韵律数字建模与韵律研究

陶建华　蔡莲红
jhtao@tsinghua.edu.cn　clh-dcs@tsinghua.edu.cn
清华大学计算机科学与技术系人机交互与媒体集成研究所

摘 要

随着近几年来，随着计算机技术、声信号处理技术和人工智能技术的飞速发展，语音合成中韵律建模越来越朝着数字化、智能化和综合化方向发展。韵律建模工作也超出了传统的单存韵律参数处理本身，而与语法分析、重音分析和预测以及其它语音声学参数的处理更多的融合在一起。各种人工智能技术也在其中得到了充分的体现，如自然语言理解中的分词、词性分析、语法和语义分析技术、人工神经网络技术、决策树技术、甚至是隐马尔可夫技术等。现代韵律建模思想，已经使得汉语的语音合成系统的自然度获得了巨大的进步。但是语音学的研究，尤其是传统语音学研究和韵律的研究对汉语韵律建模是否还存在帮助呢？本文力图从作者的韵律数字建模的工作中，包括了：音节重音的自动预测和处理，基频、音长和能量等声学物理参数的自动预测和处理，来阐述现代汉语语音学的研究在汉语韵律数字建模工作中的体现。

关键词：韵律数字建模、语音学。

引 言

汉语的韵律特征包括重音、节奏、声调和语调等特征，对其特征进行分析是建立韵律模型的基础。长期以来对汉语的韵律特征的研究多采用实验句的研究思路，其研究成果非常丰富，并从中揭示了人们发音的许多规律性韵律特征和发音机理。但这些思路直接用在韵律建模方面，往往存在着一定的局限性，这种局限性通常表现为难以形成一个系统和全面的模型解决方案。关于人机对话中普通话连读的音色变化和韵律变化，可分为两类：一类是“固然的”，另一类是“或然的”。对于"固然的"变量，说话人为了要表达某些事物，可以不假思索地说出，其中的音段和韵律自动会随语音、语法、音系三平面的规律而变；而对于“或然的”变量，则为了还要表达某种情感、口气。韵律模型不仅需要能够进行高质量的韵律参数的预测，还应该能够反应人的千差万别的特征，适应不同场合的能力，能够象人一样具有自我学习的功能。近几年来，计算机的运算速度和容量都获得了飞速的发展，从大量语料中提取连续语句的韵律特征，采用数据驱动的方式，不仅已经成为可能，以新一代计算机技术为代表的数据驱动技术和大规模语料设计和制作技术，为韵律建模的发展带来了新的契机，同时也为进一步深入研究汉语的韵律特征提供了更多的方法和材料。

以数据驱动技术为代表的韵律建模技术，融合了大量现代人工智能领域的技术，如自然语言理解中的分词、词性分析、语法和语义分析技术、人工神经网络技术、决策树技术、甚至是隐马尔可夫技术等。通过这些方法的应用，结合语料的设计，建立韵律的训练模型，从而是汉语的韵律模型获得了相当的提高。在很大程度上，计算机的自动分析和机器训练，取代了过去的大量人工统计和分析过程，似乎在一定程度上，只要通过机器自动训练，就可以解决大多数韵律模型的问题。因此，很多人乐观的认为传统的语音分析方法甚至是经验和结论，都已经不太重要了。然而，这只是表面现象，韵律建模工作牵涉的面很多，其中语料设计、语料标注、轻重音预测、韵律参数量化，甚至是模型框架设计本身都与语音学的研究密不可分。同时，不管数据驱动模型设计的算法多么巧妙和复杂，它都不可能解决韵律特征所

有可能的变化，这些变化可能是发音人固有的，也可能是随机的。本文的着重阐述了语音学研究在韵律数字建模中的重要性。并分：大规模语料设计技术与韵律代价函数、韵律节奏的预测、韵律特征间的耦合效应对韵律建模的影响等几个方面加以阐述。

1 大规模语料设计与韵律代价函数

1.1 用于语音合成的语料设计

当人在不同的语境下说话时，会有不同的韵律特征，语境与韵律特征之间具有很强的相关性。谈到语音和句法以及语义之间有密切关系的时候，林焘先生强调“绝对不能把语言的这三方面割裂开来孤立地进行研究。”V.Auberge等人[1]提出了韵律参数是受语言学参数的综合影响的观点，并作了一定阐述。并由Katherine Morton[2]在他的对话系统的语音合成模块中，根据几个基本的上下文模式，加入情感的变化，使合成出的语音变得有些生动。“语调构造由语势重音配合而形成。它是一种语音形式，它通过信息聚焦来实施超语法的功能语义。”Selkirk在1984[3][4]年就提出了一种韵律分层模型，并认为韵律结构从低到高的分层依次是音步、音节、韵律词、韵律短语和语调短语，并认为句子的韵律结构和句法结构之间存在着系统的影射关系。

