一般拓扑结构的非齐次隐含马尔科夫模型及其在中、英文语种辨识中的应用

王作英　　孙　健
(清华大学电子工程系　北京　100084)

摘　要：为了充分利用语音信号中的段长信息，该文提出了一种具有一般拓扑结构的非齐次隐含 Markov 模型(Hidden Markov Model, HMM)，并将其应用于中、英文语种辨识(Language IDentification, LID)系统。非齐次 HMM 既很好地描述了语音信号的发生过程，又准确地利用了状态的段长信息和语言中的上下文连接结构信息，对于中、英文语种辨识系统，非齐次的 HMM 系统辨识性能好于齐次的 HMM 模型。而在非齐次的 HMM 中，同段长为均匀分布相比，段长分布为正态分布时系统的辨识性能更好，表明段长确实是一种重要的语种区分信息之一，且正态分布较均匀分布更接近于真实的段长分布。

关键词：语种辨识；非齐次隐含 Markov 模型；段长分布
中图分类号：TP391.42　　**文献标识码**：A　　**文章编号**：1009-5896(2007)04-0867-03

The Inhomogeneous HMM with General Topological Structure and Its Application in Language Identification between Mandarin and English

Wang Zuo-ying　　Sun Jian
(*Department of Electronic Engineering*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: In order to use duration information in Language IDentification (LID) efficiently, the inhomogeneous Hidden Markov Model (HMM) with general topological structure is proposed, and is used to identify the language between Mandarin and English also. Because the inhomogeneous HMM with general topologic structure not only describes the duration of state more accurately than HMM, but also uses the structure information of specific language phonetics more effectively, the LID system based on the inhomogeneous HMM with general topological structure has better performance than the homogeneous HMM. For the LID system based on inhomogeneous HMM with different duration distribution, the norm distribution has better performance than the uniform distribution, it shows that the state duration is an important cue for language identification and the norm distribution can model the duration more accurately than the uniform distribution.

Key words: Language identifier; Inhomogeneous hidden Markov model; Duration distribution

1　引言

语种辨识(Language Identification, LID)是计算机利用语言间存在着的发音及语法结构区分信息，鉴别一段语音所属语言种类的过程。LID系统作为支持多语种自动系统的前端处理[1]，有着广泛的应用。例如在信息咨询或电话转接中，可以将电话自动转接到能提供该种语言服务的线路上等等。目前，随着中国与世界交流的日益深入，英文作为世界先进经济技术及文化主要载体之一，对于对话及听说系统来说，能够提供对中、英文的支持已成为必然的需求，因此能够提供准确的中、英文语种辨识前端具有重要的意义。

LID 系统的研究开始于 20 世纪 70 年代，自从HMM(Hidden Markov Model)模型被成功应用于语音识别后，HMM也被应用于LID系统中[2, 3]。但这种广泛应用于语音系统的HMM模型具有齐次性，导致模型的状态段长呈几何分布，这同实际的语音信号存在着较大的误差[4]。为了能够准确地利用语音信号中的段长信息，基于段长分布的隐含Markov模型(Duration Distribution Based Hidden Markov Mode, DDBHMM)被成功地应用于汉语的连续语音识别系统中，并取得了较好的识别性能[4]。然而，DDBHMM具有严格自左向右的状态转移结构，这自左向右的状态转移结构限制了它在某些诸如语种识别和音乐识别等要求具有更加灵活的状态跳转策略的应用。

针对这种自左向右结构的局限性，文中首先研究了具有更一般拓扑结构的非齐次 HMM 模型，并将这种非齐次 HMM 应用于语种辨识，最后在不同的段长分布情况下进行了语种辨识实验。

2　一般拓扑结构的非齐次 HMM

对于具有 N 个状态的 HMM，在系统中引入初始状态 0 及退出状态 N+1，则系统共有 N+2 个状态。图 1 所示的 HMM 模型有 4 个状态(不包括初始状态和退出状态)，图中数字表

2005-09-09 收到，2006-01-06 改回

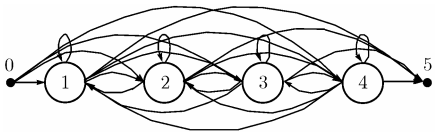

图1 可以任意跳转的4个状态的HMM模型

示系统所处状态序号。从图中可以看出，系统可以从初始状态进入任意状态，也可以从任意状态退出，同时任意两个状态间可自由跳转。

2.1 转移概率

设模型的状态转移概率与状态持续时间相关，状态 i 的段长分布为 $P_i(\tau)$，若系统在过去的 k 帧语音都驻留于状态 i，则系统在第 $k+1$ 帧语音仍驻留于状态 i 的概率为

$$a_{ii}(k+1)=P_i(\tau\geq k+1\,|\,\tau\geq k) \tag{1}$$

系统离开状态 i 跳转到状态 j ($j\neq i$)的概率为

$$\begin{aligned}a_{ij}(k+1)&=P(\tau=k\,\&\,s_{k+1}=j\neq i\,|\,\tau\geq k)\\&=(1-a_{ii}(k+1))P(s_{k+1}=j\neq i\,|\,s_k=i)\end{aligned} \tag{2}$$

