语音识别说话人自适应研究现状及发展趋势

李虎生 ,刘 加 ,刘润生
(清华大学电子工程系 ,北京 100084)

摘 要: 说话人自适应是提高非特定人语音识别系统识别性能的有效手段. 本文介绍了说话人自适应研究的现状 ,包括自适应的不同方式和不同算法 ,并详细介绍了目前应用最为广泛的 MLLR 算法和 MAP 算法. 本文还给出了对说话人自适应研究发展趋势的预测.

关键词: 语音识别; 说话人自适应
中图分类号: TN912 **文献标识码**: A **文章编号**: 0372-2112 (2003) 01-0103-06

Technology of Speaker Adaptation in Speech Recognition and Its Development Trend

LI Hu-sheng ,LIU Jia ,LIU Run-sheng
(Department of Electronic Engineering , Tsinghua University , Beijing 100084 *, China)*

Abstract: Speaker adaptation is a powerful means of improving the performance of speaker-independent speech recognition system. This paper focuses on state-of-the-art of speaker adaptation technologies that include the modes and algorithms of speaker adaptation. The details of MLLR and MAP algorithms that are widely used are also given. The trend of study on speaker adaptation is predicted in this paper.

Key words: speech recognition; speaker adaptation

1 引言

九十年代以来 ,人们在非特定人 (Speaker Independent , SI) 大词汇量连续语音识别这一研究领域获得了很大的进展 ,但与训练得较为充分的特定人 (Speaker Dependent , SD) 系统相比识别性能还是有较大的差距 ,造成这一差距的主要原因是不同说话人语音之间的差异[1]. 这一差异包括: (1) 语音学上的差异: 由于方言的存在 ,不同的地方说话人对于同一句子的发音可能有很大的不同. 汉语是一种多方言语种 ,所以方言口音的存在会对汉语语音识别系统造成严重的影响. (2) 生理上的差异: 即使人们采用标准的普通话 ,不同说话人的声道形状 ,声门特性等存在区别 ,造成产生的语音频谱特性的不同. (3) 发音习惯差异与心理状态差异: 每个人有自己发音习惯 ,说话快慢也很不一样. 说话时的心情也往往不同 ,这些习惯与心态都会对当时说话的语音频谱特征造成影响 ,从而降低识别系统的性能.

说话人之间的差异对非特定人语音识别系统造成的影响主要有两方面原因: (1) 当某一使用该系统的说话人语音与训练语音库中的所有说话人的语音都有较大的差别时 ,对该使用者的语音系统的识别性能会有严重的恶化; (2) 训练一个较好的识别系统需要采集数量很大的说话人的语音用于训练 ,让训练语音库覆盖更为广泛的语音空间 ,这样虽然可以减低 (1) 中影响 ,但同时会造成识别系统参数分布较广 ,而不是较为尖锐的分布 ,造成识别性能的普遍下降.

特定人识别系统虽然可以克服非特定人系统的以上缺点 ,但该系统需要使用者录入大量的语音用于训练 ,给使用者带来很大的不便 ,对于大词汇量的识别系统 ,所需的训练语音的数量是令人无法忍受的.

近年来 ,人们采用说话人自适应 (Speaker Adaptation , SA) 算法有效地解决了特定人和非特定人系统各自的缺点. 该方案利用系统使用者的少量训练语音 ,调整系统的参数 ,使得系统对于该使用者的性能有明显的提高. 与 SI 系统相比 ,SA 系统由于考虑了用户的特殊信息 ,因此识别性能优于 SI 系统; 而与 SD 系统相比 ,SA 系统纳入了 SI 系统的先验信息 ,需要用户的提供的训练音数量远低于 SD 系统 ,有更好的实用性. 因此非特定人 + 自适应成为当前各语音识别系统采用的实用框架 ,自适应算法也成为近年来语音识别界研究的主要热点之一.

本文主要介绍自适应算法的不同方式和主要的算法 ,并着重介绍目前最为常用的两种算法 ,最后给出作者对说话人自适应研究未来发展趋势的看法.

