语音识别中基于低秩约束的本征音子说话人自适应方法

张文林* 张连海 陈 琦 李弼程
(解放军信息工程大学信息系统工程学院 郑州 450002)

摘 要：该文提出一种基于低秩约束的本征音子(Eigenphone)说话人自适应方法。原始的本征音子说话人自适应方法在自适应语料充分时具有很好的效果，然而当自适应语料不足时，出现严重的过拟合现象，导致自适应后的系统可能比自适应前的系统还要差。首先，对协方差矩阵为对角阵的隐马尔可夫-高斯混合模型语音识别系统，推导出一种简化的本征音子矩阵估计算法；然后，对本征音子矩阵引入低秩约束，采用矩阵的核范数作为矩阵秩的凸近似，通过调节核范数的权重因子以有效控制自适应模型的复杂度；最后，给出一种加速近点梯度算法以求解新算法中引入的带有核范数正则项的数学优化问题。汉语连续语音识别的说话人自适应实验表明，引入低秩约束后，本征音子说话人自适应方法的自适应效果得到了明显提高，在5~50 s的自适应数据条件下，均取得了比最大似然线性回归后接最大后验(MLLR+MAP)自适应更佳的识别效果。

关键词：语音识别；说话人自适应；本征音子；低秩约束；近点梯度法
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2014)04-0981-07
DOI: 10.3724/SP.J.1146.2013.00848

Low-rank Constraint Eigenphone Speaker Adaptation Method for Speech Recognition

Zhang Wen-lin Zhang Lian-hai Chen Qi Li Bi-cheng
(*Institute of Information System Engineering, PLA Information Engineering University, Zhengzhou* 450002, *China*)

Abstract: A low-rank constraint eigenphone speaker adaptation method is proposed. Original eigenphone speaker adaptation method performs well when the amount of adaptation data is sufficient. However, it suffers from server overfitting when insufficient amount of adaptation data is provided, possibly resulting in lower performance than that of the unadapted system. Firstly, a simplified estimation alogrithm of the eigenphone matrix is deduced in case of hidden Markov model-Gaussian mixture model (HMM-GMM) based speech recognition system with diagonal covariance matrices. Then, a low-rank constraint is applied to estimation of the eigenphone matrix. The nuclear norm is used as a convex approximation of the rank of a matrix. The weight of the norm is adjusted to control the complexity of the adaptation model. Finally, an accelerated proximal gradient method is adopted to solve the mathematic optimization. Experiments on an Mandarin Chinese continuous speech recognition task show that, the performance of the original eigenphone method is improved remarkably. The new method outperforms the maximum likelihood linear regression followed by maximum a posterriori (MLLR+MAP) methods under 5~50 s adaptation data testing conditions.

Key words: Speech recognition; Speaker adaptation; Eigenphone; Low-rank constraint; Proximal gradient method

1 引言

在现代连续语音识别系统中，说话人自适应是一个必不可少的关键模块。对于传统的基于隐马尔可夫模型(Hidden Markov Model, HMM)-高斯混合模型(Gaussian Mixture Model, GMM)的语音识别系统，说话人自适应技术就是在给定少量说话人相关语料的条件下，根据最大似然(Maximum Likelihood, ML)或最大后验(Maximum A Posteriori, MAP)准则，对说话人无关(Speaker Independent, SI)系统中每一个GMM的高斯均值矢量进行调整，得到说话人相关(Speaker Dependent, SD)系统。

经典的说话人自适应方法可以分为3大类[1]：基于最大后验概率的方法、基于最大似然线性变换的方法和基于说话人聚类的方法，其典型代表分别是最大后验(Maximum A Posteriori, MAP)[1]自适应方法，最大似然线性回归(Maximum Likelihood Linear

2013-06-14收到，2013-12-26改回
国家自然科学基金(61175017)和国家863计划项目(2012AA011603)资助课题
*通信作者：张文林 zwlin_2004@163.com

Regression, MLLR)[2]及本征音(EigenVoice, EV)[3]说话人自适应方法。文献[1]和文献[4]提出了一种基于本征音子(EigenPhone, EP)的说话人自适应方法。与 EV 方法不同，该方法认为对于每一个说话人，其 SD 模型中不同高斯分量均值的变化(相对于 SI 模型)位于一个子空间中，称该子空间为“音子变化子空间(phone variation subspace)”，其基矢量称为“本征音子”，反映了说话人的个体特征，是说话人相关的；而不同高斯分量对应的坐标反映了不同音子之间的相关性信息，是说话人无关的。在训练阶段，可以根据训练数据得到各高斯分量的坐标矢量，在自适应阶段估计未知说话人的本征音子矩阵，即可达到说话人自适应的目的。本征音子自适应方法具有直观的物理意义，在自适应数据充分的情况下，能够得到比 MLLR 方法和 EV 方法更好的结果。然而，其缺点是在自适应数据较少的情况下，极易出现严重的过拟合现象。

