第 17 卷 第 1 期
1991 年 1 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 17, No. 1
Jan., 1991

隐 Markov 模型参数估计的一种新方法[1)]

高雨青　黄泰翼
（中国科学院自动化研究所，南京）

陈永彬
（东南大学，南京）

摘　要

本文提出一种隐 Markov 模型参数估计的新方法．该方法直接以模型作识别器时的识别率最高(或误识率最低)作为估计准则．　由该准则导出的算法的性能明显优于最大似然估计器．文中给出了该算法的一种实现形式．

实验表明，该方法的模型识别率比用最大似然方法求出的模型识别率提高 5% 左右．

关键词：随机模拟，Markov 链，参数估计，语音处理．

一、引　言

自从文献 [1] 于七十年代首次将隐 Markov 模型（简称 HMM）用于描述语音过程以来，HMM 作为时变信号的一种有效模型，在语音处理中受到广泛研究和应用．

模型参数的估计方法一直沿用 Baum 和 Eagon[2] 对离散观察概率密度函数提出的 Reestimation 算法．　Baum[3] 在对观察概率密度加以一定限制的条件下，将算法扩展到连续观察值情形，Liporace[4] 和 Juang[5] 先后放宽了对观察值概率密度的约束条件，使算法的适用范围拓广为：概率密度函数为严格对数凹函数或椭圆对称函数的有限线性组合，从而使 HMM 在语音处理中的广泛应用成为可能．

但是所有模型估计方法都是求最大似然估计器．最大似然估值器(简称 MLE) 的优点在文献 [6] 中有详尽的讨论，所有优点都只在假设的模型必须是正确的前提下才成立．而在语音信号处理，特别是语音识别中，语音信号由 Markov 模型产生这个假设并不一定成立．文献 [7] 指出，模型假设不成立时，最大似然估值器的优点就丧失，根据最大似然准则估计的模型就不是最佳模型．

近来人们不断在寻找新的估计方法[8—10]，其中以最大互信息估计器方法比较突出，当模型不正确时，其性能优于最大似然估值器[11]．

估计模型参数的目的是为了建立一个基于 HMM 的识别器或分类器，这样 HMM 的

本文于 1989 年 3 月收到．
1）本文受国家模式识别实验室基金资助．

参数估计问题就转化成识别器的训练问题．而识别器的优劣是以识别率表征的，若将识别器的训练方法直接建立在“使识别器误识率最低”这个基础上，将使训练和识别建立在相同的准则之上[1)]，由此而导出的估计方法将克服最大似然估计器的缺陷，与假设的模型是否正确无关．这就是本文提出的新的模型参数估计方法的思想．

二、基于误识率最低的模型参数估计方法

假设一个识别器的识别对象是L个单元，需要对每个单元l建立一个隐 Markov 模型 $\mathrm{M}(l)$，$l=1,\cdots,L$．模型用一个三元式表示：$M=(\pi, A, B)$．其中，π为初始状态分布概率；$A=\{a_{ij}\}_{N\times N}$为状态转移概率；$B=\{b_i(\boldsymbol{x})\}_N$为观察序列概率分布密度；$N$为状态数．设$b_i(x)$为具有多个高斯混合形式的概率密度函数

$$b_i(\boldsymbol{x})=\sum_{k=1}^{M} c_{ik}b_{ik}(\boldsymbol{x}),\quad i=1,\cdots,N.$$

其中$\sum_{k=1}^{M} c_{ik}=1$；$b_{ik}=\mathcal{N}(\boldsymbol{x}, E_{ik}, R_{ik})$，$E_{ik}$和$R_{ik}$分别为高斯密度的均值矢量和协方差矩阵．

