文章编号: 1006- 737X(2002)02- 76- 03

基于隐马尔可夫模型的语音单字识别研究

董湘君, 黄智伟

(南华大学 电气工程学院, 湖南 衡阳 421001)

摘 要: 本文针对线性模型在语音识别中的不足, 进行了隐马尔可夫模型(HMM)在语音单字识别中的研究, 主要对观察输出概率求解、最佳状态序列寻找、参数估计和模型参数的选择进行了探讨.

关键词: 隐马尔可夫模型; 语音单字识别; 参数估计

中图分类号: TN912. 34 文献标识码: A

Study of Isolated Word Recognition Based on HMM

DONG Xiang- jun, HUANG Zhi- wei

(School Of Electrical Engineering, Nanhua University, Hengyang 421001, Hunan, China)

Abstract: With respect to the deficiency of Linear Model on Speech Recognition, this paper studies of isolated word recognition based on HMM, focusing on calculating the probablity of observation sequence, searching for the optional state sequence, estimating parameters, as well as choosing model parameters.

Key words: HMM; isolated word recognition; parameter estimation

0 引言

在一语音单字形成的物理过程中, 会产生一个可观察的序列. 建立一个模型去描述这个序列的特征, 才有可能去识别它. 若分析的区间内, 信号非时变或平稳, 用线性模型描述即可. 而每个语音单字发音会因为时空和人的因素而随机变化, 在不知道语音信号的时变规律, 采用极短时间内分段线性模型, 再串接起来的作法显然不是最有效的. HMM 在本研究中对于如何识别具有不同参数的短时平稳的信号段以及如何跟踪它们之间的转化有很好的实践意义.

1 语音单字识别的 HMM

假设一组词汇中有 V 个单字要识别, 每个字有 k 次发音训练(由一人或多人发音), 则每次发音就构成一个观察序列, 并形成这个字发音特征的某个正确描述. 下面考虑进行单字识别而建立 HMM.

1) 为词汇中的每个字建立一个 HMM λ_v, 即估计模型参数(A, B, P) 使训练观察矢量与单字 V 的模型最匹配.

2) 对每个待识别的单字, 执行图 1 的过程. 用 LPC(线性预测编码) 对语音信号进行特征分析,

x 收稿日期: 2002- 03- 13

作者简介: 董湘君(1973-), 女, 湖南衡阳人, 南华大学电气工程学院助教.

以形成标准模式库(Kl , Kv) , 这是训练过程; 而后从 P (O/ K) 中判决最符合某个 HMM K, 则是识别过程.

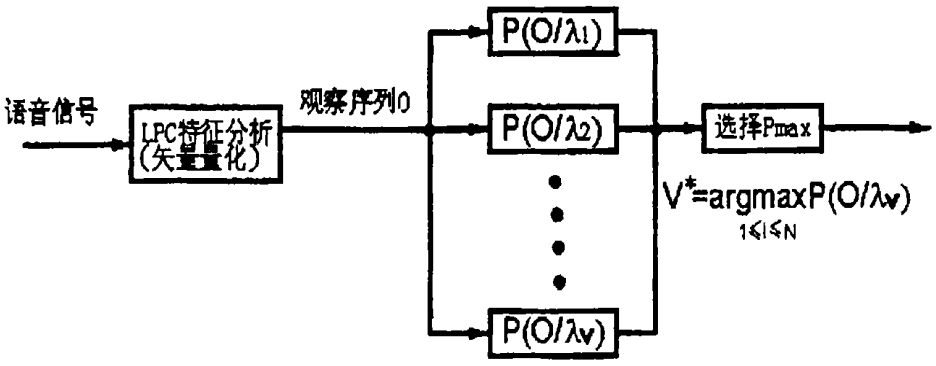

图 1　语音单字识别的框图

Fig. 1　Block diagram of isolated word recognition

2　HMM 在应用中的问题求解

运用 HMM 进行训练或识别, 要涉及几个具体问题. 那么如何着手这些问题的求解是本研究的重点运用.