在语音信号处理与语音识别领域，正则化方法近年来被越来越多地应用于解决数据稀疏问题和降低模型的复杂度。例如，利用 L1 正则化方法可以得到噪声语音信号的稀疏表达，从而提高噪声条件下的语音识别系统识别率[5]；在子空间 GMM 声学模型中，采用 L1 和 L2 正则化方法[6]可以得到具有稀疏性的模型参数，进一步提高少量数据下的声学建模能力；在基于深层神经网络的语音识别系统[7]中，采用 L1 正则化方法减少神经网络中的非零权值个数，从而在不牺牲系统识别率的情况下大大降低模型复杂度。

实验证明，在本征音子说话人自适应方法中，本征音子个数应随着自适应数据量的增加而不断增大。由于它决定了音子变化子空间的维数，与本征音子矩阵的秩密切相关，因此可以考虑将本征音子矩阵的秩作为正则项，引入低秩矩阵约束来提高本征音子说话人自适应方法的性能。直接将矩阵的秩作为正则项是无法求解的，在矩阵优化问题中，通常采用核范数作为矩阵秩的一个凸近似，从而将原问题转化为一个凸优化问题进行求解[8]。目前，基于核范数的正则化方法已被应用于矩阵恢复[8]、稳健性主分量分析[9]、图像处理[10]等领域，并取得了不错的效果。本文章节安排如下：第 2 节简要介绍了本征音子说话人自适应方法，给出了一种快速实现算法；第 3 节讨论基于核范数正则化的本征音子自适应及其数学优化算法；第 4 节给出了实验结果及分析；最后给出了本文的结论。

2 本征音子说话人自适应

假设 SI 系统中，共 M 个高斯混元，特征矢量维数为 D，令 $\boldsymbol{\mu}_m$ 为第 m 个高斯混元的均值矢量。在第 s 个说话人的 SD 系统中，第 m 个高斯混元的均值矢量用 $\boldsymbol{\mu}_m(s)$ 表示，定义音子变化矢量为 $\boldsymbol{u}_m(s)=\boldsymbol{\mu}_m(s)-\boldsymbol{\mu}_m$。在本征音子说话人自适应中，假设 $\{\boldsymbol{u}_m(s)\}_{m=1}^{M}$ 位于一个说话人相关的 N ($N<<M$) 维子空间中，称该子空间为“音子变化子空间”。设该子空间的原点为 $\boldsymbol{v}_0(s)$，基矢量为 $\{\boldsymbol{v}_n(s)\}_{n=1}^{N}$，称 $\{\boldsymbol{v}_n(s)\}_{n=0}^{N}$ 为第 s 个说话人的本征音子 EP。令第 m 个高斯混元对应的坐标矢量为 $\boldsymbol{y}_m=[y_{m1}\ y_{m2}\cdots y_{mN}]^{\mathrm{T}}$，则 $\boldsymbol{u}_m(s)$ 在音子变化子空间中可以分解为

$$\boldsymbol{u}_m(s)=\boldsymbol{v}_0(s)+\sum_{n=1}^{N}y_{mn}\boldsymbol{v}_n(s)=\boldsymbol{V}(s)\tilde{\boldsymbol{y}}_m \tag{1}$$

其中 $\boldsymbol{V}(s)=[\boldsymbol{v}_0(s)\ \boldsymbol{v}_1(s)\ \cdots\ \boldsymbol{v}_N(s)]$ 为第 s 个说话人的本征音子矩阵，其维数为 $D\times(N+1)$；$\tilde{\boldsymbol{y}}_m=[1\ \boldsymbol{y}_m^{\mathrm{T}}]^{\mathrm{T}}$ 为扩展的高斯混元坐标矢量，其维数为 $N+1$。

假设自适应数据的特征矢量序列为 $\boldsymbol{O}=\{\boldsymbol{o}(1),\boldsymbol{o}(2),\cdots,\boldsymbol{o}(T)\}$，根据最大似然准则，估计说话人相关本征音子矩阵 $\boldsymbol{V}(s)$。采用期望最大化(Expectation Maximization, EM)算法，优化的目标函数为

$$\begin{aligned}Q(\boldsymbol{V}(s))=-\frac{1}{2}\sum_{t}\sum_{m}\gamma_m(t)\big[\boldsymbol{o}(t)-\boldsymbol{\mu}_m-\boldsymbol{u}_m(s)\big]^{\mathrm{T}}\\ \cdot\boldsymbol{\Sigma}_m^{-1}\big[\boldsymbol{o}(t)-\boldsymbol{\mu}_m-\boldsymbol{u}_m(s)\big]\end{aligned} \tag{2}$$