收稿日期: 2000-04-24; 修回日期: 2002-03-29
基金项目: 国家自然科学基金 (No. 69975007); 国家 863 项目 (No. 863-306-ZD13-04-6 ,863-512-9805-10)

2 自适应方式的分类

按照训练语音获取的不同形式,自适应方式可以分为[2]:(1)批处理式:训练语音是由用户一次性录入,然后进行统一的自适应训练,更新系统参数;(2)在线式:训练语音是用户使用识别系统时所识别的语音,系统根据累积的统计量,按照一定时间间隔更新系统参数;(3)立即式:训练语音是当前正在识别的语音,该模式与在线式模式间的差别在于立即式自适应只利用当前的语音作自适应,没有在线式自适应的累积过程.

从实用角度看,在线式和立即式自适应模型由于不需要用户一次性输入一批训练语音,所以对用户的界面更具有友好性.从实现的角度看,批处理式自适应的实现难度低于在线式和立即式.而从自适应的性能看,批处理与在线式的算法本质是一致的,立即式自适应由于没有累积的过程,利用语音的信息少,所以性能劣于前两者.

按照学习过程有无监督,自适应又可以分为[1](1)有监督:自适应训练过程中训练语音的内容对于系统是已知的;(2)无监督:自适应训练过程中训练语音的说话内容对于系统是不知的,需要由系统通过识别获得.由于存在识别的错误,所以无监督自适应的性能通常要劣于有监督自适应.

各种自适应方式可以根据以上两种分类有多种组合,实用的语音识别系统可以采用批处理 + 有监督,批处理 + 无监督,在线式 + 有监督(对于识别结果需要用户验证的系统),在线式 + 无监督和立即式 + 无监督的方式.

3 自适应算法分类及简介

目前语音识别技术中使用的自适应方法主要分为两大类:(1)基于最大后验概率(Maximum a posteriori,MAP)的算法[3-7];(2)基于变换的方法[8-10].前者的基本准则是后验概率最大化,利用贝叶斯(Bayes)学习理论,将 SI 系统的先验信息与被适应人的信息相结合实现自适应;而后者则是估计 SI 系统模型与被适应人之间的变换关系,对 SI 系统的模型或输入语音特征作变换,减少 SI 系统与被适应人之间的差异.其它说话人自适应方法多数与这两种基本方法有关系,如结合最大后验概率与线性变换的自适应算法[15,30],这样可以有效地发挥各自的优点.实验证明这样的结合是有效的.以下分别介绍这两类算法:

3.1 MAP 算法[3,4,7]

基本的 MAP 算法基于以下准则[28,29]:

$$\theta_i = \arg\max_{\theta_i} P(\theta_i | \chi) \tag{1}$$

其中 χ 为训练样本,θ_i 为第 i 个语音模型的参数,θ_i 为模型参数的最大后验概率估计值.

MAP 算法采用基于最大后验概率准则,具有理论上的最优性,因此在小词表的语音识别任务中具有相当好的性能,并得到了广泛的应用.但在大词汇量语音识别系统中,MAP 算法却具有自适应速度缓慢的缺点,这是因为 MAP 算法仅对自适应训练语音中出现过的语音的模型作更新,而未出现过的语音的模型则无法实现自适应.对于大词汇量识别系统,用户的自适应语音远远无法覆盖所有的语音模型,因此有大量的模型参数没有得到自适应,造成了自适应速度的缓慢.

MAP 算法存在以上缺陷的本质原因是因为该算法没有考虑不同语音模型之间的在空间相关性,因此人们提出了多种算法,基于不同的假设,从不同角度利用不同语音间的关系,利用出现过的语音预测未出现语音,充分利用训练语音的信息,有效地加快自适应速度.这些算法包括:

® 基于线性预测的 MAP 算法[5].该算法的基本假设是不同语音模型间的关系可以用线性函数表示,其过程为:利用 SI 系统的训练语音库统计出不同语音的模型参数间的线性关系,在自适应时对于未出现的语音的模型,用已出现的语音的自适应结果以及线性关系预测其自适应结果:

$$\lambda_i = \sum_{j=1}^{M} \alpha_{ij} \lambda_j \tag{2}$$

其中,λ 为语音模型参数,i 为训练语音中未出现的某语音模型编号,j 为出现的语音模型编号,α_{ij} 为事先训练好的预测参数.

® 矢量场平滑(Vector Field Smoothing,VFS)算法[6].该算法的基本假设是:不同语音模型自适应后的变化量是一个连续函数,因此我们可以用已出现语音模型自适应后的变化量预测相邻的未出现语音的模型的变化量,从而获得未出现语音模型的自适应结果.

$$\Delta\lambda_i = \sum_{j=1}^{M} \alpha_{ij} \Delta\lambda_j \tag{3}$$

$$\lambda_i' = \lambda_i + \Delta\lambda_i \tag{4}$$

其中$\{\lambda_j, j = 1, 2, \cdots, M\}$为已出现语音的模型参数,$\lambda_i$ 为未出现语音的模型参数.这里 α_{ij} 是训练好的预测参数.

