基于三对角和共享分块对角转换矩阵的快速说话人自适应方法

丁国宏[1], 徐 波[1,2]
(1. 中国科学院自动化研究所高技术创新中心, 北京 100080; 2. 中国科学院自动化研究所模式识别国家重点实验室, 北京 100080)

摘 要: 本文提出了两种在最大似然线性回归(MLLR)框架下实现快速说话人自适应的方法. 这两种方法在本文中分别称为 Log 谱域下基于三对角转换矩阵的说话人自适应(SATD)和倒谱域下基于共享分块对角转换矩孟加拉国 说话人自适应(SASBD). 这两种方法在一定先验知识的基础上采用较少的参数来描述说话人间的差异, 因而只需要少量的自适应数据就可以得到参数的鲁棒估计. 在以整词建模的孤立词识别系统和以三音子建模的孤立词识别系统上分别进行的测试表明所提出的方法相对传统的 MLLR 自适应方法有较快的自适应性能.

关键词: 快速自适应; 转换矩阵; MLLR; 三对角矩阵; 分块对角矩阵
中图分类号: TP301.6 文献标识码: A 文章编号: 0372-2112 (2004) 10-1709-04

Fast Speaker Adaptation Based on Triple Diagonal Transform Matrices and Shared Block Matrices

DING Guo-hong[1], XU Bo[1,2]
(1. *High-Tech Innovation Center, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080, *China*;
2. *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract: In the Maximum Likelihood Linear Regression (MLLR) framework, this paper proposes two fast speaker adaptation approaches, which are called Speaker Adaptation using Triple Diagonal matrices in the log-spectral domain (SATD) and Speaker Adaptation using Shared Block Diagonal matrices (SASBD) in the cepstral domain, respectively. Based on some prior knowledge, the proposed approaches utilize fewer parameters to describe the variation between speakers, and thus fewer adaptation data are needed to give robust estimation. Experimental results in both the whole-word-modeled isolated word recognition system and the isolated word recognition system using triphones as modeling units show that the proposed approaches can provide faster performance than the traditional MLLR approaches.

Key words: fast adaptation; transfomation matrices; MLLR; triple diagonal matrices; block diagonal matrices

1 引言

在过去的十年里, 许多研究者对说话人自适应模型补偿进行了广泛的研究. 最大似然线性回归(MLLR)和最大后验概率(MAP)自适应方法是其中的两大主流. 它们给出了进行模型补偿的两个经典框架, 在其中衍生出了很多进行说话人和声学环境补偿的算法.

近些年来, 采用少量自适应数据的快速说话人自适应成为了很多研究者的研究重点. 快速说话人自适应的基本思想是充分利用先验知识, 采用尽量少的参数来描述说话人间的差异, 从而只需要少量的自适应语音就可以鲁棒估计出这些参数. 这里"少量的自适应语音"是快速说话人自适应算法的一个目标. 采用较少的参数很难全面描述说话人间的差别, 通常在有足够自适应语音时性能会降低(相对采用较多参数的自适应方法而言), 快速说话人自适应的另一个目标则是在有足够多的自适应数据时性能不降低或者降低很少. 因此, 如何正确利用先验知识就成了快速说话人自适应的关键.

最近成为主流的两类快速自适应方法是说话人选择技术[7]和基于转换的本征音(eigenvoice)方法[8]. 这两类方法都用某个参数化的形式描述说话人的特征. 说话人选择技术采用转换矩阵或者高斯模型来代表特定的说话人, 这些参数之间的距离反映了说话人间的距离, 最接近测试说话人的训练说话人的数据被用来构建新的声学模型. 在基于转换的本征音方法中, 转换矩阵被认为是描述说话人特征的参数化形式, 特征转换矩阵由对很多说话人转换矩阵的分析(求特征值)得到, 而测试说话人的转换矩阵通过对特征转换矩阵的加权求和估计出来.

