.18	0.37	2.64
峰度	-0.33	0.01	-0.42	1.00	1.37	0.04	37.69

可以看出,如果对各个声母进行统计,由于语音数据的随机性,其时长基本符合正态分布,只是有的分布峰高出标准正态,有的就稍微平坦。

2. 韵母时长分布

我们对 20 多万韵母进行同样的统计,其均值为 149.4ms,标准差 44.92ms。对于每个韵母,从其偏态系数和峰度来看,其时长也符合正态分布。

2.2 韵律属性的筛选

根据我们已有的先验知识,影响时长的因素应该有声调,音节在句中的位置,前后边界,前后声韵母等等。我们可以通过假设检验的方法计算检验统计量 F 值来判断某种环境因素对时长是否有影响,但是这种方法不能确定各个环境因素对时长影响作用的大小,为此,我们引用了统计学中 eta 平方(eta squared)的概念。eta 平方在统计学中用于检测因变量和自变量之间联系的强度,是因变量中不同组中差异所解释的方差比,其计算公式如下[7]:

$$\eta^2 = \frac{SS_{total} - SS_{within}}{SS_{total}} = \frac{SS_{between}}{SS_{total}} \quad (1)$$

令 x_{ij} 为某种属性的第 j 种可能取值中的第 i 个观测样本值,那么

$$SS_{total} = \sum_j \sum_i (x_{ij} - \bar{\bar{x}})^2 \quad (2)$$

为总体离差平方和,

$$SS_{within} = \sum_j [\sum_i (x_{ij} - \bar{x}_j)^2] \quad (3)$$

为水平内的离差平方和,

水平在统计学中指分类的一个类别,在本文中,水平指环境属性的取值,水平内的离差平方和反映组内观测样本的离散程度,

$$SS_{between} = \sum_j \sum_i (\bar{x}_j - \bar{\bar{x}})^2] \quad (4)$$

为水平间的离差平方和,反映不同属性取值分组间的差异。且在各组同为正态分布的条件下,有:

$$SS_{total} - SS_{within} = SS_{between} \quad (5)$$

因此,η^2 能反映属性与时长观测值之间的关联程度,η^2 越大,关联程度越强,反之越弱。所以本文中我们用属性的 η^2 值解释它对时长影响作用的大小。

1. 声母环境属性对时长影响作用的重要性分析

在我们使用的语料库中,句子用 5 个韵律层进行标注。L0 层为音节层,L1 层为音步层,L2 为韵律词层,L3 为主短语层,L4 为呼吸群层。在标注前后边界时,如果为多层的边界,则取最高层。

对于声母,我们从语料库中提取出来的环境属性有:声母标志,声母类型,韵母(所在音节的)标志,韵母类型,前韵母标志,调型,前调,后调,前边界,后边界,当前音节在各韵律层的位置,各韵律层长度等。我们使用统计分析工具 SPSS 中的 ANOVA 过程对 20 万个声母样本计算 eta squared,得到的结果如表 2。

表 2 声母属性 η^2 列表

环境属性	声母标志	声母类型	韵母标志	韵母类型	前韵母标志
Eta-squared	0.839	0.781	0.239	0.017	0.008
环境属性	前韵母类型	调型	前调	后调	前边界
Eta-squared	0.001	0.058	0.003	0.002	0.016
环境属性	后边界	在 L2 层位置	在 L3 层位置	L2 层长度	L3 长度
Eta-squared	0.019	0.006	0.001	0.003	0.001

很明显,对于声母的时长,除了声母自身对其起到决定性作用以外,其周围环境中的韵母标志,调型,前后边界也对当前声母的时长有较大的影响。为了防止环境属性取值过多而造成水平内样本数目过少,虽然声韵母标志的 η^2 比较大,也没有选择作为建模属性,以下对韵母的分析也同样。