® 马尔科夫随机场(Markov Random Field,MRF)算法[7].MRF 是另一种描述模型间相关性的方法.它假设码本的均值可以用二维随机场中的点来表示,"相近"的码本相互连通,两两连通的点的集合构成了一个类,类的先验概率用 Gibbs 分布来描述.自适应过程按类进行,因此可以对未出现过的语音做自适应.

3.2 基于变换的算法

这一类算法的基本假设是相近语音的 SI 系统语音空间与被适应人语音空间的变换关系也是相近的,因此可以利用训练语音中出现过的语音统计出这一变换关系,对未出现的语音的模型用该变换实现从 SI 系统到被适应人语音空间的映射,从而完成自适应过程.语音空间根据一定测度(如欧氏距离,似然度等)被划分为 R 类,各类的变换为 $T_r(\cdot)$,分别对应的训练语音集为 $X_r, r = 1, 2, \cdots, R$,模型参数为 $\lambda_r, r = 1, 2, \cdots, R$,则最优的自适应变换满足:

$$T_r = \arg\max_{T} (P(X_r | T_r)) \quad r = 1, \cdots, R \tag{5}$$

自适应后的参数 $\lambda_r, r = 1, 2, \cdots, R$ 满足:

$$\lambda_r = T_r(\lambda_r) \quad r = 1, \cdots, R \tag{6}$$

由于这一类算法充分利用了语音模型在空间中相互关系,因此具有较快的自适应速度,在大词汇量语音识别系统中得到了广泛的应用.但是由于 SI 系统模型与被适应人间的变

换关系是非常复杂的,不可能用一个简单的关系表示,而系统又必须采用一些数学上可处理的变换(如平移变换,线性变换,仿射变换等).因此无法完全得到准确的变换关系,而且算法的基本假设也是近似的,所以当训练语音较多时,该类算法无法达到很好的效果,缺乏 MAP 算法具有的渐进逼近 SD 系统的性能的优点.

目前基于变换的算法主要有:

® **最大似然线性回归**(Maximum Likelihood Linear Regression,MLLR)**算法**[8,11].该算法采用的变换形式是仿射变换,即

$$y = Ax + b \tag{7}$$

其中,x 为自适应前的参数矢量;y 为自适应后的参数矢量;A、b 分别为根据自适应训练语音,用最大似然准则估计出的变换参数.

® **随机匹配**(Stochastic Match,SM)**算法**[9].该算法采用的变换形式是平移变换:

$$y = x + b \tag{8}$$

式中各项的意义与 MLLR 算法中相同.

® **非线性变换算法**[10].以上两类算法采用的变换形式都是线性的,无法很好地模拟语音空间之间的非线性关系,因此人们采用非线性变换以弥补这一缺陷.目前基于非线性变换的算法主要采用的变换形式有分段线性变换、人工神经网络等.但由于非线性变换数学处理上的难度,其性能尚未超过基于线性变换的方法.

4　MLLR 算法与 MAP 算法

由于 MLLR 算法与 MAP 算法是目前主流的说话人自适应算法,本节仅对这两种算法作详细介绍.考虑到连续隐含马尔科夫模型(Continuous Hidden Markov Model,CHMM)[12]为目前各识别系统的主要方法,因此本文仅对 CHMM 系统的自适应做讨论.

4.1　MLLR 算法

MLLR 算法是基于变换的一种自适应算法,其变换过程采用式(7)所示的仿射变换,自适应的流程如图 1 所示.由于识别系统的训练主要由 HMM 各高斯分量的均值矢量决定,所以这里只对均值矢量进行自适应方法进行描述.