大规模语料库是现代韵律建模工作的基础。语料库的完备性和丰富性，将直接影响韵律模型的质量。影响人们发音的韵律特征最根本的因素应该是发音人的发音习惯和他（她）需要表达的意思。然而，在现实的计算处理能力中，直接分析语义是较为困难。因而，利用语境信息进行分析和设计便被大多数学者所采用，如语音特征，包括：声调、重音、音长，以及语言特征，包括：音节内部信息，如音节内的音联关系等，和超音节信息，如词性、各种位置信息、以及所含音节或词的信息等。因而，语料的设计也兼顾了语法结构上的层次信息。

语境信息的选择，因人而异，与设计者对语法、语义和韵律的理解，以及设计意图密切相关。与语料相关的语境信息的覆盖程度，即语料的完备性，是一个语料设计成功与否的非常重要的标志，当然，这与具体选用的语境信息有关。然而，不可能存在这样一种语料，它覆盖了所有的语境分布。因而，用于韵律建模而设计的语料，其主要设计原则应该是尽可能的满足影响韵律特征的语境参数分布，用于更好的体现韵律特征之间的区别特征[5]。通过对一些韵律的分布进行统计，并分析语境参数的相应分布，表明诸如：位置信息、声调信息、重音、词性等均占据了重要的作用。通过数学方法，对韵律的特征分析使我们得到了对语料设计的宝贵的资料。而语料库中语境信息的完备性，可以通过如下公式来体现：

$$CR = \sum_i \omega_i V(a_i)$$

其中，CR表示语境信息完备性，为0～1之间的一个数。ω_i为语境参数A中第i个分量a_i对韵律特征的权重，$V(a_i)$则为第i个分量a_i在语料中各种变化的覆盖程度。

1.2 韵律代价函数

很多基于多样本音库的语音合成系统，也把ω_i作为语音基元选取时代价函数的一个重要组成部分。其中代价函数为：

$S = \sum_i \gamma_i V(a_i)$，其中，$\gamma_i = f(\omega_i)$

通常情况下认为：$\gamma_i = f(\omega_i) \approx \omega_i$

如何确定函数权值的初始值，往往对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。

一种，利用神经网络的权抑制的方法，可以较为有效而迅速的确定初始权的值。如图1所示。它通过在一个传统的神经网络模型中，在输入层（语境参数）和中间隐层之间加入，权抑制层来实现。则神经网络的构成函数变为：

$$\overline{F}(w) = F(w) + \lambda \sum_{\{k|\omega_k \in w_{sel}\}} \omega_k^2$$

$$\overline{w}^{i+1} = \overline{w}^{i} - \eta \nabla \overline{F}(w) = \overline{w}^{i} - \nabla \left[\eta F(w) + \eta\lambda \sum_{\{k|\omega_k \in w_{sel}\}} \omega_k^2 \right]$$

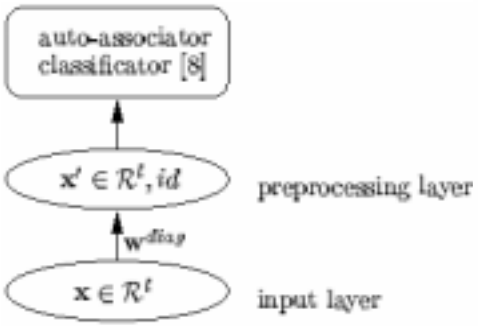

图1：具有权抑制层的神经网络模型

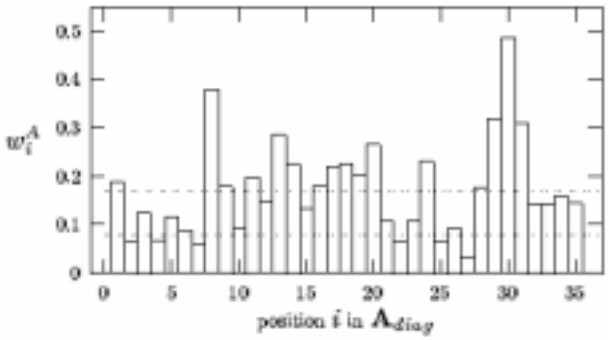

图2：权值抑制的训练结果

图2为经过训练的权值分布图，通过这些数值，可以非常方便的确定韵律代价函数的权值初始值。针对汉语，实验表明，位置信息、声调信息、重音、词性等语境信息，占据着较为重要的作用。