2. 韵母环境属性对时长影响作用的重要性分析

对于从语料库中提取的近 20 万个韵母,我们得到的环境属性有:韵母标志,韵母类型,所在音节的声母标志,声母类型,后声母标志,后声母类型,其余属性和前面声母中的相同。表 3 给出了韵母环境属性的 eta squared 值。

可见,对当前韵母时长影响作用较大的环境属性有:声母类型,调型,后边界类型,音节在

句中位置和 L3 层长度。

表 3　韵母各属性 η^2 列表

环境属性	韵母标志	韵母类型	声母标志	声母类型	后声母标志
Eta-squared	0.345	0.203	0.240	0.214	0.047
环境属性	后声母类型	调型	前调	后调	前边界
Eta-squared	0.057	0.113	0.003	0.036	0.002
环境属性	后边界	在 L2 层位置	在 L3 层位置	L2 层长度	L3 长度
Eta-squared	0.107	0.064	0.102	0.021	0.019

通过对时长影响作用的重要性分析,我们可以确定在建立时长模型时考虑 eta squared 值较大的环境属性,这样样本矢量空间就大大降低了。

2.3　属性交互作用分析

属性的交互作用体现在多个属性对时长共同作用时[8],它的存在主要是属性对时长影响特点引起的,例如三调音节时长比较长,句末位置也会对时长有增长(lengthening)的作用,但是三调音节在句末的时候并没有双重增长(double - lengthened),反而通常是比较短的。分析属性交互作用有很多方法,我们设计了残差计算的算法来定量分析交互作用的大小。

残差计算是为了衡量属性及属性取值之间交互作用大小的,以计算属性 F_1 和 F_2 的残差为例。假设这两种属性组成的特征空间为 N,其余 $n-2$ 个属性组成的特征空间为 M,这样以当前研究的属性 F_1 和 F_2 组成的特征矢量标识每行,以其余属性组成的特征矢量标识每个列,于是产生了如下的表格:

表 4　残差计算说明

	$(f_3, \cdots f_n)^1$	$(f_3, \cdots f_n)^2$	$(f_3, \cdots f_n)^3$	……	$(f_3, \cdots f_n)^M$
$(f_1 f_2)^1$	$\overline{dur_1}^1, N_1^1$ $resid_1^1$	×	$\overline{dur_1}^3, N_1^3$ $resid_1^3$	……	$\overline{dur_1}^M, N_1^M$ $resid_1^M$
$(f_1 f_2)^2$	$\overline{dur_2}^1, N_2^1$ $resid_2^1$	$\overline{dur_1}^2, N_1^2$ $resid_2^2$	×	……	$\overline{dur_2}^M, N_2^M$ $resid_2^M$
……	……	……	……	……	……
$(f_1 f_2)^N$	$\overline{dur_3}^1, N_3^1$ $resid_3^1$	$\overline{dur_2}^2, N_2^2$ $resid_2^2$	$\overline{dur_2}^3, N_2^3$ $resid_2^3$	……	×
margin	$marg._{f_3, \cdots f_n}^1$	$marg._{f_3, \cdots f_n}^2$	$marg._{f_3, \cdots f_n}^3$	……	$marg._{f_3, \cdots f_n}^M$
Squared Resid	$SQResid^1, num^1$	$SQResid^2, num^2$	$SQResid^3, num^3$	……	$SQResid^M, num^M$

由于数据稀疏性,只有少量的格子有数据。这里为了解释需要,列出了部分格子,打叉的为空数据格。格子中 $\overline{dur_1}^1$ 代表属于该格子的样本的时长平均值,下标代表同一列内的非空格号,上标为列号,N_1^1 代表该格子里样本的数目,上下标含义与时长均值的相同,$marg._{f_3, \cdots f_n}^1$ 为在某环境下属性 F_1 和 F_2 的边际值,即该列内所有样本的时长平均值,$resid_1^1$ 称为残差值,由时长均值 $\overline{dur_1}^1$ 减去 $marg._{f_3, \cdots f_n}^1$ 得到,num^1 为该列内非空格子数目。对于每一列所有 F_1 和 F_2 组合下的残差值经汇总(求平方和)得到一个 $SQResid^1$,即某环境下的残差平方和,它的大小表示在该环境下属性 F_1 和 F_2 对时长的影响大小,如果该值很大,则在该环境下,属性 F_1 和 F_2 对时长的影响作用大,否则则小。这样就得到 M 个残差平方和,由下面公