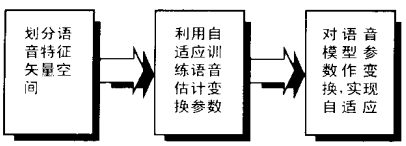

图 1　MLLR 算法流程

4.1.1　语音特征空间的划分　由于 MLLR 算法的前提假设是相近的语音共享相同的变换,因此我们需要根据一定的准则对语音空间进行划分,然后对每一类空间估计其相应的变换.划分的过程所需考虑的因素包括有:

® **划分的类数**　当将所有语音划分为一类时,我们称之为全局变换(Global Transformation);当将每一个语音模型划分为一类时,可以证明此时 MLLR 算法退化为 MAP 算法[1].一般划分的类数介于以上两个极端情况之间,并与训练语音样本数量的多少相关,并可以用算法进行控制.

® **划分的准则**

划分语音空间的准则是将相近的语音划分为一类.在 HMM 算法中,可以用高斯分布描述各语音模型的分布,因此划分语音特征空间即对不同高斯分布分量均值的进行聚类.可以用多种方法来描述语音模型之间的相似性.一种较为有效的方法是采用似然度变化[1]作为语音模型相似性的测度.

® **划分的方法**

对于不同的自适应模式,需要采用不同方式对语音空间进行划分.

对批处理自适应模式,可以根据事先设置好聚类类数,不断地对最相似的高斯分布进行两两合并,直到预定的聚类类数达到为止.

对在线式自适应模式,训练语音是逐步累积的,所以需要动态地实现对语音模型空间的划分.由于聚类过程所需的计算量较大,考虑到系统的实时性要求,因此采用决策树的方法[1,13,22]完成聚类的过程(通常采用二叉树),即在自适应前对各均值矢量进行两两聚类,直至将所有均值矢量聚为一类为止,将聚类的过程记录下来,形成一棵决策树.这棵树的每一个节点都代表一个变换类,每一个节点(即均值矢量)共享一类变换.在自适应过程中,语音空间的划分从树的根节点(即将所有均值矢量作为一个变换类)开始,当某一节点的两个子节点已积累了足够多的语音时,可以将该变换类划分为子节点所对应的两个变换类.这样,划分语音空间即成为自顶而下地遍历决策树的过程.

4.1.2　变换参数的估计　考虑到只对均值矢量作变换,可以将变换重写为:

$$\hat{\mu} = A\mu + b = W\zeta \tag{9}$$

其中 ζ 为扩展均值矢量 $[1, \mu^T]^T$,W 为扩展变换矩阵 $[b^T, A^T]^T$,需要估计的参数为扩展变换矩阵 W.设某一聚类的状态集合包含了 M 个状态,即,$\{S_m, m = 1,2,\cdots,M\}$.容易证明该状态集合类的均值矢量变换矩阵 W 可以通过求解以下方程获得[1]:

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{s_m}(t)\ \Sigma_{s_m}^{-1}x_t\zeta_{s_m}^T = \sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{s_m}(t)\ \Sigma_{s_m}^{-1}W\zeta_{s_m}\zeta_{s_m}^T \tag{10}$$

其中 $\gamma_j(t)$ 第 t 帧语音属于状态 j 的概率,Σ_m 为 S_m 状态的输出高斯概率分布函数的协方差矩阵,x_t 为第 t 帧语音的特征矢量.

当协方差矩阵 Σ 为满阵时,方程(10)没有显式解,变换矩阵 W 需要转化为矢量来解,设 $vec(\cdot)$ 表示将矩阵按行连接成为一个矢量,$Kron(\cdot)$ 表示 Kronecker 积,即令:

$$V^{(m)} = \sum_{t=1}^{T}\gamma_m(t)\ \Sigma^{(m)-1} \tag{11}$$

$$D^{(m)} = \zeta^{(m)}\zeta^{(m)T} \tag{12}$$

$$Z = \sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)\ \Sigma^{(m)-1}x_t\zeta^{(m)T} \tag{13}$$

则易证由式(10)等价于:

$$vec(Z) = \left(\sum_{m=1}^{M}kron(V^{(m)}, D^{(m)})\right)\ vec(W) \tag{14}$$

可以用各种线性方程的求解方法,如高斯消元法求解式(14). 设均值矢量的维数为 n,式(14)中方程的系数矩阵大小为 $(n^2+n)\times(n^2+n)$,因此解式(14)所需的计算量是巨大的,很难实时实现.