训练的任务是用第f个单元的样本$O(f)$估计出模型$M(f)$的参数$\pi(f)$，$A(f)$，$B(f)$（这里$B(f)$包含c_{ik}，E_{ik}，R_{ik}）．用下式能否成立来检验估值器的性能：

$$P(M(f)|O(f))\geqslant P(M(l)|O(f)),\quad l=1,\cdots,L,\ l\neq f. \tag{1}$$

为实现(1)式这个目标，就希望$P(M(f)|O(f))$尽可能大．以$P(M(f)|O(f))$作为设计估计器的目标函数，就实现了“使识别器误识率最低”这个准则．与最大似然估计器中以似然函数$P(O(f)|M(f))$作目标函数相比，$P(M(f)|O(f))$更能反映对识别器的要求．

$$\begin{aligned}P(M(f)|O(f))&=\frac{P(O(f)|M(f))\cdot P(M(f))}{P(O(f))}\\&=\frac{P(O(f)|M(f))\cdot P(M(f))}{\sum_{l=1}^{L} P(O(f)|M(l))\cdot P(M(l))}.\end{aligned} \tag{2}$$

令$F=\log P(M(f)|O(f))$，构成目标函数

$$Q=F+\sum_i \lambda_i\left(1-\sum_j a_{ij}\right)+\sum_{i,j}\mu_{ij}\left(1-\sum_t b_{ij}(O_t)\right)+\nu\left(1-\sum_i \pi_i\right). \tag{3}$$

其中λ_i，μ_{ij}，ν为拉格朗日乘子．

令$\frac{\partial Q}{\partial a_{ij}}=0$，可求出$a_{ij}$的迭代估计公式

$$\frac{\partial Q}{\partial a_{ij}}=\frac{\partial F}{\partial a_{ij}}+\lambda_i=0, \tag{4}$$

---

1) 这里有一个隐含的假设：“识别器对训练集误识率最低，则识别器误识率最低”．此假设是否成立可见文献[12]的讨论．

$$\sum_{j=1}^{N} a_{ij} \frac{\partial F}{\partial a_{ij}} = -\left[\sum_{j=1}^{N} a_{ij}\right] \lambda_i = -\lambda_i = \frac{\partial F}{\partial a_{ij}},$$

$$\bar{a}_{ij} = \frac{a_{ij} \cdot \dfrac{\partial F}{\partial c_{ij}}}{\sum\limits_{k=1}^{N} a_{ik} \cdot \dfrac{\partial F}{\partial a_{ik}}}. \tag{5}$$

其中

$$\frac{\partial F}{\partial a_{ij}} = \frac{\partial[\log P(M(f)|O(f))]}{\partial a_{ij}}$$

$$= \frac{\dfrac{\partial P(O(f)|M(f))}{\partial a_{ij}}}{P(O(f)|M(f))} - \frac{\sum\limits_{l=1}^{L} \dfrac{\partial P(O(f)|M(l))}{\partial a_{ij}} \cdot P(M(l))}{\sum\limits_{l=1}^{L} P(O(f)|M(l)) \cdot P(M(l))}. \tag{6}$$

引入向前概率

$$\alpha_t^{f,f}(i) = P(O_1(f),\ O_2(f),\cdots O_t(f),\ s_t = i | M(f)),$$

$$\alpha_t^{f,l}(i) = P(O_1(f),\ O_2(f),\cdots O_t(f),\ s_t = i | M(l))$$

和向后概率

$$\beta_t^{f,f}(i) = P(O_{t+1}(f),\cdots O_T(f) | s_t = i,\ M(f)),$$

$$\beta_t^{f,l}(i) = P(O_{t+1}(f),\cdots O_T(f) | s_t = i,\ M(l)).$$

根据文献 [14] 有

$$P(O(f)|M(f)) = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t^{f,f}(i) a_{ij}^f b_j^f(O_{t+1}(f)) \beta_{t+1}^{f,f}(j), \tag{7}$$

$$\frac{\partial P(O(f)|M(f))}{\partial a_{ij}} = \sum_{t=1}^{T} \alpha_t^{f,f}(i) b_j^f(O_{t+1}(f)) \beta_{t+1}^{f,f}(j). \tag{8}$$

