# 基于拼音统计的汉语语音理解方法

吴军　王作英　任岩松
清华大学　电子工程系（100084）

## 摘　要

语音识别系统的识别率在达到一定水平后，不通过语音理解便很难进一步提高了。本文提出了若干基于拼音的统计语言模型并用于汉语语音理解，克服了以往基于规则方法难以理解真实文本的不足。使用这种方法后，可以减少2/3的拼音识别错误，同时提高语音到汉字的识别速度，促进了汉语语音识别的实用化进程。

关键字：语音识别　语音理解　马尔可夫链　N元文法模型

## 1. 介绍

语音是人类互相传递信息的最自然和方便的形式。它不仅传递信息的速度快，而且最符合人们的习惯。因此，语音输入被视为当今第五代计算机和多媒体计算机的重要特征，是今后信息输入的主要发展方向之一，而实现它的关键技术是语音识别技术。

对汉语来说，汉字的输入问题一直是困扰我国计算机发展和普及的一大问题。由于汉语不是拼音文字，因此，任何一种键盘输入方法都难以解决提高速度和降低训练难度的矛盾。况且，用目前这些方法输入汉字时，由于精力集中在分解字上，实际难以脱稿输入。使用语音输入，则是克服上述困难的最好办法。因此，语音识别对汉语有特别重要的意义。

语音的不确定性和不稳定性使得语音识别十分困难。虽然近十年来国内外在这方面有了很大的进展，研制了一些较好的语音识别系统[1-3]，但识别率和识别速度仍未达到商品化要求。而且，语音的声学识别率达到一定水平后，通过改进声学模型再想提高即使是一个百分点都是很困难的，因此必须依靠上下文进行语音理解。

汉语由于存在一音多字的问题，对汉字的语音识别要经过单音识别和音字转换两步完成。因此语音理解可以在汉字一级进行也可以在拼音一级进行。以往的汉语语音识别系统，都是在汉字一级进行的。这主要原因是：(1) 以往自然语言理解的研究都是在汉字一级进行的；(2) 一般人认为拼音不具有明确的含义，汉字才有。在汉字一级的语音理解是在把拼音转换为汉字时，通过考察每个待识别音的很多候选字，并利用文法规则和自然语言的一些统计规律，纠正一些单音识别的错误。从效果来看，这种办法在正确率上，尤其是在速度上难以达到实用化要求。其主要原因是：

(1) 在目前单音识别率下（以后也不会有很大的提高），需很多的候选音，如THED-919系统目前要六个候选音，累计正确率才能接近100%。而汉语是一音多字，平均每个拼音对应5.87个二级国标汉字，这样一个待识别语音就要对应三十多个汉字，如果上下文的各种情况考虑得稍微多些，计算量就会大到一般微机上难以实时实现的程度，因此只能作很简单的处理，很多错误就无法纠正。即使这样，实时实现也还很勉强。

(2) 即使不考虑计算时间，由于每个待识别语音的候选字太多，会出现几种组合同时合乎文法的情况，反而引入一些转换错误。

事实上，尽管拼音不如汉字具有明确的含义，但由于汉语中二字以上的同音词比例是很低的，从纠正语音识别错误的角度来看，根本无须转换成汉字再进行理解。因此，我们提出在单音识别和音字转换两部分之间，加入拼音一级进行语音理解的思想。通过它可纠正大部分拼音错误，最后只需很少的候选，累计正确率就能接近100%。这样即使音字转换部分不变，也能提高到文字的识别率，同时大大降低计算量。从长远看，由于拼音候选数目减少，才有时间考虑更多的信息解决好一音多字的转换问题，使文字的正确率进一步提高，从根本上解决语音识别实用化的问题。

在语音理解过程中，我们采用了新的基于统计的语音理解方法。它通过对大量语料的统计，得到反映语言特点的统计信息，然后用这些信息进行理解。它可处理大规模真实文本，克服了传统的基于规则方法只对规则集内句子有效的缺陷，并且速度要快得多。

本文的第二部分介绍统计语言模型，这是用统计的方法进行语音理解的核心。第三部分介绍如何解决模型实现过程中的各种技术问题。第四部分介绍利用THED-919汉语语音识别系统提供的环境和数据进行的实验以及结果。第五部分是结束语。

