隐Markov模型(HMM)及其在语音处理中的应用

谢锦辉

华中理工大学出版社

隐 Markov 模型(HMM)及其在语音处理中的应用

谢锦辉

华中理工大学出版社

(鄂)新登字第 10 号

D235.13

图书在版编目(CIP)数据

隐 Markov 模型(HMM)及其在语音处理中的应用/谢锦辉
—武汉:华中理工大学出版社, 1995 年 4 月
ISBN 7-5609-1094-7
Ⅰ. 隐…
Ⅱ. 谢…
Ⅲ. 语音信号处理-语音识别-应用
Ⅳ. TN912.3

隐 Markov 模型(HMM)
及其在语音处理中的应用

谢锦辉

责任编辑:李凤英

*

华中理工大学出版社出版发行
(武昌喻家山 邮编:430074)
新华书店湖北发行所经销
华中理工大学出版社沔阳印刷厂印刷

*

开本:850×1168 1/32 印张:5.125 插页:2 字数:130 000
1995 年 4 月第 1 版 1995 年 4 月第 1 次印刷
印数:1-2 000
ISBN 7-5609-1094-7/TN·34
定价:5.00 元

内 容 简 介

本书全面系统地介绍了隐 Markov 模型(HMM)及其在语音处理中的应用。全书共分四章,分别讨论 HMM 的基本理论;各种有代表性的 HMM;HMM 在语音识别、增强和压缩等多方面的应用;HMM 其它相关问题,包括 HMM 与神经网络的联系,HMM 算法的 VLSI 设计以及 HMM 语音处理系统软硬件实现考虑。

本书主要是为从事语音处理,包括识别、合成和编码等方面研究的人员写的,但也可作为电子、信息、计算机、通信、自动化等相关学科的研究生和高年级大学生学习信号处理和语音处理的教学参考书。

前　言

隐 Markov 模型(Hidden Markov Models，简称为 HMM)是描述语音信号的一种统计模型。无论是它的理论还是它在语音处理中的应用，都不是很新的。早在 60 年代末和 70 年代初，HMM 的基本理论就由 Baum 等人建立了起来，并由 CMU 的 Baker 和 IBM 的 Jelinek 等人将其应用到语音识别之中，取得了很大的成功。但是，HMM 引起世界各国从事语音处理研究的学者们广泛重视，并掀起至今不衰的研究热潮，却是 80 年代中期的事。造成这种现象的原因主要有两个：首先，HMM 理论发表在数学杂志上，并未被很多从事语音处理研究的工程技术人员获悉。其次，HMM 首次应用于语音处理时，并没有提供足够的一般性介绍，从而使得多数研究人员无法理解其基本理论并将其应用到自己所从事的研究之中。直到 1983 年以后，Bell 实验室的 Rabiner 等人发表了很有影响的一系列系统介绍 HMM 的理论和应用的文章，上述状况才得以根本改善。很多研究人员，包括作者本人，正是从那时开始接触、了解并熟悉了 HMM，而且，尝试着将 HMM 用于解决自己在语音处理领域研究中所面临的问题。因此，从 80 年代中期开始，国际知名的期刊和学术会议论文集每年都推出了很多有关 HMM 理论和应用的有价值的论文，HMM 也逐渐成为语音处理领域的一个研究热点。

本书试图总结作者过去八年间在 HMM 方面所做的工作，包括：发表在《电子学报》、《通信学报》、《自动化学报》等国内外期刊和会议录上的 20 多篇学术论文；作者的博士论文“线性预测 HMM 在语音识别、压缩和增强中的应用”；作者在法国巴黎南郊的 LIMSI-CNRS(Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingenieur，Centre National de la Recherche Scientifique)做博士后的研究工作；作者负责主持的已完成的两项