将式 (6), (7), (8) 代入式 (5), 得

$$\bar{a}_{ij} =$$

$$\frac{\dfrac{\sum\limits_{t=1}^{T} \alpha_t^{f,f}(i) a_{ij}^f b_j^f(O_{t+1}(f)) \beta_{t+1}^{f,f}(j)}{P(O(f)|M(f))} - \dfrac{\sum\limits_{l=1}^{L} P(M(l)) \sum\limits_{t=1}^{T} \alpha_t^{f,l}(i) a_{ij}^l b_j^l(O_{t+1}(f)) \beta_{t+1}^{f,l}(j)}{\sum\limits_{l=1}^{L} P(O(f)|M(l)) \cdot P(M(l))}}{\dfrac{\sum\limits_{t=1}^{T} \alpha_t^{f,f}(i) \beta_t^{f,f}(i)}{P(O(f)|M(f))} - \dfrac{\sum\limits_{l=1}^{L} P(M(l)) \sum\limits_{t=1}^{T} \alpha_t^{f,l}(i) \beta_t^{f,l}(i)}{\sum\limits_{l=1}^{L} P(O(f)|M(l)) \cdot P(M(l))}}. \tag{9}$$

而最大似然方法估计 a_{ij} 的迭代公式[13]为

$$\bar{a}_{ij}=\frac{\dfrac{\sum_{t=1}^{T}\alpha_t^{f,f}(i)a_{ij}^f b_j^f(O_{t+1}(f))\beta_{t+1}^{f,f}(j)}{P(O(f)\mid M(f))}}{\dfrac{\sum_{t=1}^{T}\alpha_t^{f,f}(i)\beta_t^{f,f}(i)}{P(O(f)\mid M(f))}}=\frac{\gamma_{ij}}{\gamma_i}. \tag{10}$$

其中 γ_{ij} 为对于模型 $M(f)$，观察序列 $O(f)$ 从状态 q_i 转移到状态 q_j 的平均次数；γ_i 为 $O(f)$ 离开状态 q_i 的次数.

比较 (9)，(10) 式可发现，(9) 式的分子分母比 (10) 式的分子分母分别多了一项 γ_{ij} 和 γ_i 的加权平均，所以 (9) 式可写成

$$\bar{a}_{ij}=\frac{\gamma_{ij}(O(f)\mid M(f))-\bar{\gamma}_{ij}}{\gamma_i(O(f)\mid M(f))-\bar{\gamma}_i}. \tag{11}$$

其中

$$\bar{\gamma}_{ij}=\frac{\sum_{l=1}^{L}P(M(l))\cdot P(O(f)\mid M(l))\cdot\gamma_{ij}(O(f)\mid M(l))}{\sum_{l=1}^{L}P(M(l))\cdot P(O(f)\mid M(l))}; \tag{12}$$

$$\bar{\gamma}_i=\frac{\sum_{l=1}^{L}P(M(l))\cdot P(O(f)\mid M(l))\cdot\gamma_i(O(f)\mid M(l))}{\sum_{l=1}^{L}P(M(l))\cdot P(O(f)\mid M(l))}. \tag{13}$$

$\gamma_{ij}(O(f)\mid M(l))$ 的意义是在模型 $M(l)$ 的条件下，观察序列 $O(f)$ 从状态 q_i 转移到状态 q_j 的平均次数. $\gamma_i(O(f)\mid M(l))$ 的意义是在模型 $M(l)$ 的条件下，$O(f)$ 离开状态 q_i 的次数. 因此，$\bar{\gamma}_{ij}$ 和 $\bar{\gamma}_i$ 分别是 $\gamma_{ij}(O(f)\mid M(l))$ 和 $\gamma_i(O(f)\mid M(l))$ 的以

$$P(M(l))\cdot P(O(f)\mid M(l))$$

为权的加权平均.

