ISSN 1000-0054 清华大学学报 (自然科学版) 2004 年 第 44 卷 第 6 期 29/37
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2004, Vol. 44, No. 6 829-832

基于相关子空间本征音分析的MAP快速自适应

罗 骏, 欧智坚, 王作英

(清华大学 电子工程系, 北京 100084)

摘 要: 本征音自适应是一种快速自适应算法, 它根据对说话人矢量全空间的本征分析指导参数更新。该文提出一种基于子空间分析的本征音自适应算法, 并且不同于一般本征音自适应采用最大似然准则的做法, 该算法用最大后验准则以更好地估计参数。实验证明, 在仅有 1 句自适应数据的情况下它即能取得 6.45% 的相对误识率下降, 自适应速度远快于传统的最大后验方法, 也不存在最大似然线性回归方法在极少数据量情况下反而造成系统识别性能下降的现象。结果表明该方法并不明显依赖相关子空间的划分数量, 是一种稳健的自适应方法。

关键词: 信息处理; 语音识别; 快速自适应; 本征音; 最大似然; 最大后验; 相关子空间

中图分类号: TP 391 **文献标识码**: A

文章编号: 1000-0054(2004)06-0829-04

Eigenvoice-based MAP fast adaptation in correlation subspaces

LUO Jun, OU Zhijian, WANG Zuoying

(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: The eigenvoice approach is an efficient method for rapid speaker adaptation which directs the adaptation according to an analysis of the full speaker vector space. This article describes an algorithm for eigenspace-based adaptation restricting eigenvoices in clustered subspaces, with the maximum-likelihood (ML) criterion replaced with the maximum a posteriori (MAP) criterion for better parameter estimation. Experiments show that even with only one sentence of adaptation data, this algorithm had a 6.45% relative error ratio reduction. This method overcomes the instability of the ML linear-regression method with limited data and is much faster than the traditional MAP method. The algorithm is also not highly dependent on the number of subspace divisions, so it is a very robust adaptation algorithm.

Key words: information processing; speaker recognition; fast adaptation; eigenvoice; maximum likelihood (ML); maximum a posteriori (MAP); correlation subspaces

在语音识别领域, 快速自适应技术越来越为大家所重视。自适应技术的目标是在训练语料有限的情况下快速提升识别系统的性能。在训练语料足够充分前提下, 说话人相关(speaker dependent, SD)的识别系统总要大大好于说话人无关(speaker independent, SI)系统的表现, 然而在训练数据不足时, 传统训练方法无法有效保证系统性能的稳定提高。自适应方法的实质在于利用少量特定人的训练数据及从说话人无关训练集上得到的统计信息对用户参数进行有效的估计。自适应的结果是得到说话人自适应(speaker adaptation, SA)码本[1]。

常用的自适应方法包括最大后验(maximum a posteriori, MAP)[1,2]和最大似然线性回归(maximum likelihood linear regression, MLLR)[3]两大类, 近年来本征音方法被证明是一种新的行之有效的快速自适应算法, 不仅在孤立词识别系统上获得成功[4], 而且已经在大词汇量连续语音识别中得到应用[5]。一般的本征音自适应是对整个说话人矢量空间求解, 而本文提出的新算法则是对每个独立的子空间进行本征音分析, 并且用MAP准则取代最大似然(maximum likelihood, ML)准则实现更稳健的参数估计。实验结果表明, 新算法只需要一句适应数据就可以较大地提高识别率, 具有良好的快速自适应性能。

1 本征音算法的基本原理

本征音算法首先为每个 SD 模型构造一个说话人矢量, 该矢量包括描述 SD 模型的所有均值矢量, 将其维数记为D。假设在一个 SD 模型中共有M 个状态, 每个状态用N 维的矢量表示, 记为$\mu_1, \cdots, \mu_N$,

收稿日期: 2003-07-01
基金项目: 国家"八六三"高技术项目 (863-306-ZD03-01-2)
作者简介: 罗骏(1978-), 男(汉), 浙江, 博士研究生。
通讯联系人: 王作英, 教授,
E-mail: wzy-dee@mail.tsinghua.edu.cn