课题的主要研究成果，这两个课题是国家 863 计划项目“基于小波分析的用于连续语音识别的巨型 HMM”(1992～1993)和国家教委资助优秀年轻教师基金项目“汉语大词汇量连续语音识别系统”(1993～1994)；以及作者负责主持的正在进行的国家 863 计划项目“采用 HMM 方法的 50 000 词书面语连续语音识别系统”(1994～1995)的初步结果。同时，为了使内容完整和系统化，本书还简要介绍了国内外同行们的有关研究成果和公开发表的论文。全书共分四章，第一章，给出了 HMM 基本理论，包括 HMM 基本思想、基本算法以及算法实现中的一些问题。第二章介绍连续、半连续等等有代表性的各种形式的 HMM。第三章探讨 HMM 在语音识别、增强和压缩等多方面的应用。第四章讨论了 HMM 其它一些相关问题，包括 HMM 与神经网络的联系、HMM 算法的 VLSI 设计以及 HMM 语音处理系统软硬件实现考虑。

本书主要是为从事语音处理，包括识别、合成和编码等方面研究的人员写的，但也可作为电子、信息、计算机、通信、自动化等相关学科的研究生和高年级大学生学习信号处理和语音处理的教学参考书。

在本书完稿之际，作者首先感谢攻读博士学位时的导师万发贯教授和黄载禄教授给予的精心指导。感谢北京中科院自动化所国家模式识别实验室为我完成博士论文所提供的优良的工作环境和实验条件，感谢实验室的马颂德研究员、陈道文研究员、黄泰翼研究员等许多同志给予的关心和支持，尤其是感谢高雨青博士在 HMM 问题上与我进行的大量有成效的讨论和合作。

感谢我在法国巴黎南郊 LIMSI-CNRS 做博士后的指导老师和语音通信组(Groupé Communication Parlée)的同事们，他们是：Joseph Mariani，Françoise Néel，Gilles Adda，Martine Decker-Adda，Héléne Bonneau-Maynard，Frédéric Beaugendre 和 Phillippe Escande。也感谢该实验室许多我已拼不全姓名的那些友善的人们，他们给了我工作和生活上很多关心和支持。感谢他们

给了我或许一生之中最难忘怀的异国之旅的记忆。

作者还要感谢华中理工大学电子与信息工程系信息工程教研室的老师和同事们，在我长达十几年的学习和工作期间所给予的多方面的帮助。尤其是，我要感谢姚天任教授对本书的推荐，感谢华中理工大学出版社为本书出版付出的大量辛勤的劳动。

作者非常感谢多年来给了我巨大支持的亲友们。

最后，衷心地感谢我的妻子丑洁文，在我从事 HMM 在语音处理中的应用研究这八年时间里，我俩也走过了从相识、相爱到成家、生子的一段不平凡历程。没有她博大无私的爱，便没有本书的产生。

谨以此书献给我英年早逝的母亲。

作者

1994 年 8 月

目　录

第一章 HMM 基本理论

隐 Markov 模型(Hidden Markov Models，简称为 HMM)，作为语音信号的一种统计模型，今天正在语音处理各个领域中获得广泛的应用。而有关它的理论基础，却是在 1970 年前后由 Baum 等人建立起来的[30,31,32,33]。随后由 CMU 的 Baker 和 IBM 的 Jelinek 等人将其应用到语音识别之中[26,27,28,105,106]。由于 Bell 实验室 Rabiner 等人在 80 年代中期对 HMM 的深入浅出的介绍[144,191]，才逐渐使 HMM 为世界各国从事语音处理的研究人员所了解和熟悉，进而成为公认的一个研究热点[188,189]。本章是全书的基础，分四节介绍 HMM 的基本理论。首先在第一节中介绍 HMM 的基本思想，主要是回答什么是 HMM 这一问题，从 Markov 链着手，从分析有关 HMM 概念的实例开始，引出 HMM 的定义，并介绍 HMM 的参数。然后在第二节中给出 HMM 的基本算法，就是介绍将 HMM 应用到语音处理中经常会面临的三大基本问题的解决方案，亦即著名的 HMM 三大基本算法：前向-后向算法、Viterbi 算法和 Baum-Welch 算法。这些算法在本书以后的章节中会反复用到。接着，在第三节中介绍具体实现这些算法时应注意的一些问题，包括：初始模型选取；用多个观察值序列训练模型参数；为解决计算上的下溢问题而对算法加入比例因子的处理过程以及 Markov 链的形状。最后，第四节专门对 HMM 训练这一重要问题加以讨论，内容有：在训练数据不充分时的应付措施；怎样克服说话者的影响以及对经典训练算法加以改进的方法。