值得注意的是, 无论是说话人选择技术或者是基于转换的本征音方法, 它们都可以建立在转换矩阵的基础上. 本文的思想则是研究转换矩阵本身, 即在一定先验知识的约束下, 如何采用足够少的参数构建转换矩阵, 从而比较充分地描述说话人之间的差异. 在传统的通道长度归一化先验知识的基础

收稿日期: 2003-08-11; 修回日期: 2004-04-30

上,本文首先得到了描述说话人间差异的 log 谱域下的三对角矩阵,然后在 MLLR 框架下估计出这个转换矩阵,并用于模型补偿.此外,本文还根据动态特征由基本特征经过一阶二阶差分得到的情况,采用一个共享的矩阵来代替在分块对角矩阵情况下需要采用的三个矩阵.本文的基本思想是在声道长度归一化和动态特征由基本特征经过差分得到的先验知识的基础上采用尽量少的参数构造转换矩阵,从而实现快速说话人自适应.

本文的贡献在于提出了两个转换矩阵的形式,这两个转换矩阵受一定先验知识的约束因而只需要少量的参数就可以比较充分地描述说话人间的差异.这两个转换矩阵可以被应用到采用多回归的 MLLR 自适应,以及说话人选择技术和基于转换的本征音等方法中去.

2 声道长度归一化的线性描述

说话人之间的一个主要差别是说话人声道长度的不同,表现为语音共振峰的位置同说话人声道长度成反比,在语音识别中通常可以通过沿频率轴拉伸或压缩频谱来实现声道长度归一化.

文献[3]提出了通过修改 Mel 滤波器组系数代替对频谱的拉伸或压缩从而实现声道长度归一化的方法.在这种方法中,Mel 滤波的公式被修改为:

$$Y_\alpha(i)=\sum_{\omega=l_i(\alpha)}^{h_i(\alpha)} T_\alpha(\omega)X(\omega),\ 0\le i\le N-1$$

其中 $X(\omega)$ 是能量谱密度,α 是弯折系数.可以根据 α 调整 Mel 滤波器的上下限 $l_i(\alpha)$ 和 $h_i(\alpha)$ 及由这个上下限决定的参数 $T_\alpha(\omega)$,从而得到归一化的 Mel 能量谱 $Y_\alpha(i)$.

当 $\alpha>1$ 时,$X(\omega)$ 需要被拉伸.如果保持 $X(\omega)$ 不变,Mel 滤波器组的带宽应该被压缩,它的上下限都应该向零点移动,使得第 i 个滤波器仍然处理频谱被拉伸时对应的频率段.由于 α 通常取值在 0.88 到 1.12 之间,可以认为

$$h_{i-1}<h_i(\alpha)<h_i,\ l_{i-1}<l_i(\alpha)<l_i$$

此时,滤波得到 $Y_\alpha(i)$ 的频谱包含在滤波得到 $Y(i-1)$ 和 $Y(i)$ 的频谱中.

当 $\alpha<1$ 时,根据同样的原理,滤波得到 $Y_\alpha(i)$ 的频谱包含在滤波得到 $Y(i)$ 和 $Y(i+1)$ 的频谱中.

考虑一般情况,可以认为 $Y_\alpha(i)$ 同 $Y(i-1)$,$Y(i)$ 和 $Y(i+1)$ 是相关的.或者说,$\log Y_\alpha(i)$ 同 $\log Y(i-1)$,$\log Y(i)$ 和 $\log Y(i+1)$ 是相关的,可以把这个相关性描述成一个函数,即:

$$\log Y_\alpha(i)=f(\log Y(i-1),\log Y(i),\log Y(i+1))$$

这里 $f(*)$ 是一个无法明确描述的非线性函数.根据一阶泰勒近似的方法,可以得到下面的等式

$$\log Y_\alpha(i)\approx\theta^l_{i,i-1}\log Y(i-1)+\theta^l_{i,i}\log Y(i)+\theta^l_{i,i+1}\log Y(i+1)+b^l_i$$

其中 $\theta^l_{i,j}$ 和 b^l_i 是加权系数.弯折系数对于某个说话人是固定的,反映了该说话人的特性,这里假设加权系数对某个说话人也是固定的,也反映了该说话人本身的特征.