则说话人矢量可以表示为$(\mu_1, \cdots, \mu_M)$，并且有$D = M \cdot N$维；第二步对于训练集中的所有说话人矢量统计均值和协方差矩阵，将均值记为$e(0)$，其物理含义是SI模型的参数描述(事实上这样统计得到的$e(0)$与一般意义的SI码本有一定的区别，通常SI码本的训练方式是使得最终所有数据总体上对目标码本收敛而不是每个子训练集收敛取平均[5]。但是在训练数据足够多的情况下两者的识别效果区别不大，实验结果也证明对后一种方式训练得到的码本进行自适应同样可以取得良好的效果，因此在后面的讨论中忽略两者的差别)。随后利用主成分分析(PCA)[6]方法得到最大的K个特征值和相应的特征矢量，这些特征矢量被称之为"本征音"，记为$e(i)$，$i = 1, \cdots, K$，一般K的取值远小于D。SA模型的说话人矢量可以用SI模型和本征音的线性组合表出，即

$$p = e(0) + \sum_{i=1}^{k} w(i) \cdot e(i). \tag{1}$$

一般的本征音自适应在ML框架下求解参数集合$w(i)$，$i = 1, \cdots, K$即可完成自适应过程。

2 MAP框架下基于相关子空间分析的参数估计方法

当本征音自适应算法应用到大词汇量连续语音识别时，描述模型参数的说话人矢量维数往往很高(单Gauss情况下维数就高达几万)，与之相比，用于估计本征音的说话人矢量个数则要少得多，这有时会导致参数估计的失真。

一种解决的方案是利用说话人相关的MLLR变换矩阵取代说话人矢量[7]，本文则试图对状态划分子空间。首先，所有的状态根据相互之间的相似性(本文采用简单的欧式距离度量状态的相似性)被聚成若干类，从而说话人矢量所在的高维空间被维数小得多的子空间所代替，这些子空间被称为相关子空间。本征音分析仅在各个相关子空间内分别进行，这意味着不同相关子空间的相关性被忽略。进一步，为了避免这种忽略导致性能的不稳定，同时也为了引入更多的先验信息，参数估计的过程在MAP框架下进行，即实现MAP方法与本征音方法的组合，称之为"基于相关子空间本征音分析的MAP快速自适应"算法，实验证明这是一种稳健的参数估计方法。

假设所有语音状态被划分成H个相关子空间，对于每个相关子空间以说话人矢量p_h表示状态参数集合，h表示相关子空间索引。与全空间本征音分析的方法类似，p_h可表示为

$$p_h = e_h(0) + \sum_{i=1}^{k_h} w_h(i) \cdot e_h(i). \tag{2}$$

其中：$e_h(0)$表示SI模型中对应该空间的部分矢量，$e_h(i)$，$i = 1, \cdots, K_h$是K_h个经过PCA得到的对应子空间的本征音。

引入MAP框架取代ML准则进行参数估计，参数估计式为

$$\hat{p}_h = \arg\max_{p_h} P_0(p_h) \cdot P(X_h | p_h), \tag{3}$$

其中：X_h表示该子空间的训练数据集合，$\hat{p}_h$是对参数集合的最优估计。

根据式(2)计算$\hat{p}_h$等价于求解组合系数$w_h(i)$，因此目标函数式(3)改写为

$$(w_h(1), \cdots, w_h(K_h)) = \arg\max[f_{h1} + f_{h2}], \tag{4}$$

其中

$$f_{h1} = -\frac{1}{2}\Big[\sum_{i=1}^{K_h} w_h(i) \cdot e_h(i)\Big]^{\mathrm{T}} C_h^{-1} \Big[\sum_{i=1}^{K_h} w_h(i) \cdot e_h(i)\Big] = -\frac{1}{2}\sum_{i=1}^{K_h} \frac{(w_h(i))^2}{\lambda_h(i)}, \tag{5}$$

$$f_{h2} = -\frac{1}{2}\sum_{s \in \Psi(h)} \sum_{t \in \phi(h,s)} \Big[o_t - e_h^s(0) - \sum_{i=1}^{K_h} w_h(i) \cdot e_h^s(i)\Big]^{\mathrm{T}} C_s^{-1} \cdot \Big[o_t - e_h^s(0) - \sum_{i=1}^{K_h} w_h(i) \cdot e_h^s(i)\Big], \tag{6}$$

