

# 最大后验估计和最近邻线性回归 结合的说话人自适应方法

何 磊, 武 健, 方棣棠, 吴文虎

(清华大学计算机科学与技术系, 智能技术与系统国家重点实验语音技术中心, 北京 100084)

**摘 要:** 本文提出一种新的说话人自适应方法: 最大后验 (MAP) 估计与最近邻线性回归 (NNLR) 结合的自适应, 利用模型近邻信息和 MAP 自适应结果, 建立线性回归模型, 对没有自适应数据的模型完成模型调整. 实验证明, NNLN 要优于另一种用于 MAP 自适应框架的模型插值方法: 向量域平滑 (VFS).

**关键词:** 说话人自适应; 最大后验; 向量域平滑

**中图分类号:** TP391 **文献标识码:** A **文章编号:** 0371-2112 (2000) 11-0055-04

## A Novel speaker Adaptation Method based on Map and NNLN

HE Lei, WU Jian, FANG Di-tang, WU Wen-Hu

(Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,  
Department of Computer Science & Technology, Tsinghua University, Beijing 10084, China)

**Abstract:** This paper describes a novel speaker adaptation method that combines maximum a posteriori (MAP) estimation and nearest neighbor linear regression (NNLR). In this scheme, the relationships between speaker independent models and speaker adaptation models are trained by applying the linear regression to neighbor parameters with and without MAP adaptation. Experiments show that the less adaptation data are required in MAP/NNLR adaptation with convergence to SD model held. In addition, experiments prove that NNLN is more efficient than vector field smoothing, (VFS) which is another model interpolation technique used in MAP adaptation frame work.

**Key words:** speaker adaptation; maximum a posteriori; MAP; vector field smoothing VFS.

### 1 引言

近年来, 基于连续概率隐马尔可夫模型 (CDHMM) 的非特定人 (SI, Speaker Independent) 语音识别系统纷纷问世, 并得到了一定的应用. 然而, 研究表明, 其误识率远远高于特定人 (SD, Speaker Dependent) 识别系统. 其原因是由于不同说话人之间的个性所带来的特征空间分布的离散, 从而导致了训练条件和测试条件下的模型差异. 在非特定人的语音识别系统的训练中, 为了达到对不同说话人的适应性, 通常的方法是收集尽可能多的说话人的语料, 使训练语料覆盖范围尽可能大. 显然, 这样的方法汇集了不同说话人的特殊性, 得到的是一个平滑的模型. 因此, 相对一个充分训练的 SD 模型, 基于这样的平滑模型的语音识别系统, 其识别率的降低是难以避免的. 当然, 随着训练语料库的扩大, 对测度集而言, 系统的性能可以逐步地提高, 但随之而来的是系统成本的增加和大数据库训练的效率问题. 因此, 另一个替代的方法, 即所谓的说话人自适应方法, 得到了人们越来越多的重视.

说话人自适应的目的, 就是试图从少量的自适应数据中,

提取与说话人相关的信息, 对 SI 模型参数进行调整, 尽量消除模型差异, 从而得到说话人自适应 (SA, Speaker Adaptation) 模型. 一般地讲, 一种好的说话人自适应方法, 应该满足两个条件: 一是对 SD 模型的渐进性, 即 SA 模型应该优于 SI 模型, 并且在有足够的训练数据后收敛到充分训练后的 SD 模型; 二是自适应速度要快, 即在只有很少的自适应数据的条件下, 仍然可以有较好的自适应效果.

迄今为止, 已经有相当多的说话人自适应方法被证明是有效的. 这些方法大致可分为两类<sup>[2]</sup>: (1) 贝叶斯方法, 也即最大后验 (MAP, Maximum a Posteriori)<sup>[4,5]</sup>方法, 它根据贝叶斯准则, 将 SI 模型信息作为先验分布的估计值纳入判别式, 与自适应数据结合, 从而得到基于模型组合的自适应结果; 在先验分布的描述范围内, 这样的组合在理论上可以得到最大似然意义上的最佳. MAP 方法在理论上可以很好地满足对 SD 模型的渐进性, 实验的结果也证明了这一点. 但由于只有自适应数据对应的模型参数可以得到更新, 其自适应速度通常很慢. 因此, 提高 MAP 自适应效果的关键是如何在保持其对 SD 模

