第25卷 第3期
1999年5月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 25, No. 3
May, 1999

语音识别中统计与规则结合的语言模型[1)]

王 轩 王晓龙 张 凯

(哈尔滨工业大学计算机系 哈尔滨 150006)

摘 要 在分析语音识别系统中,基于规则方法和统计方法的语言模型,提出了一种对规则进行量化的合成语言模型.该模型既避免了规则方法无法适应大规模真实文本处理的缺点,同时也提高了统计模型处理远距离约束关系和语言递归现象的能力.合成语言模型使涵盖6万词条的非特定人孤立词的语音识别系统的准确率比单独使用词的 TRIGRAM 模型提高了4.9%(男声)和3.5%(女声).

关键词 语音识别,统计语言模型,马尔可夫模型,词网格.

LANGUAGE MODEL FOR SPEECH RECOGNITION APPLICATIONS

WANG Xuan WANG Xiaolong ZHANG Kai

(*Dept. of Computer Science and Engineering, Harbin Institute of Technology, Harbin* 150006)

Abstract In this paper a hybrid language model integrating rule-base grammar and Markov language model for speech recognition applications is described. This hybrid language model not only avoids the disadvantage of rule-base grammar in processing very large real text but also has a good performance in processing Chinese language recursive nature and long distance constrained relations, which has been applied to large vocabulary isolated work speech recognition. The male voice recognition accuracy is improved from 81.7% with Trigram language model to 86.6%, the female recognition accuracy from 87.7% to 91.2%.

Key words Speech recognition, statistical language model, Markov model, word lattice.

1 问题的提出

语音作为一种理想的人机通讯方式具有自然、方便、快速的特点.让机器能够理解人

1) 国家"八六三"高技术计划和霍英东基金资助课题.
收稿日期 1996-11-11 收修改稿日期 1997-07-15

类的语音一直是人们追求的理想.对语音识别的研究一直成为计算机科学、语音学、语言学及神经生理学研究的热点,在国内外均作为高技术重要课题加以研究.

自从70年代美国 DARPA 计划支持的语音理解系统(SUS)研究以来,世界上很多研究机构相继推出了各自的语音识别系统[4].目前的语音识别系统大都由声音信号的识别(acoustic recognition)和声音信号的理解(speech understanding)两部分组成.前者是利用语音的声学模型,把自然的声音信号转换为机器可以处理的数字表达的音节形式,可以采用的方法有矢量量化(VQ)、隐马尔可夫模型(HMM)等.声音信号的理解是利用语言模型(language model)的各种知识(如词法、句法、语义、语用和语言的统计信息)对声音进行语言解码.语言模型主要有两种:一种是传统的规则文法,称为基于规则的(rule-based)语言模型,其特点是对封闭语料处理准确,能够反映语言的远距离约束关系和递归现象,但却无法适应开放语料,鲁棒性差,知识表达的一致性、可维护性不好.近年来,另一种基于大规模真实语料的统计方法随着计算语言学的兴起得到了广泛应用,它把语言看成一个 Markov 信源,语句则是由此信源产生的文字序列,其特点是数据准备一致性好,鲁棒性强,适合处理大规模真实语料,但其只能反映语言的紧邻约束关系,对语言的递归现象无能为力.统计语言模型已被大多数语音识别系统所采用,并取得了较好的识别结果[1,3,4,9].

统计方法和规则方法各自存在着优点和不足,致使单独使用某一种方法都不会得到很高的识别率.但两种方法具有很强的互补性,因此,若把两者结合起来将会得到更好的结果[10].两者结合的关键是把规则进行量化,这样语法规则的运用可以由参数 λ_{rule} 控制,统计模型由参数 $\lambda_{\text{statistic}}$ 控制,只要找到它们之间的拟合函数 $f(\lambda_{\text{rule}},\lambda_{\text{statistic}})$,则混合系统就可由该拟和函数控制.本文正是基于这种思想提出了一种规则与统计相结合的语言模型,并把它应用到大词表非特定人语音识别系统中,取得了很好的识别效果.