其中：C_h为相关子空间h内说话人矢量的统计协方差矩阵；$\lambda_h(i)$，$i = 1, \cdots, K_h$表示该矩阵的前K_h个特征值；$e_h(i)$表示对应于特征值$\lambda_h(i)$的特征向量，也就是该子空间的第i个本征音；$\Psi(h)$为第h个相关子空间内的子状态集合；C_s为状态s的协方差矩阵；$e_h^s(i)$为相关子空间h第i个本征音矢量中与状态s相对应的部分矢量；$\phi(h, s)$为相关子空间h中利用viterbi分割方法对应到状态s的观测序列集合。

为了极大化目标函数，令目标函数对参数$w_h(i)$，$i = 1, \cdots, K$的偏导为零，有下式成立

$$\frac{\partial(f_{h1} + f_{h2})}{\partial w_h(i)} = 0, \quad i = 1, 2, \cdots, K_h. \tag{7}$$

展开即得求解参数的联立线性方程组：

$$\sum_{i=1}^{K_h} w_h(i) \cdot$$

$$\left[\sum_{s\in\Psi(h)}\left\{\sum_{t\in\Phi(h,s)}[e_h^s(i)]^{\mathrm{T}}C_s^{-1}e_h^s(j)\right\}+\delta_{ij}\frac{1}{\lambda_i}\right]=$$
$$\sum_{s\in\Psi(h)}\sum_{t\in\Phi(h,s)}[e_h^s(j)]^{\mathrm{T}}C_s^{-1}[o_t-e_h^s(0)], \quad (8)$$
$$j=1,\cdots,K_h;\qquad \delta_{ij}=\begin{cases}0, & i\neq j\\ 1, & i=j\end{cases}.$$

根据式(8)可以求得待定参数$w_h(i)$, $i=1,\cdots,K_h$, 代入式(2)即可更新语音模型参数。

3　实验及讨论

3.1　实验系统

本次实验只考察本征音技术对声学识别的改进, 因此测试只针对声学模型而没有利用语言层知识, 识别器给出多候选的拼音格结果; 基本的隐含Markov模型(hidden Markov model, HMM)单元为汉语半音节, 输出概率采用单Gauss 的分布形式。

首先利用国家“八六三”高技术计划智能计算机主题办公室提供的男声语音数据训练并提取先验信息, 共取 83 人的数据, 每人在安静环境下录制 650 句话。训练后得到 83 个 SD 码本, 即相当于获得了 83 个说话人矢量, 在进行子空间划分后对每个子空间内的说话人矢量统计均值及协方差矩阵, 均值即视为 SI 码本的一个描述, 即前面推导中的 $e_h(0)$; 对每个子空间的协方差矩阵进行 PCA 处理可求得每个子空间的本征音 $e_h(i)$ 及相应的特征值 $\lambda_h(i)$, $i=1,\cdots,K_h$。

在自适应实验阶段取训练集外的 5 个男声数据, 每人录制 120 句话(2 940 个音节), 时间长度约为 10～15 min。分别利用每个人的前 1、2、3、4、5、10、20、30、40、50、60 句话做自适应, 并用自适应后的码本识别后 60 句话比较无调音节识别率的变化情况。

在实验中所有的状态根据其距离关系聚成包含状态个数不等的类, 相关子空间的个数是可变的, 从而可以根据不同的设置比较子空间划分对识别率提高的影响。

实验的另一个重要参数是每个子空间内本征音个数的确定。由于特征值的大小表征了语音矢量在相应本征方向的变化程度, 而细小的变化可以视为噪声的影响, 因此忽略变化较小的本征方向上的分量从理论上不应该导致性能的明显下降。这一参数对实验结果的影响将在下一部分进行详细的讨论。

3.2　实验结果分析讨论

首先设定实验参数如下: 所有的状态被划分为 52 个相关子空间, 每个子空间设定本征门限 0.01。本征门限的含义是: 选取最大的前若干个特征值及相应的本征矢量, 使得余下的所有特征值的总和小于该门限。