定义

$$L_\alpha(i)=\log Y_\alpha(i)$$

$$L(i)=\log Y(i)$$

则对于 $i=1,2,\cdots,N$,有:

$$\begin{bmatrix}L_\alpha(1)\\L_\alpha(2)\\L_\alpha(3)\\\vdots\\L_\alpha(N)\end{bmatrix}=\begin{bmatrix}\theta^l_{1,1}&\theta^l_{1,2}&0&\cdots&0\\\theta^l_{2,1}&\theta^l_{2,2}&\theta^l_{2,3}&\ddots&0\\0&\theta^l_{3,2}&\theta^l_{3,3}&\ddots&0\\\vdots&\ddots&\ddots&\ddots&\vdots\\0&0&0&\cdots&\theta^l_{N,N}\end{bmatrix}\begin{bmatrix}L(1)\\L(2)\\L(3)\\\vdots\\L(N)\end{bmatrix}+\begin{bmatrix}b^l_1\\b^l_2\\b^l_3\\\vdots\\b^l_N\end{bmatrix}\tag{1}$$

这样,归一化过程可以描述为

$$\boldsymbol{L}_\alpha=\theta^l\boldsymbol{L}+\boldsymbol{b}^l\tag{2}$$

其中

$$\boldsymbol{L}_\alpha=[L_\alpha(1),L_\alpha(2),\cdots,L_\alpha(N)]^T$$

$$\boldsymbol{L}=[L(1),L(2),\cdots,L(N)]^T$$

弯折和未弯折情况下的离散余弦变换(DCT)过程为

$$\boldsymbol{C}_\alpha=\boldsymbol{D}\boldsymbol{L}_\alpha\tag{3}$$

$$\boldsymbol{C}=\boldsymbol{D}\boldsymbol{L}\tag{4}$$

其中 $\boldsymbol{C}_\alpha$ 和 $\boldsymbol{C}$ 表示弯折和未弯折时的倒谱向量.$\boldsymbol{D}$ 是 $M\times N$ DCT 矩阵,满足

$$\boldsymbol{D}\boldsymbol{D}^T=\boldsymbol{I}_M$$

其中 $\boldsymbol{I}_M$ 是 $M\times M$ 单位阵.

定义倒谱空间的转换为

$$\boldsymbol{C}_\alpha=\theta^c\boldsymbol{C}+\boldsymbol{b}^c\tag{5}$$

把式(3,4)带入上式,可以得到

$$\boldsymbol{D}\boldsymbol{L}_\alpha=\theta^c\boldsymbol{D}\boldsymbol{L}+\boldsymbol{b}^c$$

对式(2)两边同时左乘 $\boldsymbol{D}$,有

$$\boldsymbol{D}\boldsymbol{L}_\alpha=\boldsymbol{D}\theta^l\boldsymbol{L}+\boldsymbol{D}\boldsymbol{b}^l$$

由于上面两式对于所有的特征都成立,因而可以得到

$$\theta^c=\boldsymbol{D}\theta^l\boldsymbol{D}^T\tag{6}$$

$$\boldsymbol{b}^c=\boldsymbol{D}\boldsymbol{b}^l\tag{7}$$

3 log 谱域下基于三对角转换矩阵的说话人自适应(SATD)

上一部分提出了采用 log 谱域下的三对角矩阵来描述说话人间的差异,这部分给出了在 MLLR 框架下估计这个转换矩阵的公式.需要说明的是,式(4)给出了对基本 MFCC 特征进行归一化的公式,它同样可以用于模型补偿.

MLLR 自适应方法有多种不同的实现形式,有的只改变模型概率观测密度的均值[4],有的则以一种受限(constrained)的形式同时对模型均值和方差进行转换[5].为了简单起见,这里只考虑对模型均值的转换,即自适应公式满足下式

$$\hat{\mu}=\theta\mu+\boldsymbol{b},\ \hat{\Sigma}=\Sigma\tag{8}$$

特征通常由基本 MFCC 及其一阶二阶差分构成,这里考虑一个比较简单的实现,对于式(7)中的转换参数,有

$$\theta=\begin{bmatrix}\theta^c&&\\&\theta^c&\\&&\theta^c\end{bmatrix}\quad\boldsymbol{b}=\begin{bmatrix}\boldsymbol{b}^c_1\\\boldsymbol{b}^c_2\\\boldsymbol{b}^c_3\end{bmatrix}\tag{9}$$

上式对于基本MFCC和它的一阶二阶差分采用了相同的转换矩阵. 这种假设是合理的, 因为动态特征是由差分得到的, 它们间的转换能够和基本特征共享.

