

# 隐马尔可夫模型的一种有区分力的 反向传播训练方法<sup>1)</sup>

邓 伟

(苏州大学计算机工程系 苏州 215006)

(E-mail: dengw@zhz.suda.edu.cn)

赵荣椿

(西北工业大学计算机科学与工程系 西安 710072)

**摘 要** 研究隐马尔可夫模型(HMM)的一种有区分力的训练方法. 在多层前向神经网络的框架中实现了HMM的前向概率计算. 基于这一框架, 利用偏导数的反向传播计算方法, 通过梯度上升的优化过程来实现互信息的最大化, 从而对HMM进行有区分力的训练. 这一训练方法被称之为HMM的反向传播训练方法. 此外, 还设计了一个用以实现这一训练方法的在数值计算上具有强鲁棒性的算法. 语音识别的实验结果证实了这一训练方法的优越性.

**关键词** 隐马尔可夫模型, 神经网络, 区分力, 反向传播.

## BACK-PROPAGATION APPROACH TO DISCRIMINATIVE TRAINING OF HIDDEN MARKOV MODEL

DENG Wei

(Dept. of Computer Engineering, Soochow University, Suzhou 215006)

ZHAO Rongchun

(Dept. of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072)

**Abstract** In this paper, an approach to the discriminative training of hidden Markov model (HMM) is presented. The forward probability calculation of HMM is accomplished within the framework of a multilayer feedforward neural network. Based on this framework, by making use of the back-propagation method of computing partial derivatives, the maximization of mutual information can be achieved through the gradient ascent optimization process, and thus the discriminative training of HMM is performed. This approach to the training of HMM is called back-propagation training approach. Additionally, a more numerically robust algorithm is also designed for implementing this approach. The superiority of this approach is proved by the results of the speech recognition

1) 本文得到了国家航空基础科学基金的资助.

experiments.

**Key words** Hidden Markov model, neural network, discriminative, back-propagation.

## 1 引言

隐马尔可夫模型(HMM)概率统计方法目前已经成为了语音识别研究的主流方法. 尽管如此,由于训练 HMM 通常所用的 Baum-Welch 算法是基于最大似然的,而并非是基于最大后验概率的,因此 HMM 的区分力是比较弱的,这也是 HMM 语音识别方法的最大弱点<sup>[1]</sup>. 为了增强 HMM 的区分力,Bahl 等人在 80 年代后期提出了一种基于最大互信息(MMI)准则的 HMM 的参数重估训练方法<sup>[2]</sup>. 然而,对于这一有区分力的训练方法,并没有证明训练过程一定收敛,也不能保证 HMM 的参数在经过调整以后仍然服从概率分布约束<sup>[3]</sup>.

本文在多层前向神经网络的框架中实现了 HMM 的前向概率计算. 基于这一框架,本文利用偏导数的反向传播计算方法,通过梯度上升的优化过程来实现互信息的最大化,从而对 HMM 进行有区分力的训练. 这一训练方法被称之为 HMM 的反向传播训练方法. 此外,本文还设计了一个用以实现这一训练方法的在数值计算上具有强鲁棒性的算法. 实验结果表明这一算法要明显地优于基于 MMI 准则的参数重估算法.

## 2 HMM 的多层前向神经网络框架

对于给定的观测符号序列  $O=(o_1 o_2 \cdots o_T)$ ,包含  $N$  个状态、每个状态可输出  $M$  个观测符号的离散 HMM  $\lambda=(A, B, \pi)$  的前向概率可由以下公式递推计算<sup>[4]</sup>,

$$\alpha_j(1) = \pi_j b_j(o_1), \quad 1 \leq j \leq N, \quad (1)$$

$$\alpha_j(t) = \sum_{i=1}^N \alpha_i(t-1) a_{ij} b_j(o_t), \quad 1 \leq j \leq N, \quad 2 \leq t \leq T. \quad (2)$$

如果令

$$w_{ij}(t) = \begin{cases} \delta_{ij} b_j(o_1), & t = 1, \\ a_{ij} b_j(o_t), & 2 \leq t \leq T, \end{cases} \quad 1 \leq i, j \leq N, \quad (3)$$

再令

$$\alpha_j(0) = \pi_j, 1 \leq j \leq N, \quad (4)$$

则前向概率的递推计算可用以下公式统一表示为

$$\alpha_j(t) = \sum_{i=1}^N \alpha_i(t-1) w_{ij}(t), \quad 1 \leq j \leq N, 1 \leq t \leq T. \quad (5)$$