其中 $\gamma_m(t)$ 表示第 t 帧特征矢量属于 SI 模型中第 m 个高斯混元的后验概率，给定自适应数据的标注，它可以通过 Baum-Welch 前后向算法[11]计算得到。$\boldsymbol{\Sigma}_m$ 表示第 m 个高斯混元的协方差矩阵。将式(1)代入式(2)，并令其对 $\boldsymbol{V}(s)$ 的导数为 0，可以得到 $\boldsymbol{V}(s)$ 的求解公式[4]。然而文献[4]给出的求解公式中涉及 $(N+1)D\times(N+1)D$ 维矩阵的逆，对于一个典型的连续语音识别系统，特征维数(D)为 30 ~ 40 维，当 N 较大时(≥ 100)时，存储及计算这样一个高维矩阵是难以实现的。幸运的是，常用的 HMM-GMM 语音识别系统中，$\boldsymbol{\Sigma}_m$ 是一个对角阵，令其第 d 个对角线元素为 $\sigma_{m,d}$，则目标函数式(2)可以简化为

$$\begin{aligned}Q(\boldsymbol{V}(s))=-\frac{1}{2}\sum_{d}\sum_{t}\sum_{m}\gamma_m(t)\\ \cdot\sigma_{m,d}^{-1}\left[o_d(t)-\mu_{m,d}-\boldsymbol{v}_d^{\mathrm{T}}(s)\tilde{\boldsymbol{y}}_m\right]^2\end{aligned} \tag{3}$$

其中 $o_d(t)$ 及 $\mu_{m,d}$ 分别表示特征矢量 $\boldsymbol{o}(t)$ 及均值矢量 $\boldsymbol{\mu}_m$ 的第 d 维，$\boldsymbol{v}_d^{\mathrm{T}}(s)$ 表示本征音子矩阵 $\boldsymbol{V}(s)$ 的第 d 行。对式(3)进行整理可得

$$Q(\boldsymbol{V}(s))=-\frac{1}{2}\sum_{d}\left[\boldsymbol{v}_d^{\mathrm{T}}(s)\boldsymbol{A}_d\boldsymbol{v}_d(s)-\boldsymbol{b}_d^{\mathrm{T}}\boldsymbol{v}_d(s)\right]+C \tag{4}$$

其中

$$\boldsymbol{A}_d=\sum_{t}\sum_{m}\gamma_m(t)\sigma_{m,d}^{-1}\tilde{\boldsymbol{y}}_m\tilde{\boldsymbol{y}}_m^{\mathrm{T}},$$

$$\boldsymbol{b}_d=\sum_{t}\sum_{m}\gamma_m(t)\sigma_{m,d}^{-1}\left[o_d(t)-\mu_{m,d}\right]\tilde{\boldsymbol{y}}_m$$

C 为一个常数。对式(4)求关于 $\boldsymbol{v}_d(s)$ 的导数，并令导数为 0 可得其最优值 $\tilde{\boldsymbol{v}}_d(s)=\boldsymbol{A}_d^{-1}\boldsymbol{b}_d$。由于各行之间的计算相互独立，因此实际计算中，可以对 $\boldsymbol{V}(s)$ 的 D 行进行并行求解。

3　基于低秩约束的本征音子说话人自适应

3.1 基于低秩约束的本征音子说话人自适应原理

在文献[4]中，实验表明，在自适应数据量充足时，本征音子自适应方法能够取得很好的自适应效果，随着数据量的增加，本征音子的个数 N 应逐渐增大；而当数据量不足时，由于无法充分估计本征音子矩阵，会出现严重的过训练现象。为了缓解这一问题，文献[4]引入高斯先验分布，在最大后验准则下得到更为稳健的估计；然而，该方法对识别率的提高有限，只能尽量自适应之后的系统识别率不会下降。本文通过对本征音子矩阵引入低秩约束来解决这一问题。事实上，本征音子的个数 N 与本征音子矩阵的秩密切相关，引入低秩约束可以有效地限制模型的复杂度，防止过训练问题。

数学上直接将矩阵的秩作为约束条件是无法求解的，通常采用矩阵的核范数(nuclear norm)作为矩阵秩的一个凸近似，从而将原问题转化为一个凸优化问题来求解。对于本征音子矩阵 $\boldsymbol{V}(s)$，令 κ_i 为其第 i 个奇异值，则其核范数为 $\|\boldsymbol{V}(s)\|_*=\sum_i\kappa_i$。事实上，用矩阵的核范数来近似矩阵的秩，与压缩感知中常用的以矢量的 L1 范数来近似其 L0 范数是类似的：矩阵的秩等于奇异值矢量的 L0 范数，其核范数就等价于奇异值矢量的 L1 范数。对目标函数式(4)引入核范数正则项，新的优化问题可以写为

$$\begin{aligned}\widehat{\boldsymbol{V}}(s)&=\arg\min_{\boldsymbol{V}(s)}\left[\widetilde{Q}(\boldsymbol{V}(s))\right]\\&=\arg\min_{\boldsymbol{V}(s)}\left[-Q(\boldsymbol{V}(s))+\lambda\|\boldsymbol{V}(s)\|_*\right]\end{aligned}\tag{5}$$