2 语音理解系统

设一声学发现矢量序列 $\underline{A}=a_1a_2\cdots a_T$.依据汉语的声学模型可采用某种分割 $\underline{A}=A^{(1)}A^{(2)}\cdots A^{(n)}$,其中每一个声学子段 $A^{(i)}$ 可产生多个候选词 $W_1^{(i)}W_2^{(i)}\cdots W_{N_i}^{(i)}$,其声学发现的相似度可用 $p(A^{(i)}/W_n^{(i)})$ 表示,$n=1,\cdots,N_i$.当对 $\underline{A}$ 存在多种分割时,所得到的候选词可表示成词的时间序列网格(word lattice).网格中每一个节点包含起始时刻 begin、终止时刻 end、所对应的音节串 phonemes 和汉字串 token-string.所有的节点按终止时刻排序,对于任意两个节点 $W_i,W_j(i<j)$ 相当于 $\text{end}(W_i)<=\text{end}(W_j)$.$W_i,W_j$ 当满足不存在任何 W_k,使得 $\text{end}(W_i)<=\text{begin}(W_k)$ 且 $\text{begin}(W_j)>=\text{end}(W_k)$时,称 W_i,W_j 首尾相接.一个候选语句定义为从起始时刻到终止时刻一串首尾相连的节点序列$\langle W_1W_2\cdots W_m\rangle$.如图1所示,$W_1$ 和 W_5 首尾相接,$\langle W_1W_5W_9W_{10}\rangle$ 是一个候选语句.

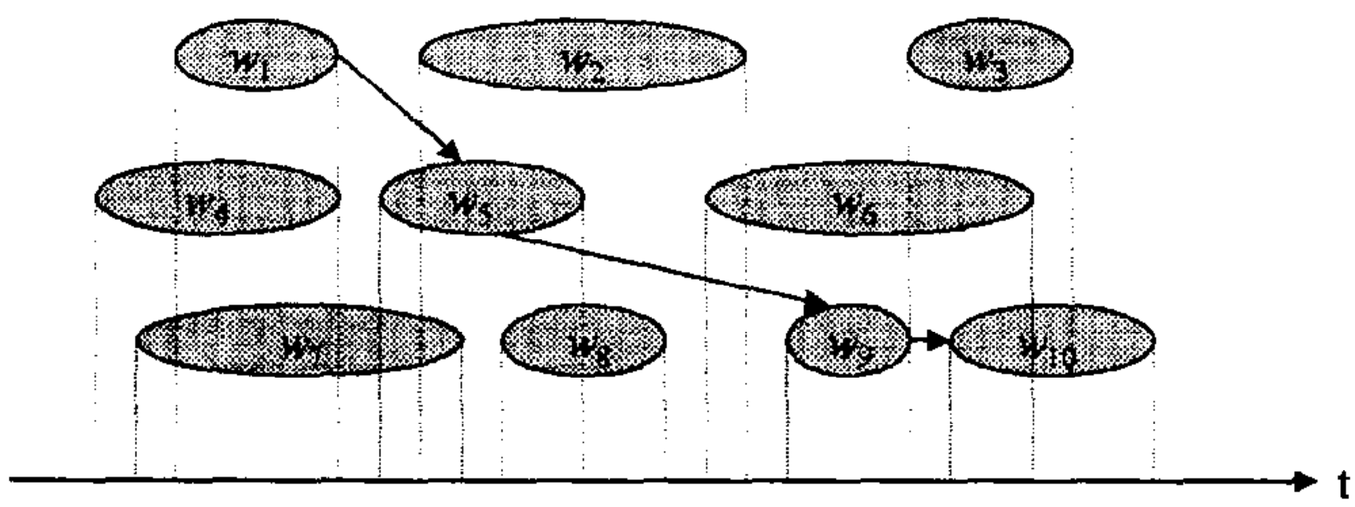

图1 词的时间序列网格示意图

语音识别的任务就是给定声学发现 $\underline{A}$ 后寻找出最可能的词序列 $\underline{W}=W_1W_2\cdots W_n$,$\underline{W}=\arg\max_{\underline{W}} p(\underline{W}/\underline{A})$. 据贝叶斯公式

$$p(\underline{W}/\underline{A})=\frac{p(\underline{A}/\underline{W})\cdot p(\underline{W})}{p(\underline{A})}, \tag{1}$$

其中 $p(\underline{A})$与 $\underline{W}$ 无关,则

$$\underline{W}=\arg\max_{\underline{W}} p(\underline{A}/\underline{W})\cdot p(\underline{W}). \tag{2}$$

由声学模型提供 $p(\underline{A}/\underline{W})$,语言模型提供 $p(\underline{W})$. 本文主要研究后一部分.

