中 文 信 息 学 报
第16卷 第1期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.16 No.1

编者按 中国中文信息学会与中国计算机学会、中国自动化学会等兄弟学会联合召开全国人机语音通讯学术会议已经是第六届了,这个系列会议已成为全国人机语音通讯界最集中的重要会议。会议的程序委员会从一百一十九篇论文中评选出十一篇优秀论文,我编辑部得到授权,出版此专集,以飨读者。

基于统计韵律模型的汉语语音合成系统的研究

陶建华 赵 晟 蔡莲红

(清华大学计算机系人机交互与媒体集成研究所 100084)

摘要:本文论述了采用统计模型进行汉语韵律层级结构分析和韵律建模的思路,在此基础上建立了汉语语音合成系统。其中,本文还仔细阐述了韵律代价函数的构造,及其参数的自动训练算法。同时,论文还分析了韵律特征间相互作用对音节基元选取的影响,并最终实现了一个连续语流中用于汉语语音合成的音节基元选取模型。测试表明了本文提出的基于统计模型的韵律层级分析和韵律建模思路,能够较好应用于汉语语音合成系统的构造,并使之具有良好的合成语音的自然度。

关键词:汉语韵律层级结构;韵律建模;韵律代价函数

中图分类号:TP391.42

Study of Chinese Speech Synthesis System Based on Statistic Prosody Model

TAO Jian-hua ZHAO Sheng CAI Lian-hong

(HCI & MI, Dep. of Computer Science and Technology of Tsinghua University, Beijing 100084)

Abstract: The paper describes the methods of Chinese Prosodic Hierarchy Analysis and Prosody Modeling, which are based on statistic algorithm. Meanwhile, the paper also describes the prosody cost function and corresponding training method for the parameters. Furthermore, the interaction among the prosodic features is analyzed in respond to its influence in speech unit selection procedure. Based on these, a Chinese Syllable Unit Selection Model was generated for the spontaneous speech synthesis system. The tests show that the method described in the paper is much suitable to the constitution of Speech Synthesis System and improves the naturalness of the synthesis result a lot.

Keywords: Chinese Prosodic Hierarchy; Prosody Modeling; Prosody Cost Function.

一、引言

进入20世纪九十年代以来,人工智能、自然语言理解、信号处理、随机过程、模式识别等领域越来越多向语音处理中渗透,并得到了非常成功的应用,导致了语音技术在多项关键性技术上的突破。同时,语音合成的研究,也变成了一个系统工程,所涉及的知识也远远超出了语音信号研究的本身,而越来越广。在这种背景下,语音合成的研究,也同样融入了许多新的概念。其中,韵律的层级结构分析和韵律预测方法的研究,受到越来越多的重视。本文也围绕着这两

本文得到国家自然科学基金(69875008)支持。

作者陶建华,男,1972年生,博士,讲师,主要研究方向为语音合成、韵律建模技术、文本分析、多媒体信息检索等。

个研究重点,分为两个主要部分:汉语韵律的层级结构分析和韵律建模方法研究,并在此基础上建立汉语音节基元选取模型,其中以统计模型的思路贯穿整个研究工作。本文首先论述了基于统计方法的韵律词分析和韵律短语分析的研究(第二节)。进而,论述了基于概率的韵律建模方法上的韵律代价函数,并详细阐述了代价函数的训练算法(第三节)。在模型的基础上,本文实现了由韵律代价函数而构筑的汉语语音合成中的选音算法,并进一步分析了韵律特征间相互作用对音节基元选取的影响。最后,论文通过实验比较分析了基于统计的韵律模型的误差分布情况,证明本文采取的统计建模思路对汉语语音合成的自然度,起到了较大的提高。

二、基于文本的汉语韵律层级结构分析

韵律特征主要包括重音、语调和韵律结构(韵律成分的边界结构)[1],在言语交流中起到非常重要的作用。它能够将语音材料组成树状层次结构,成为音系结构。在音系学里,韵律层级从小到大依次为:莫拉、音节、音步、音系词、附着语素词组、音系短语、语调短语和韵律语句。在本文的汉语语音合成的研究中,将其省略为三个基本层级:韵律词、韵律短语和语调短语。而最为关键的则是韵律词和韵律短语的分析。