2 韵律节奏的预测

2.1 韵律节奏预测的一般性问题

任何文本都存在着句法结构信息，但人说话时，却不一定按照语法结构来发音，而是按照一种有发音人特色的韵律结构来发音。 韵律节奏作为一个重要的韵律特征，对语音合成的自然度和可懂度产生这重要影响，它通过人的自然呼吸将语句按不同的语气分成一些片断，既增强了语句的节奏感、流畅性，同时也便于消除一些歧义。韵律节奏通过轻重音组合和韵律短语等信息的综合作用来体现的，因而，轻重音的预测和韵律短语边界的预测在韵律模型中有着重要的意义。韵律短语的边界在感知上体现出一系列的声学参数的特征组合，如：音节间停顿加长、短语的最末一个音节时长变长、调域的低音线下顷（沈炯）[9]等。在韵律模型中，通过对韵律短语边界位置进行精确的预测的话，就为有效的确定韵律特征中各声学参数奠定了良好的基础。 而在重音情况下，音长和基频被认为是决定音节轻重的主要因素，根据声调的不同，在不同的环境下，如短语首、短语尾、不同语气的语句中，其表现均不尽相同。如轻声音节的重音表现，在孤立词、不同语气的连续语句中，就有很大的变化[18]。在连续语流的发音中，同为重音，其声学参数的表现，却千差万别。

较早期的韵律模型，多从韵律的声学参数上考虑直接建立模型，如基频模型或模式、音长模型、停顿模型等。并对一系列现象，如不同语调中基频模式、轻重音中基频模式和音长模式等。直接建立在语境到声学参数的韵律建模思路，应该说是可行的，但语境的变化毕竟非常复杂，声学参数的任一细微变化，可能相应的语境组合的可能将非常多。因而，采取分层韵律建模的思路，将在一定程度上降低韵律建模的复杂度。其思路为，首先建立语境到韵律节奏的模型，包括重音预测、韵律短语预测等。继而，进一步通过重音和韵律短语信息，并结合语境参数实现声学参数的预测。

2.2 基于 CART 模型的韵律短语预测

目前对韵律短语预测的主要方法有：采用统计模型，利用韵律边界相对语境信息的统计结果，来进行韵律短语边界的预测；另一种方法，采用分类决策树技术的规则学习算，该方法利用语料中手工标注的韵律边界信息建立语境到韵律短语边界的预测模型。

其中，规则学习法，如决策树方法，是一种非常实用、有效的方法。该方法在各个领域

得到了广泛的应用，由于其方法简捷，且可以融入一定的人工经验，因此，采用决策树方法对韵律重音和短语界定是一种较好而可行的方法。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要了较大的改善。图 2 所示为利用分类决策树法实现韵律短语预测的示意图。

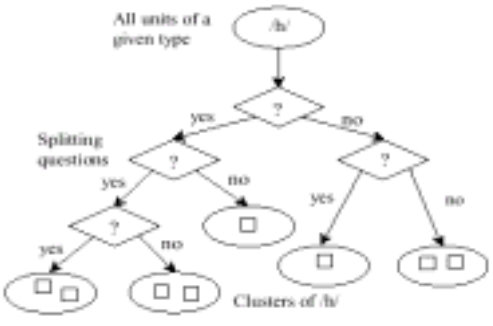

图 3：韵律重音和短语预测决策树示意图

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：	|yi1　yi1|feng4　pei2| chi1 he1|

表 1：用于韵律短语边界预测的文本及相应标注实例

而利用的规则信息则根据不同的上下文信息来确定，如：词性、位置、音节声调等。决策树的叶子评估函数，可以采用输出目标的距离来衡量。该算法的特点要求设计的规则简练并多样化。同时，由于汉语语法结构的独特性，一些研究表明[20]，汉语的一些虚词，如：在、和、的、地、于等，以及一些常用的关键词，如：如果、但是、虽然等，在汉语的韵律短语或语法从句起着重要作用。因而，进一步分析这些词在韵律节律中的体现，对最终改善规则学习的算法，将起到很大的作用。

2.3 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：包括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。
- ✧ 词性相关类：预留了 67 个词性的判断，实际语料的词性标注只使用了 50 个词性。
- ✧ 位置信息类：包括词内字位置，句内词位置和字数、词数的判断。