2. 1　观察输出概率 P(O | K) 的计算

在图 1 中, 观察序列 O = (O1 O2 , OT) 是可见的, 而其可能经历的状态序列却是隐藏的, 对某一状态序列 q = (q1 q2 , qT) 可以计算出输出序列 O 的概率 P (o | q, K) .

基于观察的统计独立性, 有

$$P(o \mid q, K) = \prod_{t=1}^{T} P(o_t \mid q_t, K) \qquad (1)$$

而 P (o_t | q_t, K) = $b_{q_t}(o_t)$

{* 涉及到 B 参数 * }, 所以

$$P(o \mid q, K) = \prod_{t=1}^{T} b_{q_t}(o_t)$$

若设 K下, 状态序列和观察输出序列同时出现的概率为 P (o, q | K), 则

$$P(o, q \mid K) = P(o \mid q, K)\, P(q \mid K)$$

而

$$P(q \mid K) = P_{q1}\, a_{q1q2} a_{q2\,q3}, \ a_{qT-1qT}$$

{* 涉及到 A 参数和 P 参数 * }

不难得到, 在给定模型下, 观察输出 O 的概率就是对所有可能状态序列 q 与输出序列 O 同时出现的概率求和

$$P(o \mid K) = \sum_{所有q} P(o \mid q, K) P(q \mid K)$$

$$= \sum_{q1q2,\ q_T} P_{q1}\, b_{q1(o1)} a_{q1q2} b_{q2(o2)}, \ a_{qT-1qT} b_{qT}(o_T)$$

2. 2　最佳状态序列寻找

在计算 P (o | K) 时, 其中有个概率计算涉及到状态序列与观察序列同时出现, 那个状态序列也就是最佳状态序列. 不难发现, 最佳状态序列的寻找实质是揭示 HMM 的/ 隐藏0 部分, 找出所有状态序列中与输出序列相吻合的/ 正确0 序列. 这有非常实际的物理意义, 反映的恰恰是识别过程. 如果我们选择状态 qT, 使它们在各个 t 时刻均为最可能状态. 则定义给定输出 O 和模型 K后, t 时刻出现 Si 状态的概率就是

$$r_t(i) = P(q_t = s_i \mid o, K) = \frac{p(0, q_t = S_i \mid K)}{P(0 \mid K)} \qquad (2)$$

易知, 有了 2. 1 中的数学准备, (2) 式并不难求解.

所以 t 时刻最可能出现的状态 q_t* 为

$$q_t^* = \arg \max_{1 \le i \le N} [r_t(i)] \qquad 1 \le t \le T$$

当我们将所有 q_t* (1 [t [T) 都求解出来, 也就找到了最相吻合的观察序列, 从而能正确的识别出待识别的语音单字.

2. 3　参数估计

在图 1 中, 建立参数模型参数 K= (A, B, P) 时要求模型与语音单字最匹配; 另外在计算 P (O | K) 和寻找最佳状态序列时, 其给定的模型也是与语音单字最吻合的. 由于 HMM 的随机性, 最初的模型不可能是最佳的, 则获得这个最佳 K意义重大, 而这个过程也就是参数估计. 这就表明,

HMM 的模型参数 K在每组语音单字识别中, 都能根据其具体的参数特征趋于完善. 而线形模型匹配法则是按照模板分析参数精度, 选择合适的失真测度进行参数的匹配, 对语音动态特性的利用远不如 HMM.