2.1 基于统计和规则相结合的韵律词分析模型

1 基于 Ngram 的分词模型建立、参数平滑和消歧处理

定义待处理文本由 $C=C_1\,C_2\cdots C_n$ 组成,W_i^k 表示第 k 种可能中第 i 个词,T^k 和 T_i^k 表示与 C_i 和 W_i^k 分别对应的词性序列和词性,可以得到分词和词性标注决策规则:

$$W^{opt}=\arg\max(P(W^k,T^k\mid O))=\arg\max(P(W^k,T^k))=\arg\max(P(T^k)P(W^k\mid T^k)) \quad (1)$$

在确定分词和词性的同时,确定文字的读音。在系统中,则通过对事件次数的折扣进行模型参数的平滑,同时,采用汉语复合词的规则来进行分词消歧处理。

2 韵律词合并规则的引入

单纯的统计语言模型并不完全适用于语音合成,其主要原因就是韵律词和语法词之间存在着一定的差异,语法词无法完全体现语气特性。同时汉语中独特的变调问题,如:上音变调,“一”、“七”、“八”、“不”等词的变调,姓氏的读音等一系列语音合成中非常关心的问题,也无法用统计模型完全解决。

研究表明,汉语的虚词,如:“在、和、的、地、于”等,以及一些常用的关键词,如:“先生、所长、市长、爷爷、博士”等称谓名词、“市、县”等地名名词、“年、月、日”等时间名词、“吃、偷”等后向动词,以及“虽然、因为、所以”等等,在构成韵律词中,起到前向合并或后向合并的诱导左右。因而,本文为此引入了汉语韵律词合并词典,通过对可能产生前向和后向合并诱导的词汇,建立索引词典,根据词的类型,制定前向和后向合并的规则,如“年”前面如果是数字,则向前合并,构成“年代”的韵律词。通过韵律词的修正,增加合成语音的韵律平滑程度。

2.2 韵律短语的预测

1 韵律短语预测的一般性问题

韵律短语作为一个重要的韵律特征,它在韵律词的基础上,通过人的自然呼吸将语句按不同的语气分成一些片断,即增强了语句的节奏感、流畅性。

2 基于分类决策树的汉语韵律短语预测

由于可以融入一定的人工经验,本文采用了决策树方法对韵律短语进行界定。这种方法基于大量的已标注了分词、拼音和词性的语料的统计,通过实现根据经验提出的大量问题集而自动训练决策树,并使用训练后的决策树对给定的文本进行节奏的预测,其结果相较传统的方

法有了较大的改善。图 1 所示为利用分类决策树法实现韵律短语预测的示意图。

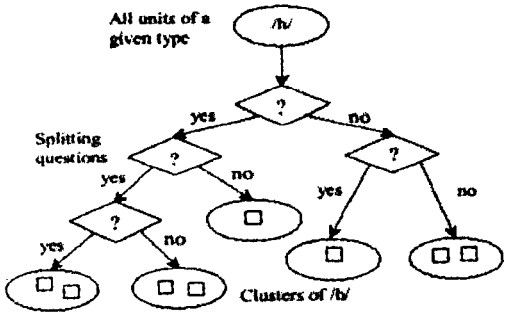

图 1　韵律重音和短语预测决策树示意图

表 1　用于韵律短语边界预测的文本及相应标注实例

已分词的文本：	| 一 | 一 | 奉　陪 | 吃　喝 |
词性标注：	z　z　v　v \
韵律边界标注：	/ 一　一 / 奉　陪　吃　喝 /
拼音信息：	| yil　yil | feng4　pei2 | chil　he1 |

利用的规则信息则根据不同的上下文信息来确定，其中包括：词性、位置、音节声调等。决策树的叶子评估函数，采用输出目标的距离来衡量。

3　问题集的提出

在汉语韵律短语的预测中，决策树节点的特征向量包括了 13 个数值型或分类型的变量，从这些变量中提出了 95 个问题。这些问题包括了两种类型，即枚举型和数值型。无论何种问题，其答案都是二值化的。这些问题的合理选择能够使得构造的决策树生长平衡，即一定程度上防止决策树右偏现象的发生。

选择的问题分为三大类：

* 拼音相关类：如：当前字和前后字音调的判断，当前字的声母、韵母和前接的韵母和后接的声母的判断等。
* 词性相关类：预留了 50 个词性的判断，实际语料的词性标注只使用了 29 个词性。
* 位置信息类：如：词内字位置，句内词位置和字数、词数的判断等。

