基于本征音子说话人子空间的说话人自适应算法

屈 丹* 张文林
(信息工程大学信息系统工程学院 郑州 450000)

摘 要：本征音子说话人自适应算法在自适应数据量充足时可以取得很好的自适应效果，但在自适应数据量不足时会出现严重的过拟合现象。为此该文提出一种基于本征音子说话人子空间的说话人自适应算法来克服这一问题。首先给出基于隐马尔可夫模型-高斯混合模型(HMM-GMM)的语音识别系统中本征音子说话人自适应的基本原理。其次通过引入说话人子空间对不同说话人的本征音子矩阵间的相关性信息进行建模；然后通过估计说话人相关坐标矢量得到一种新的本征音子说话人子空间自适应算法。最后将本征音子说话人子空间自适应算法与传统说话人子空间自适应算法进行了对比。基于微软语料库的汉语连续语音识别实验表明，与本征音子说话人自适应算法相比，该算法在自适应数据量极少时能大幅提升性能，较好地克服过拟合现象。与本征音自适应算法相比，该算法以较小的性能牺牲代价获得了更低的空间复杂度而更具实用性。

关键词：语音信号处理；说话人自适应；本征音子；本征音子说话人子空间；低秩约束；本征音
中图分类号：TN912.34 **文献标识码**：A **文章编号**：1009-5896(2015)06-1350-07
DOI:10.11999/JEIT141264

Speaker Adaptation Method Based on Eigenphone Speaker Subspace for Speech Recognition

Qu Dan Zhang Wen-lin
(*Institute of Information System Engineering*, *PLA Information Engineering University*, *Zhengzhou* 450000, *China*)

Abstract: The eigenphone speaker adaptation method performs well when the amount of adaptation data is sufficient. However, it suffers from severe over-fitting when insufficient amount of adaptation data is provided. A speaker adaptation method based on eigenphone speaker subspace is proposed to overcome this problem. Firstly, a brief overview of the eigenphone speaker adaptation method is presented in case of Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) based speech recognition system. Secondly, speaker subspace is introduced to model the inter-speaker correlation information among different speakers' eigenphones. Thirdly, a new speaker adaptation method based on eigenphone speaker subspace is derived from estimation of a speaker dependent coordinate vector for each speaker. Finally, a comparison between the new method and traditional speaker subspace based method is discussed in detail. Experimental results on a Mandarin Chinese continuous speech recognition task show that compared with original eigenphone speaker adaptation method, the performance of the eigenphone speaker subspace method can be improved significantly when insufficient amount of adaptation data is provided. Compared with eigenvoice method, eigenphone speaker subspace method can save a great amount of storage space only at the expense of minor performance degradation.

Key words: Speech signal processing; Speaker adaptation; Eigenphone; Eigenphones' speaker subspace; Low-rank constraint; Eigenvoice

1 引言

连续语音识别系统中，训练数据与测试数据不匹配会造成系统性能的急剧下降。声学模型自适应技术就是根据少量的测试数据对声学模型进行调整，增加其与测试数据的匹配程度，从而提高系统的识别性能。造成训练与测试数据不匹配的因素包括说话人、传输信道或说话噪声环境的不同，相应的自适应技术分别称为“说话人自适应”[1]、“信道自适应”[2]或“环境自适应”[3]。说话人自适应技术的方法也可以应用于信道自适应或环境自适应。说话人自适应通常包括特征层自适应[4,5]和声学模型自适应，因此，声学模型的说话人自适应[1]是当前语音识别系统一个必不可少的重要组成部分。

声学模型的说话人自适应就是利用少量的未知

2014-09-30 收到，2014-12-29 改回
国家自然科学基金(61175017, 61302107 和 61403415)资助课题
*通信作者：屈丹 qudanqudan@sina.com