实验表明，辅以助词分析的决策树模型，获得了 86%（相对于人工标注）的韵律短语预测的精确度。这一精确度，基本能够满足语音合成系统的韵律处理的要求。

3 韵律特征间的耦合效应

传统的语音合成系统，其韵律模型的做法是，通过人工分析大量的语句或篇章，总结出一系列典型的韵律特征，如：音节基频曲线、词的基频曲线、句调特征等。然后找出语境信息到这些韵律特征的映射规律，即总结出韵律模式：

$$\vec{Y} = \varphi(\vec{A})$$

较早期的语音规则合成系统，大多采取了此设计模式。该模式设计的中心思想就是认为语境信息到韵律特征的关系，是一个类似函数的映射关系，因而能够用规则或标注信息，加以完全确定。现代较为常用的决策树韵律模型中，采用有指导性的训练方法，其最后实现韵

律生成的过程，依然是一个在人工指导下的类似函数映射。然而，韵律特征参数的分布同时受着语境信息的影响和韵律本身的相互影响，这种影响满足一定的概率关联关系，而不是一个简单的函数映射。即，不能简单的说，在某种情况下，人发音的韵律特征一定会怎么样；而只能认为，在一种情况下，人发音的韵律特征可能是什么样的，当然也可能是别的样子，只是它们出现的概率不一样。正如，沈炯先生在“从轻音现象看语音与语法研究的关系”[9]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。离散成分是平常人都能把握的形式——在言语使用过程和语言教学中分别把握它。只要它能被把握，语音研究就可以在离散基础上进行，可以用文字或其他表音手段来列举语言样品，讨论问题就方便了。”

对于一个已知的语句参数，与之相对应的韵律特征参数，应为所有韵律特征参数中出现概率最大的一组，即为：

$$\vec{Y} = \arg\max_n P(\vec{Y}_n \mid \vec{A}) = \arg\max_n \frac{P(\vec{A} \mid \vec{Y}_n)P(\vec{Y}_n)}{P(\vec{A})} = \arg\max_n P(\vec{A} \mid \vec{Y}_n)P(\vec{Y}_n) \qquad (2)$$

假设：对于 $\vec{A}$ 存在 $\vec{Y}_j$，使得 $\begin{cases} P(\vec{Y}_{i=j} \mid \vec{A}) = 1 \\ P(\vec{Y}_{i\neq j} \mid \vec{A}) = 0 \end{cases}$ 其中：$(j \in I, n \in N)$

则 $\vec{Y}$ 可以表示成 $\vec{A}$ 的函数形式，即：

$$\vec{Y} = \varphi(\vec{A})$$

由此，韵律模型被简化成了简单的函数形式。而很多情况下，$P(\vec{Y}_n)$ 所导致的韵律特征本身的相互作用又是非常明显。例如：汉语韵律中的逆同化和顺同化效应，上音连读变调等。这一结论，在吴宗济先生的“普通话三字组变调规律”[6]，以及林茂灿先生的“北京话轻声的声学性质”[13]、“普通话轻声与轻重音”[14]的论文中均提到了不同程度的体现。如：林茂灿先生提到的“轻声音节 F_0 曲线的形成，是由于它跟前面重读音节发生声调协同发音所致”，就是这一现象的典型反映。因而，一个好的韵律模型必须要能够反应韵律受语境的制约情况，而且还因能够体现韵律自身相互作用的现象。

因而，在采用韵律代价函数进行音节基元选取的韵律模型中，其整句评估因子可以表示为：$Q = \sum_j (\sum_i \gamma_i V_j(a_i)) P_{j-1,j}$，其中，$P_{j-1,j}$ 表示韵律单元的转移出现概率，它体现了韵律特征之间的相互作用。