公式(5)表明前向概率的计算可以通过图 1 所示的前向神经网络来实现. 这个网络包含  $T+1$  层神经元节点,每一层的节点数均为  $N$ ,其中的每一个节点都与 HMM 的一个状态对应. 第  $t-1$  层节点与第  $t$  层节点之间的连接权值为  $w_{ij}(t)$ ,每个节点的活动函数为线



性函数. 网络的输入为 0 时刻的前向概率  $\alpha_j(0)$ , 第  $t$  层上的各个节点的输出为对应的  $t$  时刻的前向概率  $\alpha_j(t)$ , 网络的输出为最后时刻的前向概率  $\alpha_j(T)$ .

### 3 HMM 的反向传播训练方法

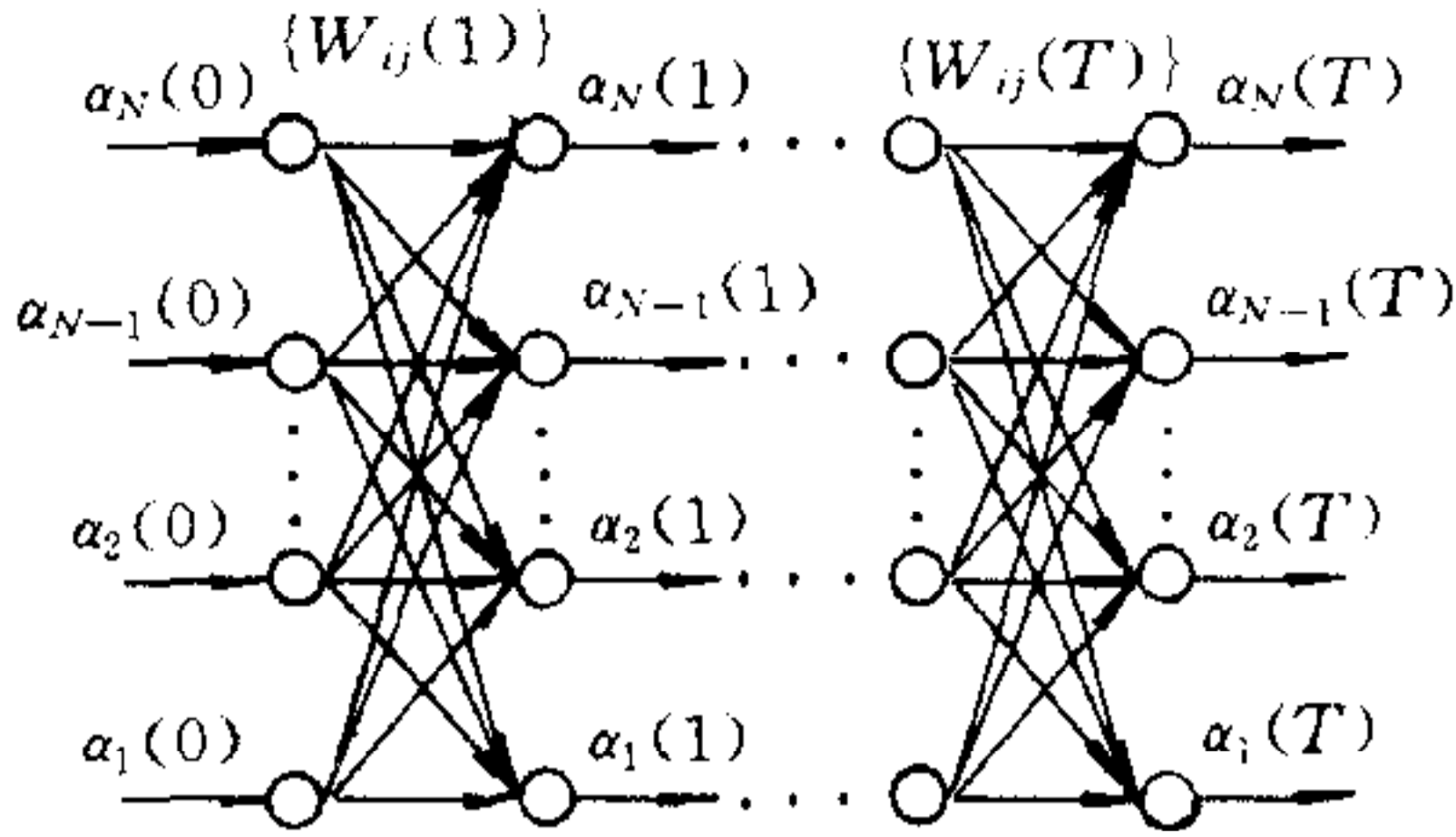


图 1 实现 HMM 前向概率计算的多层前向神经网络

文章的前一部分在多层前向神经网络的框架中实现了 HMM 的前向概率计算. 基于这一框架, 下面将给出 HMM 的反向传播训练方法.