型的渐进性的前提下,提高其自适应速度.目前这方面已有的工作包括 MAP/VFS<sup>[7,8]</sup>和 RMP (Regression-based model prediction)<sup>[1]</sup>方法等等.(2)变换法,其中最具有代表性的是最大似然线性回归(MLLR, Maximum Likelihood Linear Regression)<sup>[3,6]</sup>方法,它假定 SI 模型和 SA 模型之间的差异可以用一系列的线性回归模型来描述,并利用自适应数据来估算对应的回归参数,其最优化目标为最大似然.进而对所有的模型参数用对应的回归模型来完成自适应调整.在 MLLR 自适应中,由于不同的模型分布可以共享同一个回归模型(极端的情况下可以采用一个全局的回归模型来完成自适应),可以得到快速的自适应.但理论上 MLLR 无法象 MAP 一样直接满足对 SD 模型的渐进性;此外,经典的 MLLR 方法没有对不同分布共享一个回归模型的可信度进行量化衡量,这也可能影响到自适应效果.所以,MLLR 的改进需要得到一个高效的模型聚类方法,其聚类结果可以较好地符合同一线性回归模型;此外,必须满足在自适应数据充足的条件下对 SD 模型的收敛要求.

从上面的分析可以看到,MAP 方法和 MLLR 方法具有相当好的互补性.因此一个比较自然的思路就如何将这两种方法的优点结合起来,得到一种更好的自适应方法.本文就是基于这一思路,提出了 MAP 和 NNLR 结合的自适应方法.首先,用 MAP 方法对自适应数据所对应的模型分布完成自适应,得到初始的 SA 模型,然后对没有自适应数据的模型,利用近邻信息(即采用模型距离聚类),建立对应的回归模型,进一步完成自适应.此外,用相关系数来度量回归模型的可信度,避免不可靠的估计.

## 2 MAP 方法

作为一种参数估计方法,MAP 提供了一条途径来把与應用有关的数据以一种最优的方式组合到初始的模型中,由于在训练过程中结合了先验信息,适用于训练数据稀疏的情况,而这正是最大似然(MI, Maximum Likelihood)估计无法克服的问题,因而被广泛的运用到各种自适应环境下.

对于一个混合高斯分布的 CDHMM,给定自适应观察序列  $X = (x_1, \dots, x_T)$ ,对状态  $i$  的第  $k$  个高斯混合成分,MAP 估计式如下<sup>[4]</sup>:

$$\tilde{\mu}_{ik} = (\mu_{ik} + \sum_{t=1}^T c_{ikt} x_t) / (\mu_{ik} + \sum_{t=1}^T c_{ikt}) \quad (1)$$

其中,  $\mu_{ik}$  和  $\tilde{\mu}_{ik}$  分别为自适应前后的模型分布均值,  $\mu_{ik}$  是先验参数和自适应数据之间的相对权重参数;而  $c_{ikt} = P(x_t = x_i, i, ik)$  表示时刻  $t$  在状态  $i$  用第  $k$  个混合成分产生  $x_t$  的概率密度.

从式(1)中可以看到,MAP 方法结合了先验 SI 模型参数和自适应数据两方面的信息.随着自适应数据的增加,  $\sum_{t=1}^T c_{ikt}$  将趋向无穷大,而 MAP 的结果也将收敛到 ML 估计式:

$$\tilde{\mu}_{ik} = \sum_{t=1}^T c_{ikt} x_t / \sum_{t=1}^T c_{ikt} \quad (2)$$

从而直接证明了 MAP 的渐进性.

分布的协方差矩阵也有类似的结论,但由于实验表明,协方差矩阵的自适应效果并不明显,所以本文中的自适应都只针对均值向量.

## 3 MAP/NNLR 方法

如第 1 节所述,MAP 方法只能对有自适应数据的模型完成自适应,而在基元数较多时,要求所有的模型都有自适应数据,将需要相当大的自适应数据量.表 1 是本文采用的识别系统的自适应数据覆盖状况:(识别基元为 417 个汉语无调音节)

表 1 自适应数据覆盖率

自适应数据(句)	20	50	100	250
模型覆盖率(%)	33	53	63	90

从表 1 中可以看到,即使自适应数据达到 250 句,仍然有 10% 的模型在 MAP 方法中无法得到更新;而在实际的应用中,这样大的自适应数据是很难得到的.因此,必须对 MAP 方法进行改进.

在这里,需要完成的是 SI 模型和 SA 模型之间差异的描述方式和对应的参数估计方法的设计;而在前面的讨论中已经提到,在基于变换的自适应方法中,(以 MLLR 为代表),不同的模型分布从 SI 空间到 SA 空间的模型差异可以用同一相对固定的线性回归模型来描述,从而减少了自适应的数据的要求.本文采用了与之类似的思路,假设 SI 空间聚集在一起的模型可以用同一个回归模型映射到 SA 空间.这里的假设一方面来自 VFS 方法的启示;另一方面,对 MAP 自适应结果的统计,也证明了假设的合理性(表 2).