其中 $\lambda>0$ 为核范数权重，λ 越大所得到的矩阵 $\widehat{\boldsymbol{V}}(s)$ 的核范数越小，其秩也就越低。

3.2 基于低秩约束的本征音子矩阵优化算法

式(5)是一个凸优化问题，已有多种求解算法，如快速迭代收缩-阈值算法(Fast Iterative Shrinkage-Thresholding Algorithm, FISTA)[12]，ADMiRA算法[13]，奇异值阈值法(Singular Value Thresholding, SVT)[14] 等。本文采用一种加速近点梯度法(Accelerated Proximal Gradient, APG)[15,16]对其进行求解。

对于一个凸函数 $R(\boldsymbol{X})$ ($\boldsymbol{X}\in R^{m\times n}$)，其在矩阵 $\boldsymbol{X}$ 的近点映射(proximal mapping)[16]为

$$\text{prox}_R(\boldsymbol{X})=\arg\min_{\boldsymbol{Y}}\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2+R(\boldsymbol{Y})\tag{6}$$

其中，$\|\boldsymbol{X}\|_F$ 表示矩阵的 Frobenius 范数。式(6)可以理解为在矩阵 $\boldsymbol{X}$ 附近找一个离其尽量近的矩阵 $\boldsymbol{Y}$，使得 $R(\boldsymbol{Y})$ 尽量小。当 R 为某一个集合 $\mathbb{C}$ 的指示函数时，$\text{prox}_R(\boldsymbol{X})$ 就是点 $\boldsymbol{X}$ 向 $\mathbb{C}$ 的一个投影。因此，一个函数的近点映射算子可以视为集合投影的一个推广。对于矩阵的核范数，其近点映射是一种收缩算子，其定义如下：若矩阵 $\boldsymbol{X}$ 的奇异值分解为 $\boldsymbol{X}=\boldsymbol{P}\text{diag}(\kappa_1,\kappa_2,\cdots,\kappa_n)\boldsymbol{Q}^{\mathrm{T}}$，则核范数 $\|\cdot\|_*$ 的近点映射算子为

$$\text{prox}_{\lambda\|\cdot\|_*}(\boldsymbol{X})=\boldsymbol{P}\text{diag}\left((\kappa_i-\lambda)_+\right)\boldsymbol{Q}^{\mathrm{T}}\tag{7}$$

其中 $\text{diag}(\kappa_1,\kappa_2,\cdots,\kappa_n)$ 表示对角线元素为 $\kappa_1,\kappa_2,\cdots,\kappa_n$ 的对角矩阵，$(\kappa_i-\lambda)_+=\max\{\kappa_i-\lambda,0\}$。

对于凸优化问题 $\min\widetilde{Q}(\boldsymbol{X})=Q(\boldsymbol{X})+R(\boldsymbol{X})$，其中 $Q(\boldsymbol{X})$ 为一个可微函数，其在 $\boldsymbol{X}$ 处的梯度记为 $\nabla Q(\boldsymbol{X})$，则近点梯度算法可以归结为迭代公式：

$$\boldsymbol{X}^{(k+1)}=\text{prox}_{\eta^{(k)}R}\left[\boldsymbol{X}^{(k)}-\eta^{(k)}\nabla Q\left(\boldsymbol{X}^{(k)}\right)\right]\tag{8}$$

其中，$\boldsymbol{X}^{(k)}$ 为第 k 次迭代前的结果，$\eta^{(k)}$ 是一个正的不增序列，且 $\lim_{k\to\infty}\eta^{(k)}=0$，$\sum_k\eta^{(k)}=\infty$。事实上，近点梯度算法可以视为投影梯度算法的一个推广：在每一步迭代过程中，首先进行步长为 $\eta^{(k)}$ 的梯度下降得到 $\boldsymbol{X}^{(k)}-\eta^{(k)}\nabla Q(\boldsymbol{X}^{(k)})$，然后再对其作函数 $\eta^{(k)}R$ 的近点映射；当 R 为某一个集合 $\mathbb{C}$ 的指示函数时，该算法即等价为求解带有 $\boldsymbol{X}\in\mathbb{C}$ 约束问题的投影梯度算法。针对本文提出的低秩约束下的本征音子自适应问题式(5)，由于原始问题的目标函数 $Q(\boldsymbol{V}(s))$ (式(4))是可微的，其在矩阵 $\boldsymbol{V}(s)$ 处的梯度为

$$\nabla Q(\boldsymbol{V}(s))=\left[\frac{\partial Q(\boldsymbol{V}(s))}{\partial\boldsymbol{v}_1(s)},\ \frac{\partial Q(\boldsymbol{V}(s))}{\partial\boldsymbol{v}_2(s)},\cdots,\frac{\partial Q(\boldsymbol{V}(s))}{\partial\boldsymbol{v}_d(s)}\right]^{\mathrm{T}}$$