根据 Shannon 的信息理论[6],通常可把自然语言看成一个离散的马尔可夫模型,用模型 M 表示. 当前字词 w_i 只与前 $n-1$个词 $w_{i-N},w_{i-N+1},\cdots,w_{i-1}$相关时,则称为 N 元文法模型(N-GRAM),记为 M_N. 若词网格中的一候选语句为 $\underline{w}=w_1w_2\cdots w_k$,则

$$p_{M_N}(\underline{w})=p(w_1w_2\cdots w_k)=\prod_{i=1}^{N-1} p(w_i/w_{i-N}w_{i-N+1}\cdots w_{i-1}). \tag{3}$$

假如词典为 V,则建立 M_N 时需要统计的参数个数为 $|V|^N$. 当 N 较大时参数空间十分庞大,一般取 $N=2$或3,此时的文法分别称为 BIGRAM 文法和 TRIGRAM 文法.

当训练样本空间不足够大时,许多合法语言现象在训练集中未出现,使得很多参数为0,这就是所谓的数据稀疏问题. 为此要对参数进行平滑处理,处理的方法很多[2],本文采用的是删除插值方法[7].

单纯的统计求解方法由于只向前看有限的词,对于常用的二元和三元文法只能向前看一个或两个词,因此对超出此范围的语言约束关系就无能为力了.

3　规则与统计结合的语言模型

传统的语法分析都是进行完整的语法分析,最终都要形成一棵完整的语法分析树. 当语法树不能形成时,语法分析就宣告失败. 其实,在分析失败时分析器已经做了许多局部的正确分析,全局语法分析的失败往往是因为知识库中的规则不足导致的. 图2(a)中由于知识库中包含 rule1,rule2和 rule3,则形成一棵完整语法树. 如果规则库中不包含 rule3,则句法分析在形成图2(b)时因无法进行而失败. 这时形成一个森林,如果能够运用统计关联信息,则仍然可以得到统计意义上最佳的分析结果. 在大规模真实文本的转换中,极大多数情况下不能进行完整的语法分析,图2(b)的出现情况要远远大于图2(a)的情况.