假定观测符号序列  $O = (o_1 o_2 \dots o_T)$  是由 HMM 集合  $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_V\}$  中的  $\lambda_c$  输出的, 那么  $O$  与  $\lambda_c$  之间的互信息定义为<sup>[2]</sup>

$$D = I_{\Lambda}(O, \lambda_c) = \log \frac{P(O, \lambda_c)}{P(O)P(\lambda_c)} = \log P(O | \lambda_c) - \log \sum_{v=1}^V P(O | \lambda_v) P(\lambda_v). \quad (6)$$

对于互信息关于  $\Lambda$  中各个模型的概率参数的偏导数, 有以下计算公式

$$\begin{aligned} \frac{\partial D}{\partial a_{ij}} &= \sum_{t=2}^T \frac{\partial D}{\partial w_{ij}(t)} \frac{\partial w_{ij}(t)}{\partial a_{ij}} = \sum_{t=2}^T \frac{\partial D}{\partial w_{ij}(t)} b_j(o_t), 1 \leq i, j \leq N, \quad (7) \\ \frac{\partial D}{\partial b_{jk}} &= \sum_{t=1}^T \sum_{i=1}^N \frac{\partial D}{\partial w_{ij}(t)} \frac{\partial w_{ij}(t)}{\partial b_j(o_t)} \delta_{o_t, v_k} = \\ &= \frac{\partial D}{\partial w_{jj}(1)} \delta_{o_1, v_k} + \sum_{t=2}^T \sum_{i=1}^N \frac{\partial D}{\partial w_{ij}(t)} a_{ij} \delta_{o_t, v_k}, 1 \leq j \leq N, 1 \leq k \leq M, \quad (8) \end{aligned}$$

而

$$\frac{\partial D}{\partial w_{ij}(t)} = \frac{\partial D}{\partial \alpha_j(t)} \frac{\partial \alpha_j(t)}{\partial w_{ij}(t)} = \alpha_i(t-1) \frac{\partial D}{\partial \alpha_j(t)}, \quad 1 \leq i, j \leq N, 1 \leq t \leq T. \quad (9)$$

至于互信息关于前向概率的偏导数, 有以下表达式

$$\begin{aligned} \frac{\partial D}{\partial \alpha_j(t)} &= \sum_{i=1}^N \frac{\partial D}{\partial \alpha_i(t+1)} \frac{\partial \alpha_i(t+1)}{\partial \alpha_j(t)} = \\ &= \sum_{i=1}^N \frac{\partial D}{\partial \alpha_i(t+1)} w_{ji}(t+1), \quad 1 \leq j \leq N, 1 \leq t \leq T-1. \quad (10) \end{aligned}$$

基于偏导数的以上反向递推计算公式, 我们定义以下反向传播项

$$e_j(t) = \frac{\partial D}{\partial \alpha_j(t)}, \quad 1 \leq j \leq N, 1 \leq t \leq T. \quad (11)$$

将(10)式中的偏导数用相应的反向传播项来代替, 即可得到以下反向传播公式

$$e_j(t) = \sum_{i=1}^N e_i(t+1) w_{ji}(t+1), \quad 1 \leq j \leq N, \quad 1 \leq t \leq T-1. \quad (12)$$

对于  $\Lambda$  中的  $\lambda_v, 1 \leq v \leq V$ , 反向传播项  $e_j(t)$  在  $T$  时刻的值为

$$e_j(T) = \frac{\partial D}{\partial \alpha_j(T)} = \frac{\delta_{v,c} \sum_{u=1}^V P(O | \lambda_u) - P(O | \lambda_c)}{P(O | \lambda_c) \sum_{u=1}^V P(O | \lambda_u)}, \quad 1 \leq j \leq N. \quad (13)$$

将(11)式代入(9)式,然后将结果分别代入(7),(8)两式,即可利用反向传播项来计算互信息关于概率参数的偏导数

$$\frac{\partial D}{\partial a_{ij}} = \sum_{t=2}^T \alpha_i(t-1) b_j(o_t) e_j(t), \quad 1 \leq i, j \leq N, \quad (14)$$

$$\frac{\partial D}{\partial b_{jk}} = \pi_j \delta_{o_1, \nu_k} e_j(1) + \sum_{t=2}^T \sum_{i=1}^N \alpha_i(t-1) a_{ij} \delta_{o_t, \nu_k} e_j(t), \quad 1 \leq j \leq N, 1 \leq k \leq M. \quad (15)$$