下面介绍输出概率函数为混合高斯分布的 NNLR 方法:

对 SI 空间的模型分布  $ik = (\mu_{ik}, \Sigma_{ik})$ , 设其在 SA 空间的对应分布为  $\tilde{ik} = (\tilde{\mu}_{ik}, \tilde{\Sigma}_{ik})$ . 其中  $\mu_{ik}$ ,  $\Sigma_{ik}$  分别为均值矢量和协方差矩阵.则可以假设存在线性回归模型:

$$\tilde{\mu}_{ik} = B\mu_{ik} + b_o \quad (3)$$

这里  $ik$  满足  $O$  均值的正态分布.

对变换矩阵  $B$  和平移向量  $b_o$  的估计可以采用最小平方估计(LSE, Least Squares Estimation),其最小化目标函数为:

$$J_m^2 = \sum_{m \in A} (\tilde{\mu}_m - B\mu_m - b_o)^T (\mu_m - B\mu_m - b_o) \quad (4)$$

其中  $A$  是  $ik$  在 SI 空间中的近邻集合.这里已经假定集合  $A$  中的模型分布和  $ik$  从 SI 空间到 SA 空间遵循同一个回归模型,对应的回归参数为变换矩阵  $B$  和平移向量  $b_o$ .在大多数情况下,可以得到的有自适应数据的近邻模型是比较有限的.因此,如果  $B$  矩阵中的自由参数太多,可能无法得到合理的估计值.在本文中,假设  $B$  矩阵为对角阵,其对角元素可以用向量  $b = (b^1, b^2, \dots, b^P)^T$  ( $P$  为维数)来表示.这样,对参数的估计,可以分为独立的各维来处理,即采用了一元线性回归模型.于是式(4)成为:

$$J_m^d = \sum_{m \in A} (\tilde{\mu}_m^d - b^d \mu_m^d - b_o^d)^T (\tilde{\mu}_m^d - b^d \mu_m^d - b_o^d) \quad (5)$$

其中,上标  $d$  表示为第  $d$  维,  $d = 1, 2, \dots, P$ ; 以式(5)作为最小化的对象,得到估计式:

$$b^d = \sum_{m \in A} (\mu_m^d - \tilde{\mu}_m^d) (\tilde{\mu}_m^d - \tilde{\mu}_m^d) / \sum_{m \in A} (\mu_m^d - \tilde{\mu}_m^d)^2 \quad (6)$$

$$b_o^d = \tilde{\mu}_m^d - b^d \mu_m^d \quad (7)$$

其中,  $\bar{\mu}_m$  和  $\tilde{\mu}_m$  是集合  $A$  中各分布的总均值:  $\bar{\mu}_m = \sum_{m \in A} \mu_m / M$ ,  $\tilde{\mu}_m = \sum_{m \in A} \tilde{\mu}_m / M$ , 这里的  $M$  是集合  $A$  中的近邻数.

为了验证上面的估计结果的合理性, 可以用相关系数来衡量:

$$d = \frac{(\mu_m^d - \bar{\mu}_m^d)(\tilde{\mu}_m^d - \tilde{\bar{\mu}}_m^d)}{\sqrt{\sum_{m \in A} (\mu_m^d - \bar{\mu}_m^d)^2 \sum_{m \in A} (\tilde{\mu}_m^d - \tilde{\bar{\mu}}_m^d)^2}} \quad (8)$$

这里  $d$  为 0 到 1 之间的数, 越接近 1, 说明线性回归的可信度越高. 对某一分布估计的合理性衡量, 可以用各维相关系数的平均值来表示:

$$d = \frac{\sum_{d=1}^p d_j / p}{p} \quad (9)$$

对得到比较充分的 MAP 自适应的 SA 模型的统计表明, 大多数的分布和其近邻可以很好的符合同一回归模型. 表 2 是 16 个高斯混合数为 16 时的统计结果 (模型近邻数取 20):

表 2 回归模型的置信度统计

>	0.6	0.7	0.8	0.9
满足条件的模型分布百分比	98	96	92	82

因此, 给定初始 SI 模型和对应的初始 SA 模型 (经过了 MAP 自适应), 对没有自适应数据的模型分布, 可以从有自适应数据的近邻分布中训练出对应的回归模型, 从而完成相应的自适应, 得到最后的 SA 模型. 这里, 为了避免不合理的估计, 可以对相关系数加上域值限制, 只有满足条件的回归模型可以被接受.