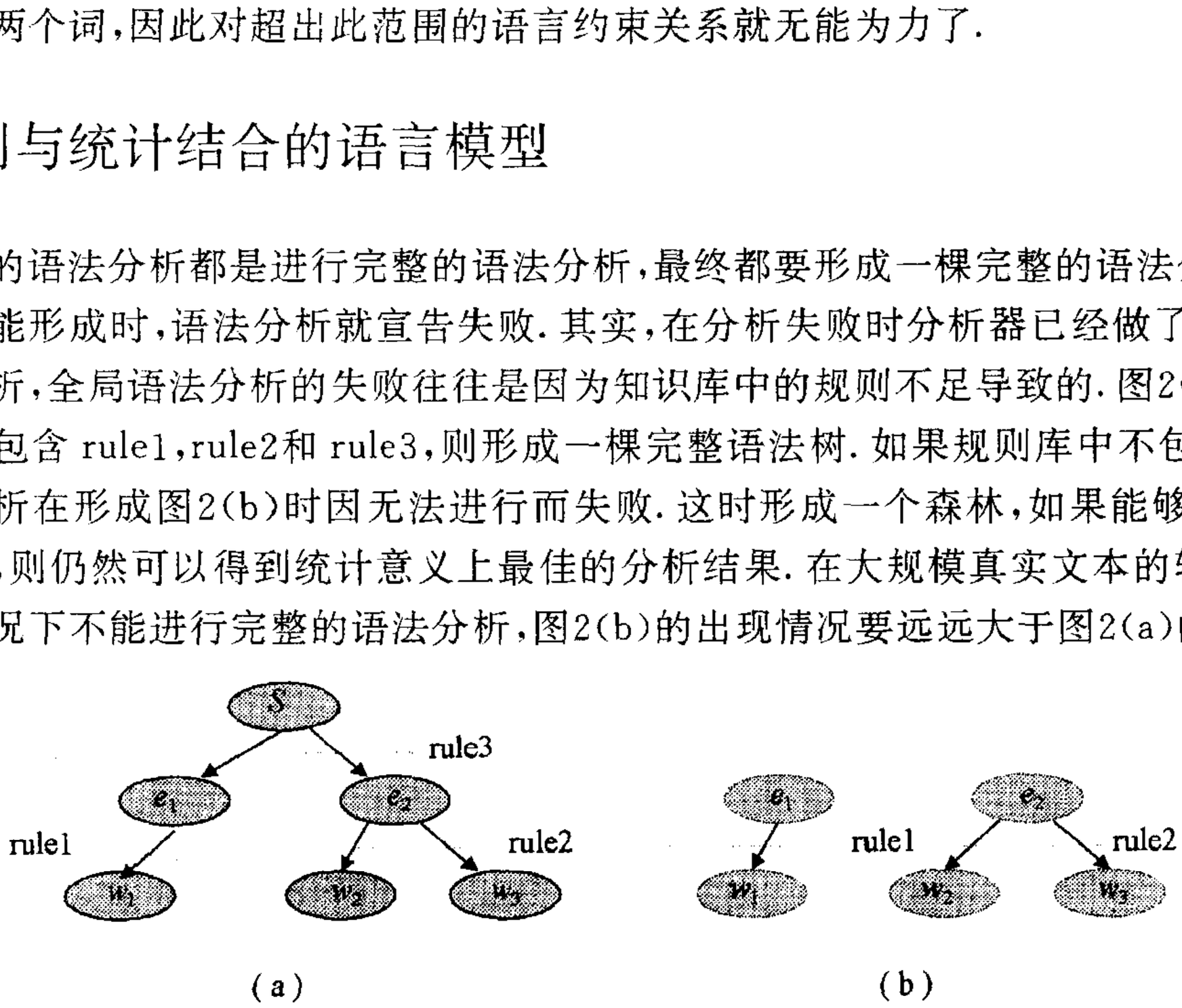

图2　完全和不完全语法分析示意图

规则文法与统计语言模型相结合的方式有两种,一种是在语法分析的过程中加入统计关联信息,指导语法分析的进行;另一种是把语法分析的中间结果表示成词网格中的节

点形式，在求解最优路径时对符合语法规则的路段进行加权处理. 我们采用的是后一种方式. 为了讨论方便，先给出以下定义.

定义1. 所有规则库中可表示成 $A_1+A_2+\cdots+A_t\rightarrow B$ 的规则，称为 t 元规则.

定义2. 一个元素为五元组〈begin，end，$R(t)$，sublist，attr〉，其中 $R(t)$ 为一个 t 元规则 $A_1+A_2+\cdots+A_t\rightarrow B$，sublist 为生成该元素的元素列表 $\{e_1,e_2,\cdots,e_t\}$，$e_i(i=1,\cdots,t)$ 的属性为 A_i，begin 和 eng 分别为该元素的起始时刻和终止时刻，attr 为该元素的属性，等于 B.

一个元素唯一对应着一颗树，其根节点为其本身，后代节点为生成它的元素以及它们的子孙. 叶节点全部为词典中的词组成，它们的列表记为 wordlist $=\{w_1,w_2,\cdots,w_n\}$. 元素是对完整的语法分析树中子树的一种描述.

定义3. 当词网格中包含元素节点时称为元素网格.