实验表明，基于决策树的韵律短语分析模型，获得了 84 %(相对于人工标注，在本文的语料标注中，语料标注为经过训练的听音人，其人工标注统一性在 88 %左右) 的韵律短语预测的精确度。因而该模型的运算精确度，基本满足了语音合成系统的韵律处理的要求。

三、韵律代价函数

3.1　韵律建模的概率描述

汉语中的语境信息，根据的文本的上下文信息，按照其对汉语韵律特征不同层次的影响，共分为五组：当前音节信息(声母类型、韵母类型、声调类型、在词中位置、与前音节的耦合度)；相邻前音节信息(韵母类型和声调类型)；相邻后音节信息(声母类型和声调类型)；音节所在韵律短语信息(音节数、在句中位置、重音类型、距前一个重音距离和距后一个重音距离) 以及语句信息(语句类型和韵律短语个数)。共 17 个参数[6]。

语境与韵律特征之间具有很强的相关性，韵律特征参数的分布受着语境信息的影响，这种影响又满足一定的概率关联关系，而不是一个简单的函数映射。正如，在汉语轻声的研究工作中，沈炯先生在“从轻音现象看语音与语法研究的关系”[4]一文中阐述：“当我们说一种语音现象很明显的时候，主要是指它的离散成分很容易把握，一般并不指它的音理是否容易认知。”从概率的角度，对于一个已知的语境参数，与之相对应的韵律特征参数可表述为：

$$\overset{7}{Y}_n = \arg\max_m P(\overset{7}{Y}_{n,m} | \overset{7}{A}_n) \tag{2}$$

由 Bayesian 公式可以得到：

$$\vec{Y}_n = \arg\max_m P(\vec{Y}_{n,m} | \vec{A}_n) = \arg\max_m \frac{P(\vec{A}_n | \vec{Y}_{n,m})\, P(\vec{Y}_{n,m})}{P(\vec{A}_n)} \tag{3}$$

由于 $P(\vec{A}_n)$表示语境信息的统计分布,可视为常数,将其忽略,公式(2)将进一步转换为:

$$\vec{Y}_n = \arg\max_m P(\vec{Y}_{n,m} | \vec{A}_n) = \arg\max_m P(\vec{A}_n | \vec{Y}_{n,m})\, P(\vec{Y}_{n,m}) \tag{4}$$

3.2 韵律代价函数

当公式(4)中 $P(\vec{Y}_{n,m}) =$常数 C时,它可以转换为:

$$\vec{Y}_n = \arg\max_m P(\vec{A}_n | \vec{Y}_{n,m}) = \arg\max_m (S_{n,m}) = \arg\max_m \left| \sum_i \gamma_i V(a_{n,m,i}) \right| \tag{5}$$

而,$S_{n,m} = \sum_i \gamma_i V(a_{n,m,i})$ (6)

则为反应语境对韵律特征作用的韵律代价函数。其中 $\gamma_i = f(\omega_i)$,$a_{n,m,i}$,表示待合成语句中第 n 个音节中第 m 个候选样本中第 i 个语境参数值,它是语境信息的数值化表示,通常取非负整数。在系统实现过程,则假设:$\gamma_i = f(\omega_i) \approx \omega_i$,其中 ω_i 为不同语境参数产生贡献的影响因子或称权值。函数权值 ω_i 的初始值确定,韵律处理的影响很大,虽然进一步的权值调整可以通过训练机制来实现。

函数 $V(a_{n,m,i})$表示候选音节样本的语境参数 $a_{n,m,i}$与目标语境参数的逼近度,它为 0~1 之间的归一化值。本文将语境参数按其数学特性分为有限量化和分级量化两类。有限量化类包括:词性、声韵母类型等,不同参数不代表参数之间的层次关系,只反映参数的类型;分级量化类包括:重音级别、边界特性、所有位置信息和距离信息等,这些参数的值具有量化可比性。