说话人语料(自适应语料)，在最大似然或最大后验准则下，将说话人无关(Speaker-Independent, SI)声学模型调整至说话人相关(Speaker-Dependent, SD)声学模型，使得语音识别系统更具说话人针对性，从而提高系统的识别率。在隐马尔可夫模型的连续语音识别系统框架下，主流的说话人自适应技术可分为三大类[1]：基于最大后验概率的方法、基于变换的自适应方法和基于说话人子空间的自适应方法，分别以最大后验(Maximum A Posteriori, MAP)自适应方法、最大似然线性回归(Maximum Likelihood Linear Regression, MLLR)及本征音(Eigen Voice, EV)方法[6]及其拓展算法为代表。2004年，文献[7]通过对 SD 声学模型中各高斯混元均值矢量相对于 SI 声学模型的变化量进行子空间分析，得到一种新的子空间分析方法。该方法与说话人子空间中的“本征音”相类似，因此称该子空间的基矢量为“本征音子(Eigen Phone, EP)”，该空间为“音子变化子空间”。但文献[7]提出的方法是一种“多说话人”声学建模技术，只能得到训练集中说话人相关的声学模型，对于测试集中的未知说话人没有给出其声学模型的自适应方法。

2011 年，文献[8]提出了一种基于本征音子的说话人自适应方法，克服了文献[7]本征音子模型的不足，能够对测试集未知说话人进行自适应。由于该方法对于每个未知说话人需要估计一个扩展的本征音子矩阵，其参数较多，在自适应数据量较少时，极易出现过拟合现象；即使对参数估计过程引入各种正则化方法，其自适应效果仍达不到基于说话人子空间的方法[9,10]。对于传统 MLLR 说话人自适应方法，为了提高其在少量自适应数据条件下的性能，有学者提出在训练阶段寻找MLLR线性变换矩阵的一组基，在自适应阶段利用这组基估计新的变换矩阵的线性组合，从而减少待估参数数量。这种方法称为“本征空间 MLLR”自适应方法[11−14]。该方法本质上是将说话人子空间的思想用于说话人相关变换矩阵的估计，对变换矩阵建立了一个说话人子空间。

为此，本文将上述思想引入本征音子说话人自适应方法中，提出了基于本征音子说话人子空间的说话人自适应算法。新方法充分利用了扩展的本征音子矩阵也是说话人相关的这一特点，对本征音子的说话人子空间进行建模。与本征音子自适应方法相比，该方法在少量自适应数据量下具有良好的性能，很大程度克服了过拟合现象。与说话人子空间自适应方法相比，新方法的子空间基矢量的维数大大降低，具有更低的空间复杂度。本文章节安排如下：第 2 节给出了本征音子说话人自适应方法；第 3 节讨论基于本征音子说话人子空间的自适应方法的数学优化算法及与说话人子空间自适应方法的比较；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算法

2.1 音子变化子空间及本征音子

本文仅讨论基于隐马尔可夫模型的连续语音识别系统的说话人自适应。假设在 SI 声学模型中，共有 M 个高斯混元，特征矢量维数为 D，训练集合中共有 S 个说话人。令 $\boldsymbol{\mu}_m$ 和 $\boldsymbol{\mu}_m^{(s)}$ 分别为 SI 模型和第 s 个说话人 SD 模型中第 m 个高斯混元的均值矢量。定义音子变化矢量 $\boldsymbol{u}_m^{(s)}$ 为 $\boldsymbol{u}_m^{(s)}=\boldsymbol{\mu}_m^{(s)}-\boldsymbol{\mu}_m$。在本征音子说话人自适应中，对于第 s 个说话人，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 $\boldsymbol{\Pi}^{(s)}$ 中，称 $\boldsymbol{\Pi}^{(s)}$ 为说话人相关的“音子变化子空间”。设 $\boldsymbol{\Pi}^{(s)}$ 的原点为 $\boldsymbol{v}_0^{(s)}$，基矢量为 $\{\boldsymbol{v}_n^{(s)}\}_{n=1}^{N}$，称 $\{\boldsymbol{v}_n^{(s)}\}_{n=1}^{N}$ 为第 s 个说话人的本征音子(Eigen Phone, 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

