第 32 卷 第 20 期 计 算 机 工 程 2006 年 10 月
Vol.32 № 20 Computer Engineering October 2006
·人工智能及识别技术· 文章编号：1000—3428(2006)20—0203—02 文献标识码：A 中图分类号：TP391.42;TN912.34

应用 MAP 方差估计的话者自适应训练方法

黄盈椿[1]，王欢良[2]，冯 涛[2]
(1. 中国科学院电子学研究所，北京 100080；2. 哈尔滨工业大学计算机科学与技术学院，哈尔滨 150001)

摘 要：近年来话者自适应训练(SAT)方法日益受到重视。然而在实际中此方法通常因为部分方差的估计失误而导致识别性能下降。该文提出了一种应用最大后验概率(MAP)估计方差的全新 SAT 方法，它能够根据后验概率动态地调整模型的方差，从而解决上述问题。在 Switchboard 数据库上的实验显示，新方法能够显著地提高识别性能，并且有效地提升系统的稳定性。

关键词：语音识别；话者自适应；话者自适应训练；MAP

Speaker Adaptive Training of Appling MAP Estimation for Covariance

HUANG Yingchun[1], WANG Huanliang[2], FENG Tao[2]
(1. Institute of Electronics, Chinese Academy of Sciences, Beijing 100080;
2. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

【Abstract】 Recently there has been a growing interest in speaker adaptive training(SAT). However, errors can often arise when estimating covariance matrices in the original SAT framework due to the lack of observations in some Gauss components. This paper presents a novel approach which applies maximum a posteriori (MAP) covariance-estimating into original SAT. Experimental results in Switchboard corpus demonstrate that the proposed method can deliver significant reductions in word error rate (WER) and raise the robustness of SAT process.

【Key words】 Speech recognition; Speaker adaptation; Speaker adaptive training(SAT); Maximum a posteriori(MAP)

近些年来，非特定人大词表连续语音识别研究取得了长足的进展，但要将该识别系统应用到实际中还有许多困难需要解决。不同话者之间的差异性是造成这些困难的主要原因[1]。因此，若要基于传统方法训练一个较好的话者无关(Speaker Independent, SI)模型，就需要采集数量足够多的来自不同说话人的语音数据。对于大词表识别系统来说，这种方法所需训练样本的数量往往是巨大的，训练所需的时间也是无法忍受的。1996 年Anastasakos等人提出的SAT方法[2,5]则有效地缓解了系统性能对训练样本总量和话者覆盖面的严重依赖。本文介绍与MLLR(Maximum Likelihood Linear Regression)自适应[3]相结合的SAT方法，进一步指出该方法中存在着方差估计的缺陷并提出一种全新的应用MAP准则[4]估计方差的改进方案，从而解决了原始SAT中存在的问题。

1 SAT 算法简述

现实生活中，一个语音序列的产生可抽象成两个阶段：第 1 阶段产生了只包含语意信息的初始语音序列，该阶段独立于任何说话人，并且可以用HMM来建模描述；第 2 阶段加入了当前话者的语音特性，此过程可抽象成将第 1 阶段产生的初始语音序列通过一个含有当前话者信息的滤波器[2]。MLLR自适应正体现了上述思想，它采用线性变换来刻画当前话者的语音特征，通过将SI模型中的均值经线性变换后生成话者自适应（Speaker Adaptation, SA）模型，达到利用少量数据获得稳健的话者相关模型的目的。SAT方法则进一步将MLLR技术融入到模型训练中。假设训练集包含R个说话人，每人有T_R长度的样本序列 $O^{(r)}=(o_1^r,\ldots,o_{T_r}^r),1\le r\le R$。若采用传统的SI模型训练方式，其极大似然下准则函数可以表示成：

$$\overline{\lambda}=\arg\max_{\lambda} P(O\mid\lambda)=\arg\max_{\lambda}\prod_{r=1}^{R}P(O^{(r)}\mid\lambda) \tag{1}$$

这种训练方法忽略了不同说话人或外部环境对语音数据造成的偏差，不可避免地使得估计出的 SI 模型与训练人之间存在一定程度的耦合，当测试集与训练集不匹配时会导致自适应后识别效果不理想。而 SAT 方法则充分考虑到这一问题，其目标函数变为

$$(\overline{\lambda},\overline{G})=\arg\max_{(\lambda,G)}\prod_{r=1}^{R}P(O^{(r)}\mid G^{(r)}\lambda) \tag{2}$$