韵律代价函数的结果为一组语境信息的加权统计值。通过权值的记忆和训练达到适应不同语料库的目的。

3.3 韵律代价函数的权值训练

假设:韵律代价函数中初始权值向量为:$\vec{\omega}^0 = \{\omega_1^0, \omega_2^0, \cdots, \omega_p^0, \}$,经过 $j-1$ 次训练后,权重系数为 $\vec{\omega}^j = \{\omega_1^j, \omega_2^j, \cdots, \omega_p^j\}$,其中 p 为权值矢量的维数,j 为非负整数。它们满足约束条件:

$$\sum_{i=1}^{P} \omega_i^j = 1 \tag{7}$$

约束条件的目的是使得训练尽量能够产生收敛,同时使得权值的调整在整个向量空间保持均衡。

训练所采用的样本集为:$\{\vec{Y}_1, \vec{Y}_2, \cdots, \vec{Y}_N\}$,通过韵律代价函数选音得到的样本集为:$\{\vec{Y}_1, \vec{Y}_2, \cdots, \vec{Y}_N\}$ 其输出和训练语料的误差为:

$$E(\vec{\omega}^j) = E(\vec{\omega}^j, \vec{Y}) = \frac{1}{N} \Sigma_{n=1}^{N} (\vec{Y}_n - \vec{Y}_n)^2 \tag{8}$$

其中,Y 由样本的声学参数构成向量空间,它由样本的 $SPiS$ 参数[6]、能量归一化均值、音节音长以及音节的头部、中部和尾部的 $MFCC$ 参数平均值组成。

第 j 步训练过程中,函数的权重调节则通过下式进行:

$$\omega_i^{j+1} = \omega_i^j + \eta^j \cdot d_i^j \tag{9}$$

其中 η^j 为第 j 步权重调节的步长,d^j 则表示第 j 步权值调节的方向,通过如下步骤来实现:

定义：ΔY_n 为第 n 个音节，选音的结果，与目标音节之间的韵律特征误差，它可以表示为，$\Delta \overset{7}{Y}_n = \left| \arg \max_m \left| \sum_i \gamma_i V(a_{n,m,i}) \right| - \overset{7}{Y}_n \right|$，与之相对应的语境参数为：$A_n^s = (a_{n,1}^s, a_{n,2}^s, \cdots, a_{n,p}^s)$，相应的语境参数各个分量与目标语音的语境参数各个分量之间的误差为：$\Delta(V(a_{n,i}^s)) = [V(a_{n,i}^s) - V(a_{n,i})]$

定义：连续语音第 n 个音节中，所有候选样本中，与目标音节韵律特征误差最小的样本的其语境参数为 $A_n^{\min} = (a_{n,1}^{\min}, a_{n,2}^{\min}, \cdots, a_{n,p}^{\min})$，最小韵律特征误差为：$(\Delta \overset{7}{Y}_n)_{\min}$，相应的语境参数各个分量与目标语音的语境参数各个分量之间的误差为：$\Delta(V(a_{n,i}^{\min})) = [V(a_{n,i}^{\min}) - V(a_{n,i})]$

一般情况下，语境参数已经被归一化到 0 和 1 之间，因而可以计算得到：

$$d_i^j = \left| 1 - \frac{(\Delta \overset{7}{Y}_n)_{\min}}{\Delta \overset{7}{Y}_n} \right| \left| \Delta(V(a_{n,i}^{\min})) - \Delta(V(a_{n,i}^s)) \right| + C \tag{10}$$

由约束条件(7)，并结合式(9)可以得到：

$$\sum_{i=1}^{p} \omega_i^{j+1} = \sum_{i=1}^{p} \omega_i^j + \sum_{i=1}^{p} \eta^j \cdot d_i^j = 1$$

进而可以求出：

$$\sum_{i=1}^{p} \eta^j \cdot d_i^j = 0$$

通常为了简化运算过程，η^j 被假定为 0～1 之间的一个常数 η。则可以得到：

$$\sum_{i=1}^{p} d_i^j = 0 \qquad \text{从而：}$$

$$C = \frac{1}{P} \left| \frac{(\Delta \overset{7}{Y}_n)_{\min}}{\Delta \overset{7}{Y}_n} - 1 \right| \sum_{i=1}^{p} \left| \Delta(V(a_{n,i}^{\min})) - \Delta(V(a_{n,i}^s)) \right| \tag{11}$$

利用公式(9)、(10)和(11)可以实现训练的整个过程。

四、语音合成中选音模型的实现

由于公式(5)是在假设 $P(Y_{n,m}^7)$ 为常数的前提下所得。而很多情况下，$P(Y_{n,m}^7)$ 能够反映韵律特征本身的相互作用，且非常明显。例如：当一个音节本身被重读时，通常会影响后续音节的发音等。这一结论，在吴宗济先生的“普通话三字组变调规律”[3]，以及林茂灿的“北京话轻声的声学性质”[7]、“普通话轻声与轻重音”[5]的论文中均得到了不同程度的论证。

