汉语语音合成中的文本分析和韵律处理

陶建华　蔡莲红　赵晟

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，北京，100084

摘要：本文通过阐述新一代汉语语音合成系统中文本分析、短语合并、韵律代价函数及韵律代价函数在语音基元选取中的体现，从而较为详细的分析了现代汉语语音合成系统的工作流程和有关的关键技术。指出传统意义上文本分析模型，无法体现语音合成中韵律节奏，并使韵律参数的预测变得困难。本文通过引入在文本分析中引入韵律节奏预测机制，从而使文本分析和韵律处理有机的结合起来，并阐明汉语语音合成性能进一步提高的方法。

关键词：汉语语音合成，文本分析，韵律处理

Text Analysis and Prosody Processing for Chinese Speech Synthesis

Tao Jianhua*　Cai Lianhong**　Zhao Sheng

HCI&MI, Dep. of CS, Tsinghua University, Beijing, 100084

* jhtao@tsinghua.edu.cn　** clh-dcs@tsinghua.edu.cn

Abstract: The paper explains the text analysis, phrase determination, prosody assessment function and unit selection method of newest Modern Chinese Speech Synthesis System, and also explains the running procedure and some key technologies of the System. It is pointed out that the results of traditional text analysis cannot offer the information of prosody rhythm for speech synthesis and makes it difficult in prosodic parameters prediction. The paper introduces a prosody rhythm detection method to speech synthesis, and combines the text analysis with prosody processing smoothly. Furthermore, the paper also presents the ideas on how to improve the quality of Chinese Speech Synthesis.

Keyword: Chinese Speech Synthesis, Text Analysis, Prosody Processing

一、 引言

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，并采用数据驱动的方式实现语音合成系统，不仅已经成为现实，而且为语音合成技术的发展带来了新的契机。以数据驱动技术为代表的语音合成技术，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，从而是汉语语音的合成质量获得了相当的提高。同时

这些方法的成功应用，在很大程度上也改变了汉语语音合成研究的研究重点，使汉语语音合成的研究突破早期重点在单纯算法的研究上，而变成一个系统工程的研究。语音合成的整体研究和开发，迈上了一个全新的台阶。

一般意义上的文本分析主要是指分词、注音和词性标注这些部分，现代文本处理的研究已经从语法结构和语义一层考虑，但是由于人在自然语流中发音虽然与语法结构有很大的关联性，但又具有很大的任意性，其韵律节奏中的韵律短语，并不等同于语法意义上的短语信息。而当韵律短语能够很好的得到体现，在重音一级和语调一级才可能更好的进行语音的韵律处理。因而在语音合成系统中，文本分析的结果并不能直接用来进行韵律处理，本文的重点就是通过在文本分析结果的基础上，引入韵律节奏的预测机制，从而实现文本处理和韵律处理的融合，并且并从更深的层次上分析了韵律的特性，及韵律处理方法提高的途经。由于韵律在语音合成系统占有非常重要的位置，它对基于数据驱动的系统中基频参数预测、选音及声学模型均有很重要的影响，因而更好的研究韵律模型以及研究韵律和其它相关信息处理的关系，将有着现实的意义。

二、 统计和规则相结合的文本分析

近年来，基于统计的文本分析模型越来越成熟，并被广泛的应用，然而在语音合成系统中，是否只使用统计语言模型就可以解决问题？由于，有相当一批读音中的变调问题，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，无法用统计模型完全解决，因而与规则的结合是必须的。同时，本文从另一个侧面有提出了应用规则简化统计模型中网络的复杂度，在保证分析质量的同时，使之便于计算，提高了运行速度。尤其在网络服务器一级中应用时，这一问题变得非常突出。

1. 基于 Ngram 的统计模型建立、训练和参数平滑

分词、注音和词性标注是语音合成体系结构最为初始的重要组成部分，也是语言模型需要解决的问题。汉语的分词，可以看成一个优化问题——在多种分词可能中，按照某种决策规则选取最优的分词结果。图 1 表示了句子“只有为人民工作”的全切分网络：（其中粗线条表示了要求解的路径，即正确的分词结果）