MAP 与 NNLR 结合的自适应方法具体实现如图 1:

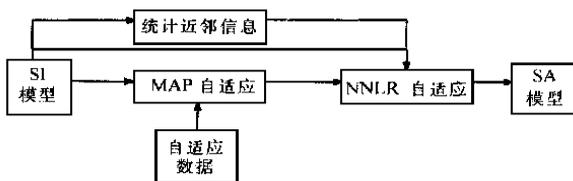


图 1 MAP/NNLR 自适应框图

其中, 有两点需要注意: 一是对于高斯混合数比较大的模型, 每次自适应都重新统计模型近邻信息, 将需要相当长的计算时间. 本文采取的方法是将 SI 模型的模型近邻统计结果预先存储, 在 NNLR 自适应时只需要从中选择有自适应数据的那部分. 其次是对于自适应数据很少的模型, 同样可以用 NNLR 方法来平滑.

### 4 NNLR 与 VFS 的比较

作为一种模型插值/平滑方法, VFS 被广泛的应用在基于 MAP 的自适应框架中的. 其核心是假设 SI 空间中的模型近邻关系在 SA 空间中得到了保持, SI 空间可以平滑地过渡到 SA 空间. 在实现插值时, 用模型近邻在 MAP 自适应前后的平移向量的加权和来作为对应模型的平移向量的估计值, 即:

$$\mu_{ik} = \sum_{m \in A} W_m \mu_m / \sum_{m \in A} W_m \quad (10)$$

其中,  $\mu_m$  为模型分布  $m$  在 MAP 自适应前后的均值平移向

量,  $\mu_{ik}$  为无自适应数据的模型, 则由上式可以得到其对应的均值平移向量的估计值, 进而完成自适应. 这里的权重  $W_m = e^{(-d_m^2/s)}$ , 其中的  $s$  为平滑因子, 而模型距离  $d_m^2$  定义为:

$$d_m^2 = \sum_{p=1}^p \frac{(\mu_m^p - \mu_{ik}^p)^2}{p_{ik}} \quad (11)$$

其中的  $p_{ik}$  为协方差矩阵  $\Sigma_{ik}$  的第  $p$  个对角元素 (这里假设  $\Sigma_{ik}$  为对角矩阵).

注: NNLR 的模型距离度量同样用式 (11).

从式 (10) 可以看到, VFS 仅仅用自适应前后的模型均值平移向量来描述 SI 模型和 SA 模型之间的模型差异, 而 NNLR 方法则用线性回归模型来描述模型差异, 描述参数除了平移向量, 还有一个比例因子, 有更强的描述能力. 因此, 在回归假设合理的条件下, NNLR 在理论上可以更好地描述模型差异. 实验结果也表明, MAP/NNLR 可以有更好的自适应效果.

### 5 实验结果比较

本文所采用的实验平台是基于分段的混合高斯分布的音节识别系统. 与传统的 CDHMM 相比, 去掉了状态转移矩阵  $A$ , 只保留了输出概率密度函数. 由于本文提出的的自适应方法, 是针对输出概率密度函数设计的, 因此, 在该系统上得到的结论在传统的 CDHMM 中同样适用. 模型基元是汉语的 417 个无调音节, 采用自左向右无跳越的 6 个状态描述, 输出概率密度函数为混合的高斯分布, 混合数为 16.

采用的数据库取自 863 语料库, 共 37 人, 每人 520 句左右. SI 模型的训练采用了 30 人的数据. 用于自适应训练和测试的是其余的 7 人的数据. 每人的自适应数据分 4 种 (20、50、100、250 句); 测试集为在自适应集外抽取的 100 句. 特征参数为 16 维的 MFCC 和对应的 16 维自回归系数.

本文安排了 MAP、MAP/VFS 和 MAP/NNLR 方法的对比实验, 最大模型近邻数都取 20 (实验最优值).

表 3 MAP、MAP/VFS 和 MAP/NNLR 自适应方法对比 (首选误识率下降 %)

自适应数据 / 方法	20	50	100	250
MAP	2.0	6.2	13.3	30.4
MAP/VFS	5.3	12.4	20.1	43.1
MAP/NNLR	9.4	17.8	27.0	45.2

从表 3 可以看到, NNLR 方法有效的弥补了 MAP 自适应速度慢的缺点. 在自适应数据比较少时, NNLR 方法也可以很好的提高自适应效果 (少于 50 句时提高了三倍以上). 而随着自适应数据量的提高, 误识率的降低逐渐向 MAP 方法收敛, 保持了 MAP 方法对 SD 模型的渐进性. 同时在收敛的过程中也一直优于 MAP 方法: 即使在自适应为 250 句时, NNLR 方法可以对自适应数据较少的模型进行有效的平滑.