对转换参数 $\lambda=\{\theta^c, b_i^c, i=1, \cdots, 3\}$ 的估计可以采用 EM 算法来求解, 定义辅助函数为

$$Q(\lambda, \lambda_0)=\sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_t(i)\left[K_i-\frac{1}{2}(\boldsymbol{y}_t-\theta\mu_i-\boldsymbol{b})^T\Sigma_i^{-1}(\boldsymbol{y}_t-\theta\mu_i-b)\right]$$
$$=\sum_{i=1}^{N}\frac{n_i}{2}[-(\bar{\mu}_i-\theta\mu_i-\boldsymbol{b})^T\Sigma_i^{-1}(\bar{\mu}_i-\theta\mu_i-\boldsymbol{b})]+K$$

其中 $\gamma_t(i)$ 是 t 时刻的特征属于第 i 个输出的概率, K_i 和 K 是与转换参数无关的常数, 并且

$$n_i=\sum_{t=1}^{T}\gamma_t(i)$$
$$\bar{\mu}_i=\frac{1}{n_i}\sum_{t=1}^{T}\gamma_t(i)\boldsymbol{y}_t$$

Σ_i 是对角阵, 则 Σ_i, $\bar{\mu}_i$ 和 μ_i 可以分块为

$$\Sigma_i=\mathrm{diag}\{\Sigma_{i1}^c, \Sigma_{i2}^c, \Sigma_{i3}^c\}$$
$$\mu_i=[\mu_{i1}^{cT}, \mu_{i2}^{dT}, \mu_{i3}^{cT}]^T$$
$$\bar{\mu}_i=[\bar{\mu}_{i1}^{cT}, \bar{\mu}_{i2}^{dT}, \bar{\mu}_{i3}^{cT}]^T$$

这样, 改写辅助函数, 去掉与转换参数无关的常数, 并把式(6, 7)代进来, 有

$$Q(\lambda, \lambda_0)=-\sum_{i=1}^{N}\sum_{j=1}^{3}\frac{n_i}{2}[(\bar{\mu}_{ij}^c-\theta^c\mu_{ij}^c-\boldsymbol{b}_j^c)^T(\Sigma_{ij}^c)^{-1}(\bar{\mu}_{ij}^c-\theta^c\mu_{ij}^c-\boldsymbol{b}_j^c)]$$
$$=-\sum_{i=1}^{N}\sum_{j=1}^{3}\frac{n_i}{2}[(\boldsymbol{D}\boldsymbol{D}^T\bar{\mu}_{ij}^c-\boldsymbol{D}\theta^l\boldsymbol{D}^T\mu_{ij}^c-\boldsymbol{D}\boldsymbol{b}_j^l)^T(\Sigma_{ij}^c)^{-1}$$
$$(\boldsymbol{D}\boldsymbol{D}^T\bar{\mu}_{ij}^c-\boldsymbol{D}\theta^l\boldsymbol{D}^T\mu_{ij}^c-\boldsymbol{D}\boldsymbol{b}_j^l)]$$
$$=-\sum_{i=1}^{N}\sum_{j=1}^{3}\frac{n_i}{2}[(\bar{\mu}_{ij}^l-\theta^l\mu_{ij}^l-\boldsymbol{b}_j^l)^T(\Sigma_{ij}^l)^{-1}(\bar{\mu}_{ij}^l-\theta^l\mu_{ij}^l-\boldsymbol{b}_j^l)]$$

其中

$$(\Sigma_{ij}^l)^{-1}=\boldsymbol{D}^T(\Sigma_{ij}^c)^{-1}\boldsymbol{D} \quad (10)$$
$$\bar{\mu}_{ij}^l=\boldsymbol{D}^T\bar{\mu}_{ij}^c$$
$$\mu_{ij}^l=\boldsymbol{D}^T\mu_{ij}^c$$

定义 $\boldsymbol{W}^l$ 和 μ_{ij}^l 如下

$$\boldsymbol{W}^l=[\theta^l, \boldsymbol{b}_1^l, \boldsymbol{b}_2^l, \boldsymbol{b}_3^l], \mu_{i1}^l=\begin{bmatrix}\mu_{i1}^l\\1\\0\\0\end{bmatrix}, \mu_{i2}^l=\begin{bmatrix}\mu_{i2}^l\\0\\1\\0\end{bmatrix}, \mu_{i3}^l=\begin{bmatrix}\mu_{i3}^l\\0\\0\\1\end{bmatrix} \quad (11)$$