现在还没有一种闭合解析式进行参数估计, 本研究采用迭代法来优化参数. 在已知观察输出。O 和模型 K的情况下, 定义 t 时刻状态为 Si, t+ 1 时刻状态为 Sj 的概率为 N(i , j), 则

$$N(i,j) = P(q_t = S_i, q_{t+1} = S_j \mid o, K) \quad (3)$$

联系(2) 式和(3) 式, 有

$$r_t(i) = \sum_{j=1}^{N} N(i,j)$$

于是我们得到模型参数重估公式如下:

1) $P_i = r(i)$;

2) $\overline{a_{ij}} = \sum_{t=1}^{T-1} N(i,j) / \sum_{t=1}^{T-1} r_t(i)$;

3) $\overline{b_j(K)} = \sum_{\substack{t=1 \\ O_t = v_{ki}}}^{T-1} r_t(j) / \sum_{t=1}^{T} r_t(j)$.

其中:

$\sum_{t=1}^{T-1} N(i,j)$ 是状态 Si 向 Sj 转移次数的均值;

$\sum_{t=1}^{T-1} r_t(i)$ 是状态 Si 开始转移次数的均值;

$\sum_{\substack{t=1 \\ O_t = v_{ki}}}^{T-1} r_t(j)$ 是从状态 Sj 得到输出 Vk 的次数的均值;

$\sum_{t=1}^{T} r_t(j)$ 是出现状态 Sj 次数的均值.

当重估参数 P_i、$\overline{a_{ij}}$、$\overline{b_j(K)}$ 构成新的 $\overline{K}$与上一次的 K几乎不变时, 此时参数就达到最佳的某个极限, 也就完成了参数估计. 此时的模型也是最佳的.

2.4 模型参数的选择

HMM 中, A, B, P 是最能表征模型的参数, 但对于其它参数的选择, 也应有足够的重视. 语音单字识别中, 忽略HMM中以下两个模型参数会增加误识率, 而对它们进行合适的选择将更益于模型的优化. 下面我们用实例来说明.

2.4.1 模型中状态数N 的确定

状态数 N 有两种确定方法, 一种是根据对这个单字发音方式的观察数来确定, 别一种则是根据字的音素数来确定. 很明显, 根据音素数更客观, 更符合实际情况. 一般, 每个字有 2 ~ 10 个音素, 如果让状态数大致等于发音数, 则2~ 10 个状态数的模型将比较适合. 对一组词汇, 为求得模型的统一, 应当限制每个字的模型有相同的状态数, 则当模型代表相同音素数的字时, 吻合情况最佳; 而模型表示不同音素数的字时, 势必有误差, 从则带来一定的误识率. 图 2 反映了一组词汇的字模型中, 确定为不同状态数对字的误识率产生的影响. 可见, 只要对一组词汇进行语音识别, 因为每个字的音素数不一定相同, 误识的风险总是存在. 即便我们在前面的参数估计中已经做得很完善, 都不能保证 100% 的正识率.

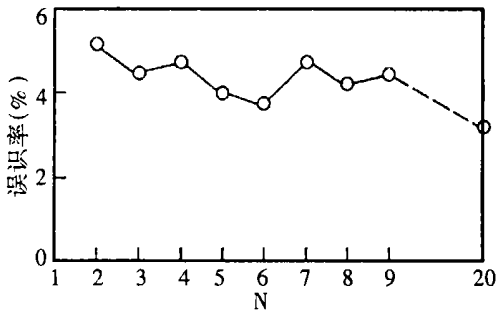

图 2 HMM 中状态数为 N 时, 产生的字平均错误率(0- 9 构成的数字词汇)

Fig. 2 Average word error rate versus the number of states N in HMM(for a digits vocabulary from 0~ 9)

2.4.2 限制某些参数估计

由 2.3, 我们知道最终得到的模型参数 K= (A, B, P) 是相对稳定的, 而某些模型参数可能在优化过程中会趋于零. 某个模型参数为零意味着这个参数将在后面的识别中失去其继续存在的意义. 实验也表明, 这样会增加单字的误识率. 例如, 对于模型参数 B, 应限定 bj(k) \ Emin 即在训练观察集中, 某个状态 Sj 在第K 次观察未出现, 也应保证 Sj 有一确定的出现概率, 否则将增加误识率.