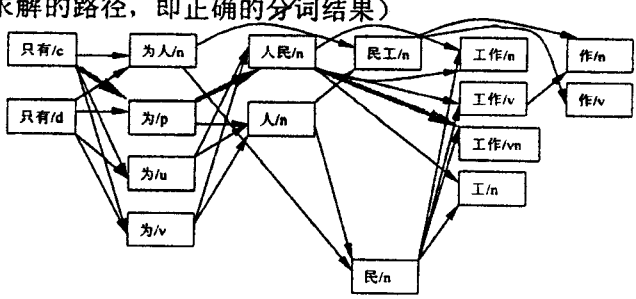

图 1：汉语分词和词性标注网格图

定义待处理文本由 $C = C_1C_2...C_n$ 组成，其中 C_i 是组成句子的汉字字符，可能的分词结果有 K 种，C_i 表示第 k 中可能，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。如果把原来的统计语言模型和最大概率决策综合起来，可以得到分词和词性标注决策规则：

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

其中假设字符C_i的出现概率为常数。假定T^k词性序列中，每个词性标记T_i^k的出现只决定于它前面 n-1 个词性标记，而且W_i^k的出现只和T_i^k有关。

模型参数通过下式从语料中统计计算得到：

$$P(T_B|T_A)=\frac{C(T_AT_B)}{C(T_A)},\qquad P(W_A|T_A)=\frac{C(W_A,T_A)}{C(T_A)},\qquad P(T_c|T_AT_B)=\frac{C(T_AT_BT_c)}{C(T_AT_B)}$$

$C(T_A)$，$C(T_AT_B)$，$C(T_AT_BT_c)$代表词性标记串T_A，T_AT_B，$T_AT_BT_c$在语料中出现的次数。$C(W_A,T_A)$表示词W_A标记为T_A的次数。计算出来的参数数据保存到文件中，用于对模型的测试。其中词性同现概率存到独立的文件中，$P(W_A|T_A)$作为词条属性存到词典中。

通过把出现的事件次数打打折扣，多出来的次数分布到未出现的事件上，由此实现对事件概率的平滑。对于 ngram 参数的平滑，本文采用的算法由 Good_Turing Discounting 和 Backoff 先后阐述过[8]。

2．网络的简化和分词规则的引入

考察全切分网络，可以很容易的发现分词歧义。其中典型的组合歧义为，“他将来上海”中“将来”应该分解为“将”和“来”，而不是象多数情况下一样合成一个词，这是典型的组合歧义。句子“他们上台北去了”中“上台”和“台北”都是词，这是典型的交集歧义。

有关交集歧义的排除方法已有一些论述[9]，本文则在发现交集歧义后，使用一些类似的规则消除局部歧义，从而简化网络。譬如“他的确切菜了”，发现“的确切”为交集歧义字段后，可以查询相关的消歧信息：当“切”后面跟的是“菜”、“肉”等词，歧义字段切分为“的确 菜”。这样就可以不用再利用路径求解消歧了。

对于组合歧义，除了可以采取象交集歧义一样的规则外，本文还采用了基于词性的策略。组合歧义字段既可以看作一个复合词，也可以看成两个词的组合。汉语复合词的规则有陈述式、支配式、补充式、联合式、偏正式等。文[10]详细的研究了汉语两字复合词的构词规则。如果一个组合歧义字段的各个组成部分的词性搭配符合复合词的构词规则，可以认为该组合歧义可以消去。设组合歧义字段为 AB 和其组合部分 A、B，词性分别为T_{AB}、T_A、T_B。实验中采取的词性消歧规则有：

（1）T_A=T_B，认为是联合式复合词。如“人民“，“开关”，“满足”等

（2）T_A=v，T_B=n，认为是支配式复合词。如“主席“，“顶针”等

（3）T_A=a，T_B=n 或 v，认为是偏正式复合词。如“黑板“，“美观”等

（4）T_A=n，T_B=v，$T_{AB}\neq$n，认为是偏正式复合词。如“御用”，“人为”等，而“物理学”则不能消去歧义。（比较：“物理学起来很难”，“物理学很有意思”）