则有

$$Q(\lambda, \lambda_0)=-\sum_{i=1}^{N}\sum_{j=1}^{3}\frac{n_i}{2}[(\bar{\mu}_{ij}^l-\boldsymbol{W}^l\mu_{ij}^l)^T(\Sigma_{ij}^l)^{-1}(\bar{\mu}_{ij}^l-\boldsymbol{W}^l\mu_{ij}^l)]$$

由式(11)可知, $\boldsymbol{W}^l$ 不是全矩阵, 其中的某些项恒为零. 由 $Q(\lambda, \lambda_0)$ 对 $\boldsymbol{W}^l$ 取导数, 去掉对应 $\boldsymbol{W}^l$ 单元为零的部分, 并令得到的导数为零, 有

$$\sum_{i=1}^{N}\sum_{j=1}^{3}n_i[-(\Sigma_{ij}^l)^{-1}\boldsymbol{W}^l\mu_{ij}^l\mu_{ij}^{lT}+(\Sigma_{ij}^l)^{-1}\bar{\mu}_{ij}^l\mu_{ij}^{lT})]_W=0 \quad (12)$$

上式中 $[*]_W$ 表示把矩阵 $*$ 限制成 $\boldsymbol{W}^l$ 的形式, 即相对应的单元恒为零.

由式(10)知, $(\Sigma_{ij}^l)^{-1}$ 不是对角阵, 文献[1]给出了对这种情况的计算公式. 定义

$$\boldsymbol{A}_{ij}=(\Sigma_{ij}^l)^{-1}$$
$$\boldsymbol{B}_{ij}=\mu_{ij}^l\mu_{ij}^{lT}$$
$$\boldsymbol{C}=\sum_{i=1}^{N}\sum_{j=1}^{3}n_i(\Sigma_{ij}^l)^{-1}\bar{\mu}_{ij}^l\mu_{ij}^{lT}$$

则式(12)可以改写为

$$\sum_{i,j}n_i[\boldsymbol{A}_{ij}\boldsymbol{W}^l\boldsymbol{B}_{ij}]_W=[\boldsymbol{C}]_W$$

把上式两边按行拉成列向量, 则上式左边可以写成直积的形式[6], 有

$$\sum_{i,j}n_i[\boldsymbol{A}_{ij}\otimes\boldsymbol{B}_{ij}^T]_{[]}\vec{\boldsymbol{W}}^l=\vec{\boldsymbol{C}}$$

上式中 $\otimes$ 是直积符号, $\vec{}$ 表示去掉其中对应 $\boldsymbol{W}^l$ 中恒定为零的单元, 并按行把矩阵拉成列向量, $[*]_{[]}$ 表示去掉矩阵 $*$ 与 $\vec{\boldsymbol{W}}^l$ 和 $\vec{\boldsymbol{C}}$ 相对应的行和列后得到的矩阵.

这样就可以计算出 $\boldsymbol{W}^l$, 然后根据式(11)和式(6, 7)得到转换参数 θ^c 和 $\boldsymbol{b}_i^c$, $i=1, 2, 3$, 最后实现模型补偿.

为了叙述的方便, 这里没有考虑包含能量特征的情况, 读者可以参考文献[1]中的分析和讨论.

4　倒谱域下基于共享分块对角转换矩阵的说话人自适应(SASBD)

前一部分假设倒谱域下动态特征和基本特征的转换矩阵是共享的, 实际上这可以被推广到采用分块对角矩阵作为转换矩阵的情况, 这就是本文要介绍的第二种方法. 它对模型修正的基本公式为

$$\mu=\theta\mu+\boldsymbol{b}, \Sigma=\Sigma \quad (13)$$

其中

$$\theta=\begin{bmatrix}\theta^c & & \\ & \theta^c & \\ & & \theta^c\end{bmatrix} \quad \boldsymbol{b}=\begin{bmatrix}\boldsymbol{b}_1^c\\\boldsymbol{b}_2^c\\\boldsymbol{b}_3^c\end{bmatrix} \quad (14)$$

这个公式同式(8, 9)相仿, 但式(9)中 θ^c 受三对角矩阵 θ^l 的约束, 而在上式中, θ^c 是没有任何先验约束的全矩阵.