其中，一个较为典型的例子就是，轻声音节在连续语句中受前后音节不同声调和重音情况的影响时，具有不同的韵律特征表现形式。

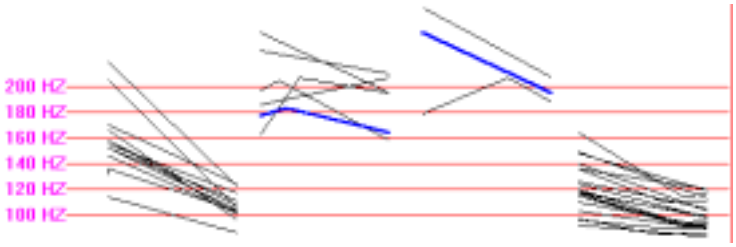

图 4：反映了轻声音节受重音的影响时的基频和音长的情况。

由图 4 可以看出，在重音条件下，前音节调型为阴平时，轻声音节的调型多表现为中起降调，一般音域较大。当前音节为阳平时，轻声音节的调型多表现为高起，呈降调或平调，音域较窄。当前音节为上声时，调型则多表现为高起，基本呈平调，音域很窄，而且音域整体变高，这与其他情况下出现在上声后的轻声音节有较大的区别。当前音节为去声时，轻声音节的调型则多表现为轻微下降，音域较窄，调型是低起还是中起与轻声音节与前音节的音域范围密切相关。另外，在不受重音影响时，除上声外的轻声音节的模式，与重音条件下的模式比较接近。上声后的轻声音节却主要表现为低起平调或升调，音域较窄。在前音节四种不同的调型下，轻声音节的音长与它们的比例关系分别为：1:0.59、1:0.66、1:0.61、1:0.73。总的来说，在重音条件下，轻声音节的基频相对于其他情况，往往受重音的影响被不同程度

的抬高。而在去声的重音音节后的轻声音节的音长较其他略长。

由于韵律参数预测，除了要有较好的语境信息到韵律参数之间的预测模型外，必须将整个模型放在整句甚至是篇章的整体环境中考虑。而在连续的语句中，韵律特征间的耦合效应则不可避免的出现，因而在韵律建模中，引入反应韵律特征转移的函数，对从整体上完善韵律模型，并提高韵律参数预测质量，起到了较大的作用。

4 总结

本文从作者多年来在韵律建模中的几个小侧面反应了语音学研究在韵律数字建模中的影响和推动作用。说明了，无论采用何种方法进行韵律建模工作，它都与语音学各个方面研究的结合是密不可分的，并体现在韵律建模的各个方面。同时，已现代计算机技术为特征的韵律建模工作，也给语音学的研究提出了许多新的需求，新的技术和算法的引入，在多个方面促使着该领域的快速发展。下一代的语音合成系统，必将是一个具有主动思维能力的，有情感的系统，无论是韵律模型的体系结构本身，还是语音学中韵律的研究，都在朝着更加口语化、情感化的方向发展。现代的技术发展很快，当语音学的研究，较多的应用现代测量仪器，并融入现代分析手段，不仅将会给语音学的发展引入新的推动力，也必将是韵律建模的工作走向新的高度。

引 文

[1]V.Auberge etc, “Generation of Intonation: A Global approach”, European Conference on Speech Communication and Technology, 1995, P2065-2068

[2]Katherine Morton, “Adding Emotion to Synthetic to Synthetic Speech Dialogue Systems”, ICSLP97, P675-678

[3]Selkirk, E. (1984) Phonology and syntax: the relation between sound and structure. Cambridge, MA: MIT press.

[4]Selkirk, E. & Shen Tong (1988) Prosodic domains in Shanghai Chinese, The phonology - syntax connection, edited by Sharon Inkelas & Draga Zec. Chicago: The University of Chicago Press.Shen,X-N.S. (1992) A pilot study on the relation between the temporal and syntactic structure in Mandarin, Journal of the International Phonetic Association 22.1/2: 35-43.

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

[6]吴宗济，普通话三字组变调规律，中国语言学报，第二期，1985

[7]吴宗济，从声调与乐律的关系提出普通话语调处理的新方法，中国语文，1997，P243-258

[8]郭锦桴，汉语声调语调阐要与探索，北京语言学院出版社

[9]沈炯，汉语语调模型议，语文研究，1992，VOL 4，P16-24

[10]沈炯，北京话上声连读的调型组合和节奏形式，中国语文，1994年第4期

[11]沈炯，汉语语调构造和语调类型方言，1994年第四期

[12]沈炯，从轻音现象看语音与语法研究的关系，吕叔湘等著，马庆株编《语法研究入门》，商务印书馆1999，158页

[13]林茂灿、颜景助，北京话轻声的声学性质，方言，1980年第3期

[14]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。

[15]许毅，Production and perception of coarticulated tones. Journal of the Acoustical Society of America 95:2240-2253.

[16]林焘、王理嘉，“语音学教程”，北京大学出版社

[17]陶建华、蔡莲红、赵世霞，“汉语TTS系统中可训练韵律模型的研究”，声学学报，2001,1,第26卷，P67-72

[18]陶建华、周同春、蔡莲红，“连续语流中轻声音节的模式”，第四届现代语音学学术会议论文集

[19]Shaw-Hwa Hwang and Sin-Horng Chen, “A Prosodic Model of Mandarin Speech and Its Application To Pitch Level Generation for Text-to-Speech”, IEEE Tran. 1995, P616-619

[20]应宏，蔡莲红，“结构助词在韵律短语界定中的作用”，第四届语音学会议，1999.8，北京