当协方差矩阵 Σ 为对角阵时,易证 W 的第 i 行满足:

$$G^{(i)}w_i^T = k^{(i)},\ i=1,2,3,\cdots,n \tag{15}$$

其中:

$$G^{(i)} = \sum_{m=1}^{M}\frac{1}{\sigma_i^{(m)2}}\zeta^{(m)}\zeta^{(m)T}\sum_t\gamma_m(t) \tag{16}$$

$$k^{(i)} = \sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)\frac{1}{\sigma_i^{(m)2}}x_t\zeta^{(m)T} \tag{17}$$

其中 σ_i 为协方差矩阵对角线上第 i 个元素. 求解式(15)可以用高斯消元法或广义逆矩阵的求解法,其计算量远小于求解式(14)运算量,因此更具有实用性.

4.1.3 模型均值矢量的变换 在估计出该类最优的线性变换矩阵 W 后,就可以根据式(9)对该类所包含每个模型的均值矢量进行线性变换. 使识别模型适应于输入的说话人语音特征.

4.2 MAP 算法

MAP 算法以最大后验概率为基本准则,根据式(1)可以推导出某状态自适应后的均值矢量估值 μ[20]:

$$\mu = \frac{\sum_{t=1}^{T}\gamma(t)\,x_t + \tau\mu}{\sum_{t=1}^{T}\gamma(t) + \tau} \tag{18}$$

其中 μ 为自适应前的均值矢量,$\gamma(t)$ 为第 t 帧语音属于该状态的概率,τ 为先验参数,一般根据实验结果取为某一正常数. 式(18)可改写为:

$$\mu = \beta\mu_{SD} + (1-\beta)\mu \tag{19}$$

其中 μ_{SD} 为特定人训练结果:

$$\mu_{SD} = \frac{\sum_{t=1}^{T}\gamma(t)\,x_t}{\sum_{t=1}^{T}\gamma(t)} \tag{20}$$

$$\beta = \frac{\sum_{t=1}^{T}\gamma(t)}{\sum_{t=1}^{T}\gamma(t) + \tau} \tag{21}$$

式(19)表明 MAP 算法估计结果为 SD 参数与 SI 参数的线性组合,但加权系数随训练语音的变化而变化. 当没有训练语音时,估计结果即为 SI 参数;当训练语音较少时,当 β 取较小值时,估计结果接近于 SI 参数,可以防止过训练现象;当训练语音增加时,当 β 取较大值时,SD 参数在结果中的比重加大,使系统性能随被适应人训练语音数据增加逐步提高;当训练语音很多时,$\beta\to1$,系统渐进地逼近于 SD 系统. 可以看到 MAP 算法的实现过程比 MLLR 算法要简单.

5 自适应算法性能

文献[14]中对 MAP 算法和 MLLR 算法的性能均作了测试与比较.

5.1 MAP 算法

对 MAP 算法性能的测试采用小词表语音识别中具有重要实际意义的非特定人汉语数码串连续语音识别任务. 汉语数码串语音识别系统采用 MFCC 参数和连续 HMM 作为声学模型,模型参数强化训练采用 MCE 算法[23]. 识别过程采用多候选帧同步搜索算法. 图 2 给出了有监督、批处理模式下 MAP 算法自适应后的性能以及非特定人的误识率. 误识率由 10 个说话人平均所得,其中 SI 表示非特定人误识率,SA 表示自适应后的误识率. 由图可见,自适应算法的性能随着自适应语音的增加逐步上升,在采用 40 个数字串进行自适应时,误识率由 6.8 %下降到 3.8 %,相对下降 44 %,有效地提高了识别性能.

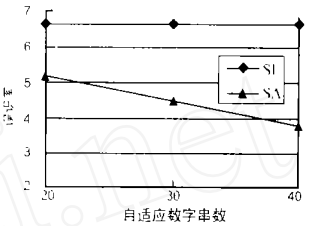

图 2 MAP 算法性能

5.2 MLLR 算法

测试 MLLR 算法性能使用的是邮包语音校核系统[16]. 该系统实现包含全国 4000 多个地名的连续语音命令自动识别,其声学模型为基于半音节 HMM,识别过程采用了多子树帧同步维特比搜索算法. 图 3 给出了在线式 MLLR 算法性能的测试结果.