至此，我们得到了 a_{ij} 的迭代估计公式. 按同样的方法，令

$$\frac{\partial Q}{\partial b_j(O_{t+1})}=0,\quad \frac{\partial Q}{\partial \pi_i}=0,$$

可求出 $b_j(O_{t+1})$ 和 π_i 的估计公式

$$\bar{c}_{ik}=\frac{\lambda_{ik}(O(f)\mid M(f))-\bar{\lambda}_{ik}}{\gamma_i(O(f)\mid M(f))-\bar{\gamma}_i}, \tag{14}$$

$$\bar{E}_{ik}=\frac{\xi_{ik}(O(f)\mid M(f))-\bar{\xi}_{ik}}{\lambda_{ik}(O(f)\mid M(f))-\bar{\lambda}_{ik}}, \tag{15}$$

$$\bar{R}_{ik}=\frac{\eta_{ik}(O(f)\mid M(f))-\bar{\eta}_{ik}}{\lambda_{ik}(O(f)\mid M(f))-\bar{\lambda}_{ik}}, \tag{16}$$

$$\bar{\pi}_i=\zeta_i-\bar{\zeta}_i. \tag{17}$$

其中

$$\lambda_{ik}(O(f)|M(f)) = \frac{\sum_{t=1}^{T}\left[\sum_{j=1}^{N}\alpha_t^{f,f}(j)a_{ji}^f\right]c_{ik}^f b_{ik}^f(O_{t+1}(f))\beta_{t+1}^{f,f}(i)}{P(O(f)|M(f))}; \tag{18}$$

$$\xi_{ik}(O(f)|M(f)) = \frac{\sum_{t=1}^{T}\left[\sum_{j=1}^{N}\alpha_t^{f,f}(j)a_{ji}^f\right]c_{ik}^f b_{ik}^f(O_{t+1}(f))\beta_{t+1}^{f,f}(i)O_t(f)}{P(O(f)|M(f))}; \tag{19}$$

$$\eta_{ik}(O(f)|M(f)) = \frac{\sum_{t=1}^{T}\left[\sum_{j=1}^{N}\alpha_t^{f,f}(j)a_{ji}^f\right]c_{ik}^f b_{ik}^f(O_{t+1}(f))\beta_{t+1}^{f,f}(i)[O(f)-E_{ik}^f][O(f)-E_{ik}^f]^T}{P(O(f)|M(f))}; \tag{20}$$

$$\zeta_i = \frac{\alpha_1^{f,f}(i)\beta_1^{f,f}(i)}{P(O(f)|M(f))}. \tag{21}$$

而 $\bar{\lambda}_{ik}$, $\bar{\xi}_{ik}$, $\bar{\eta}_{ik}$ 和 $\bar{\zeta}_i$ 分别是 $\lambda_{ik}(O(f)|M(l))$, $\xi_{ik}(O(f)|M(l))$, $\eta_{ik}(O(f)|M(l))$ 和 $\zeta_i(O(f)|M(l))$ 以 $P(M(l))\cdot P(O(f)|M(l))$ 为权的加权平均.

本文是以 $b_i(\boldsymbol{x})$ 为混合高斯密度为例给出算法的,该方法同样适用于概率密度为其它形式的模型参数估计.

对于"左至右"模型,每个模型的训练数据不能是一个观察序列,而必须是由若干序列组成的训练集,即 $M(f)$ 的训练集 $\boldsymbol{O}(f)$ 为

$$\boldsymbol{O}(f) = \{O^1(f), O^2(f), \cdots O^Q(f)\}.$$

这时可将 (11) 式修改为

$$\bar{a}_{ij} = \frac{\sum_{q=1}^{Q}[\gamma_{ij}(O^q(f)|M(f)) - \bar{\gamma}_{ij}^{(q)}]}{\sum_{q=1}^{Q}[\gamma_i(O^q(f)|M(f)) - \bar{\gamma}_i^{(q)}]}. \tag{22}$$

(14)—(17) 式也作类似的修改,这里不再列出.