## 2. 统计语言模型

每一种语言包括汉语都有很多的统计规律，用基于统计的方法进行语音理解关键是要用一种简单有效的数学模型来描述自然语言的统计规律。这种数学模型称为统计语言模型。

### 2.1 统计语音识别的方法

用统计的方法识别一句话的语音，就是要找出与输入的待识别语音序列A最相符一个候选拼音句子S。根据最大后验概率准则有:

$$S = \underset{j}{\mathrm{Argmax}}\ P(S^{(j)}|A) \tag{1}$$

其中j为候选句子的序号。

根据Bayes公式，并利用P(A)与j无关性，可推得:

$$S = \underset{j}{\mathrm{Argmax}}\ P(A|S^{(j)})\cdot P(S^{(j)}) \tag{1'}$$

其中估计$P(A|S^{(j)})$的模型称为声学模型，本文不作讨论。计算先验概率$P(S^{(j)})$的模型称为语言模型，是本文研究的中心。一个语音识别系统没有加入语音理解时，实际上是假设每句话$S^{(j)}$的先验概率相同，即使用的是最大似然准则:

$$S = \underset{j}{\mathrm{Argmax}}\ P(A|S^{(j)}) \tag{2}$$

相比公式(1)它舍去了$P(S^{(j)})$的作用。而实际上不同的$S^{(j)}$的先验概率相差是极大的，因此它的作用不应忽略。用统计的方法进行语音理解就是要设法求出它来，并用它来重新计算各候选的概率。

### 2.2 N元文法模型[4]

为了计算$P(S^{(j)})$，假设在公式(1)中:

$S^{(j)} = (W_1^{(j)}, W_2^{(j)}, \ldots, W_M^{(j)})$，

$A = (A_1, A_2, \ldots, A_M)$，

其中i=1, 2, ..., M为时间序列号，j=1, 2, ..., L为候选序列号，$W_i^{(j)}$为某个拼音，$A_i$为某个待识别的语音。

根据条件概率的公式，有:

$$P(S^{(j)}) = P(W_1^{(j)}, W_2^{(j)}, \ldots, W_M^{(j)})$$

$$= \prod_{i=1}^{M} P(W_i^{(j)}|W_1^{(j)}, \ldots, W_{i-1}^{(j)}) \tag{3}$$

在实际语音理解时，用公式(3)计算$P(S^{(j)})$是不可能的，因为它的复杂度为$O(L^M)$。当M很大时，这是一个NP问题。但在实际遇到的语句中，每个字只和相邻的几个字关系较密切，而和较远的关联较小。因此，我们可以假设自然语言是一个马尔可夫链，即一句话中的每个音$W_i$只和前面N-1个有限的音有关，和更前面的无关。我们假定每一句话前面存在N-1个空音，这时$P(S^{(j)})$的计算就可以用下面的公式进行:

$$P(S^{(j)}) = \prod_{i=1}^{M} P(W_i^{(j)}|W_{i-N+1}^{(j)}, \ldots, W_{i-1}^{(j)}) \tag{4}$$

我们把$W_{i-N+1}^{(j)}, \ldots, W_{i-1}^{(j)}$称为$W_i$的上下文，简记为$C_i$。此时，公式(4)可写为:

$$P(S^{(j)}) = \prod_{i=1}^{M} P(W_i^{(j)}|C_i) \tag{4'}$$

它可用Viterbi算法快速地求出来。

这种利用一句话中待识别语音的前N-1个已经确定的音来推测当前这个音的马尔可夫模型又称为N元文法模型(N-Gram Model)。

## 3. 模型的实现[5]

### 3.1 THED-919汉语语音识别与理解系统

本文的工作是利用THED-919汉语语音识别与理解系统的环境进行的。该系统是针对特定人的、孤立语音识别与理解系统，目前包括单音识别和音字转换两部分。其中单音识别部分从话筒输入的语音中识别出拼音，并将每个待识别语音的前六选拼音送入音字转换部分。音字转换部分根据词典、语法规则和统计信息，结合上下文，从六选对应的所有汉字中找出最可能的。目前单音识别部分的首选正确率可达90%。音字转换部分对正确的拼音(首选完全正确)的转换率可达97%；在首选正确率为90%时的转换率降到93%，且计算时间要延长很多。这部分目前在时间和空间上开销已很大，使更细致的处理无法进行。为此，我们在现有两部分之间嵌入拼音理解部分，目的是使：(1) 首选(也包括其它各选)正确率大大提高。(2) 所需候选数目减小平均两个以内。这样将明显提高系统的识别率和速度。另外，也为今后作非特定人的和连续语音的系统作准备。