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

其中，$\boldsymbol{V}^{(s)}=[\boldsymbol{v}_1^{(s)}\ \boldsymbol{v}_2^{(s)}\ \cdots\ \boldsymbol{v}_N^{(s)}]$ 和 $\widetilde{\boldsymbol{V}}^{(s)}=[\boldsymbol{v}_0^{(s)}\ \ \boldsymbol{V}^{(s)}]$ 分别为第 s 个说话人的本征音子矩阵和扩展本征音子矩阵，其维数分别为 $D\times N$ 和 $D\times(N+1)$；$\boldsymbol{y}_m$ 和 $\tilde{\boldsymbol{y}}_m=[1\ \boldsymbol{y}_m^{\mathrm{T}}]^{\mathrm{T}}$ 为高斯混元坐标矢量和扩展高斯混元坐标矢量，其维数分别为 N 和 $N+1$。在训练阶段，通过对训练说话人相关声学模型的音子变化超矢量 $\boldsymbol{u}_m=[\boldsymbol{u}_m^{(1)\mathrm{T}}\ \boldsymbol{u}_m^{(2)\mathrm{T}}\cdots\ \boldsymbol{u}_m^{(S)\mathrm{T}}]^{\mathrm{T}}$ 进行主分量分析可以得到各高斯混元的坐标矢量 $\{\boldsymbol{y}_m\}_{m=1}^{M}$ [8]，即根据式(1)，$\boldsymbol{u}_m$ 可以分解为

$$\boldsymbol{u}_m=\begin{bmatrix}\boldsymbol{u}_m^{(1)}\\ \boldsymbol{u}_m^{(2)}\\ \vdots\\ \boldsymbol{u}_m^{(S)}\end{bmatrix}=\begin{bmatrix}\boldsymbol{v}_0^{(1)}+\boldsymbol{V}^{(1)}\boldsymbol{y}_m\\ \boldsymbol{v}_0^{(2)}+\boldsymbol{V}^{(2)}\boldsymbol{y}_m\\ \vdots\\ \boldsymbol{v}_0^{(S)}+\boldsymbol{V}^{(S)}\boldsymbol{y}_m\end{bmatrix}=\boldsymbol{u}_0+\boldsymbol{V}\boldsymbol{y}_m=\widetilde{\boldsymbol{V}}\tilde{\boldsymbol{y}}_m \quad (2)$$

其中，$\boldsymbol{u}_0=\begin{bmatrix}\boldsymbol{v}_0^{(1)}\\ \boldsymbol{v}_0^{(2)}\\ \vdots\\ \boldsymbol{v}_0^{(S)}\end{bmatrix}$，$\boldsymbol{V}=\begin{bmatrix}\boldsymbol{V}^{(1)}\\ \boldsymbol{V}^{(2)}\\ \vdots\\ \boldsymbol{V}^{(S)}\end{bmatrix}$，$\widetilde{\boldsymbol{V}}=\begin{bmatrix}\widetilde{\boldsymbol{V}}^{(1)}\\ \widetilde{\boldsymbol{V}}^{(2)}\\ \vdots\\ \widetilde{\boldsymbol{V}}^{(S)}\end{bmatrix}$。

其中 $\boldsymbol{u}_0=\dfrac{1}{M}\sum_{m=1}^{M}\boldsymbol{u}_m$，$\boldsymbol{v}_0^{(s)}=\dfrac{1}{M}\sum_{m=1}^{M}\boldsymbol{u}_m^{(s)}$，矩阵 $\boldsymbol{V}$ 的列为协方差矩阵的前 N 个特征值所对应的特征矢量。