式中 $P(s_{k+1}=j\neq i\,|\,s_k=i)$ 表示系统在过去的 k 帧语音处于状态 i，在第 $k+1$ 帧语音发生状态跳转的条件下转移到状态 j 的概率，它只与语言的结构有关。因此只与源状态和目标状态相关，而与系统发生状态跳转的时间无关。令

$$c_{ij}=P(s_{k+1}=j\neq i\,|\,s_k=i) \tag{3}$$

则由式(2)可以得到系统的转移概率为

$$a_{ij}(k+1)=(1-a_{ii}(k+1))\cdot c_{ij},\quad \forall j\neq i \tag{4}$$

由式(1)及式(4)可以看出，HMM 的状态转移概率随着状态的持续时间变化而变化，因此这种 HMM 具有非齐次性。

2.2 最优状态序列

系统输出的 T 帧语音观测量序列为 $\boldsymbol{O}=(\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_T)$，其所对应的状态序列为 $S=(s_1,\cdots,s_T)$，则观测序列所对应的最优状态序列为

$$\hat{S}=\max_S P(\boldsymbol{O},S)=\max_S P(S)P(\boldsymbol{O}\,|\,S) \tag{5}$$

设状态变量 s_i 有 N 个取值：$s_i\in\{q_1,q_2,\cdots,q_N\}$，引入系统的初始状态 q_0 和退出状态 q_{N+1}，序列 S 的取值表示为 $Q=\left(q_0,\underbrace{q_1,\cdots,q_1}_{d_1\text{个}},\cdots,\underbrace{q_i,\cdots,q_i}_{d_i\text{个}},\cdots,\underbrace{q_N,\cdots,q_N}_{d_N\text{个}},q_{N+1}\right)$，其中 d_i 表示状态 q_i 的驻留长度，满足 $\sum_{i=1}^{N}d_i=T$。设事件 $u_{i,i+1}$ 表示从状态 q_i 跳转到状态 q_{i+1}，由式(3)知 $P(u_{i,i+1})=c_{q_iq_{i+1}}$。取 $U=(u_{01},u_{12},\cdots,u_{N,N+1})$ 表示由状态跳转事件构成的事件序列，则 U 反映了状态间的连接结构信息。事件 v_i 表示系统在第 d_i 个时间段内驻留于状态 q_i，$P(v_0)=1$，令 $V=(v_0,v_1,\cdots,v_{N+1})$，$U$ 和 V 相互独立，则

$$\begin{aligned}P(Q)&=P(UV)=P(U)P(V\,|\,U)\\&=P(u_{01},u_{12},\cdots,u_{N-1,N})P(v_0,\cdots v_{N+1})\\&=\prod_{i=1}^{N}P(u_{i-1,i}\,|\,u_{01},\cdots,u_{i-2,i-1})P(v_i\,|\,v_0,\cdots,v_{i-1})\end{aligned} \tag{6}$$

$P(u_{i-1,i}\,|\,u_{01},\cdots,u_{i-2,i-1})$ 的值取决于语言的上下文结构信息，不同语言的结构信息存在着巨大的差别，因此它是语种辨识的重要信息。$P(v_i\,|\,v_0,\cdots,v_{i-1})$ 取决于语言的发音规则以及发声器官发声时的运动惯性，属于声学模型的部分。根据式(5)、式(6)有

$$\begin{aligned}\hat{S}=\arg\max_{(q_1,q_1,\cdots,q_K)}\Bigg\{&\max_{(d_1\cdots,d_K)}\prod_{i=1}^{K}P(u_{i-1,i}\,|\,u_{01},\cdots,u_{i-2,i-1})\\&\cdot\left[P(v_i\,|\,v_0,\cdots,v_{i-1})\prod_{\tau=t_i+1}^{t_{i+1}}b(\boldsymbol{o}_\tau\,|\,q_i)\right]\Bigg\}\end{aligned} \tag{7}$$