### 3.2 N元文法模型的简化实现

在识别到汉字的语音识别系统如THED-919系统中，由于只有给出了确定的拼音后，才能开始进行音字转换，因此需要及时地给出语音理解后的结果，至少延时不能太长，否则就无法及时进行汉字转换，实时性得不到保障。而利用公式(4)计算一句话$S^{(j)}$的概率要等到这句话讲完才能得到，不能做到每读入一个音，就给音字转换部分一个拼音。因此我们实际使用时对这种模型进行简化，使它能用于THED-919这样的系统，并达到实时的要求。

假设我们要识别一句话中的第i个待识别音$A_i$，它的上下文$C_i$，则按下公式计算$W_i$：

$$W_i = \underset{j}{\mathrm{Argmax}}\ P(W_i^{(j)} | A_i, C_i)$$

$$= \underset{j}{\mathrm{Argmax}}\ P(Ai | W_i^{(j)}, C_i) * P(W_i^{(j)} | C_i) / P(A_i | C_i) \qquad (5)$$

其中$A_i$，$C_i$与j无关，故：

$$W_i = \underset{j}{\mathrm{ArgMax}}\ P(A_i | W_i^{(j)}) * P(W_i^{(j)} | C_i) \qquad (5')$$

式中的因子$P(W_i^{(j)} | C_i)$体现了语言模型的作用。

用公式(5')直接输出理解后的语音，而不考虑一句话后面对前面的影响，就不必等这句话说完才确定每个拼音，可保证实时性。这样得到的$W_i$虽然是局部最优解，但从实验结果看，接近全局最优解，而且拼音理解可以输出不止一个候选，在音字转换时再对整句话进行搜索求出最佳解，还能纠正一些错误。这种安排在识别到汉字时从时间和效果上综合考虑，无疑是最佳的。

公式(5')的这种模型由于是从N元文法模型中简化来的，我们因此称之为简化的N元文法模型。在不会混淆时我们仍称它为N元文法模型。由于N元文法模型的时间和空间复杂度都是$o(L^N)$，因此N不能太大，一般为2或3。此时的N元文法模型又专门地称为二元文法模型(Bigram Model)和三元文法模型(Trigram Model)，同时把不考虑上下文，只考虑每个音本身概率的模型称为一元文法模型(Unigram Model)。我们建立了基于拼音的二元文法模型、三元文法模型并针对汉语特点提出了前向—后向的二元文法模型。分别介绍如下：

#### 3.2.1 二元文法模型

二元文法模型是根据前一个语音来预测当前待识别的语音，即$C_i=W_{i-1}$。此时：

$$W_i = \underset{j}{\mathrm{Argmax}}\ P(A_i | W_i^{(j)}) * P(W_i^{(j)} | W_{i-1}) \qquad (6)$$

二元文法模型简便、高效、无延时、易实现，但是利用的上下文信息较为有限。作为它的改进，我们提出了一种新的文法模型，称为"前向—后向的二元文法模型"。

#### 3.2.2 前向—后向的二元文法模型

前向—后向的二元文法模型是根据一个音$W_i$的前后各一个音$W_{i-1}$和$W_{i+1}$来确定当前的音$W_i$，故称为前向—后向的。根据汉语一、二字词占大多数的特点，我们假定$W_{i-2}$和$W_{i+1}$是无关的，此时：

$$W_i = \underset{j}{\mathrm{Argmax}}\ P(A_i|W_i^{(j)}) * P(W_i^{(j)}|W_{i-1}, W_{i+1}) \quad (7)$$

其中

$$P(W_i^{(j)}|W_{i-1}, W_{i+1}) = P(W_i^{(j)}|W_{i-1}) * P(W_{i+1}|W_i^{(j)}) / P(W_{i+1}) \quad (8)$$

上面的推导中利用了$W_{i-1}$和$W_{i+1}$的无关性。又因为$P(W_{i+1})$与j无关，公式5可变为:

$$W_i = \underset{j}{\mathrm{Argmax}}\ P(A_i|W_i^{(j)}) * P(W_i^{(j)}|W_{i-1}) * P(W_{i+1}|W_i^{(j)}) \quad (7')$$