在元素网格中，一个语句候选定义为从起始时刻到终止时刻一串首尾相连的元素节点序列 $s=\langle e_1,e_2,\cdots,e_n\rangle$. 若 wordlist$(e_i)=\{w_{i1},\cdots,w_{it_i}\}$，$i=1,\cdots,n$，则语句候选的评价函数为

$$f(s)=\lambda p(e_1,e_2,\cdots,e_i,\cdots,e_n)=$$
$$\lambda_1 p(e_1)\lambda_2 p(e_2/e_1)\cdots\lambda_n p(e_n/e_1,e_2,\cdots,e_{n-1}).$$

上式中

$$p(e_i/e_1\cdots e_{i-1})=p(w_{i-1,1})\prod_{k=2}^{t_{i-1}}p(w_{i-1,k}/w_{i-1,1}\cdots w_{i-1,k-1})p(w_{i,1}/w_{i-1,1}\cdots w_{i-1,t_{i-1}})\cdot$$
$$p(w_{i,1})\prod_{k=2}^{t_i}p(w_{i,k}/w_{i,1}\cdots w_{i,k-1});$$

λ_i 是规则的调整函数，生成元素 e_i 的规则的可信度越高，λ_i 的取值越大；wordlist(e_i)越大，λ_i 的取值越大；λ 即是规则与统计的拟合参数，从树库语料库中训练得到；现在的系统中，λ_i 的取值与生成元素 e_i 的规则的可信度和 wordlist(e_i)成正比. 其它形式的拟合函数正在实验中. 当统计语言模型为三元模型时，

$$p(e_i/e_1\cdots e_{i-1})=p(w_{i-1,1})p(w_{i-1,2})\prod_{k=3}^{t_{i-1}}p(w_{i-1,k}/w_{i-1,k-2}w_{i-1,k-1})\cdot$$
$$p(w_{i,1}/w_{i-1,t_{i-1}-1}w_{i-1,t_{i-1}})p(w_{i,1})p(w_{i,2})\prod_{k=3}^{t_i}p(w_{i,k}/w_{i,k-2}w_{i,k-1}).$$

规则和统计相结合的语言模型可以处理语言中的远距离搭配关系，如“一只可爱的小花猫”和“一枝可爱的小花”的名词与数量词的修饰关系. 同时也能处理语言中的递归现象，如“一只非常漂亮可爱的小花猫”和“九百八十七万六千五百四十三元两角壹分的支票”等. 对于语法规则未能描述的语言现象，模型退化成单纯的二元、三元统计文法. 在多条语句候选都可得到完整的语法分析树时，统计信息还可以作为选择最优方案的评价标准.

下面通过一个例子“庄严热烈的技术鉴定会正在进行着”，进一步阐述前面的方法.

为简单起见，假设规则库中只包含以下规则

RULE1： $A+$“的”$+N$(或 N_p)$\rightarrow N_p$ [1.3]，

RULE2： $V+$“着”$\rightarrow N_p$ [1.4]，

RULE3:　$N+N\rightarrow N_p$　[0.7].

其中 A 表示形容语,V 表示动词,V_p 表示动词短语,N_p 表示名词短语,N 表示名词. 每一条规则后面括号中的数字为规则的可信度.

经语音识别器得到的音节的时间网格如图3所示,每一音节下面的数字是该音节的声学评价. 经过词典查询得到每一个候选音节的同音汉字候选及其词性,按音节的终止时刻排列形成词的时间网格如图4所示. 这里给出的是孤立词语音识别数据,因此音节间没有交叉,对于连续语音识别就会出现音节交叉现象. 图中标有"*"的节点为由规则推出的元素. 如,"热烈的技术鉴定会"其生成规则为 RULE1和 RULE3;"进行着"其生成规则为 RULE2. 因为在计算元素的权值时要考虑到规则的可信度和元素的长度等因素,所以在用 VITERBI[8]算法或是 A^*[5,7]算法求解最优路径时,经过权值较高的元素的路径的评估值也较高. 本例中求得的最优路径为"庄严热烈的鉴定会正在进行着"(如图5所示).