其中 $b(\boldsymbol{o}_\tau\,|\,q_i)$ 表示系统处于状态 q_i 的条件下，观测量为 $\boldsymbol{o}_\tau$ 的概率。若假设 $u_{i,i+1}$ 之间相互独立，将式(1)和式(4)代入式(7)中，得

$$\hat{S}=\arg\max_{(q_1,\cdots,q_N)}\left\{\max_{(d_1\cdots,d_N)}\prod_{i=0}^{N+1}\left[c_{q_i,q_{i+1}}\cdot D_{q_i}(d_i)\prod_{\tau=t_i+1}^{t_{i+1}}b(\boldsymbol{o}_\tau\,|\,q_i)\right]\right\} \tag{8}$$

式中 $D_{q_i}(d_i)$ 表示状态的段长分布。

2.3 训练和解码算法

非齐次HMM模型是一个有后效的过程，Viterbi或其它基于Bellman动态规划的解码算法都不再适用。为了解决这一问题，在具有自左向右拓扑结构的DDBHMM中采用了最优状态序列分割算法(Maximum Likelihood States Sequence，MLSS)进行快速的训练和解码[5]。在DDBHMM的MLSS算法中，要求状态的段长分布为凸函数。由式(8)知，当段长分布为正态分布时，因为 $c_{q_{i-1}q_i}$ 为常数，当 $D_{q_i}(d_i)$ 满足MLSS算法中段长分布为凸性条件时[5]，$c_{q_{i-1}q_i}D_{q_i}(d_i)$ 也仍然为凸函数，因此具有一般拓扑结构的非齐次HMM可以采用MLSS算法进行快速的训练和解码。

3 中、英文语种辨识

对中、英文分别建立一个具有 N 状态一般拓扑结构的非齐次 HMM 模型，其中各个状态之间可自由跳转。辨识过程就是将语音分别与中文和英文的模型进行分配，取匹配距离最优的语言作为辨识结果。当取似然对数值作为匹配得分时，即

$$\hat{L}=\arg\max_L D(\boldsymbol{O}\,|\,\lambda_L)=\arg\max_L\left[\log P(\boldsymbol{O}\,|\,\lambda_L)\right] \tag{9}$$

式中 λ_L 为语言 L 的模型参数。采用 MLSS 算法时，根据式(8)有

$$\begin{aligned}\hat{L}=\arg\max_L\Bigg[&\max_{(q_{L,1},\cdots,q_{L,N})}\Bigg\{\max_{(d_1\cdots,d_N)}\sum_{i=0}^{N+1}\Bigg[\log c_{q_{L,i},q_{L,i+1}}\\&+\log D_{q_{L,i}}(d_i)+\sum_{\tau=t_i+1}^{t_{i+1}}\log b(\boldsymbol{o}_\tau\,|\,q_{L,i})\Bigg]\Bigg\}\Bigg]\end{aligned} \tag{10}$$

式中 $Q=(q_{L,0},q_{L,1},\cdots,q_{L,i},\cdots,q_{L,N+1})$ 表示由语言 L 模型中的状态所组成的状态序列，t_i 为状态 $q_{L,i-1}$ 的终止帧，其满足 $t_1=0$，$t_{N+1}=T$。$D_{q_i}(t_{i+1}-t_i)$ 表示状态 $q_{L,i}$ 段长为 $t_{i+1}-t_i$ 时的概率，$b(\boldsymbol{o}_\tau\,|\,q_{L,i})$ 表示系统处于状态 $q_{L,i}$ 的条件下，观测量为 $\boldsymbol{o}_\tau$ 的概率。

由式(10)知系统的模型参数包括状态观测量概率分布 $b(\boldsymbol{o}\,|\,s)$、状态的段长分布 $D_s(\tau)$ 及状态跳转概率矩阵 $\{c_{ij}\}$。

表 1 中、英文语种辨识错误率(%)

语种	中文			英文		
模型	齐次 HMM	非齐次 HMM		齐次 HMM	非齐次 HMM	
段长分布		均匀	正态		均匀	正态
1 状态	13.6			13.7		
4 状态	24.9	16.9	16.0	24.9	16.2	16.2
8 状态	24.7	15.6	15.0	24.8	15.5	15.0
16 状态	24.6	14.0	13.9	24.5	13.8	13.2
32 状态	24.6	12.3	11.9	24.6	12.0	11.7
64 状态	24.4	10.4	9.7	24.2	10.2	9.5
96 状态	24.5	10.1	9.3	24.3	9.9	9.2
128 状态	24.4	9.8	9.2	24.2	9.7	9.1

其中 $b(\boldsymbol{o}\mid s)$ 利用高斯混合分布来逼近，即系统处于状态 s 的情况下观测矢量分布概率密度函数为

$$b(\boldsymbol{o}\mid s)=\sum_{k=1}^{K}\alpha_k N_s(\boldsymbol{o},\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k^2) \tag{11}$$