实验结果表明, 基于子空间本征音分析的MAP 自适应算法即便是在自适应数据很少的情况下也能迅速提高系统的性能, 在测试集合内每个测试人的误识率均随自适应数据的增加一致下降。平均的误识率变化与经典MAP、MLLR 方法的实验结果比较如表 1 所示。

表 1　本征音法与经典MAP、MLLR 法的自适应后误识率比较

自适应句数	一候选误识率×100			五候选误识率×100		
	MAP 法	MLLR 法	本征音法	MAP 法	MLLR 法	本征音法
0	30.96	30.96	30.96	9.17	9.17	9.17
1	30.90	103.22	28.96	9.06	101.01	8.81
2	30.75	73.26	28.64	9.05	46.78	8.47
3	30.72	49.03	28.12	8.99	20.44	8.47
4	30.57	40.84	27.99	8.98	14.61	8.31
5	30.55	33.64	27.46	8.94	10.74	8.34
10	30.01	28.34	26.83	8.74	8.38	7.60
20	28.89	26.33	24.97	8.51	7.55	6.86
30	27.42	24.61	23.95	8.03	7.16	6.47
40	26.07	25.49	23.28	7.69	7.50	6.38
50	24.92	24.00	22.84	7.51	6.99	6.39
60	24.01	23.83	21.81	7.20	7.00	6.11

可看到, 当自适应数据从 1 句话逐渐增长到 60 句话, 基于子空间本征音分析的MAP 自适应算法首选误识率由 30.96% 下降到 21.81%, 最终的误识率相对降低 29.55%, 性能稳定地优于传统MAP 和MLLR 方法, 五候选性能也有一致的提升。在数据量较少的情况下, 传统MAP 自适应的效果不明显, 而MLLR 由于不能稳定地估计变换矩阵甚至会造成识别率灾难性地上升, 本征音方法显然更适合需要快速自适应的场合。

图 1 用于比较本征门限对实验结果的影响。相关子空间的数目设为 52, 本征门限分别取 0.01 和 0.1。根据统计, 当门限为 0.01 时 52 个子空间内的本征音总数为 3 720, 而门限为 0.1 时, 总数减少到 1 976 个, 少了将近一半。而从图 1 中可以看到误识率几乎没有变化。因此可以在不丧失计算精度的前提下减少每个子空间的本征音数量, 从而降低计算量和存储空间。

图 2 是比较不同子空间数目设置的结果, 本征门限固定为 0.01。在实验中所有状态分别被聚类成 89、52 和 39 个相关子空间, 尽管由此导致每个子空间的特性大相径庭, 但它们一致地使得误识率得到

下降。因此这种自适应方法并不明显依赖于子空间划分的数量。

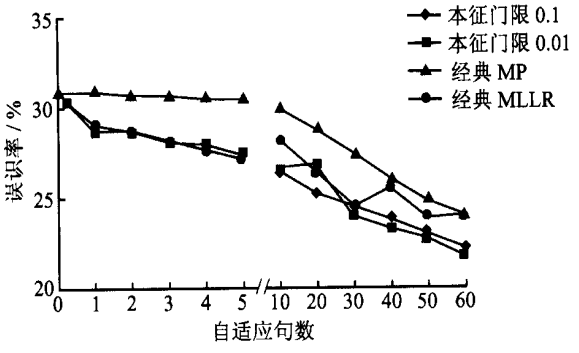

图 1 本征门限对误识率的影响(一候选)

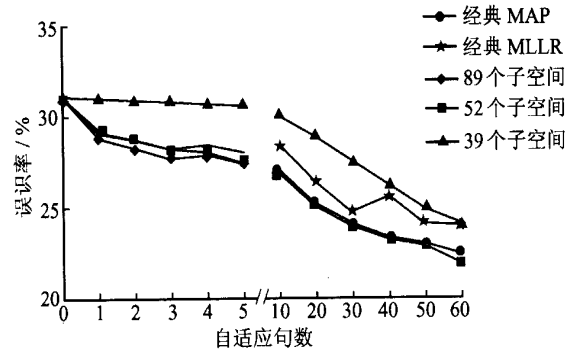

图 2 子空间划分对识别率的影响(一候选)

4 结 论

基于子空间本征音分析的MAP 自适应算法是一种快速稳健的算法,在自适应数据较少的情况下对性能的改善尤其明显。它的另一个显著优点是只需要极少的本征音矢量即可估计出所有参数的分布,大大降低了计算的时间和空间复杂度。该方法与Gauss 混合模型相结合将是大词汇量连续语音识别技术实用化中极有前途的一种方法。

参考文献 (References)

[1] Lee C-H, Lin C-H, Juang B-H. A study on speaker adaptation of the parameters of continuous density hidden Markov models [J]. *IEEE Trans on Signal Processing*, 1991, **39**(4): 806-814.