利用以上偏导数计算公式,直接对概率参数进行调整,并不能保证经过调整后的参数仍然服从概率分布约束.为此,这里对参数进行以下形式的变换<sup>[5]</sup>

$$a_{ij} = e^{h_{ij}} / \sum_{q=1}^N e^{h_{iq}}, \quad 1 \leq i, j \leq N, \quad (16)$$

$$b_{jk} = e^{g_{jk}} / \sum_{r=1}^M e^{g_{jr}}, \quad 1 \leq j \leq N, 1 \leq k \leq M. \quad (17)$$

互信息  $D$  关于无约束参数  $h_{ij}$  和  $g_{jk}$  的偏导数,可由以下公式来计算

$$\frac{\partial D}{\partial h_{ij}} = \sum_{q=1}^N \sum_{t=2}^T (\delta_{q,j} - a_{ij}) a_{iq} b_q(o_t) \alpha_i(t-1) e_q(t), \quad 1 \leq i, j \leq N, \quad (18)$$

$$\frac{\partial D}{\partial g_{jk}} = b_{jk} \sum_{r=1}^M (\delta_{r,k} - b_{jr}) \left[ \pi_j \delta_{o_1, \nu_r} e_j(1) + \sum_{t=2}^T \sum_{i=1}^N \alpha_i(t-1) a_{ij} \delta_{o_t, \nu_r} e_j(t) \right], \quad (19)$$

$$1 \leq j \leq N, \quad 1 \leq k \leq M.$$

利用由以上这两个公式计算所得的偏导数,沿着梯度上升的方向调整参数  $h_{ij}$  和  $g_{jk}$ ,进而调整参数  $a_{ij}$  和  $b_{jk}$ ,即可使互信息达到最大.这就是 HMM 的反向传播训练方法.然而,由于前向概率计算过程中的下溢问题,直接实现这一方法并不可行.

## 4 反向传播训练方法的实现

在前向概率的计算过程中,为了避免数值计算下溢,要对计算所得的各个时刻的前向概率进行定标.带有定标的前向概率的递推计算公式如下<sup>[4]</sup>:

$$\hat{\alpha}_j(0) = \pi_j, \quad 1 \leq j \leq N, \quad (20)$$

$$\gamma_j(t) = \sum_{i=1}^N \hat{\alpha}_i(t-1) w_{ij}(t), \quad 1 \leq j \leq N, \quad 1 \leq t \leq T, \quad (21)$$

$$c(t) = 1 / \sum_{i=1}^N \gamma_i(t), \quad 1 \leq t \leq T, \quad (22)$$

$$\hat{\alpha}_j(t) = c(t) \gamma_j(t), \quad 1 \leq j \leq N, \quad 1 \leq t \leq T. \quad (23)$$

此时,反向传播项应重新定义为

$$e_j(t) = \frac{\partial D}{\partial \gamma_j(t)}, \quad 1 \leq j \leq N, \quad 1 \leq t \leq T. \quad (24)$$

反向传播项的递推计算公式相应地变为

$$e_j(t-1) = \frac{1}{\sum_{l=1}^N \gamma_l(t-1)} \sum_{i=1}^N e_i(t) [w_{ji}(t) - \gamma_i(t)], \quad 1 \leq j \leq N, 2 \leq t \leq T, \quad (25)$$



$$e_j(T) = \frac{1}{\prod_{t=1}^{T-1} c(t)} \frac{\delta_{v,c} \sum_{u=1}^V P(O|\lambda_u) - P(O|\lambda_c)}{P(O|\lambda_c) \sum_{u=1}^V P(O|\lambda_u)}, 1 \leq j \leq N. \quad (26)$$