其中 $\dfrac{\partial Q(\boldsymbol{V}(s))}{\partial\boldsymbol{v}_d(s)}=-\boldsymbol{A}_d\boldsymbol{v}_d(s)+\boldsymbol{b}_d$；则本文针对问题式(5)的一个加速近点投影算法流程如下(为了简洁起见，将 $\boldsymbol{V}(s)$ 简记为 $\boldsymbol{V}$)：

(1)初始化　$k=0$，$t^{(0)}=t^{(-1)}=1$，$\boldsymbol{V}^{(0)}=\boldsymbol{V}^{(-1)}=0$，$\eta^{(0)}=1.0$，计算 $\widetilde{Q}(\boldsymbol{V}^{(0)})$；

(2)设置 $\boldsymbol{Y}^{(k)}=\boldsymbol{V}^{(k)}+\dfrac{t^{(k-1)}-1}{t^{(k)}}\left(\boldsymbol{V}^{(k)}-\boldsymbol{V}^{(k-1)}\right)$；

(3)计算

$$\boldsymbol{V}^{(k+1)}=\text{prox}_{\eta^{(k)}\lambda\|\cdot\|_*}\left[\boldsymbol{Y}^{(k)}-\eta^{(k)}\nabla Q\left(\boldsymbol{Y}^{(k)}\right)\right]$$

(4) 若 $\widetilde{Q}(\boldsymbol{V}^{(k+1)})>\widetilde{Q}(\boldsymbol{V}^{(k)})$，则设置 $\eta^{(k)}\leftarrow 0.8\eta^{(k)}$，转步骤(3)；

(5)若 $\dfrac{\left|\widetilde{Q}(\boldsymbol{V}^{(k+1)})-\widetilde{Q}(\boldsymbol{V}^{(k)})\right|}{\left|\widetilde{Q}(\boldsymbol{V}^{(k)})\right|}<10^{-5}$，则停止迭代

过程，令 $\widehat{\boldsymbol{V}}=\boldsymbol{V}^{(k+1)}$ ；否则，设置 $t^{(k+1)}=\dfrac{1+\sqrt{1+4(t^{(k)})^2}}{2}$, $\eta^{(k+1)}=\eta^{(k)}$, $k\leftarrow k+1$，转步骤(2)。

上述算法中，第(2)步采用动量法(momentum method)加快迭代收敛过程。其中，$t^{(k)}$ 的计算采用了文献[12]中给出的一个经验公式(第(5)步)；在迭代初始时刻($k=0$ 时)，$(t^{(k-1)}-1)/t^{(k)}=0$ ；而在 $k\to\infty$ 时，$(t^{(k-1)}-1)/t^{(k)}\to 1$。实验证明，这一方法可以明显加快算法的迭代收敛过程。第(3)步即是近点梯度下降法的迭代公式。其中，λ 是核范数的权重，$\eta^{(k)}$ 是第 k 步的下降步长。这里对 $\eta^{(k)}$ 采用了一种 1 维线性搜索的方法：在第(4)步当检测到迭代前后目标函数 $\tilde{Q}$ 变大时，按 0.8 的系数减小步长 $\eta^{(k)}$，重新回到第(3)步进行迭代过程。不难证明，这一步长选择方法满足迭代收敛条件。最后一步，检查迭代前后 $\tilde{Q}$ 的相对减少量是否小于 10^{-5}，若“是”则停止迭代过程，否则重新回到步骤(2)进行迭代。

4 实验结果及分析

为了验证本文算法的有效性，本节针对一个典型的汉语连续语音识别系统进行了实验。实验数据采用微软语料库[17]，训练集共有 100 个说话人，共 19688 句话，约为 33 个小时的数据；测试集共 25 个说话人，每人 20 句话，每句话的平均时长约为 5 s。采用典型的 13 维美尔频率倒谱系数(Mel-Frequency Cepstrum Coefficients, MFCC)及其一阶和二阶差分系数，总的特征矢量维数为 39。基线系统中的说话人无关模型利用 HTK 工具包(3.4.1 版本)[11]训练得到，采用三音子有调声韵母作为声学建模单元，每个 HMM 模型含有 3 个输出状态，每个状态共 8 个高斯混元，三音子聚类后总的高斯混元数为 19136。训练阶段采用基于回归树(32 个回归类)的 MLLR 自适应方法得到 100 个训练说话人相关模型。识别阶段，以 HTK 中的 HVite 工具为解码器进行连续语音识别，采用有调音节全连接的解码网络，不采用语法模型。这种解码网络的系统对声学模型的要求最高，可以更好地测试声学模型自适应的效果。在说话人自适应实验中，对每个测试说话人随机抽取 10 句话作为自适应数据，用于对 SI 声学模型进行有监督说话人自适应；将剩下的 10 句话作为测试数据，在其上统计有调音节的平均正确识别率[18]作为实验结果。在测试数据上，SI 模型的平均正确识别率为 53.04%(文献[17]中结果为 51.21%)。实验中，本文针对下列说话人自适应算法进行对比实验：