同时, 表 3 的结果也证明, 在同等实验条件下, NNLR 方法要优于 VFS 方法. 尤其在自适应数据较少时, NNLR 的效果明显好于 VFS. 关于这一点, 在第 4 节对 NNLR 和 VFS 的分析比较中, 已经在理论上进行了解释, 这里的实验结果则直接证明

了用线性回归方法来描述和量化模型差异的合理性.

## 6 总结和展望

本文分析了当前具有代表性的几种说话人自适应方法的优缺点,结合 MAP 方法对 SD 模型的渐进性和变换方法的快速自适应,提出用 NNLN 方法来对 MAP 估计的结果完成进一步的自适应,并提出用相关系数来度量和确保自适应结果的可信度.实验表明,NNLN 方法充分地利用了 MAP 自适应结果和模型近邻信息,对无自适应数据的模型完成了模型更新,有效地提高了自适应速度.与 VFS 方法相比,由于 NNLN 方法用线性回归模型取代了简单的加权插值来描述 SI 模型和 SA 模型的差异,在回归假设具有较高可信度的条件下,NNLN 有更强的模型描述能力,因此取得了更好的自适应效果.

任何一种自适应方法,都有两个根本问题要解决:模型差异的描述方式及其对应参数的估计.例如 MAP 方法是用一种最优意义下模型组合来描述模型差异,而变换方法则用映射关系来描述模型差异.此外,对自适应性能和速度的综合考虑也是研究自适应方法的最终目标.因此,研究新的模型差异描述方法和对应的参数估计方法,以及综合考虑自适应性能和速度的要求,研究有较高自主性的自适应方法,将成为下一步研究的重点.而对于本文提出的 NNLN 方法,可以用多项式回归来进一步提高模型差异的描述能力.这也将在下一步的工作中去验证.

### 参考文献:

- [ 1 ] Ahadi S. M. Woodland P. C. Combined bayesian and predictive technipued for rapid speaker adaptation of continuous density hidden Markov models [J]. Computer Speech and Language ,1997 ,11 :187 - 206.
- [ 2 ] Diglakis V. V. Rtischev D. Neumeyet L. G. Speaker adaptation using contrained estimation of Gaussian mixtures [J]. IEEE Trans. SAP , 1995 ,3(5) :357 - 365.

- [ 3 ] Gales M. J. F. Woodland P. C. Mean and variance adaptation within the MLLR eramework [J]. Computer Speech and Language ,1996 ,10 : 249 - 246.
- [ 4 ] Gauvain J. L. Lee C. H. Maximum a posteriori estimation for multi-variate Gaussian mixture observations of Markov Chains [J]. IEEE Trans. SAP ,1994 ,2(2) :291 - 298.
- [ 5 ] Lee C. H. ,Gauvain J. L. Speaker adaptation bases on MAP estimation of HMM Parameters [J]. Proc. IEEE ICASSP ,1993 ,2 :652 - 655.
- [ 6 ] Leggetter C. J. , Woodland P. C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models [J]. Computer Speech and Language ,1995 ,9 :171 - 185.
- [ 7 ] Takahashi J. Sagayama S. Vector-firld-smoothed bayesian learning for fast and Incremental speaker/ telephone-channel adapation [J]. Computer Speech and Language ,1997 ,11 :121 - 146.
- [ 8 ] Tonomura M. , Kosaka T. Matsunage S. , (Tonomura96) . Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation [J]. Computer Speech and Language , 1996 ,10 :117.

### 作者简介:



何 磊 1974 年出生. 1992 年考入清华大学计算机科学与技术系, 1997 年获工学学士学位和经济学学士学位. 并保送为清华大学计算机科学与技术系计算机应用专业博士研究生. 目前主要研究方向为声学模型的自适应和高鲁棒性的特征参数提取.

武 健 1975 年出生. 1993 年考入清华大学计算机科学与技术系, 1998 年获清华大学工学学士学位和经济学学士学位, 同年保送为清华大学计算机科学与技术系计算机应用专业硕士研究生. 研究兴趣主要包括关键词识别、大词表连续语音识别中的声学建模、自适应和语言模型, 目前从事语言模型连续语音识别中的搜索策略方面的研究工作.