式得出最后的残差值:

$$Resid = \sum_{i=1}^{M} \frac{SQResid^i}{num^i} \quad (6)$$

总的来说,我们所采用的残差值是以单个格子为单元的所有时长平均的一种空间距离算法,在某种程度上解决了样本数目不足以及特征空间不平衡等问题,但是由于语言的分布不均衡性空格子问题依然无法得到彻底解决。

2.4 残差计算结果及建模

由于计算过程是针对每个声韵母的,每种属性组合都对应一个残差值,根据其残差值的分布进行了聚类,目的是确定不同残差分布的声韵母的模型表达式,以简化模型表达式数目。

声母残差分布非常相似,都是配对韵母与调型,前后边界的交互作用,所以对于声母不予以分类,而是采用同一种模型表达式。为了考虑各个韵律属性对时长在非交互情况下的作用,模型内加入全加部分,即各个属性取值对时长作用的独立的叠加效果。那么,声母的多项式模型表达式为:全加 + 配对韵母 *调型 + 配对韵母 *前边界 + 配对韵母 *后边界。

对于韵母,我们同样根据残差分布进行聚类,得到三类。另外考虑到做非线性回归对样本数目的要求,将样本数不够充足的韵母分为一类,为第四类。在韵母中同样加入全加部分,另外为了考虑所有属性之间的交互作用,加入了全乘部分,即假设所有属性之间都有对时长的交互作用。

1. 第一类的韵母有:i, u, ai, ao, ian, ing, uan, a, e, ie, uo, ou, iao, iou, uei, an, en, ang, eng, in, iang, uen. 其残差分布大小依次是后边界类型与 L3 层位置,配对声母类型与调型,调型与 L3 层位置,配对声母类型与 L3 层位置,因此采用的模型为:全乘 + 全加 + 后边界 *音节在 L3 层的位置 + 配对声母类型 *调型 *音节在 L3 层的位置;

2. 第二类的韵母有:ii, iii, ei, o, iong, ua, uai, van, ia, v, ong, ve. 其残差分布较大的情况在于配对声母和其他影响属性之间,因此采用的模型为:全乘 + 全加 + 配对声母 *调型 + 配对声母类型 *后边界 *音节在 L3 层位置;

3. 第三类的韵母有:er. 对于该韵母只有后边界类型与 L3 层位置的交互作用相对较大,采用的模型为:全加 + 后边界 *音节在 L3 层位置;

4. 第四类的韵母有:uang, vn. 采用的模型为:全加 + 配对声母类型 *调型 + 配对声母 *后边界 *音节在 L3 层的位置(其实是简化的第二个模型)。

下面用非线性回归的方法对这 5 种模型进行训练。

3 模型回归训练及评测

3.1 声韵母时长预测模型回归训练

对上面确定的声韵母模型用非线性回归的方法进行训练,每个声韵母一个模型,模型的表达式参照上述聚类结果。训练结果如下:

所有声母时长预测值和观测值的集内相关系数为 0.952,RMSE 12.21ms;集外相关系数 0.952,RMSE 12.33ms。对每个声母,计算其集内和集外的相关系数及 RMSE,见图 1,2。

所有韵母时长预测值和观测值的集内相关系数为 0.844,RMSE 24.01ms;集外相关系数 0.826,RMSE 25.44ms。对每个韵母,也计算其集内和集外的相关系数及 RMSE,见图 3,4。