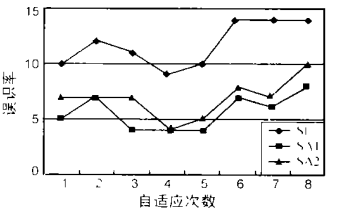

图 3 MLLR 自适应性能

变换类的划分采用了基于决策树的方法. 测试集包括 4 个说话人的各 420 句语音,误识率由这 4 个人的误识率平均所得,自适应过程为每隔 40 句语音进行一次参数的更新,然后用后面的 100 句语音测试"即时"误识率,SI 表示非特定人误识率,SA1 表示有监督自适应后的误识率,SA2 表示无监督自适应后的误识率. 由测试结果可见,无论是有监督模式还是无监督模式,MLLR 算法可以明显的降低误识率,但当自适应语音数量很多时,误识率无法进一步降低,同时也可以看到,有监督模式自适应的性能略优于无监督模式.

6 自适应算法发展趋势

目前 MLLR 算法和 MAP 算法分别在大词汇量语音识别和小词汇量语音识别中获得了很大的成功,但仍有很多问题需要解决. 针对这两种算法中存在问题,许多研究者进行了改进,也提出了一些新的方法. 作者认为,对于一个语音识别系

统,最优的自适应过程应该是如图 4 所示的过程:在用户使用系统前,先进行有监督、批处理的自适应,使识别性能有迅速的提高,然后在识别过程中进行无监督、在线式的自适应,使系统的性能渐进式的逼近于特定人系统.

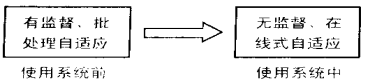

图 4　最优自适应过程

在整个自适应过程中的主要困难在于:

® 在有监督、批处理自适应时,自适应语音数量需要尽可能地少,否则会使用户感觉不方便,因此如何从有限的语音中尽量地提高识别性能是该阶段的主要问题.

® 在无监督、在线式自适应时,自适应语音的数量是很大的,主要问题在于如何减少识别错误对无监督自适应的影响和如何使系统能充分利用大量的训练语音.

因此针对说话人自适应问题主要需要研究内容包括:

® 在训练语音不够充分的条件下,如何实现稳健参数自适应调整. 例如在估计 MLLR 算法的线性变换矩阵时,采用满阵的协方差矩阵模型需要的参数太多,难以实现充分的稳健估计. 这样可能反而造成识别性能的恶化,而采用对角阵的形式又会因为变换形式过于简单而难以充分提高识别性能,因此可以采用对角块矩阵的形式,这样可以大大减少所需估计的参数,提高变换矩阵的统计可靠性. 而如何确定各块的位置和大小即成为尚需解决的问题[25].

® 利用置信度[19-20]判断无监督自适应中识别结果的可靠性,将不可靠的识别结果从自适应语音中删去,从而减少无监督自适应和有监督自适应间的差距. 采用什么形式的置信度计算方法和对什么语音单位作拒绝是需要研究的重点. 文献[18]在这一方面做出了初步的探讨.

® 采用具有层次结构变换类划分方法[30]. 当决策树划分到叶子节点时,基于变换的算法可以近似地退化为 MAP 算法,从而实现渐进的自适应过程. 但如何判断一个节点是否需要划分(即该节点的子节点的参数已估计充分)时,文献[1,13]是根据该节点的训练语音数据是否超过一个阈值,这个方法是非常粗糙的. 可以考虑将假设检验(原假设为当前参数估计充分)引入该判决过程,而用于作假设检验的语音可以是在线式自适应中下一个节拍的自适应语音.

® MLLR 算法与 MAP 算法有各自的优点. 如何将两种算法更好的结合,这需要有好的理论框架,早期的方法是简单地将 MLLR 算法与 MAP 算法的结果作加权组合[15]. 新的具有代表性方法有:(1)采用最大后验概率准则确定最佳的变换,而不是最大似然准则[30,35,36],由于增加了 $p(W)$ 信息,因此使性能得到改进;(2)结合最大互信息准则(Maximum Mutual Information,MMI)与线性回归算法[32,33].

® 采用基于非线性变换的自适应[10]. 线性变换的限制始终是 MLLR 算法的一个缺陷,这使得 MLLR 算法在自适应语音数量很大时无法进一步提高性能,因此可以用非线性变换来解决这一缺陷. 如何找出高效实时的训练算法和如何解决训练数据的不足是采用非线性变换时需要解决的首要问题.

® 利用说话人自适应实现说话人归一化[13,26]. 经典的非特定人识别系统使用来自多个说话人的大量数据进行训练,这样会导致模型参数的分散,使模型的鉴别能力较低. 近年来人们发展了说话人归一化的技术,通过声道长度归一化[26]或线性变换[27]等方法,将训练集内的语音映射到模型空间,减小训练集语音的分散程度,对于用户语音则用自适应提高其识别性能. 实验证明,进行了说话人归一化的 SI + SA 系统性能优于普通的 SI + SA 系统.