2.2 本征音子的最大似然估计

在自适应阶段，假设未知说话人自适应数据的特征矢量序列为 $\boldsymbol{O}=\{\boldsymbol{o}(t)\}_{t=1}^{\mathrm{T}}$，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

$$Q(\boldsymbol{V}(s))=-\frac{1}{2}\sum_{t}\sum_{m}\gamma_m(t)\left[\boldsymbol{o}(t)-\boldsymbol{\mu}_m-\boldsymbol{u}_m^{(s)}\right]^{\mathrm{T}}\cdot\boldsymbol{\Sigma}_m^{-1}\left[\boldsymbol{o}(t)-\boldsymbol{\mu}_m-\boldsymbol{u}_m^{(s)}\right] \tag{3}$$

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，则可以通过 Baum-Welch 前后向算法[15]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(3)，并令其对 $\widetilde{\boldsymbol{V}}^{(s)}$ 的导数为 0，可以得到 $\widetilde{\boldsymbol{V}}^{(s)}$ 的求解公式[8]。然而文献[8]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，当音子变化子空间 N 较大时(≥ 100)时，存储及求逆计算都非常消耗内存和计算时间。但传统 HMM-GMM 的声学模型中，$\boldsymbol{\Sigma}_m$ 通常是一个对角阵，令其第 d 个对角线元素为 $\sigma_{m,d}$，则目标函数式(3)可以简化为

$$Q\left(\widetilde{\boldsymbol{V}}^{(s)}\right)=-\frac{1}{2}\sum_{d}\sum_{t}\sum_{m}\gamma_m(t)\sigma_{m,d}^{-1}\cdot\left[\boldsymbol{o}_d(t)-\boldsymbol{\mu}_{m,d}-\tilde{\boldsymbol{\nu}}_d^{(s)\mathrm{T}}\tilde{\boldsymbol{y}}_m\right]^2 \tag{4}$$

其中，$\boldsymbol{o}_d(t)$ 及 $\boldsymbol{\mu}_{m,d}$ 分别为特征矢量 $\boldsymbol{o}(t)$ 及均值矢量 $\boldsymbol{\mu}_m$ 的第 d 维元素，$\tilde{\boldsymbol{\nu}}_d^{(s)\mathrm{T}}$ 表示本征音子矩阵 $\widetilde{\boldsymbol{V}}^{(s)}$ 的第 d 行。对式(4)进行整理可得

$$Q\left(\widetilde{\boldsymbol{V}}^{(s)}\right)=-\frac{1}{2}\sum_{d}\left[\tilde{\boldsymbol{\nu}}_d^{(s)\mathrm{T}}\boldsymbol{A}_d\tilde{\boldsymbol{\nu}}_d^{(s)}-\boldsymbol{b}_d^{\mathrm{T}}\tilde{\boldsymbol{\nu}}_d^{(s)}\right]+C \tag{5}$$

其中，$\boldsymbol{A}_d=\sum_t\sum_m\gamma_m(t)\sigma_{m,d}^{-1}\tilde{\boldsymbol{y}}_m\tilde{\boldsymbol{y}}_m^{\mathrm{T}}$，$\boldsymbol{b}_d=\sum_t\sum_m\gamma_m(t)\sigma_{m,d}^{-1}\left[\boldsymbol{o}_d(t)-\boldsymbol{\mu}_{m,d}\right]\tilde{\boldsymbol{y}}_m$，$C$ 为一个常数。对式(5)求关于 $\tilde{\boldsymbol{\nu}}_d^{(s)}$ 的导数，并令导数为 0 可得其最优值 $(\tilde{\boldsymbol{\nu}}_d^{(s)})_{ML}=\boldsymbol{A}_d^{-1}\boldsymbol{b}_d$。由于各行之间的计算相互独立，因此实际计算中，可以对 $\widetilde{\boldsymbol{V}}^{(s)}$ 的 D 行进行并行求解，因此求解时间很快。