§1.1 HMM 基本思想

1.1.1 Markov 链

Markov 链是 Markov 随机过程的特殊情况，即 Markov 链是状态和时间参数都离散的 Markov 过程。从数学上，可以给出如下定义：

随机序列 X_n，在任一时刻 n，它可以处在状态 $\theta_1,\cdots,\theta_N$，且它在 $m+k$ 时刻所处的状态为 q_{m+k}的概率，只与它在 m 时刻的状态 q_m 有关，而与 m 时刻以前它所处状态无关，即有：

$$P(X_{m+k}=q_{m+k}/X_m=q_m,X_{m-1}=q_{m-1},\cdots,X_1=q_1)$$
$$=P(X_{m+k}=q_{m+k}/X_m=q_m)$$

其中，
$$q_1,q_2,\cdots,q_m,q_{m+k}\in(\theta_1,\theta_2,\cdots,\theta_N) \tag{1-1}$$

则称 X_n 为 Markov 链，并且称

$$P_{ij}(m,m+k)=P(q_{m+k}=\theta_j/q_m=\theta_i),$$
$$1\leqslant i,j\leqslant N,\quad m,k\text{ 为正整数} \tag{1-2}$$

为 k 步转移概率，当 $P_{ij}(m,m+k)$与 m 无关时，称这个 Markov 链为齐次 Markov 链，此时

$$P_{ij}(m,m+k)=P_{ij}(k) \tag{1-3}$$

以后若无特别申明，Markov 链就是指齐次 Markov 链。当 $k=1$ 时，$P_{ij}(1)$称为一步转移概率，简称为转移概率，记为 a_{ij}，所有转移概率 a_{ij}，$1\leqslant i,j\leqslant N$ 可以构成一个转移概率矩阵，即

$$\boldsymbol{A}=\begin{bmatrix}a_{11} & \cdots & a_{1N}\\ \cdots & & \cdots\\ a_{N1} & \cdots & a_{NN}\end{bmatrix} \tag{1-4}$$

且有
$$0\leqslant a_{ij}\leqslant 1 \tag{1-5}$$

$$\sum_{j=1}^{N}a_{ij}=1 \tag{1-6}$$

由于 k 步转移概率 $P_{ij}(k)$ 可由转移概率 a_{ij} 得到，因此，描述 Markov 链的最重要参数就是转移概率矩阵 $\boldsymbol{A}$。但 $\boldsymbol{A}$ 矩阵还决定不了初始分布，即由 $\boldsymbol{A}$ 求不出 $q_1=\theta_i$ 的概率，这样，完全描述 Markov 链，除 $\boldsymbol{A}$ 矩阵之外，还必须引进初始概率矢量 $\boldsymbol{\pi}=(\pi_1,\cdots,\pi_N)$，其中

$$\pi_i = P(q_1=\theta_i),\quad 1\leqslant i\leqslant N \tag{1-7}$$

显然有

$$0\leqslant \pi_i \leqslant 1 \tag{1-8}$$

$$\sum_i \pi_i = 1 \tag{1-9}$$

实际中，Markov 链的每一状态可以对应于一个可观测到的物理事件。比如天气预测中的雨、晴、雪等，那么，这时它可称之为天气预报的 Markov 链模型。根据这个模型，可以算出各种天气(状态)在某一时刻出现的概率。

1.1.2 HMM 基本概念

HMM 是在 Markov 链的基础之上发展起来的。由于实际问题比 Markov 链模型所描述的更为复杂，观察到的事件并不是与状态一一对应，而是通过一组概率分布相联系，这样的模型就称为 HMM。它是一个双重随机过程，其中之一是 Markov 链，这是基本随机过程，它描述状态的转移。另一个随机过程描述状态和观察值之间的统计对应关系。这样，站在观察者的角度，只能看到观察值，不像 Markov 链模型中的观察值和状态一一对应；因此，不能直接看到状态，而是通过一个随机过程去感知状态的存在及其特性。因而称之为"隐"Markov 模型，即 HMM。现在来看一个著名的说明 HMM 概念的例子——球和缸(Ball and Urn)实验。