® 快速说话人自适应算法也是近年来一个研究热点,如何用少量的训练语音(几秒语音数据)快速的调整模型实现自适应,这在口语对话信息查询系统有着重要作用. 比较典型的方法有:(1)基于特征语音(Eigenvoice)模型的变换方法[31],该方法采用一组特定人识别模型来快速自适应新的说话人模型;(2)基于扩展的最大后验概率(Extended Maximum a Posteriori,EMAP)自适应算法,该方法考虑模型之间的相关性,采用这些相关信息来变换在训练语音数据中不包括的模型参数[33].

7　结论

本文介绍了目前语音识别中的主要的说话人自适应算法,并对说话人自适应的发展趋势作了探讨. 相信经过人们不懈的努力,说话人自适应的性能可以有进一步的提高.

参考文献:

[1] C J Leggetter. Improved acoustic modelling for HMMs using linear transformations [D]. Cambridge University,1995.

[2] G Zavaliagkost,R Schwatz,J Makhoul. Batch,incremental,and instantaneous adaptation techniques for speech recognition [A]. ICASSP [C]. 1995.

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

[4] J L Gauvain,C H Lee. Maximum a posteriori estimation for multivariate Gaussian observations [J]. IEEE Trans. on Speech and Audio Processing,1994,2(2):291-298.

[5] S M Ahadi,P C Woodland. Rapid speaker adaptation using model prediction [A]. ICASSP [C]. 1995.

[6] J Takahashi,S Sagayama. Vector-field-smoothed Bayesian learning for fast and incremental speaker/telephone-channel adaptation [J]. Computer Speech and Language,1997,11(2):127-146.

[7] B M Shahshahani. A Markov random field approach to Bayesian speaker adaptation [J]. IEEE Trans. on Speech and Audio Processing,1997,5(2):183-191.

[8] C J Leggetter,P C Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models [J]. Computer Speech and Language,1995,9(2):171-185.

[9] A Sankar,C H Lee. Maximum likelihood approach to stochastic matching for robust speech recognition [J]. IEEE Trans. on Speech and Audio Processing,1996,4(1):190-202.

[10] A C Surendran,C H Lee,M Rahim. Nonlinear compensation for stochastic matching [J]. IEEE Trans. on Speech and Audio Process-

ing ,1999 ,7(6) :643 - 655.

[11] V V Digalakis ,D Rtischev ,L G Neumeyer. Speaker adaptation using constrained estimation of Gaussian mixtures[J]. IEEE Trans. on Speech and Audio Processing ,1995 ,3 (5) :357 - 365.

[12] 杨行峻 ,迟惠生等. 语音信号数字处理 [M]. 北京 :电子工业出版社 ,1995.

[13] 郑榕. 在线学习汉语连续语音识别系统的研究[D]. 北京 :清华大学电子工程系 ,1998.

[14] 李虎生. 汉语数码串语音识别及说话人自适应[D]. 北京 :清华大学电子工程系 ,2000.

[15] V V Digalakis ,L G Neumeyer. Speaker adaptation using combined transformation and Bayesian Methods [J]. IEEE Trans. on Speech and Audio Processing ,1996 ,4(4) :294 - 300.

[16] 张昊天. 邮包核对语音识别系统的开发和实现[D]. 北京 :清华大学电子工程系 ,2000.

[17] 李虎生 ,杨明杰 ,刘润生. 汉语数码语音识别自适应算法[J]. 电路与系统学报 ,1999 ,4(2) :1 - 6.

[18] Husheng Li ,Jia Liu ,Runsheng Liu. Confidence measure based unsupervised adaptation[A]. Proc. of ICSLP [C]. Beijing : ICSIP ,2000.

[19] G Williams. A study of the use and evaluation of confidence measures in automatic speech recognition [R]. Technical Report ,Dept. of Computer Science ,University of Sheffield ,Feb. 1998.

[20] L D Colton. Confidence and rejection in automatic speech recognition [D]. Oregon Graduate Institute. 1997.

[21] Q Huo ,C H Lee. On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate [J]. IEEE Trans on Speech Audio Processing. 1997 ,5(1) :161 - 172.