同样,基于前向概率的计算公式,利用反向传播项,可以得到以下偏导数计算公式

$$\frac{\partial D}{\partial h_{ij}} = \sum_{q=1}^N \sum_{t=2}^T (\delta_{q,j} - a_{ij}) a_{iq} b_q(o_t) \hat{\alpha}_i(t-1) e_q(t), 1 \leq i, j \leq N, \quad (27)$$

$$\frac{\partial D}{\partial g_{jk}} = b_{jk} \sum_{r=1}^M (\delta_{r,k} - b_{jr}) \left[ \pi_j \delta_{o_1, v_r} e_j(1) + \sum_{t=2}^T \sum_{i=1}^N \hat{\alpha}_i(t-1) a_{ij} \delta_{o_t, v_r} e_j(t) \right],$$

$$1 \leq j \leq N, 1 \leq k \leq M. \quad (28)$$

由于以上这两个偏导数计算公式是基于带有定标的前向概率计算公式的,因此在数值计算上这两个公式具有很强的鲁棒性.利用这两个公式,可以得到以下用于调整  $\Lambda$  中各个模型的概率参数,使互信息达到最大的反向传播训练算法:

- 1) 按照公式(20)~(23)计算各个模型的前向概率  $\hat{\alpha}_j(t)$ .
- 2) 按照公式(26)和(25)计算各个模型的反向传播项  $e_j(t)$ .
- 3) 按照公式(27)和(28)计算  $D$  关于各个模型的参数  $h_{ij}$  和  $g_{jk}$  的偏导数.
- 4) 利用计算所得的偏导数,按照公式

$$\Delta x(n) = \eta \frac{\partial D}{\partial x(n-1)} + \zeta \Delta x(n-1) \quad (29)$$

调整各个模型的参数  $h_{ij}$  和  $g_{jk}$ , 然后按照公式(16)和(17)调整参数  $a_{ij}$  和  $b_{jk}$ . (29)式中  $x$  表示要调整的参数,  $n$  是迭代次数,  $\eta$  和  $\zeta$  表示学习速率和动量.

5) 将参数调整后的新的互信息与原有的互信息进行比较,判断其是否会随着参数的调整而进一步增大. 如果将进一步增大,则返回 1) 进行下一次迭代,否则终止整个迭代过程.

## 5 反向传播训练方法的实验结果

下面通过两个语音识别实验,对本文提出的 HMM 的反向传播训练方法作出评价.

第一个实验是识别特定说话人的 0~9 这 10 个孤立数字语音. 每一个数字都让同一个人发音 30 次, 11kHz 的频率采样. 采集的语音信号中一半用于 HMM 的训练, 另一半用于 HMM 的测试. 对各个语音信号进行 12 阶的 LPC 前端谱分析<sup>[4]</sup>, 对产生的每一个谱特征向量都用一个 64 码字的码书进行矢量量化, 以得到各个语音信号的观测符号序列.

每一个数字都用一个包含 5 个状态、64 个观测符号的 HMM 来建模. 首先用 Baum-Welch 算法来训练这些 HMM, 然后把经过训练的 HMM 作为初始模型, 用本文提出的反向传播训练算法再进行训练. 为了便于比较, 对于用 Baum-Welch 算法训练过的初始 HMM, 也用基于 MMI 准则的参数重估算法再进行训练. 表 1 列出了以上这三组通过不同的训练算法所获得的 HMM 的语音识别性能测试结果.

从测试结果来看, 反向传播训练算法使 HMM 的区分力有了更大程度的增强. 在用反向传播训练算法所进行的再训练过程中, 共进行了 34,589 次迭代, 在 SUN ULTRA 600 工作站上耗时 1h23min. 在基于 MMI 准则的参数重估再训练过程中, 共进行了 36,712 次迭代, 耗时 1h31min. 虽然反向传播训练算法的一次迭代过程的计算量, 与基于

MMI 准则的重估算法的一次迭代过程的计算量基本相当,但是反向传播训练过程的收敛速度通常比较快,因此使用反向传播训练算法往往会耗费较少的时间.

表 1 三种训练算法所产生的三组 HMM 的数字语音识别性能测试结果

	Baum-Welch 算法	基于 MMI 准则的重估算法	反向传播训练算法
误识率	4%(6/150)	2%(3/150)	0.67%(1/150)
平均互信息	-3.852	-1.147	-0.903

第二个实验是识别非特定说话人的 50 个常用的孤立汉语有调音节语音. 这 50 个音节包括(括号内是声调):“ni (3)”和“li (3)”、“qi (3)”和“qing (3)”、“shang (4)”和“xiang (4)”这三对容易混淆的音节,此外还包括不易识别的轻声音节“de ()”和“le ()”. 这 50 个音节由 10 个人(男女各 5 人)各发音 10 次,11kHz 的频率采样,采集的 5000 个语音信号用于 HMM 的训练. 另外 10 个人(男女各 5 人)对这 50 个音节各发音 10 次,同样的频率采样,这 5000 个语音信号用于 HMM 的测试. 与第一个实验相同,用三种训练算法产生三组 HMM,对它们进行测试. 表 2 列出了测试结果.