设有 N 个缸，每个缸中装有很多彩色的球，球的颜色由一组概率分布描述，如图 1.1 所示。实验是这样进行的，根据某个初始概率分布，随机地选择 N 个缸中的一个，例如第 i 个缸，再根据这个缸中彩色球颜色的概率分布，随机地选择一个球，记下球的颜

缸 1	缸 2	缸 N
$P(红)=b_{11}$	$P(红)=b_{21}$	$P(红)=b_{N1}$
$P(蓝)=b_{12}$	$P(蓝)=b_{22}$	$P(蓝)=b_{N2}$
$P(绿)=b_{13}$	$P(绿)=b_{23}$	$P(绿)=b_{N3}$
⋮	⋮	⋮
$P(黄)=b_{1M}$	$P(黄)=b_{2M}$	$P(黄)=b_{NM}$

图 1.1 球和缸实验

色，记为 O_1，再把球放回缸中，又根据描述缸的转移的概率分布，随机选择下一个缸，例如，第 j 个缸，再从缸中随机选一个球，记下球的颜色，记为 O_2，一直进行下去。可以得到一个描述球的颜色的序列 $O_1, O_2, \cdots$；由于这是观察到的事件，因而称之为观察值序列。但缸之间的转移以及每次选取的缸被隐藏起来了，并不能直接观察到。而且，从每个缸中选取球的颜色并不是与缸一一对应，而是由该缸中彩球颜色概率分布随机决定的。此外，每次选取哪个缸则由一组转移概率所决定。

1.1.3 HMM 定义

有了前面讨论的 Markov 链以及球和缸实验，就可以给出 HMM 的定义，或者说，一个 HMM 可以由下列参数描述：

a. N：模型中 Markov 链状态数目。记 N 个状态为 $\theta_1, \cdots, \theta_N$，记 t 时刻 Markov 链所处状态为 q_t，显然 $q_t \in (\theta_1, \cdots, \theta_N)$。在球与缸实验中的缸就相当于状态。

b. M：每个状态对应的可能的观察值数目。记 M 个观察值为 $V_1, \cdots, V_M$，记 t 时刻观察到的观察值为 O_t，其中 $O_t \in (V_1, \cdots, V_M)$。在球与缸实验中所选彩球的颜色，就是观察值。

c. π：初始状态概率矢量，$\pi = (\pi_1, \cdots, \pi_N)$，其中

$$\pi_i = P(q_1 = \theta_i), \quad 1 \leqslant i \leqslant N \tag{1-10}$$

在球与缸实验中指开始时选取某个缸的概率。

d. $\boldsymbol{A}$：状态转移概率矩阵，$\boldsymbol{A}=(a_{ij})_{N\times N}$，其中

$$a_{ij} = P(q_{t+1} = \theta_j / q_t = \theta_i), \quad 1 \leqslant i, j \leqslant N \tag{1-11}$$

在球与缸实验中指描述每次在当前选取的缸的条件下选取下一个缸的概率。

e. $\boldsymbol{B}$：观察值概率矩阵，$\boldsymbol{B}=(b_{jk})_{N\times M}$，其中

$$b_{jk} = P(O_t = V_k / q_t = \theta_j), \quad 1 \leqslant j \leqslant N, 1 \leqslant k \leqslant M \tag{1-12}$$

在球与缸实验中，b_{jk}就是第 j 个缸中球的颜色 k 出现的概率。

这样，可以记一个 HMM 为

$$\lambda = (N, \boldsymbol{M}, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}) \tag{1-13}$$