三、一种实现形式

从 (11), (14), (15), (16), (17) 式可看出,该算法在估计第 f 个模型的参数 $M(f)$ 时,要用到其它所有 $L-1$ 个模型的信息. 这有两点不便,一是当 L 比较大时, 算法实现起来相当复杂;二是不能直接用该算法开始训练,需要有初始模型估计.

用最大似然估计器先对训练集数据进行预估计,以得到初始估计模型 $\{m(l)\}_L$. 用 $\{m(l)\}_L$ 对训练样本进行识别测试. 设测试结果为: 有 $L_1(f)$ 个识别单元易与第 f 个单元混淆或发生误识,即

$$P(O(f)|M(f)) \leqslant \delta + P(O(f)|M(l)), \quad l = 1, \cdots, L_1(f). \tag{23}$$

δ 为衡量易混淆程度的门限值.

然后再用本文提出的方法, 根据 (11), (14), (15), (16), (17) 式对 $m(f)$ 进行迭代估计. 与原始算法的另一个区别在于: 以 $m(f)$ 为初始估计, 求 $M(f)$ 时,只用到与第

f 个单元易混淆的 $L_1(f)$ 个模型参数. 即 (12) 式变成

$$\bar{\gamma}_{ij}=\frac{\sum_{l=1}^{L_1(f)}\gamma_{ij}(O(f)|M(l))\cdot P(M(l))\cdot P(O(f)|M(l))}{\sum_{l=1}^{L_1(f)}P(M(l))\cdot P(O(f)|M(l))}.$$

$\bar{\gamma}_i,\bar{\lambda}_{ik},\cdots$等也按类似的方式简化.

这种实现方式实际上是在加权平均中忽略权值 $P(O(f)|M(l))\cdot P(M(l))$ 比较小的那些项. $P(M(l))$ 表示第 l 个单元的出现概率,若假设 L 个识别单元的出现是独立、等概率的,则 $P(M(l))=1/L$, 权值的大小仅由 $P(O(f)|M(l))$ 表示. 事实上,

$$D_{fl}=P(O(f)|M(f))-P(O(f)|M(l)) \tag{24}$$

可看成是模型 $M(f)$ 和 $M(l)$ 之间的距离. 由式 (23) 和 (24) 可看出, 本文介绍的这种实现方式是只考虑与 $M(f)$ 的距离 $D\leqslant\delta$ 的那些模型,而忽略 $D>\delta$ 的模型对 $M(f)$ 训练时的影响.

四、实　　验

作者在汉语全音节识别系统[1)]的实现中,采用了本文提出的训练 HMM 的新方法,下面给出一些实验结果.

实验在国家模式识别实验室 VAX11-750 语音系统上进行. 词汇表为 38 个韵母(即 $L=38$), 实验材料是汉语全部音节(共 1131 个)中的韵母. 特定发音人(女)共发 3200 个音节,从中分割出韵母段组成训练集,平均每个韵母的训练集有 84 个发音序列.

实验样本的采集条件为将来自话筒或录音机的语音信号经过预放大和 0.1—5.9 kHz 带通滤波之后进行 12 kHz A/D 变换. 对信号进行预加重 $(1-0.98Z^{-1})$, 用于高频提升,以消除唇辐射的影响. 256 点(10.66 ms) 为一帧,帧移为 128 点,对信号加 Hamming 窗后,求 12 阶 LPC 系数,并转换成 12 阶倒谱系数[13].

训练集对由最大似然估计器求出的初始模型 $m(l)(l=1,\cdots L)$ 作识别测试,有 71 次误识.

用新算法迭代估计时, $L_1(f)$ 的平均取值为 5, $L_1(f)_{\max}=10$, $L_1(f)_{\min}=3$. 求出模型 $M(l)$, $l=1,\cdots,L$. 再对训练集作识别测试,只有 18 个误识.