其中，$\overline{G}=(\overline{G}^{(1)},\ldots,\overline{G}^{(R)})$ 为训练集中不同话者 MLLR 矩阵的集合，$\overline{G}^{(r)}\overline{\lambda}$ 表示经过 MLLR 变换后生成的 SA 模型。从式(2)看出，SAT 方法在训练阶段将参数估计的对象由原先的 SI 模型转变为 SA 模型，在模型空间中补偿了各个话者的语音特征，间接地消除了 SAT 模型与训练话者之间的耦合，使得模型更具有话者无关特性。

2 SAT 改进方案

SAT方法消除了说话人因素造成的偏差，使得样本的特征空间更加紧凑，在SAT模型中分布将变得更加尖锐，从数值上看即方差将大幅度地降低。理论上说，一个模型的分布越尖锐，该模型就越能快速而准确地获取识别结果。然而通常的训练过程中，落到每个高斯元上的样本数往往是不均匀的，而SAT模型的尖锐分布更加重了这种不均匀性。进一步

作者简介：黄盈椿(1981－)，男，硕士生，主研方向：模式识别，信号处理；王欢良，博士生、讲师；冯 涛，博士、副教授

收稿日期：2006-01-03 **E-mail：**huangyingc@mails.gucas.ac.cn

发现，SAT中更新方差表现出很强的不稳定性。具体表现在：当方差因为观测样本不足而被高估时，它与SAT即将获得的尖锐分布相抵触；而方差被低估时，其过低的数值又将极大地影响该高斯元的输出概率，两方面的因素都直接地导致最后识别效果的下降。针对这一问题，本文尝试将MAP准则引入到SAT方差的更新过程中，即在训练阶段将考虑方差先验分布的影响。由此得到基于MAP准则的SAT改进方法，则其目标函数变为[4]

$$(\overline{\lambda},\overline{G}) = \arg\max_{(\lambda,G)} \prod_{r=1}^{R} P(O^{(r)} \mid G^{(r)}\lambda) \cdot g(\lambda) \tag{3}$$

式(3)中 $g(\lambda)$ 为方差的先验分布，其分布形式为正态维希特分布(Normal-Wishart)。文献[4]中证明，当训练样本数量趋于无穷时，基于 MAP 与基于 ML 准则得出的估计是一致的。借助最大化(Expectation-Maximization, EM)算法导出改进 SAT 的 Q-function 为

$$Q_a(\theta,\overline{\theta}) = \sum_{r,t,k}^{R,T_r,K} \gamma_k^{(r)}(t) \cdot \log N(o_t^{(r)} \mid \overline{A}^{(r)}\overline{\mu}_k^{(r)} + \overline{\beta}^{(r)}, \overline{\Sigma}_k) \cdot g(\overline{\Sigma}_k) \tag{4}$$

其中，$\gamma_k^{(r)}(t)$ 为 t 时刻训练样本落在第 k 个高斯元上的后验概率。

算法的第 1 步对式(4)中的 $\overline{G}^{(r)}$ 求导，得到每个话者的 MLLR 矩阵。由于该步和 MLLR 自适应完全相同因此不在此陈述，参看文献[3]。第 2 步依照 EM 算法的步骤将更新后的变换矩阵代入(4)式来求解模型均值，推出其更新公式为

$$\overline{\mu}_k = \left(\sum_{r,t}^{R,T_r} \gamma_k^{(r)}(t) \overline{A}^{(r)T} \Sigma_k^{-1} \overline{A}^{(r)} \right)^{-1} \left(\sum_{r,t}^{R,T_r} \gamma_k^{(r)} \overline{A}^{(r)T} \Sigma_k^{-1} (o^{(r)}(t) - \overline{\beta}^{(r)}) \right) \tag{5}$$

同理，第 3 步将更新后的变换矩阵和均值代入式(4)，可求出模型协方差矩阵的更新公式。因为实验中协方差矩阵选用对角阵，故可以简化为

$$\overline{\sigma}_{kl}^2 = \left(\tau_k + \sum_{r,t}^{R,T_r} \gamma_k^{(r)} \right)^{-1} \left(\tau_k \cdot \sigma_{kl}^2 + \sum_{r,t}^{R,T_r} \gamma_k^{(r)} (o_l^{(r)}(t) - \overline{\mu}_{kl}^{(r)})^2 \right) \tag{6}$$

式(6)中 $\overline{\mu}_{kl}^{(r)}$ 为相应 SA 模型均值的第 L 维元素。式(6)中引入了新的估计量 τ_k，它是方差先验分布的一个参数，推出更新公式为

$$\overline{\tau}_k = \tau_k + \sum_{r,t}^{R,T_r} \gamma_k^{(r)} \tag{7}$$

完成以上 4 步即结束一次迭代。若一次迭代结束后目标函数不收敛，则返回第 1 步继续迭代直到收敛为止。从方差的更新式(6)看出，该计算结果等效为标准 SAT 方差更新前后的一个插值。当训练样本落在某个高斯元的后验概率较小时，即认为该高斯元上训练不充分，方差将自动偏向更新前的初值；反之，偏向标准 SAT 的估计结果。实验证明，改进 SAT 可以根据后验概率动态修正模型的方差，有效地修正标准 SAT 中的缺陷。