[22] J T Chien. Online hierarchical transformation of hidden Markov models for speech recognition[J]. IEEE Trans. on Speech and Audio Processing ,1999 ,7(6) :656 - 667.

[23] B H Juang ,W Chou ,C H Lee. Minimum classification error rate methods for speech recognition [J]. IEEE Trans. on Speech and Audio Processing ,1997 ,3(3) :257 - 265.

[24] V V Digalakis. Online adaptation of hidden Markov models using incremental estimation algorithms [J]. IEEE Trans. on Speech and Audio Processing. 1999 ,7(3) :253 - 261.

[25] E L Bocchieri ,V Digalakis ,A Corduneanu ,C Boulis. Correlation modeling of MLLR transform biases for rapid HMM adaptation to new Speakers[A]. ICASSP [C]. 1999.

[26] Li Lee ,R Rose. A frequency warping approach to speaker normalization [J]. IEEE Trans. on Speech and Audio Processing ,1998 ,6(1) :49 - 60.

[27] T Anastasakos et al. Speaker adaptive training : a maximum likelihood approach to speaker normalization [A]. ICASSP [C]. 1997. 1043 - 1046.

[28] 张金槐 ,唐雪梅. BAYES 方法 [M]. 长沙 :国防科技大学出版社 ,1993.

[29] R O Duda ,P E Hart. Pattern Classification and Scene Analysis[M]. New York :John Wiley ,1973.

[30] O Siohan ,C Chesta ,C H Lee. Joint maximum a posteriori adaptation of transformation and HMM parameters [J]. IEEE Trans on Speech and Audio Processing ,2001 ,9(4) :417 - 428.

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

[32] F Wallhoff ,D Willett ,G Rigoll ,Frame discrimination and confidence-driven adaptation for LVCSR [A]. ICASSP [C]. 2000. 1835 - 1838.

[33] F Wallhoff ,D Willett ,G Rigoll. Scaled likelihood linear regression for hidden Markov model adaptation[A]. Eurospeech [C]. Scandinavia , 2001. 1229 - 1232.

[34] R Kuhn ,F Perronnin ,P Nguyen ,J C Junqua ,L Rigazio. Very fast adaptation with a compact context-dependent eigenvoice model[A]. ICASSP [C]. 2001. 373 - 376.

[35] K Shinoda ,C H Lee. A structural Bayes approach to speaker adaptation [J]. IEEE Trans. on Speech and Audio Processing ,2001 ,9(3) :276 - 287.

[36] T A Myrvoll ,K K Paliwal ,T Sveensen. Fast adaptation using Affine transformation with Hierarchical priors [A]. Eurospeech [C]. Scandinavia ,2001. 1233 - 1236.

作者简介:

李虎生 男 ,1975 年出生于四川宜宾 ,1998 年 7 月获得清华大学电子工程系无线电技术专业学士学位 ,2000 年 7 月获得清华大学电子工程系通信与电子系统硕士学位 ,目前在美国 Princeton 大学攻读博士学位.

刘 加 男 ,1954 年出生于福建福州 ,1983 年 7 月获得清华大学电子工程系无线电技术专业学士学位 ,1986 年 7 月获得清华大学电子工程系通信与电子系统硕士学位 ,1990 年 4 月获得清华大学电子工程系通信与电子系统博士学位 ,1990 年 - 1992 年 中科院遥感卫星地面站从事美国 6 号陆地卫星图像处理系统开发工作 ,1992 年 - 1994 年在英国剑桥大学作博士后 ,从事语音识别系统研究工作 ,现任清华大学电子系教授 ,博士生导师 ,中国电子学会高级会员 ,美国 IEEE 会员 ,近 5 年来在国内外期刊及学术会议上发表论文 40 多篇 ,目前研究方向包括 :语音识别、语音合成、语音编码、语音识别专用芯片设计 ,多传感器融合技术 ,以及多媒体数字通信系统.

刘润生 男 ,1958 年清华大学无线电专业毕业留校任教至今 ,现任清华大学电子工程系教授 ,博士生导师 ,长期从事脉冲数字电路、模拟电路、集成电路、集成电路设计、电子电路 CAD、通信与信号处理等方面的教学和科研工作 ,近十年来在国内外期刊及学术会议上发表论文 100 余篇 ,目前的主要研究方向包括 :汉语数码及中小词汇量语音识别及其应用 ,数字信号处理及模数混合集成电路设计 ,电子电路快速模拟算法的研究.