前向—后向的二元文法模型要比一般的二元文法模型更有效，而且可完全利用原二元文法模型的参数，因此很好实现。虽然它有一个音的延时，但在实现时先用其向前的因子$P(W_i^{(j)}|W_{i-1})$进行理解，给出一个结果。在下一个音进来后，再利用向后因子$P(W_{i+1}|W_i^{(j)})$进行补充理解，只有两个结果不同时，才回过来修改前一个音，便可基本上消除延时的影响。

### 3.2.3 三元文法模型

三元文法模型是根据前两个语音来预测当前待识别的语音，即$C_i=W_{i-2}, W_{i-1}$。此时:

$$W_i = \underset{j}{\mathrm{Argmax}}\ P(A_i|W_i^{(j)}) * P(W_i^{(j)}|W_{i-2}, W_{i-1}) \quad (9)$$

三元文法模型相比前向—后向的二元文法模型，也是利用两个音对当前的拼音进行理解，由于$W_{i-2}$对$W_i$的约束力不如$W_{i+1}$，故理解效果略差，但它没有延时。

### 3.3 模型的训练

建立N元文法模型的关键在于获得$P(W_i|C_i)$，它是通过大量统计得到的。在统计过程中唯一要注意的问题是零概率问题，我们以二元文法模型为例说明如何解决它。二元文法模型的参数$P(W_i|W_{i-1})$可按下公式近似计算:

$$P(W_i|W_{i-1}) = f(W_i|W_{i-1}) = N(W_{i-1}, W_i)/N(W_{i-1}) \quad (10)$$

其中$N(W_{i-1}, W_i)$是语音对$(W_{i-1}, W_i)$在统计文本中出现的次数，$N(W_{i-1})$是音$W_{i-1}$出现的次数。不幸的是即使对很大量的语料进行统计，对于许多语音对来说，出现的次数仍是零，然而有些这样的语音对在识别时恰恰会遇到，这就是所谓的零概率问题，需要进行平滑处理。我们采用了删除插值的方法(Deleted Interpolation)[4]，将模型的统计公式改进为:

$$P(W_i|W_{i-1}) = \lambda_2 * f(W_i|W_{i-1}) + \lambda_1 * P(W_i) \quad (10')$$
$$\lambda_1 + \lambda_2 = 1,\ 0 \leqslant \lambda_1, \lambda_2 \leqslant 1$$

得到插值二元文法模型。

其中$\lambda_1$和$\lambda_2$分别为一元文法和单纯二元文法在插值模型中的比重，可以通过实验得到。

同样，三元文法模型用一元和二元文法模型参数进行了平滑处理。

## 4. 实验及结果

本文工作所用的训练语料是新华社1990—1991年的二百万字的新闻稿。测试的语料均在训练集以外，包括七届人大五次会议的政府工作报告和一些新闻稿（包括经济、政治、外交、体育和生活等方面的短讯和一些较长的介绍性文章），共计三万字左右。选取这些文章进行测试，是因为它们的内容覆盖面广，有代表性。单音识别的数据是用我们一个工作人员1992年初在THED-919系统上的识别的结果，那时的识别率低于年底鉴定的。使用识别率稍低的数据目的在于看较坏情况下能理解到什么程度，以便为今后作非特定人的系统作准备。用各种文法模型进行理解，实验结果如下:

| 候 选 数目 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 未加理解 | 86.3 | 95.0 | 97.1 | 97.4 | 97.7 | 98.8 |
| UNIGRAM | 88.4 | 96.9 | 97.6 | 98.1 | 98.6 | 98.8 |
| BIGRAM* | 93.3 | 98.2 | 98.6 | 98.8 | 98.8 | 98.8 |
| FB-BIGRAM* | 95.3 | 98.2 | 98.6 | 98.8 | 98.8 | 98.8 |
| TRIGRAM* | 94.6 | 98.2 | 98.6 | 98.8 | 98.8 | 98.8 |

* 由于未较正的拼音数据只有六个候选音，校正后达到未校正的第六候选的概率之后，以后的候选字概率不会再提高。

表2. 几种模型的校正结果(%)
(对二元文法模型 $\lambda_2$=0.3-0.7，对三元文法模型 $\lambda_3$=0.2-0.8)

从表1 中可以看出，用二元文法模型、前向—后向的二元文法模型和三元文法模型进行理解，首选正确率分别提高了7%、9%和8.3%，能减少51%、66%和61%的首选错误， 比简单的一元文法模型优越得多，而前向—后向的二元文法模型比传统的二元文法模型又有明显的优越性，三元文法模型接近前向—后向的二元文法模型的水平，但不存在延时，两者各有千秋。本文实验是在20MHz的386微机上进行的，用C语言实现。理解速度分别为每秒100、70和30个拼音左右。