定义辅助函数如下

$$Q(\lambda, \lambda_0)=-\sum_{i=1}^{N}\sum_{j=1}^{3}\frac{n_i}{2}[(\bar{\mu}_{ij}^c-\theta^c\mu_{ij}^c-\boldsymbol{b}_j^c)^T(\Sigma_{ij}^c)^{-1}(\bar{\mu}_{ij}^c-\theta^c\mu_{ij}^c-\boldsymbol{b}_j^c)]$$

与 SATD 相类似, 定义 $\boldsymbol{W}^c$ 和 μ_{ij}^c 为

$$\boldsymbol{W}^c=[\theta^c, \boldsymbol{b}_1^c, \boldsymbol{b}_2^c, \boldsymbol{b}_3^c], \mu_{i1}^c=\begin{bmatrix}\mu_{i1}^c\\1\\0\\0\end{bmatrix}, \mu_{i2}^c=\begin{bmatrix}\mu_{i2}^c\\0\\1\\0\end{bmatrix}, \mu_{i3}^c=\begin{bmatrix}\mu_{i3}^c\\0\\0\\1\end{bmatrix} \quad (15)$$

则辅助函数可以改写为

$$Q(\lambda, \lambda_0)=-\sum_{i=1}^{N}\sum_{j=1}^{3}\frac{n_i}{2}[(\bar{\mu}_{ij}^c-\boldsymbol{W}^c\mu_{ij}^c)^T(\Sigma_{ij}^c)^{-1}(\bar{\mu}_{ij}^c-\boldsymbol{W}^c\mu_{ij}^c)]$$

由 $Q(\lambda, \lambda_0)$ 对 $\boldsymbol{W}^c$ 求导数, 并令其为零, 有

$$\sum_{i=1}^{N}\sum_{j=1}^{3}n_i[-(\Sigma_{ij}^c)^{-1}\boldsymbol{W}^c\mu_{ij}^c\mu_{ij}^{cT}+(\Sigma_{ij}^c)^{-1}\bar{\mu}_{ij}^c\mu_{ij}^{cT}]=0$$

这个方程可以简单地由文献[4]给出的按行计算的方法求解.

5 实验评测

为了充分反映所提出的自适应算法的性能, 实验分别在以整词建模的孤立词识别系统和以三音子建模的孤立词识别系统上完成. 在所有的测试中, 信号采样率均为 8k, 采用包括 12 维 MFCC 及其一阶二阶差分构成的 36 维特征.

除了本文提出的两种方法外, 进行评测的还包括 MLLR 框架下采用分块对角矩阵和全矩阵的自适应方法. 所有的自适应都在监督方式下完成.

5.1 在整词建模的孤立词识别系统上的测试

实验数据包括 138 人(68 女, 70 男) 的孤立词语音, 每人说 61 个孤立词. 由于该系统识别率相当高, 采用 80 人(40 男, 40 女) 的语音建模, 剩余的语音进行测试时, 识别率达到 97.57%, 在基线识别率已经很高的情况下很难充分反映自适应的效果. 这里构造了一个说话人声道特性强不匹配的测试平台: 采用女声建模, 男声识别. 在这个平台上可以比较出不同自适应方法在声学模型和测试语音间存在不匹配情况下的性能.

图 1 给出了采用不同方法时识别性能随自适应语音数目变化的曲线. 显然在以 20 个孤立词作为自适应数据的情况下, 在采用全矩阵的自适应方法中转换参数还没有得到鲁棒估计. 由图 1 可以看出, SATD 自适应速度最快, SASBD 其次. 当仅仅采用 3 个孤立命令作为自适应语音时 SATD 就可以得到比较满意的效果. 由于这里采用的训练数据和测试数据都是在办公室环境下采用相同的设备录制的, 训练和测试间存在的不匹配主要是说话人间的差异. 实验结果充分说明了 SATD 和 SASBD 在说话人和声学模型间存在不匹配时快速自适应的性能.