用同一发音人的另一次发音作测试集 (1131 个音节中的韵母段), 分别对 $m(l)$ 和

表　1

误识率 / 测试数据 / 模　型	训练集 3200 个样本	测试集 1131 个样本
m(最大似然法)	2.2%	15.2%
M(本文新方法)	0.57%	10.5%

1) 高雨青等,汉语全音节语言识别理论及系统研究,第七届全国模式识别与机器智能学术会议, 1989, pp4-36—4-41.

$M(l)$，$(l = 1, \cdots, L)$ 作识别测试，结果分别只有 15.2% 和 10.5% 的误识率，新算法的识别率比最大似然算法提高了5%左右(见表 1)。

致谢：本文在国家模式识别实验室支持下完成，在此对陈道文研究员等人给予的支持和帮助表示感谢。

参 考 文 献

[1] Jelinek, F., Continuous Speech Recognition by Statistical Methods, Proceeding of IEEE, **64**(1976), 532—556.

[2] Baum, L. E., Eagon, J. A., An Inequality with Application to Statistical Prediction for Functions of Markov Processes and to a Model for Ecology, *Bull Amer can Mathmatics Society,* **73**(1976), 360—363.

[3] Baum, L. E., et al., A Maximization Technique Occuring in the Statistical Analysis of Probabilitic Functions of Markov Chains, *Annual Mathmatical Statistics,* **41**(1970), 164—171.

[4] Liporace, L. R., Maximum Likelihood Estimation for Multivariate Observations of Markov Source, *IEEE Transaction on Information Theory,* **IT**-28(1982), 729—734.

[5] Juang, H., et al., Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains, *IEEE Transaction on Information Theory,* **IT-32**(1986), 307—309.

[6] Kendall, M., et al., Advanced Theory of Statistical, Macmillam (1979), 38—81.

[7] Nadas, A., A Decision Theoretic Formulation of a Training Problem in Speech Recognition and a Comparison of Training by Unconditional Versus Conditional Maximum Likelihood, *IEEE Transaction on Acoustics, Speech and Signal Processing,* **ASSP-31**(1983).

[8] Bahl, L., et al., MMI of HMM Parameters, Proceeding of International Conference on ASSP, 1(1986), 49—52.

[9] Ephraim, Y., et al., On the Relations Between Modeling Approaches for Information Sources, Proceedings of International Conference on ASSP, 1(1988), 27—30.

[10] Bahl, L. R., et al., A New Algorithm for the Estimation of HMM Parameters, Proceedings of International Confwerence on ASSP, **1**(1988), 436—439.

[11] Nadas, A., et al., On a Model-robust Training Method for Speech Recognition, *IEEE Transaction on Acoustics, Spech and Signal Processing,* **ASSP-36**(1988), 1432—1436.

[12] Gao Y. Q., et al., Dynamic Adaptation of HMM for Robust Speech Recognition, Proceedings of International Symposium on Circuit and System, 1989, 1336—1339.

[13] Levison, S. E., et al., An Introduction to the Application of the Theory of Probubilistic Functions of a Markov Processes to Automatic Speech Recognition, *Bell System Technique Journal,* 62(1983), 1035—1074.

A NEW METHOD FOR THE ESTIMATION OF HIDDEN MARKOV MODEL PARAMETERS

GAO YUQING HUANG TAIYI

(*National Lab. of Pattern Recognition, Institute of Automation, Academia Sinica*)

CHEN YONGBIN

(*Department of Radio Engineering, Southeast University*)

ABSTRACT

The authors present a new method for estimating the parameter values of hidden Markov models in speech processing. In stead of relying on the traditional maximum likelihood rule, this method is based on the rule of maximum recognition accuracy when the model is used directly as recognizer. The performance of this method is better than maximum likelihood estimation (MLE). Experiment has shown that the accuracy of speech recognition is about 5% higher than that of MLE.

Key words: Stochastic modelling; markov chain; parameter estimation; speech processing.