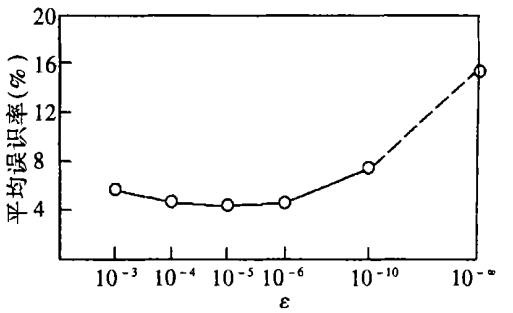

图 3 对最小密度值 E 的字平均误识率

Fig. 3 Average word error rate versus the minimum discrete density value E

(下转第 87 页)

支座的最小锚固长度 La_E.

按照(3)、(4) 式进行计算(详细计算从略), 表 1 列出计算结果.

4　结束语:

从上面的分析中可以看到, 在地震作用下, 多层框架结构梁下, 纵筋伸入支座内的最小锚固长度 La_E 可根据支座梁下纵筋实际应力或构造要求配筋的设计强度值中较大值与梁下纵筋的强度设计值之比进行折减, 这样不但减少了支座的用钢量, 降低造价, 而且方便施工, 工程质量也得到保证.

参考文献:

[1] 建筑抗震设计规范[S]. GB50011- - 2001.

[2] 砼结构设计规范[S]. GBJ10- 89.

[3] 建筑结构构造资料集[C]. 建筑工业出版社, 1990.

表 1　不同抗震等级钢盘锚固长度比较表

Table 1　The compere table of the anchoring length of the reinforing steel bars different aseismatic degree

抗震等级	支座处纵筋最小配筋计算的钢筋面积 (mm^2)	支座底面与顶面配筋量比值计算出来的最小钢筋面积 (mm^2)	最小配筋根数与直径	比较前面三者取较大面积 (mm^2)	实际计算拉应力值(N)	3522 的强度设计值 (N)	折减系数	折减后 3522 在支座内锚固长度 La_E	不折减时的锚固长度 La_E
一	0.4% * 250* 600= 600	760* 0.5* 310/ 210= 561	2514	600	126000	353344.2	0.357	20.7d	45d
二	0.3% * 250* 600= 450	760* 0.3* 310/ 210= 337	2514	450	94500	353344.2	0.267	13d	40d
三	0.25% * 250* 600= 375	760* 0.3* 310/ 210= 337	2512	375	78750	353344.2	0.223	6.7d	35d
四	0.25% * 250* 600= 375		2512	375	78750	353344.2	0.223	6.7d	35d

(上接第 78 页)

由图 3 知, 平均错误率在一很宽范围(10^{210} F E F 10^{-3}) 大约保持为一个常数, 而当 E- > 0 时, 错误率陡然增加. 可见, 我们必须防止这样的参数在优化过程中变得太小.

3　结束语

运用 HMM 进行单字语音识别, 主要解决的问题是前 3 个. 而其中最关键是进行参数估计, 因为最佳模型参数 K的获得是寻找与观察输出相符合的最佳状态序列以及计算观察输出概率的基础. 另外, HMM 也广泛应用在连续语音识别、说话人识别方面.

参考文献:

[1] 胡航. 语音信号处理[M], 哈尔滨工业大学出版社, 2000: 94~ 103.

[2] L. R. Rabiner and B. H. Juang. Fundamentals of Speech Recgnition[M], Prentice- Hall, 1993: 321~ 390.

[3] Y. Ephraim and L. Rabiner. On the ralations between model2 ing approaches for speech recognition[J]. IEEE Trans. In2 formation Theory, 1990: 372~ 380.

[4] C. - H. Lee, C. - H. Lin, and B. H. Juang. A study on speeker adaption of the parameters of continuous density hid2 den markov models[J]. IEEE Trans. Signal Processing, 1991: 806~ 841.