(1)MLLR+MAP：最大似然线性回归(MLLR)后接最大后验估计(MAP)的自适应算法，根据文献[5]中的实验结果，采其最好的实验设置：即对 MLLR 采用包含 32 个回归类的回归树及分块对角变换矩阵，对 MAP 其先验权重设置为 10；

(2)ML-EP：基于最大似然估计的本征音子自适应算法，本征音子个数(N)取 50 或 100；

(3)MAP-EP：基于最大后验估计的本征音子自适应算法，先验高斯分布的方差的倒数(σ^{-2})从 10 调整到 2000；

(4)LR-EP：本文提出的基于低秩约束的本征音子自适应算法，核范数权重(λ)从 10 调整到 200。

其中(1)是目前常用的说话人自适应算法，(2)和(3)是本文作者近期提出的本征音子说话人自适应算法，(4)是本文提出的基于低秩约束的本征音子自适应算法。为了比较各方法在不同自适应数据量下的自适应效果，分别对 1 句话，2 句话，4 句话，6 句话，8 句话和 10 句话的自适应数据进行了有监督说话人自适应实验。

表 1 中给出了前 3 种自适应算法的典型实验结果，每种算法的最好结果在表中以黑体标明。由表 1 可见，在自适应数据量充足(≥ 4 句话)的情况下，当 $N=50$ 时，ML-EP 算法的性能优于 MLLR+MAP 算法；而当 $N=100$ 时，其性能下降明显，在 4~10 句话时均达不到 MLLR+MAP 的自适应性能。这是由于对于 ML-EP 算法，$N=100$ 时要估计的参数数量比 $N=50$ 时多出一倍，即使数据量为 10 句话，仍无法得到本征音子矩阵的充分估计，出现过拟合现象。这种现象在数据量少时(≤ 2 句话)尤为明显：无论是 $N=50$ 还是 $N=100$, ML-EP 算法的识别率甚至低于自适应前的 SI 系统。这一严重的过拟合现象在引入高斯先验分布后得到了一定的缓解。由表 1 结果可见，MAP-EP 算法在 1~2 句话时通过调整高斯先验分布的方差(σ^2)可以大大提高自适应后的系统正识率：在 $N=50, \sigma^{-2}=1000$ 时，平均正识率分别达到 53.92%和 54.28%；在 $N=100$, $\sigma^{-2}=2000$ 时，平均正识率也能达到 53.69%和 54.28%。这些结果已经接近甚至超出了 MLLR+MAP 算法的最好结果(分别为 53.32%和 54.93%)。然而，在该参数设置下，当自适应数据量充足时(≥ 4 句话)，却制约了 MAP-EP 算法的性能；此时，减少 σ^{-2} 的值，可以提高系统的正识率：在 $N=100$, $\sigma^{-2}=10$ 时，10 句话时平均正识率达到 60.70%，其结果略高于 MLLR+MAP 算法的最好结果。该现象说明，在 $N=100$ 时，通过引入适当的约束，可以提高系统的自适应性能。

根据上述结论，本文针对 $N=50$ 和 $N=100$ 时的本征音子自适应算法，引入低秩矩阵约束，将本征音子矩阵核范数的权重(λ)从 10 调整到 200，典型的实验结果如表 2 所示。在每种测试条件下，对 25 个测试说话人，计算了其本征音子矩阵秩的平均值，在表 2 中以括号中的数字表示。

由表 2 可见，在引入低秩约束后，通过调整核范数的权重，可以使得本征音子自适应算法的效果

表 1 3 种已有自适应算法的正确识别率(%)

自适应方法	参数设置	自适应数据量					
		1	2	4	6	8	10
MLLR+MAP	–	**53.32**	**54.93**	**57.83**	**58.50**	**59.65**	**60.16**
ML-EP	$N=50$	**33.74**	**51.38**	**58.16**	**59.00**	**59.84**	**60.62**
	$N=100$	19.14	41.46	54.30	57.91	59.44	60.13
MAP-EP ($N=50$)	$\sigma^{-2}=10$	43.26	53.67	**58.18**	**59.11**	**59.78**	**60.45**
	$\sigma^{-2}=100$	50.08	53.69	56.71	58.35	59.21	59.80
	$\sigma^{-2}=1000$	**53.92**	**54.28**	55.35	56.13	56.95	57.41
	$\sigma^{-2}=2000$	53.63	54.13	54.80	55.43	56.27	56.69
MAP-EP ($N=100$)	$\sigma^{-2}=10$	27.91	44.63	53.78	57.39	**59.61**	**60.70**
	$\sigma^{-2}=100$	45.24	50.31	**55.77**	**57.55**	59.34	60.30
	$\sigma^{-2}=1000$	53.29	54.22	55.75	56.78	57.41	58.29
	$\sigma^{-2}=2000$	**53.69**	**54.28**	55.52	56.34	56.55	57.74