这些规则既消除局部歧义，也简化了路径求解的复杂度。从而能够较大的提高文本处理的运算速度。

三、 韵律节奏的预测

1．韵律节奏预测的一般性问题

文本分析的结果，是得到了分词、注音和词性等基本信息，在此基础上可以获得一定层次上的语法结构，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）[4]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。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。建立韵律预测机制，较为完美的方法当然应该是从语义的分析入手，然而现代文本分析技术还很能从深层次中获取语句或篇章的完整语义信息，即便是这样，直接建立在语法结构到声学参数的韵律建模思路，也是可行的，但由于语法结构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，首先利用文本分析得到的分词、注音和词性等一系列结果，建立语法结构到韵律节奏的模型，包括韵律短语预测、重音预测等。继而，进一步通过重音和韵律短语信息，并结合成统一的语境信息实现韵律声学参数的预测和进行选音的步骤。

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法结构的统计结果，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法信息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。

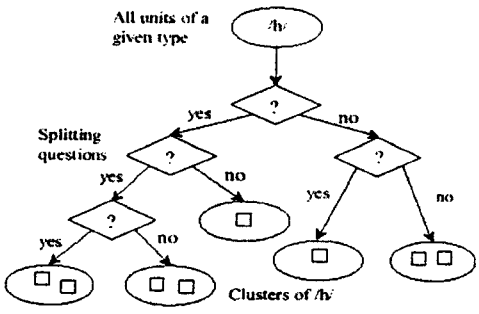

图 2：韵律重音和短语预测决策树示意图

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。

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。

由于汉语语法结构的独特性，一些研究表明，汉语的虚词，如：在、和、的、地、于等，以及一些常用的关键词，如：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。本文通过实验表明，辅以助词分析的决策树模型，获得了 86%（相对于人工标注，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，基本能够满足语音合成系统的韵律处理的要求。

四、 韵律代价函数

1．音节选取中的韵律代价函数

韵律的声学参数与语境参数紧密相连，语境信息由文本分析的结果和韵律节奏（包括韵律短语和重音等信息）等信息组成，它是文本处理和韵律节奏处理的后续的有机组成部分，语境的选择，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。针对汉语，常用的语境信息包括：语音特征，其中有声调、重音、音长等信息；以及语言特征，其中有音节内部信息，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。根据Selkirk在1984[1]年提出的一种韵律分层模型，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。

直接绕过韵律声学参数的预测，而利用从语境信息而设计的韵律代价函数来确定语音的基元选取，正在被目前许多成功的汉语语音合成系统所采用，并取得了非常好的合成效果。其中代价函数为：

$$S=\sum_i \gamma_i V(a_i)，\ 其中，\ \gamma_i = f(\omega_i) \tag{2}$$

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

其中 ω_i 为不同语境参数产生贡献的影响因子或称权值，如何确定函数权值 ω_i 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。

本文介绍一种利用神经网络的权抑制的方法，较为有效而迅速的确定初始权的值。如图3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：

$$\overline{F}(w) = F(w) + \lambda \sum_{\{k|\omega_k \in w_{sel}\}} \omega_k^2 \tag{5}$$

$$\overline{w}^{i+1} = \overline{w}^{i} - \eta \nabla \overline{F}(w) = \overline{w}^{i} - \nabla \left[\eta F(w) + \eta\lambda \sum_{\{k|\omega_k \in w_{sel}\}} \omega_k^2 \right] \tag{6}$$

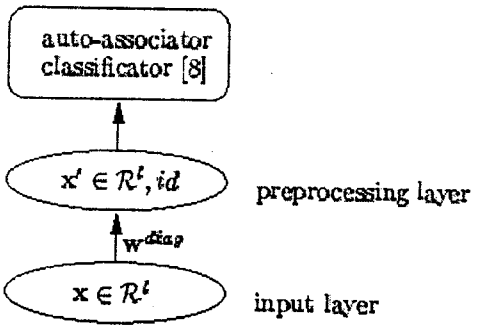

图3：具有权抑制层的神经网络模型

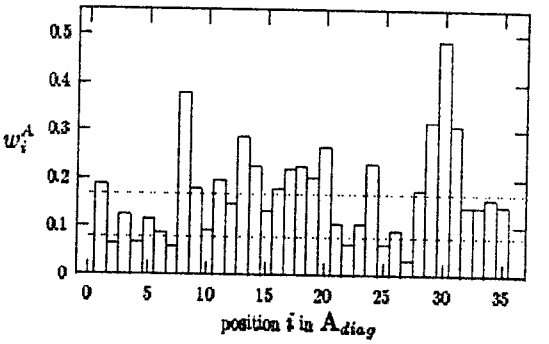

图4：权值抑制的训练结果

图4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，占据着较为重音的作用。

2．韵律特征间的耦合效应对基元选取的影响

通过前面的论述，语境信息到韵律特征存在者映射关系，然而这种关系是确定的吗？其实“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。” [4]