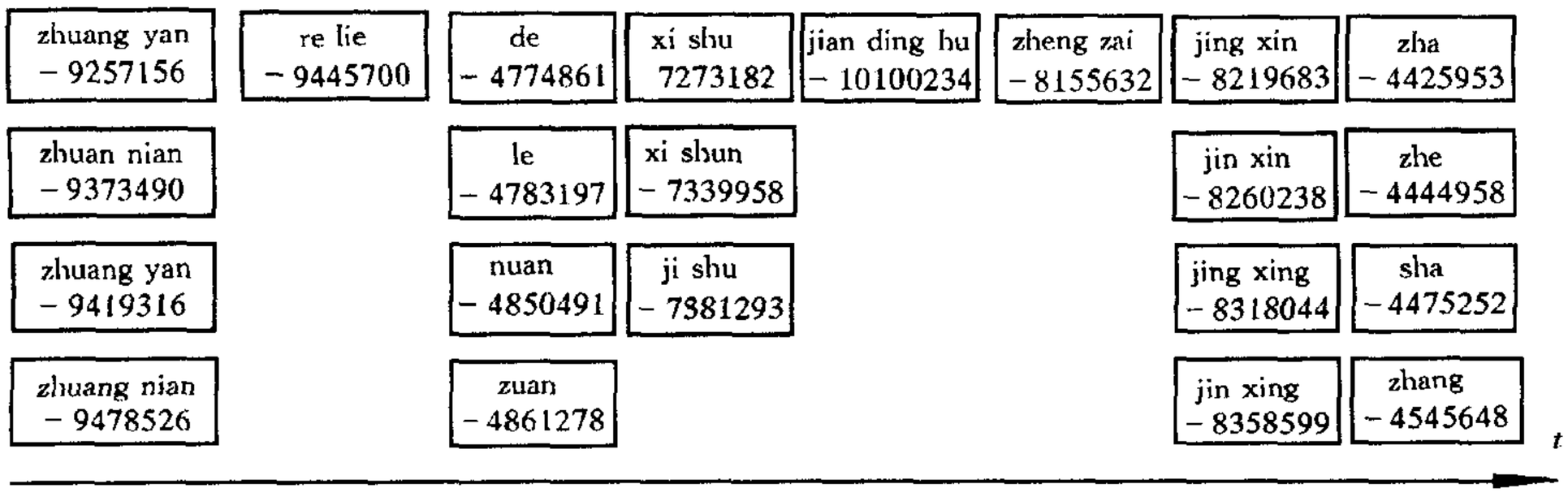

图3　语音识别得到的音节候选

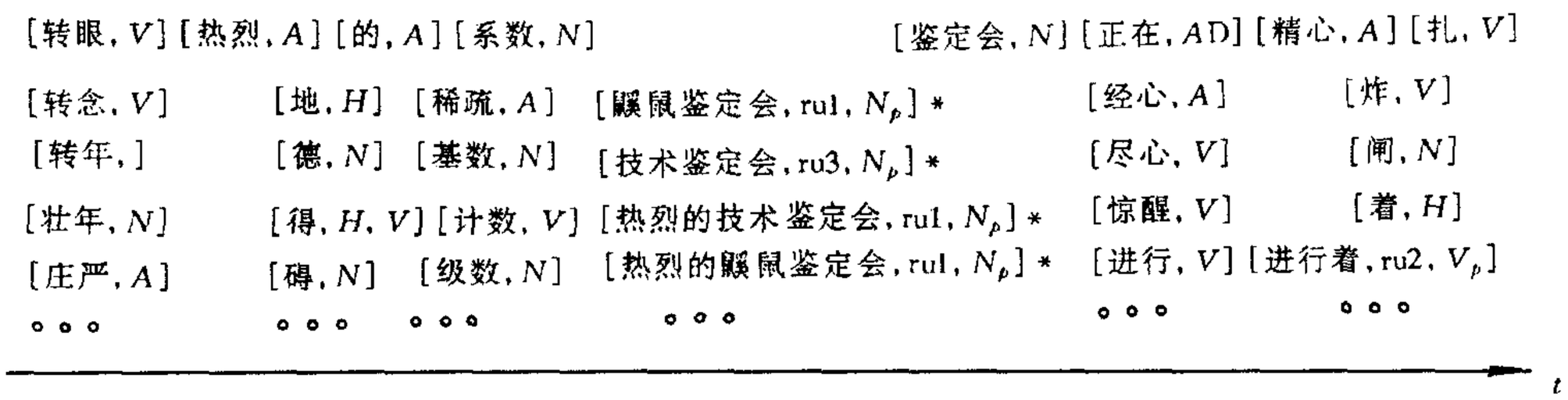

图4　元素时间网格

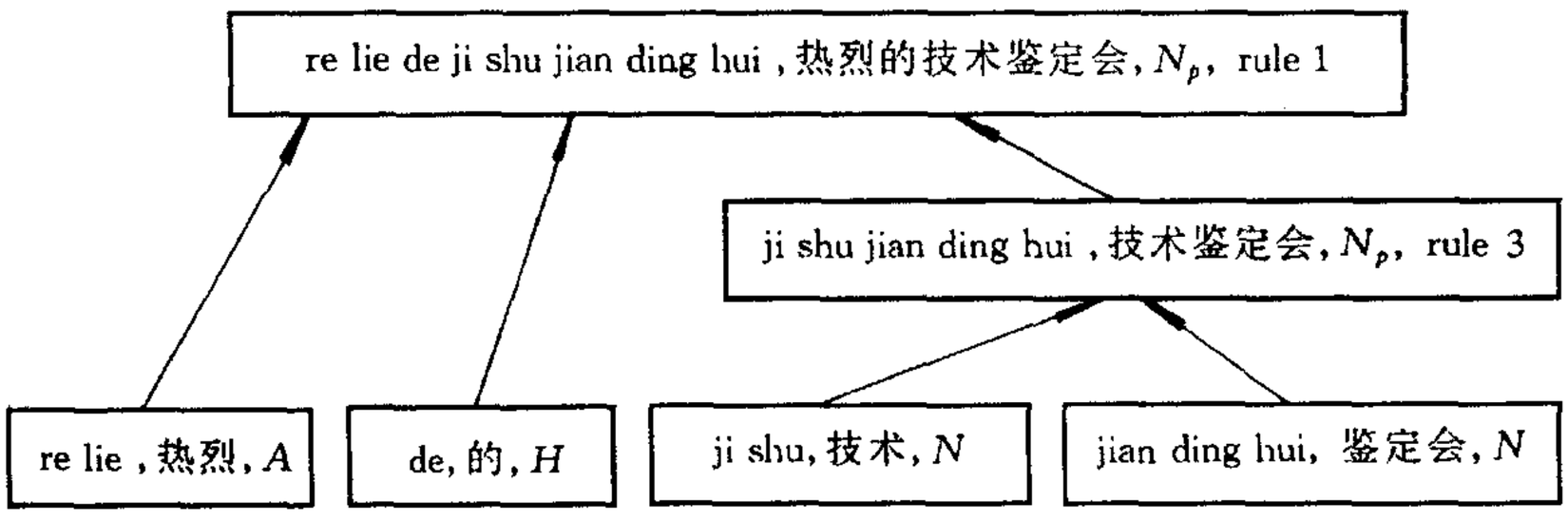

图5　元素对应的语法分析树示例

4　实验与结论

实验中选取了六组男声和六组女声,每组五句话的语音数据进行测试,语音识别采用

CDHMM 模型，统计语言模型采用基于词的 TRIGRAM，统计数据库约20M. 词典收词65 000条，采用1993，1994年《人民日报》作为训练语料. 实验平台为 PC586/120，32M 内存. 在单独使用 TRIGRAM 的情况下，男声平均识别率为81.7%，女声平均识别率为87.7%；加入规则模块后，男声平均识别率为86.6%，女声平均识别率为91.2%. 分别提高了4.9%和3.5%，如表1和表2所示.

表1 TRIGRAM 语言模型下的识别结果

测试人	第一人	第二人	第三人	第四人	第五人	第六人	平 均	时间
女声	93.5%	83.4%	78.1%	93.7%	91%	86.9%	87.7%	1.38h
男声	77.5%	88.3%	85.2%	74.8%	83.3%	81.5%	81.7%	1.33h

表2 TRIGRAM+RULE 语言模型的识别结果

测试人	第一人	第二人	第三人	第四人	第五人	第六人	平 均	时间
女声	94.7%	88.2%	85.4%	94.7%	94.1%	90.3%	91.2%	1.38h
男声	82.2%	91.5%	87.6%	81.3%	89.7%	87.6%	86.6%	1.73h

实验结果表明，混合语言模型有效地把规则文法处理语言中长距离的约束关系和递归现象的能力嵌入到统计语言模型中，使其具有比单纯的统计语言模型和规则文法更强的约束能力. 目前，基于规则和统计相结合的语言模型已经应用于大词表非特定人语音识别与理解系统中，取得了比较好的识别效果.