表 2 基于低秩约束的本征音子说话人自适应后平均正确识别率(%)(括号中数字为所有测试说话人本征音子矩阵秩的平均值)

自适应方法	参数设置	自适应数据量					
		1	2	4	6	8	10
LR-EP($N=50$)	$\lambda=60$	53.55	55.52	58.18	58.85	59.86	60.53
		(24.6)	(32.5)	(37.0)	(37.2)	(37.4)	(37.4)
	$\lambda=70$	53.68	55.56	58.14	59.13	59.76	60.55
		(21.8)	(30.6)	(36.4)	(37.0)	(37.2)	(37.4)
	$\lambda=80$	**53.92**	**55.62**	**58.29**	**59.21**	**59.95**	**60.57**
		(19.6)	**(29.2)**	**(35.2)**	**(36.1)**	**(37.1)**	**(37.4)**
	$\lambda=90$	53.80	55.56	58.10	59.00	59.80	60.37
		(17.2)	(27.4)	(33.4)	(35.0)	(36.2)	(37.4)
	$\lambda=100$	53.73	55.48	57.87	58.37	59.59	60.34
		(15.4)	(26.3)	(32.2)	(35.0)	(36.2)	(37.4)
LR-EP($N=100$)	$\lambda=100$	53.59	54.64	57.32	59.17	59.82	61.02
		(24.0)	(31.5)	(35.6)	(37.0)	(37.3)	(37.6)
	$\lambda=110$	53.94	54.60	57.87	59.04	60.18	61.27
		(22.5)	(30.4)	(35.0)	(36.8)	(37.2)	(37.4)
	$\lambda=120$	**54.26**	**55.32**	**58.02**	**59.40**	**60.21**	**61.32**
		(21.0)	**(29.4)**	**(34.8)**	**(36.3)**	**(37.2)**	**(37.4)**
	$\lambda=130$	54.24	55.10	57.62	59.27	60.09	61.29
		(19.6)	(28.4)	(33.8)	(36.0)	(36.8)	(37.1)
	$\lambda=140$	54.11	55.01	57.26	59.25	60.13	61.20
		(18.5)	(27.4)	(33.0)	(35.4)	(36.6)	(37.0)

得到明显提升。当 $N=50,\lambda=80$ 时，在 1,2,4,6,8 和 10 句话自适应数据量下，平均正确识别率分别为 53.92%, 55.62%, 58.29%, 59.21%,59.95%和 60.57%，这些结果均优于 ML-EP, MAP-EP 及 MLLR+MAP 的最好结果。当 $N=100,\lambda=120$ 时，在 2 句话和 4 句话自适应数据量下，平均正确识别率分别为 55.32%和 58.02%，仅略低于 $N=50,\lambda=80$ 时的结果；在 1,6,8 和 10 句话自适应数据量下，平均正确识别率分别为 54.26%, 59.40%, 60.21%和 61.32%，这是相同自适应数据量下所有测试系统中的最好结果。与 MAP-EP 算法不同，LR-EP 算法可以在同一参数设置下(如 $N=50$ 时，λ 取为 80; $N=100$ 时，λ 取为 120)应对不同的自适应数据量，既不会在数据量少时出现过拟合现象，也不会在数据量充分时出现欠拟合现象。

由表 2 可以看出，本征音子矩阵的秩随着核范数权重 λ 的变化而不同。在相同的自适应数据量下，随着 λ 的增大，本征音子矩阵的秩随之增大；而在相同的权重(λ)设置下，随着自适应数据量的增加，本征音子矩阵的秩也随之增大，此结果与理论分析相吻合；随着自适应数据量的增加，更多的参数可以得到稳健的估计，应该增大音子变化子空间的维数，而音子变化子空间的维数与本征音子矩阵的秩是等价的。对比 $N=50,\lambda=80$ 和 $N=100,\lambda=120$ 时的结果可见，两种参数设置下，对于各种自适应数据量所得到的本征音子矩阵的秩几乎是相同的，两种参数设置均找到了各种自适应数据量下的本征音子矩阵秩的最佳值；相比而言，$N=100,\lambda=120$ 时的结果比 $N=50,\lambda=80$ 时的结果略好，这可能是由于随着 N 的增大，每个高斯混元的坐标矢量 $\boldsymbol{y}_m$ 的维数增加，对音子变化子空间的描述更为精确，从而使得系统具有更强的自适应能力。