由此，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：

$$\hat{Y} = \arg\max_{n} P(\vec{Y}_n | \vec{A}) = \arg\max_{n} \frac{P(\vec{A} | \vec{Y}_n) P(\vec{Y}_n)}{P(\vec{A})} = \arg\max_{n} P(\vec{A} | \vec{Y}_n) P(\vec{Y}_n) \tag{3}$$

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

则 $\vec{Y}$ 可以表示成 $\vec{A}$ 的函数形式，即：

$$Y = \varphi(A)$$

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”[3]，以及林茂灿先生的“普通话轻声与轻重音”[6]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，当只考虑相邻音节的韵律特征关系时，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。

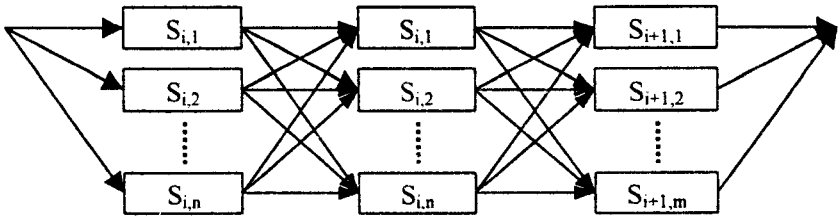

图 5：考虑韵律转移概率后的韵律代价函数在基元选取时的示意图

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，将会起到较大的作用。

五、 总结

本文总结了汉语语音合成系统的主要核心框架中文本处理和韵律处理部分，并着重阐述了它们之间在语音合成系统中的有机结合。由于采用韵律代价函数，并辅以大规模语料设计技术，已经使得语音合成的质量在很大程度上接近自然人说话。事实上，采取韵律代价函数使得研究和设计人员在韵律和语音学领域的研究水平，对系统设计的成败将产生较大的影响，尤其是这种方法将较大的依赖于语料的设计，同时语料也变得非常庞大。随着语音合成系统的进一步发展，无论是语音合成系统本身，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，为了实现特色语音合成以及说话人模拟等一系列的研究目标，韵律模型中韵律声学参数的预测和语音合成算法的研究，反而会变得越来越重要。语音合成系统正朝着概念到语音(Concept to Speech—CTS)和意念到语音(Intention to Speech—ITS)方向发展，而这些目标的实现，与文本的分析能力，深层次的语法结构和语义信息的获取能力，紧密联系在一起。随着新的技术和算法的不断引入，语音合成中的文本分析也越来越朝着分析参数多样化和复杂化的方向发展，其词性划分也越来越细，同时语法结构的深层次分析也将会对语音韵律处理中语境参数的组成产生重大的影响，从而整个韵律模型也会作相应的调整，并为韵律预测，和语音合成研究的总体飞跃带来更大的契机。

参考文献

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

[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，第二期，1985

[4] 沈炯，汉语语调模型议，语文研究，1992，VOL 4，P16-24

[5] 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

[6] 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。

[7] 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,P67-72

[8] Katz,S.M(1987),Estimation of probabilities from sparse data for the language model component of a spech recognizer. IEEE transactions on ASSP,35,400-401

[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，Vol 39, No.5

[10] 王政红，论双音复合词的构成格式，南京理工大学学报，Vol10，6，1997

[11] 黄昌宁，中文信息处理中的分词问题，语言文字应用 1997 年第 1 期(总第 21 期)

[12] Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-sate word-segmentation algorithm for Chinese. Computational Linguistics, 22(3).

[13] David Palmer, 1997.A Trainable Rule-Based Algorithm for Word Segmentation Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL '97), Madrid, 1997.