[2] Chengalvarayan R, LI Deng. A maximum a posteriori approach to speaker adaptation using the trended hidden Markov model [J]. *IEEE Trans on Speech and Audio Processing*, 2001, **9**(5): 549-557.

[3] Lee C-H, Lin C-H, Juang B-H. Speaker adaptation of continuous density HMM's using linear regression [A]. Proc 3rd Int Conf on Spoken Language Processing (ICSLP'94) [C]. Yokohama: IEEE Press, 1994. 451-454.

[4] Kuhn R, Junqua J-C, Nguyen P, et al. Rapid speaker adaptation in eigenvoice space [J]. *IEEE Trans on Speech and Audio Processing*, 2000, **8**(6): 695-707.

[5] Botterweck H. Very fast adaptation for large vocabulary continuous speech *recognition* using eigenvoices [A]. Proc 6th Int Conf on Spoken Language Processing (ICSLP'00) [C]. Piscataway, NJ, USA: IEEE Press, 2000. 354-357.

[6] Jolliffe I T. Principal Component Analysis [M]. Berlin: Springer-Verlag, 1986.

[7] CHEN Kuan-ting, LIAU Wen-wei, WANG Hsin-min, et al. Fast speaker adaptation using eigenspace-based maximum likelihood linear regression [A]. Proc 6th Int Conf on Spoken Language Processing (ICSLP'00) [C]. Piscataway, NJ, USA: IEEE Press, 2000. 742-745.

(上接第 828 页)

5 结 论

对于 H. 264 中树状结构的增强运动预测,应用 SEA 算法,与全搜索性能完全相同,可有效减少运算量,对不同运动程度的图像序列,整像素运动矢量搜索块匹配运算量可减少到全搜索的 1%～ 20%。利用不同大小块运动矢量部分重叠相关性,简单有效地确定运动矢量搜索初始位置,应用两个宽松的条件快速判定目标运动矢量,改进 SEA 算法可在性能损失极小情况下,进一步将运动矢量搜索速度提高 3～ 5 倍,块匹配运算量约为全搜索的 0. 3%～ 4%。

参考文献 (References)

[1] ITU-T Recommendation H. 263. Video Coding for Low Bit Rate Communication [S]. 1998.

[2] ITU-T Recommendation H. 264/ISO/IEC 11496-10. Advanced Video Coding Final Committee Draft, Document JVT F100d2 [S]. 2002.

[3] ITU-T Recommendation H. 262 ISO/IEC 13818-2 (MPEG-II Video) International Standard. Generic Coding of Moving Pictures and Associated Audio Information: Video [S]. 2000.

[4] ISO/IEC 14469-2 (MPEG-IV Visual). Coding of Audio Visual Objects—Part 2: Visual [S]. 1999.

[5] LI Wenhua, Salari E. Successive elimination algorithm for motion estimation [J]. *IEEE Trans Image Process*, 1995, **4**(1): 105-107.

[6] Jung S M, Shin S C, Baik H, et al. New fast successive elimination algorithm [A]. Circuits and Systems, Proc 43rd IEEE Midwest Symp, Vol 2 [C]. 2000. 616-619.

[7] Gao X Q, Duanmu C J, Zou C R. A multilevel successive elimination algorithm for block matching motion estimation [J]. *IEEE Trans Image Process*, 2000, **9**(3): 501-504.

[8] Jung S M, Shin S C, Baik H, et al. Efficient multilevel successive elimination algorithms for block matching motion estimation [J]. *Vision, Image and Signal Processing, IEE Proc*, 2002, **149**(2): 73-84.