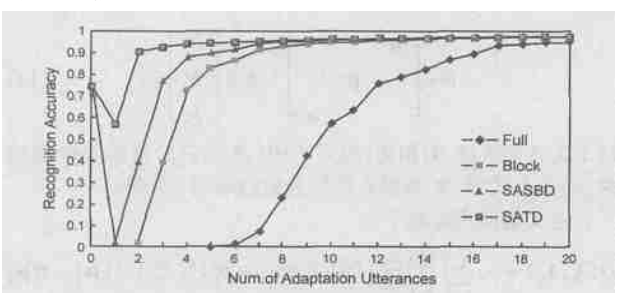

图 1 整词建模的孤立词识别系统上的测试结果(Full 和 Block 分别表示采用全矩阵和分块对角矩阵作为转换矩阵的方法)

5.2 在以三音子为单元建模的孤立词识别系统上的测试

以三音子为单元建模的孤立词识别系统采用与性别无关的非特定人声学模型, 它由 863 数据以及实验室录制的其它语音经降采样后的 8k 数据训练得到. 这个声学模型包括 1426 个混合高斯输出, 每个输出包含 16 个混合单元. 测试数据包括 10 个说话人的 3000 个孤立命令(识别引擎的词表就是这 3000 词).

构造这个测试平台是为了反映所提出的自适应方法在三音子建模的大词汇量语音识别情况下的性能. 由于采用全矩阵的自适应方法速度很慢, 这里进行测试比较的仅包括本文提出的两种方法和采用分块对角矩阵的自适应方法.

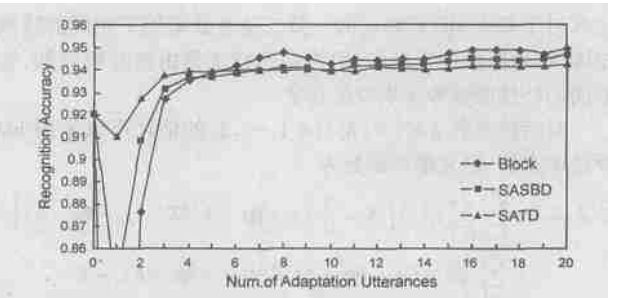

图 2 在以三音子为建模单元的孤立词识别系统上的测试结果(Block 表示采用分块对角矩阵作为转换矩阵的方法)

图 2 给出了这三种自适应方法的识别性能随自适应语音数目变化的曲线. 结果表明 SATD 仍然具有最快的自适应性能, SASBD 的自适应速度较采用分块对角矩阵的方法也要快一些. 不过值得注意的是, 当有充足的自适应数据时, SASBD 和 SATD 相对采用分块对角矩阵的方法性能要差一点. 这是因为采用较少的参数尽管能比较充分地反映说话人间的差异, 但仍然不全面. 对于本文中所提出的两种方法, 尤其是 SATD, 能保证有较快的自适应效果, 并且在有充足自适应数据时性能仍然比较理想.

6 结论

本文提出了两种快速自适应方法, 它们分别由 Log 谱域下的三对角矩阵和倒谱域下的共享分块对角矩阵来描述说话人间的差异. 这两种方法都在最大 在线性回归框架下实现, 同传统的采用倒谱域下的分块对角矩阵和全矩阵作为转换矩阵相比, 所提出的方法能够采用足够少的参数来描述说话人间的差异. 实验结果充分证明了这一点. 此外, 所提出转换矩阵的形式可以被应用到采用多回归的 MLLR 自适应, 以及说话人选择技术和基于转换的本征音等方法中去.

作者简介:

丁国宏 男, 1977 年 3 月出生于湖北应城, 1998 年 7 月获西北工业大学工业自动化专业学士学位, 2001 年 4 月获西北工业大学控制理论与控制工程硕士学位, 现攻读中国科学院自动化研究所模式识别与智能系统专业博士学位, 目前的研究方向主要包括: 说话人自适应, 语音识别噪声环境鲁棒性研究等.

徐 波 男, 1966 年 7 月出生于浙江鄞县, 1988 年毕业于浙江大学电机工程系并获学士学位, 此后在中科院自动化所从事语音、语言信息的处理、识别等方面的学习和研究, 并分别于 1992 年和 1997 年获工学硕士、博士学位, 现为模式识别国家重点实验室副主任, 口语信息处理研究组组长; 清华大学信息学院兼职教授, 国家自然科学基金委评委; 中国声学学会和中国自动化学会委员,《自动化学报》编委, 目前的主要研究方向包括: 语音识别, 自然语言理解, 语音识别嵌入式系统.