表 2 三种训练算法所产生的三组 HMM 的音节语音识别性能测试结果

	Baum-Welch 算法	基于 MMI 准则的重估算法	反向传播训练算法
误识率	15.66%(783/5000)	6.34%(317/5000)	3.70%(185/5000)
平均互信息	-14.852	-3.008	-1.526

这些测试结果表明仍然是反向传播训练算法使 HMM 的区分力有了更大程度的增强. 在同样的工作站上完成以上 HMM 的各个训练过程. 在用反向传播训练算法所进行的再训练过程中,共进行了 1,241,437 次迭代,耗时 48h39min. 在用基于 MMI 准则的参数重估算法所进行的再训练过程中,共进行了 1,264,108 次迭代,耗时 50h17min. 使用反向传播训练算法的再训练过程仍然耗时较少.

## 6 结束语

反向传播训练方法是以 MMI 准则为训练准则的,因此它是 HMM 的一种有区分力的训练方法. 此外,这一训练方法还具有以下优点:利用偏导数的反向传播计算方法,通过梯度上升的优化过程来实现互信息的最大化;通过变换将 HMM 的参数所必须服从的概率分布约束条件松弛化. 因此,反向传播训练方法很好地克服了基于 MMI 准则的参数重估训练方法的两个缺点——训练迭代过程不能保证收敛和经过调整的 HMM 的参数不能保证仍然服从概率分布约束. 实验结果还表明:反向传播训练方法可使 HMM 的区分力有更大程度的增强;使用反向传播训练方法的 HMM 的再训练过程往往耗时较少.

然而,反向传播训练方法的计算量相当大,也不能保证所求得的解一定是全局最优解. 因此,进一步的研究工作应当是寻找更为有效的实现算法,以及探索将这一算法应用于大词汇量、连续语音识别的途径.

## 参 考 文 献

- 1 Rabiner L R, Juang B H. Hidden Markov models for speech recognition-strengths and limitations. *NATO ASI*



*Series*, 1992, **F75**: 3~29

- 2 Bahl L R, Brown P F, de Souza P V *et al.* Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: IEEE ICASSP'86, Tokyo, Japan, 1986: 49~52
- 3 Normandin Y. Maximum mutual information estimation of hidden Markov models. In: Automatic Speech and Speaker Recognition, Norwell, MA: Kluwer Academic Publishers, 1996, 57~81
- 4 Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. In: Proceedings of the IEEE, 1989, **77** (2): 257~285
- 5 Bridle J S. Alpha-nets: A recurrent neural network architecture with a hidden Markov model interpretation. *Speech Communication*, 1990, 9 (1): 83~92

**邓 伟** 1999年10月在西北工业大学计算机科学与工程系获博士学位, 现在苏州大学计算机工程系任教. 研究领域有语音识别, 神经网络, 人工智能. 已发表论文多篇.

**赵荣椿** 西北工业大学计算机科学与工程系教授, 博士生导师. 研究领域包括信号处理, 图像处理, 模式识别, 计算机视觉.

(上接第 491 页)

序号	项目名称	主要内容	时间	人数	会期(天)	地点	联系人	备注
19	第十一届全国遥测遥控技术年会	二十一世纪遥测、遥感、遥控技术发展展望; 多目标综合测量; 分包遥测、遥感、遥控技术及设备研制; 高码率调制、解调、同步处理设备研制; 高精度 GNSS 外测处理技术; 快速实时数据处理技术; 测控中多数据流、多协议技术; 遥测、遥感、遥控可显示技术; 动态仿真、自动化检测技术; 遥测、遥感、遥控新设备, 新工艺, 新系统	3 季度	80	5	成都或绵阳	朱成勤 北京 9200 信箱 74 分箱 邮编: 100076 电话: 68382508 传真: 98382304	由遥测专业委员会与中国宇航学会遥测专业委员会联合举办
20	全国第五届低成本自动化技术学术交流会	低成本自动化技术的发展、应用、典型事例的介绍等	4 月	100	3	云南丽江	张振华 北京 919 信箱 邮编: 100071 电话: 63812255-3382	由应用专业委员会和中国金属学会冶金自动化专业委员会联合举办
21	全国第四届 DCS、PLC 和现场总线学术交流会	DCS、PLC、现场总线的应用与发展	3 季度	100	4	浙江舟山	同上	同上

(下转第 546 页)