其中 $\boldsymbol{o}$ 表示语音帧的观测矢量，$\boldsymbol{\sigma}_k$ 表示第 k 个高斯分布的加权系数，其满足 $\sum_{k=1}^{K}\alpha_k=1$，$N_s(\boldsymbol{o},\boldsymbol{\mu}_k,\boldsymbol{\sigma}_k^2)$ 表示均值矢量 $\boldsymbol{\mu}_k$，协方差阵 $\boldsymbol{\sigma}_k^2$ 为对角阵的多维矢量高斯分布概率密度函数。段长分布 $D_s(\tau)$ 可以取正态分布或均匀分布等。

首先对训练语料进行聚类得到系统的状态，然后利用 MLSS 分割算法得到训练语料所对应的状态序列，对分割的结果利用 K 均值聚类算法得到观测量分布参数。当利用正态分布逼近段长分布时，其分布参数为

$$\mu_i=\frac{1}{K}\sum_{j=1}^{K}d_j,\quad \sigma_i^2=\frac{1}{K}\sum_{j=1}^{K}(d_j-\mu_i)^2 \tag{12}$$

其中 μ_i 表示状态 i 段长分布的均值，σ_i^2 表示方差，K 表示状态 i 在训练语料中出现的总次数，第 j 次出现时持续的段长为 d_j。状态的跳转概率为

$$c_{ij}=M_{ij}\Big/\sum_{\substack{k=1\\k\neq i}}^{N}M_{ik} \tag{13}$$

其中 M_{ik} 表示在训练语料的分割结果中出现由状态 i 跳转至状态 k 的总次数。

4　实验

用于实验的数据共分为两部分，一部分用来训练中、英文的语言相关的模型参数，一部分用来测试系统的性能。863 智能计算机主体办公室提过的中文语料被用来训练中文的模型参数，共 169 个文件(男、女声各 83 个)。英文的训练语料为 WSJ0 说话人无关训练数据集，共有 84 个说话人，7236 句话构成。测试数据为 1997 年的 Broadcast News Speech Corpus 中分别随机选取了 10 个中、英文语音文件作为测试数据。实验中的语音帧长为 20ms，帧叠为 10ms，声学特征为 14 维 MFCCs 及能量，加上它们的一阶和二阶差分，共 45 维。段长分布为正态分布，测试结果如表 1 所示。

5　讨论

在目前广泛应用于语音系统中的齐次HMM模型中，其转移概率为常数，段长呈几何分布，这与实际的语音信号误差较大，导致多状态系统的性能与单状态系统相比，系统的辨识性能并不能得到提高[2]。而非齐次HMM模型能够正确地描述状态段长信息，当系统的状态数较少时，由于模型描述的状态跳转概率与语言相关的连接结构信息误差较大，系统的性能没有单状态系统的性能好。而随着系统状态数增加，系统性能逐步得到改善。试验结果表明当段长分布为均匀分布，当系统的状态数多于 32 个时，非齐次HMM系统的辨识性能将好于单状态的齐次HMM模型的性能。同时，段长分布为正态分布时有最好的辨识性能，说明正态分布较均匀分布更接近于真实的段长分布。由以上分析知，语音信号中的段长信息反映了发音变化速率，可以作为语种间的区分信息之一，具有一般拓扑结构的非齐次HMM模型准确地利用了状态段长信息，同时又能够充分利用语言中的声学连接结构信息。

参 考 文 献

[1] Zissman M A and Berkling K M. Automatic language identification. *Speech Communication*, 2001, 35(1-2): 115–124.

[2] Zissman M A. Automatic language identification using Gauss mixture and hidden Markov models, In: 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-93, Minneapolis, Minnesota, USA, 1993, 2: 399–402.

[3] House A S and Neuburg E P. Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *J. Acoust. Soc. Amer*, 1977, 62(3): 708–713.

[4] 王作英，肖熙. 基于段长分布的 HMM 语音识别模型. 电子学报, 2004, 32(1): 46–50.
Wang Zuo-ying and Xiao Xi. Duration distribution based HMM speech recognition models. *Acta Electronica Sinica*, 2004, 32(1): 46–50.

[5] Wang Z Y and Gao H G. An inhomogeneous HMM speech recognition algorithm. *Chinese Journal of Electronics*, 1998, 7(1): 73–77.

王作英：男，1935 年生，教授，博士生导师，研究方向为语音信号处理与模式识别.

孙　健：男，1976 年生，博士生，研究方向为语种辨识与语音识别.