或简写为

$$\lambda = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}) \tag{1-14}$$

更形象地说，HMM 可分为两部分，一个是 Markov 链，由 $\boldsymbol{\pi}$、$\boldsymbol{A}$ 描述，产生的输出为状态序列，另一个是一个随机过程，由 $\boldsymbol{B}$ 描述，产生的输出为观察值序列，如图 1.2 所示。T 为观察值时间长度。

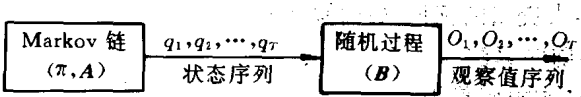

图 1.2 HMM 组成示意图

§1.2 HMM 基本算法

1.2.1 前向-后向算法

这个算法是用来计算给定一个观察值序列 $O=O_1, O_2, \cdots, O_T$ 以及一个模型 $\lambda=(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ 时，由模型 λ 产生出 O 的概率 $P(O/$

$\lambda)$。

根据图 1.2 所示 HMM 的组成，$P(O/\lambda)$最直接的求取方法如下：

对一个固定的状态序列 $S=q_1,q_2,\cdots,q_T$，有

$$P(O/S,\lambda)=\prod_{t=1}^{T}P(O_t/q_t,\lambda)=b_{q_1}(O_1)b_{q_2}(O_2)\cdots b_{q_T}(O_T) \tag{1-15}$$

其中

$$b_{q_t}(O_t)=b_{jk}|_{q_t=\theta_j,O_t=V_k},1\leqslant t\leqslant T \tag{1-16}$$

而对给定 λ，产生 S 的概率为

$$P(S/\lambda)=\pi_{q_1}a_{q_1q_2}\cdots a_{q_{T-1}q_T} \tag{1-17}$$

因此，所求概率为

$$\begin{aligned}P(O/\lambda)&=\sum_{\text{所有}S}P(O/S,\lambda)P(S/\lambda)\\&=\sum_{q_1,q_2,\cdots,q_T}\pi_{q_1}b_{q_1}(O_1)a_{q_1q_2}b_{q_2}(O_2)\cdots a_{q_{T-1}q_T}b_{q_T}(O_T)\end{aligned} \tag{1-18}$$

显而易见，上式的计算量是十分惊人的，大约为 $2TN^T$ 数量级，当 $N=5,T=100$ 时，计算量达 10^{72}，这是完全不能接受的。在此情况下，要求出 $P(O/\lambda)$还必须寻求更有效的算法，这就是 Baum 等人提出的前向-后向算法：

(1)前向算法

定义前向变量为

$$\alpha_t(i)=P(O_1,O_2,\cdots,O_t,q_t=\theta_i/\lambda),\quad 1\leqslant t\leqslant T \tag{1-19}$$

那么，有

a. 初始化：

$$\alpha_1(i)=\pi_ib_i(O_1),\quad 1\leqslant i\leqslant N \tag{1-20}$$

b. 递归：

$$\alpha_{t+1}(j)=\Big[\sum_{i=1}^{N}\alpha_t(i)a_{ij}\Big]b_j(O_{t+1}),$$
$$1\leqslant t\leqslant T-1,1\leqslant j\leqslant N \tag{1-21}$$

c. 终结：

$$P(O/\lambda)=\sum_{i=1}^{N}\alpha_T(i) \tag{1-22}$$

其中 $$b_j(O_{t+1}) = b_{jk}|_{o_{t+1}=v_k} \tag{1-23}$$

这种算法计算量大为减少，变为 $N(N+1)(T-1)+N$ 次乘法和 $N(N-1)(T-1)$ 次加法。同样，$N=5, T=100$ 时，只需大约 3 000 次计算(乘法)。这种算法是一种典型的格型结构，如图 1.3 所示。