参 考 文 献

1 Jelinek F. The development of an experimental discrete dictation recognizer. In: Proceedings of the IEEE, 1973, 1616—1624

2 Nadas A. On turing's formula for word probabilities, *IEEE Trans. Acoustic, Speech* and *Signal Processing*, 1985, **33**: 1414—1416

3 Derouault A M, Mermelstein P. Fast search strategy in a large vocabulary word recognizer. *Journal of the Acoustical Society of America*, 1988, **84**: 2007—2017

4 Lee K F, Hon H W, Reddy R. An overview of the SPHINX speech recognition system. *IEEE Trans. Acoustics, Speech and Signal Processing*, 1990, **38**: 35—44

5 Nagata M. A stochastic Japanese morphological analyzer using a forward-DP backward A* N-Best search algorithm. In: Proceedings of 15th International Conference on Computational Linguistics, Japan Kyoto: 1994, 201—207

6 Shannon C. Prediction and entropy of printed english. *Bell System Technical Journal*, 1951, **30**: 50—64

7 Jelinek F. Self-organized language modeling for speech recognition. In: IEEE ICASSP'89, 1989, 587—595

8 Viterbi A J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transaction on Information Theory*, 1967, **13**: 260—269

9 Rohlicek J R, Chow Y L, Roucos S. Statistical language modeling using a small corpus from an application domain. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, New York: 1988, 267—270

10 Chien Lee-Feng, Chen Keh-Jiann, Lee Lin-Shan. A best-first language processing model integrating the unification for speech recognition applications. *IEEE Transactions on Speech and Audio Processing*, 1993, **1**(2): 221—240

王 轩 1969年出生. 获哈尔滨工业大学计算机系硕士学位，现为哈尔滨工业大学计算机系博士研究生，研究方向为语音识别、中文信息处理、自然语言理解.

王晓龙　1955年出生.哈尔滨工业大学计算系教授、博士生导师,香港理工大学高级研究员.主要研究方向为人工智能、中文信息处理、语音文字识别.

第三届全球华人智能控制与智能自动化大会(CWC ICIA'2000)征文通知

第三届全球华人智能控制与智能自动化大会将于2000年6月28日至7月2日在中国科学技术大学召开,现开始征文.

征文范围包括:·智能系统与专家系统·智能控制·神经网络与模糊控制·机器人学与机器人控制·大系统·调度、规划、管理与决策系统·自治、容错和故障诊断系统·制造系统和 DEDS·计算机辅助设计·信息处理与信息系统·系统理论和控制理论·建模、辨识和估计·自适应控制·变结构控制·非线性系统及其控制·遗传算法·混合动力学系统·H_∞控制和鲁棒控制·最优化与最优控制·分布式计算机控制系统·仪器仪表·各领域中的应用.

请于1999年12月1日前提交3份用中文或英文写成的论文全文复印件,并在第一页附上:文章标题;作者姓名和单位;详细通讯地址(包括电话、传真和电子邮件地址);摘要;3—5个关键字以及所属征文范围.

会议将于2000年1月15日前发出录用通知,接通知后请在2000年3月15日前提交正式激光打印论文.

论文请寄:

安徽合肥 中国科学技术大学自动化系

CWC ICIA'2000秘书处　丛爽　博士

联系电话:0551—3601513　传真:0551—3603244

联系人:薛美盛　　邮编:230027

Dr. Cong Shuang

The secretariat of the CWC ICIA'2000

c/o Department of Automation

University of Science & Technology of China

Hefei, Anhui 230027, P. R. China

Tel:+86-551-3601513　　Fax:+86-551-3603244

E-mail:cwcicia@ustc.edu.cn

http://cwcicia.ustc.edu.cn

会议将组织代表游览黄山、九华山.