3 实验结果

3.1 实验数据及设置

实验平台是基于高斯混合分布的英语音节识别系统，识别单元为约 10 万个英文 Tri-Phone，平均每个 Tri-Phone 由 5 个状态描述。所有的训练数据均采自 Switchboard Mini-train 数据库，该数据库收集了大量自然生成的电话对话，数据采样率为 8KHz，特征参数为 39 维 PLP(PLP_E_D_A)。完整的 Mini-train 包含了 23 小时的训练数据，由 674 个说话人约 32 000 句英语对话组成。测试集分别选用 NIST Eval2000 与 RT03s。其中 Eval2000 包含 40 个说话人，约 2 000 句对话；RT03s 包含 72 个说话人，约 4 000 句对话。

3.2 Mini-train 实验

本次实验所用的初始模型为两个经 Mini-Train 充分训练的传统 SI 模型，其高斯混合数分别为 8 和 12。经过改进 SAT 算法的一次迭代后，将它们各步更新后的结果列出，并与标准 SAT 方法的结果相比较，如表 1、表 2 所示。由于 SAT 算法的性能体现在自适应之后，因此自适应前的结果将不再列出(表中 AFTER_M 表示仅完成了更新均值一步；AFTER_MV 表示更新了均值和方差，结束一次迭代)。

表 1 8 高斯元模型上标准 SAT 和改进 SAT 非监督自适应后识别错误率(WER)

迭代步数	标准 SAT		改进 SAT	
	Eval2000	RT03s	Eval2000	RT03s
SI	38.8%	43.6%	38.8%	43.6%
AFTER_M	37.9%	42.9%	37.9%	42.9%
AFTER_MV	38.6%	43.8%	37.6%	42.5%

表 2 12 高斯元模型上标准 SAT 和改进 SAT 非监督自适应后识别错误率(WER)

迭代步数	标准 SAT		改进 SAT	
	Eval2000	RT03s	Eval2000	RT03s
SI	36.1%	42.2%	36.1%	42.2%
AFTER_M	35.5%	41.2%	35.5%	41.2%
AFTER_MV	36.0%	42.0%	35.3%	40.7%

从表 1、表 2 中看出，在 8 高斯模型中，应用标准 SAT 更新方差后，Eval2000 上的识别错误率由 37.9%上升到 38.6%；RT03s 则由 42.9%上升为 43.8%，在另一个 12 高斯模型中应用标准 SAT 也表现出同样的上升趋势。这些数据充分说明了标准 SAT 更新方差的敏感性和不稳定性。而改进后的 SAT 方法基于 MAP 从本质上消除了该问题。在 8 高斯模型中，应用改进的 SAT 更新模型方差前后，Eval2000 上的 WER 由 37.9%降为 37.6%；RT03s 由 42.9%降为 42.5%。而在 12 高斯模型上，Eval2000 上的 WER 由 35.5%降为 35.3%；RT03s 上由 41.2%降为 40.7%。可见与标准 SAT 相比，改进的 SAT 方法稳步地提升系统的性能。在改进的 SAT 方法的初始，需要为统计量 τ_k 提供一个初始值，该值的选取直接影响着后续 MAP 插值的效果。本次实验中选取出现机率最大的后验概率值作为其初始值，取得了很好的效果。

另外，将改进 SAT 方法再次应用于 12 高斯的模型中，经 3 次迭代后模型参数趋于收敛。用 Eval2000 和 RT03s 分别进行非监督自适应后得出的测试结果如表 3 和表 4 所示。

表 3、表 4 的数据显示，在 Eval2000 和 RT03s 中，改进 SAT 最高获得了 8.95%和 8.49%的 ERR 的提升。甚至在 50 句话自适应的 RT03s 测试集上，改进 SAT 将 ERR 提高了 5.05%，明显优于 MLLR 所贡献的 3.44%的 ERR。当取 20 句话做自适应时，改进 SAT 在 Eval2000 和 RT03s 两个测试集上分别将 ERR 提升了 3.16%和 3.9%(除去 MLLR 的影响)；当取 50 句话做自适应时，两个测试集的 ERR 分别提升了 3.42%和 4.05%。虽然两个测试集的结果不完全相同，但改进 SAT 方法所表现出的性能是一致的。改进 SAT 方法不但继承了原有 SAT 的优点，而且在迭代过程中始终保持着很强的稳定性。

(下转第 212 页)