在上述实验中，由于实验语料的限制，本文在测试集上通过调整参数 N 及 λ 以得到最佳的识别效果，并将其与其它方法的最佳结果相比较以体现新方法的优越性。在实际应用中，应建立一个独立于训练集数据之外的开发集数据，调整参数 N 及 λ，将开发集上识别率最高时对应的参数作为其最佳取值。

5 结论

本文提出了一种基于低秩约束的本征音说话人自适应方法。新方法在本征音子矩阵估计过程中，引入低秩约束，用矩阵的核范数作为其秩的一个凸近似，对优化的目标函数引入带有核范数的正则项，并采用加速近点梯度算法得到本征音子矩阵的迭代优化算法。引入低秩约束后，可以有效地对自适应模型的复杂度进行控制，在数据量少时得到低维音子变化子空间，在数据量充足时得到高维音子变化子空间。实验证明，新算法在各种自适应数据量下均优于经典的 MLLR+MAP 自适应算法及原始的本征音子自适应算法。

参 考 文 献

[1] Zhang Wen-lin, Zhang Wei-qiang, Li Bi-cheng, *et al*.. Bayesian speaker adaptation based on a new hierarchical probabilistic model[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2012, 20(7): 2002–2015.

[2] Zhang Shi-lei and Qin Yong. Model dimensionality selection in bilinear transformation for feature space MLLR rapid speaker adaptation[C]. Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Kyoto, Japan, 2012: 4353–4356.

[3] 张文林, 牛铜, 张连海, 等. 基于最大似然可变子空间的快速说话人自适应方法[J]. 电子与信息学报, 2012, 34(3): 571–575. Zhang Wen-lin, Niu Tong, Zhang Lian-hai, *et al*.. Rapid speaker adaptation based on maximum-likelihood variable subspace[J]. *Journal of Electronics & Information Technology*, 2012, 34(3): 571–575.

[4] Zhang Wen-lin, Zhang Wei-qiang, and Li Bi-cheng. Speaker adaptation based on speaker-dependent eigenphone estimation[C]. Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, Hawaii, USA, 2011: 48–52.

[5] Gemmeke J and Van hamme H. Advances in noise robust digit recognition using hybrid exemplar-based techniques [C]. Proceedings of Interspeech, Oregon, Portland, 2012.

[6] Lu L, Ghoshal A, and Renals S. Regularized subspace Gaussian mixture models for speech recognition[J]. *IEEE Signal Processing Letters*, 2011, 18(7): 419–422.

[7] Yu D, Seide F, Li G, *et al*.. Exploiting sparseness in deep neural networks for large vocabulary speech recognition[C]. Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Kyoto, Japan, 2012: 4409–4412.

[8] Deledalle C, Vaiter S, Peyre G, *et al*.. Risk estimation for matrix recovery with spectral regularization[OL]. http://arxiv.org/abs/1205.1482, 2012.

[9] Candes E J, Li X, Ma Y, *et al*.. Robust principal component analysis?[J]. *Journal of ACM*, 2011, 58(3): DOI: 10.1145/1970392. 1970395.

[10] Chen Chih-fan, Wei Chia-po, and Wang Y C F. Low-rank matrix recovery with structural incoherence for robust face recognition[C]. Proceedings of Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 2012: 2618–2625.

[11] Young S, Evermann G, Gales M, *et al*.. The HTK book (for

HTK version 3.4)[OL]. http://htk.eng.cam.ac.uk/docs/docs.shtml. 2009.

[12] Beck A and Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems[J]. *SIAM Journal on Imaging Sciences*, 2009, 2: 183–202.

[13] Lee K and Bresler Y A. Atomic decomposition for minimum rank approximation[J]. *IEEE Transactions on Information Theory*, 2010, 56(9): 4402–4416.

[14] Cai J, Candes E, and Shen Z. A singular value thresholding algorithm for matrix completion[J]. *SIAM Journal on Optimization*, 2010, 20(4): 1956–1982.

[15] Toh K C and Yun S. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares prolems[J]. *Pacific Journal of Optimization*, 2010, 6(3): 615–640.

[16] Parikh N and Boyd S. Proximal algorithms[J]. *Foundations and Trends in Optimization*, 2013, 1(3): 123–231.

[17] Chang E, Shi Y, Zhou J, *et al.*. Speech lab in a box: a Mandarin speech toolbox to jumpstart speech related research[C]. Proceedings of Eurospeech, Scandinavia, 2001: 2799–2802.

张文林： 男，1982 年生，博士生，研究方向为语音信号处理、语音识别、机器学习.

张连海： 男，1974 年生，副教授，研究方向为语音信号处理、语音编码、语音识别.

陈　琦： 男，1974 年生，讲师，研究方向为语音信号处理、语音识别、音频水印.