从图 1～4 中我们可以看出,在不影响训练结果的情况下,该方法的集内外测试结果相差很小,这主要是因为训练过程中使用的属性个数比较少的原因。另外,声韵母时长模型的训练

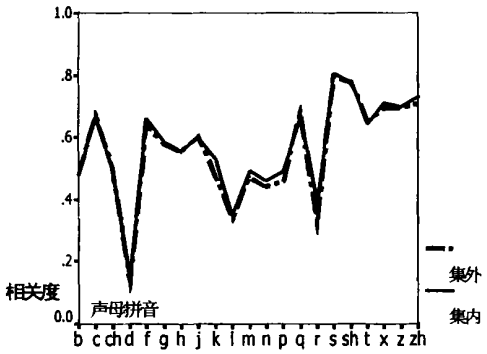

图 1　声母集内外相关系数对比

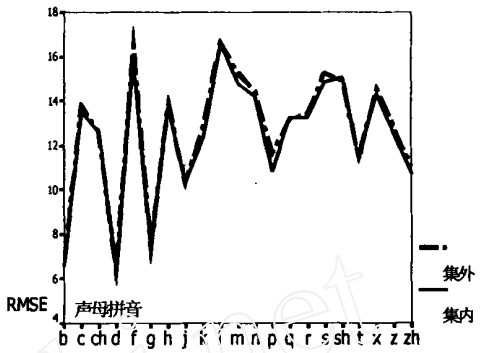

图 2　声母集内外 RMSE 对比

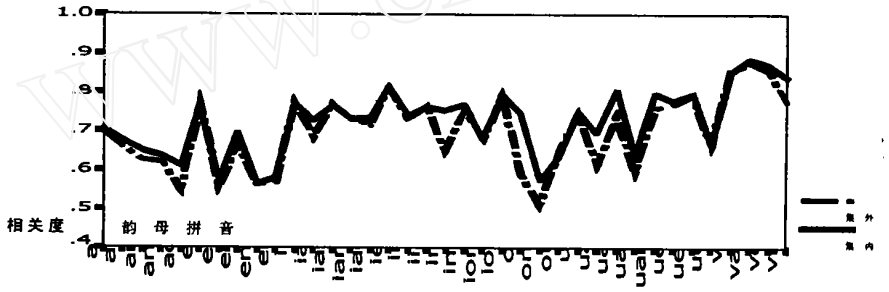

图 3　韵母集内外相关系数对比

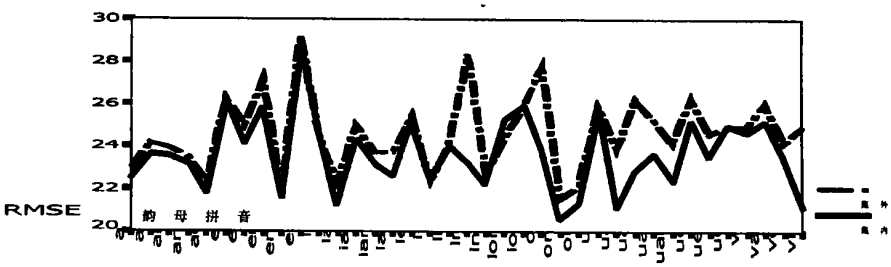

图 4　声母集内外相关系数对比

结果和其时长分布有很大的相关,如果其时长较长,而且其分布方差较小,预测的结果就好,反之则差。如声母中的 d, b, r 以及韵母中的 er, in, iou。除了时长分布之外,还有样本数目的关系,由于自然语言的样本不平衡性,有的声韵母的样本空间很小,训练的结果就不够理想。

3.2　模型评测

我们建立了两个时长预测模型:PR 模型,使用了相同属性的 Wagon 模型,选择了 1000 句,对每句计算预测的 RMSE 和相关度,如表 5 所示:

表 6　模型客观评测

模型	PR	Wagon
平均 RMSE	26.0ms	26.4ms
平均相关度	0.858	0.856

表 6　模型主观评测

模型	PR	Wagon	Wav
平均分	4.27	4.08	4.45

从客观评分(见表 6)上来看,PR 模型比 wagon 有所提高,但不是很明显。为了进行听感评

测,我们用上述三个预测器分别预测合成了 50 句话,然后请 6 个人以 5 分制打分,最终汇总的结果如表,其中 wav 是原始录音文件。

可以看出,从听感上比较,PR 模型比 Wagon 性能明显要好。主客观的评测差别比较大,根据我们的研究发现,这是因为 PR 模型的预测比较稳定,没有过大的误差,所以听感上会优于 wagon 模型。我们于是随机选择了一些句子观察两者对时长预测的结果,发现使用 Wagon 虽然对某些韵律环境下的声韵母能够预测得到很好的结果,但是由于在训练时,容易造成过适应(over fitting),对于另外一些韵律环境下的声韵母会有较大的误差,根据我们以前研究的结论[9],句子中的几个特别差的预测会给整个句子的听感带来比较大的损失。其实,无论是从时长建模还是从语音合成角度来说,PR 模型相对于 wagon 都是比较有优势的选择。

4 结论

本文在研究时长建模的同时,引入了统计学中 eta squared 的概念,设计了计算属性交互作用的算法,提出了通过聚类确定各个声韵母模型表达式的方法。交互作用在其他如基频和重音等方面也有体现,因此这样一套分析理论同样适用于别的韵律分析。

从上面分析过程来看,PR 模型实现比较简单,属性分析过程稍微复杂,预测结果和 wagon 相比较稍为稳定,模型应用时所需资源很少。

在分析了 PR 模型和 wagon 对时长的预测后,发现 PR 模型相对于 wagon 模型基本上都能更好地预测阅读状态的音段时长。我们在建立时长模型时候遇到的另外一个问题是,对于声母时长的预测做得还不够满意,根据我们的分析,是因为声母在切分时容易产生误差,时长普遍较短,而方差又相对较大。对于预测效果差的声母如 b, d, r 需要通过更多的语音学以及发音机理方面的工作进行改进。

参 考 文 献:

[1] S. H. Chen, S. H. Hwang, et al. An ANN-based prosodic information synthesizer for Mandarin text - to - speech [A]. IEEE trans. on Speech Audio Processing[C], 1998,6(3):226 - 239.

[2] Riedi M. A Neural - Network - Based Model of Segmental Duration for Speech Synthesis[A], in: Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH 95)[C], Madrid, Spain, 1995,1:599 - 602.

[3] Hyunsong Chung. Duration Models and Perceptual Evaluation of Spoken Karean[A]. Proceedings of Speech Prosody 2002, Aix - en - Provence[C], France, 2002. 219 - 222.

[4] Van Santen. J. P. H. Assignment of segmental duration in text - to - speech synthesis[J]. Computer Speech and Language, 1994, 8: 95 - 128.

[5] Riedi M. Modeling Segmental Duration with Multivariate Adaptive Regression Splines[A], in 5th European Conference on Speech Communication and Technology (EUROSPEECH 97)[C], Rhodes, Greece, 1997, 5: 2627 - 2630.

[6] S. H. Chen, W. H. Lai, et al. A New Duration Modeling Approach for Mandarin Speech[J]. IEEE Trans. Speech Audio Processing, 2003, 11(4):308 - 320.

[7] 贾俊平,何晓群. 统计学[M]. 北京:中国人民大学出版社,第一版,2001.

[8] Van Santen, J. P. H. Contextual Effects on Vowel Duration[J]. Speech Communication. 1992, 11,513 - 546.

[9] Sun Lu, Hu Yu, Ren - hua Wang. 2004, Perceptual Analysis of Duration Evaluation in Mandarin[A]. TAL2004 [C]. Beijing, China.