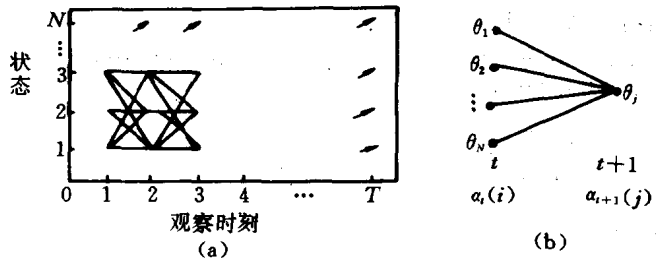

图 1.3 前向算法示意图

(a) 格型结构； (b) t 时刻递归关系

(2)后向算法

与前向算法类似，定义后向变量为

$$\beta_t(i) = P(O_{t+1}, O_{t-2}, \cdots, O_T / q_t = \theta_i, \lambda), \quad 1 \leqslant t \leqslant T-1,$$

其中 $$\beta_T(i) = 1 \tag{1-24}$$

类似，有

a. 初始化： $$\beta_T(i)=1, \quad 1 \leqslant i \leqslant N \tag{1-25}$$

b. 递归： $$\beta_t(i)=\sum_{j=1}^{N} \alpha_{ij} b_j(O_{t+1}) \beta_{t+1}(j),$$

$$t=T-1, T-2, \cdots, 1, \quad 1 \leqslant i \leqslant N \tag{1-26}$$

c. 终结： $$P(O/\lambda)=\sum_{i=1}^{N} \beta_1(i) \tag{1-27}$$

后向算法的计算量大约在 N^2T 数量级，也是一种格型结构。

1.2.2 Viterbi 算法

这个算法解决了给定一个观察值序列 $O=O_1, O_2, \cdots, O_T$ 和一

个模型 $\lambda=(\pi,A,B)$，在最佳的意义上确定一个状态序列 $Q^*=q_1^*,q_2^*,\cdots,q_T^*$ 的问题。

"最佳"的意义有很多种，由不同的定义可得到不同的结论。这里讨论的最佳意义上的状态序列 Q^*，是指使 $P(Q,O/\lambda)$[①] 最大时确定的状态序列 Q^*。Viterbi 算法可以叙述如下：

定义 $\delta_t(i)$为时刻 t 时沿一条路径 $q_1,q_2,\cdots,q_t$，且 $q_t=\theta_i$，产生出 $O_1,O_2,\cdots,O_t$ 的最大概率，即有

$$\delta_t(i)=\max_{q_1,q_2,\cdots,q_{t-1}} P(q_1,q_2,\cdots,q_t,q_t=\theta_i,O_1,O_2,\cdots,O_t/\lambda) \tag{1-28}$$

那么，求取最佳状态序列 Q^* 的过程为

a. 初始化：
$$\delta_1(i)=\pi_i b_i(O_1),\quad 1\leqslant i\leqslant N \tag{1-29}$$
$$\varphi_1(i)=0,\quad 1\leqslant i\leqslant N \tag{1-30}$$

b. 递归：
$$\delta_t(j)=\max_{1\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}]b_j(O_t),2\leqslant t\leqslant T,1\leqslant j\leqslant N \tag{1-31}$$
$$\varphi_t(j)=\operatorname{argmax}_{1\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}]^{②},2\leqslant t\leqslant T,1\leqslant j\leqslant N \tag{1-32}$$

c. 终结：
$$P^*=\max_{1\leqslant i\leqslant N}[\delta_T(i)] \tag{1-33}$$
$$q_T^*=\operatorname{argmax}_{1\leqslant i\leqslant N}[\delta_T(i)] \tag{1-34}$$

d. 状态序列求取：
$$q_t^*=\varphi_{t+1}(q_{t+1}^*),\qquad t=T-1,T-2,\cdots,1 \tag{1-35}$$

应当指出，Viterbi 算法的一个副产品 $P^*=\max_Q P(Q,O/\lambda)$和前向-后向算法计算出的$P(O/\lambda)=\sum_Q P(Q,O/\lambda)$之间的关系为

对语音处理应用而言，$P(Q,O/\lambda)$动态范围很大，或者说不同

① 此处及本书以后部分，为方便起见，将 $P(Q/O,\lambda)$记为 $P(Q,O/\lambda)$。

② 符号说明：如果 $i=I$ 时，$f(i)$达到最大值，那么 $I=\operatorname{argmax}_{1\leqslant i\leqslant N}[f(i)]$。