我们已将二元文法模型加入到THED-919系统中，由于理解后首选正确率很高，当理解后首选概率较大时，我们向音字转换部分只提供一个候选，只有在首选正确率较小时，才提供2-4 个候选，这样平均只要1.7个候选就能达到98.4%的累计正确率。此时系统到文字的正确率如下：

| | 拼音首选正确率 | 提供的平均候选数 | 累计拼音正确率 | 文字正确率 |
|---|---|---|---|---|
| 加入理解前 | 86.3 | 6 | 98.8 | 91.4 |
| 用BIGRAM理解后 | 93.3 | 1.76 | 98.4 | 93.3 |
| 直接输入正确拼音 | 100 | 1 | 100 | 97 |

表2. 用二元文法模型理解后对文字正确率的影响

从上表中可以看出，二元文法模型理解后，整个系统识别到文字正确率提高了1.9%，相当于减少了22%的文字错误，虽然看上去没有拼音提高明显，这主要原因是受限于音字转换本身的正确率。 另外这样识别每个语音所用时间比未加语音理解前快得多。

## 5. 结束语

从上述实验中，我们得出可以下结论:

(1) 用二元文法模型、前向—后向的二元文法模型和三元文法模型对单音识别结果校正后使首选识别率大大提高，而前向—后向的二元文法模型和三元文法模型作用更为明显。

(2) 用本文的方法和模型进行拼音校正后，平均1.76选时的正确率几乎达到校正前六选的水平。拼音理解本身速度很快，在系统中占的开销几乎可以忽略不计，使用后可大大减少识别每个字的时间。

(3) 使用拼音理解后，识别系统对单音识别率的要求可以降低，对减小和进行非特定人的连续语音的识别都有好处。

由于上述特点，这种拼音层的语音理解对语音识别实用化有特别重要的意义。

我们今后将在这方面进行进一步深入的工作，我们的重点将放在以下几方面:

(1) 进行语料的分类统计，建立专业数据库。

(2) 进行自适应的研究，针对被识别文本的内容，在理解过程中自动调整模型参数[6]。

(3) 增大训练量，尤其是三元文法模型的训练量。

(4) 改进现有的音字转换方法，进一步提高整个系统的识别率。

鸣谢:

在本文的工作中，我们得到教研组陆大金教授的热情关心和支持，得到了清华大学计算机系黄昌宁教授热心的指导和帮助，得到了教研组王侠、肖熙和计天颖同志的热情帮助以及协作单位——中国电子器件工业公司声控部王政贤、马国华等同志的大力协助。在此，我们对他们表示最衷心的感谢。

参考文献:

1. F. Jelinek, The development of an Experimental Discrete Dictation Recognizer IEEE ICASSP'91, pp.587-595, 1991.

2. K.F. Lee, H. Hon, R. Reddy, An Overview of the SPHINX Speech RecognitionSystem, IEEE Trans. on ASSP Vol.38, No.1, pp.35-45, 1990.

3. 赛德919(THED-919)语音识别与理解系统鉴定测试报告，赛德919 语音识别与理解系统鉴定委员会测试组，1992，12，北京

4. F.Jelinek, Self-Organized Language Modeling for Speech Recognition, ICASSP'91, pp.450-506, 1991

5. 吴军， 基于拼音的汉语语音理解方法的研究与实现， 清华大学电子工程系硕士论文，(导师: 王作英)，1993.6

## *STOCHASTIC LANGUAGE MODELS FOR CHINESE SPEECH RECOGNITION BASED ON CHINESE SPELLING*

Wu Jun, Wang Zuoying, Ren yansong
Department of Electronic Engineering, Tsinghua University

Abstract

The rate of speech recognition could be hardly improved when it is as high as 90%, unless speech understanding is used. In this paper, a new approachof Chinese speech understanding—— spelling based stochastic language model approach is proposed and has been used to solve the problem of unrestricted speech understanging, which classical method —— rule based approach could not. It can be used to eliminate and two thirds syllable errors as well as reduce processing time tremendously, so that make Chinese speech recognition systemcommercially available.

Keywords: Speech Recognition, Speech Understanding, Markov Chain, N-Gram Model.