当只考虑相邻音节的韵律特征关系时 $P(Y_{n,m}^7)$，则可以通过韵律单元的韵律特征转移概率来反应：

$$P(\overset{7}{Y}_{n,m}) = \sum_{g=1}^{M} P(\overset{7}{Y}_{n,m} | \overset{7}{Y}_{n-1,g}) P(\overset{7}{Y}_{n-1,g}) \tag{12}$$

由图 2 所示，利用该韵律单元的韵律特征转移概率，构筑成音节单元的选取矩阵，通过路径搜索来获取最符合韵律特性的音节候选单元，采用波形拼接的方法合成出语音并输出。

图 3 和图 4 显示了考虑韵律特征相互作用前后的选音结果图。其中“是”、“保”、“护”、“环”等音节在考虑韵律特征相互作用后，其调域出现了下降或一些其它变化，从而使该句发音更为平滑，并体现了更好的节奏感。

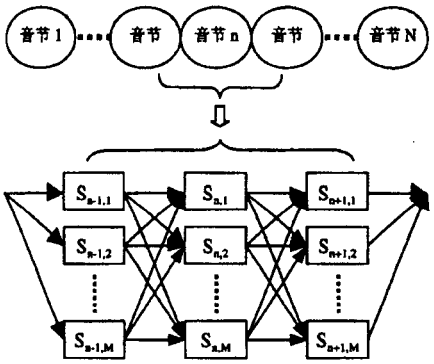

图 2　考虑韵律特征相互作用的音节单元选取矩阵

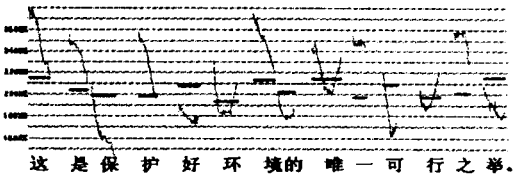

图 3　不考虑韵律特征相互作用选音的结果

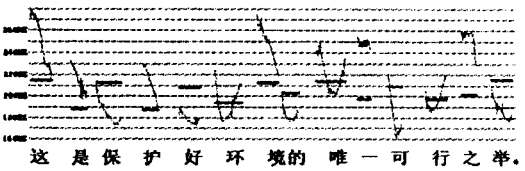

图 4　考虑韵律特征相互作用选音的结果

五、总结

本文采用统计中概率的方法对韵律层级结构分析和韵律建模思路进行了阐述，同时，本文进一步分析了采用韵律代价函数进行直接基元选取的算法及其参数的训练算法。在此基础上，实现一个利用选音矩阵构筑的汉语语音合成声学模型，并对合成语音输出结果进行了测试。本文的工作，虽已在语音合成系统上实现，但最主要的目的是希望通过本文的工作，将语音合成的研究引入更多的数值化、可计算化和可训练化的方法。新一代的语音合成系统正向着概念到语音或意念到语音的方向发展，其语料库的组成也越来越庞大，信息含量也变得更为丰富。语音合成需要逐步摆脱过渡依赖设计人经验，变得规范化，在这种发展趋势中，有关整个语音的研究可计算化，尤其是语音合成各组成核心模组的数值模型化，将会变得越来越重要。

参 考 文 献

[1] Selkirk, E. Phonology and syntax: the relation between sound and structure. MIT press, 1984

[2] Achim Mueller, Jianhua Tao, Ruediger Hoffmann, Data-driven importance analysis of linguistic and phonetic information, ICSLP2000

[3] 吴宗济. 普通话三字组变调规律. 中国语言学报, 1985, 第二期

[4] 沈炯. 从轻音现象看语音与语法研究的关系. 吕淑湘等.《语法研究入门》. 商务印书馆, 1999, 158 页

[5] 林茂灿, 颜景助. 普通话轻声与轻重音. 语言教学与研究, 1990 年第 3 期

[6] 陶建华, 蔡莲红等. 汉语 TTS 系统中可训练韵律模型的研究. 声学学报, 第 26 卷: 67 - 72

[7] 林茂灿, 颜景助. 北京话轻声的声学性质. 方言. 1980 年第 3 期

[8] Andrew J. Hunt and Alan W. Black, " Unit selection in a concatenative speech synthesis system using a large speech database", ICASSP 96

[9] 孙茂松等. 消解中文三字长交集型分词歧义的算法. 清华大学学报, 1999, 39(5)

[10] 王政红. 论双音复合词的构成格式. 南京理工大学学报, 1997, 10(6)