(下转第 1719 页)

machine vision[A]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C], Miami Beach: IEEE Press, 1986. 364 – 374.

[2] Zhang Z Y. A Flexible New Technique for Camera Calibration[J]. IEEE Trans., 2000, PAMI-22(11): 1330 – 1334.

[3] Faugras O, et al. Camera Self-calibration: Theory and experiments[A]. Proceedings of the Second European Conference on Computer Vision [C]. London: Springer-Verlag, 1992. 321 – 334.

[4] Anders H. A new approach to hand-eye calibration[A]. Internation Conference on Pattern Recognition[C]. Barcelona: IEEE Press, 2000. 1525 – 1529.

[5] Zhao T, et al. Self-calibration of a camera from video of a walking human[A]. Proceedings of international conference on Pattern Recognition[C]. Quebec City: IEEE Press, 2002. 562 – 567.

[6] Ha J E. 3D structure recovery and calibration under varying intrinsic parameters using known angles[J]. Pattern Recognition, 2001, 34(2): 351 – 359.

[7] Hartley R. An algorithm for self-calibration from several views[A]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C]. Seattle: IEEE Press, 1994. 737 – 744.

[8] Heyden A, et al. Euclidean reconstruction from constant intrinsic parameters[A]. Proceedings 13th international conference on Pattern Recognition[C]. Vienna: IEEE Press, 1996. 339 – 343.

[9] Lourdes A, et al. Linear self-calibration of a rotating and zooming camera[A]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C]. Corfu, Greece: IEEE Press, 1999. 1015 – 1021.

[10] Ma S D. A self-calibration technique for active vision system[J]. IEEE Trans., 1996, Robotics and Automation-12(1): 114 – 120.

[11] Heyden A, et al. An iterative factorization method for projective structure and motion from image sequences[J]. Image and Vision Computing 1999, 17(13): 981 – 991.

作者简介:

刘侍刚　男, 1973 年 11 月生于江西省峡江县, 1997 年获哈尔滨工程大学学士学位, 2001 年获哈尔滨工程大学硕士学位, 现为西安电子科技大学综合业务网国家重点实验室博士研究生, 主要研究方向为计算机视觉、图像处理、三维重建、虚拟现实等方面. Email: xdlsg@tom.com

吴成柯　男, 1938 年出生于安徽黄山, 教授, 博士生导师, 主要研究方向为计算机视觉、三维重建、图形图像处理、视频编码和图像通信等方面, 发表科技论著 4 本, 在国内外杂志上发表论文 100 余篇.

(上接第 1712 页)

参考文献:

[1] Ding G-H, et al. Implementing vocal length normalization in the MLLR framework[A]. Proceedings of International Conference on Spoken Language Processing[C]. Denver: Causal Productions, 2002. 1389 – 1392.

[2] Ding G-H, et al. Transform-based fast speaker adaptation using triple diagonal and shared block diagonal matrices[A]. Proceedings of International Conference on Acoustics, Speech and Signal Processing[C]. Hong Kong: IEEE Signal Processing Society, 2003, 1. 300 – 303.

[3] Lee L, et al. A frequency warping approach to speaker normalization [J]. IEEE Transactions on Speech and Audio Processing, 1998, 6(1): 49 – 60.

[4] Gales M J F, et al. Mean and variance adaptation within the MLLR framework[J]. Computer Speech and Language, 1996, 10(4): 249 – 264.

[5] Digalakis V V, et al. Speaker adaptation using constrained estimation of Gaussian mixtures[J]. IEEE transactions on speech and audio processing, 1995, 3(5): 357 – 366.

[6] 程运鹏. 矩阵论[M]. 西安: 西北工业大学出版社, 1989. 306 – 313.

[7] Huang C, et al. Speaker selection training for large vocabulary continuous speech recognition[A]. Proceedings of International Conference on Acoustics, Speech and Signal Processing[C]. Orlando: IEEE Signal Processing Society, 2002, 1. 609 – 672.

[8] Chen K T, et al. Fast speaker adaptation using eigenspace-based maximum likelihood linear regression[A]. Proceedings of International Conference on Spoken Language Processing[C]. Beijing: China Military Friendship Publish, 2000, 3. 742 – 745.