3 基于本征音子说话人子空间的自适应方法

3.1 本征音子说话人子空间

将 2.2 节扩展本征音子矩阵 $\widetilde{\boldsymbol{V}}^{(s)}$ 视为 $D\times(N+1)$ 维矩阵空间中的一个点，假设其位于一个 K ($K<S$)维的子空间中，我们将该子空间称为“本征音子说话人子空间”。令 $\{\widetilde{\boldsymbol{V}}_k\}_{k=1}^{K}$ 为该子空间的一组基，假设 $\widetilde{\boldsymbol{V}}^{(s)}$ 在这组基下的坐标矢量为 $\boldsymbol{x}^{(s)}=[x_1^{(s)}\ x_2^{(s)}\ \cdots\ x_K^{(s)}]$，则 $\widetilde{\boldsymbol{V}}^{(s)}$ 可以表示为

$$\widetilde{\boldsymbol{V}}^{(s)}=\sum_{k=1}^{K}x_k^{(s)}\widetilde{\boldsymbol{V}}_k \tag{6}$$

其中，$\{\widetilde{\boldsymbol{V}}_k\}_{k=1}^{K}$ 可以通过对训练说话人的扩展本征音子矩阵 $\{\widetilde{\boldsymbol{V}}^{(s)}\}_{s=1}^{S}$ 进行主分量分析得到，具体实现方法为：

定义扩展本征音子超矢量 $\tilde{\boldsymbol{v}}^{(s)}$ 为扩展本征音子矩阵 $\widetilde{\boldsymbol{V}}^{(s)}$ 的列矢量化操作结果，即

$$\tilde{\boldsymbol{v}}^{(s)}=\mathrm{vec}\left(\widetilde{\boldsymbol{V}}^{(s)}\right)=\begin{bmatrix}\boldsymbol{v}_0^{(s)}\\ \boldsymbol{v}_1^{(s)}\\ \vdots\\ \boldsymbol{v}_N^{(s)}\end{bmatrix} \tag{7}$$

则 $\tilde{\boldsymbol{v}}^{(s)}$ 为一个 $D\times(N+1)$ 维的矢量。

令训练说话人扩展本征音子超矢量的均值为 0，对 $\{\tilde{\boldsymbol{v}}^{(s)}\}_{s=1}^{S}$ 进行主分量分析，设其协方差矩阵前 K 个最大的特征值对应的特征矢量为 $\tilde{\boldsymbol{v}}_1,\tilde{\boldsymbol{v}}_2,\cdots,\tilde{\boldsymbol{v}}_K$。与本征音类似，$\{\tilde{\boldsymbol{v}}_k\}_{k=1}^{K}$ 构成扩展本征音子超矢量子空间中的一组基，由于它们的维数只有 $D\times(N+1)$ 维(本征音维数为 $M\times D$ 维)，本文将称其为“紧致本征音(Compact Eigen Voice, CEV)”。

扩展本征音子矩阵对应的基矢量 $\{\widetilde{\boldsymbol{V}}_k\}_{k=1}^{K}$ 为

$$\widetilde{\boldsymbol{V}}_k=\mathrm{unvec}_{D,N+1}(\tilde{\boldsymbol{v}}_k),\quad k=1,2,\cdots,K \tag{8}$$

其中，$\mathrm{unvec}_{D,N+1}(\bullet)$ 表示矩阵化函数，它将一个 $D\times(N+1)$ 维列矢量的元素依次按列排列成一个 $D\times(N+1)$ 维的矩阵。

3.2 自适应算法具体描述

给定本征音子子空间中的扩展高斯混元坐标矢量 $\{\tilde{\boldsymbol{y}}_m\}_{m=1}^{M}$ 和本征音子说话人子空间的一组基 $\{\widetilde{\boldsymbol{V}}_k\}_{k=1}^{K}$，基于本征音子说话人子空间的自适应过程为：对于一个未知说话人 s'，利用自适应数据估计其在本征音子说话人子空间中的坐标矢量 $\boldsymbol{x}^{(s')}$，将式(6)得到的扩展本征音子矩阵 $\widetilde{\boldsymbol{V}}^{(s')}$ 代入音子变化矢量定义式，可得说话人相关高斯均值矢量 $\boldsymbol{\mu}_m^{(s')}$ 为

$$\boldsymbol{\mu}_m^{(s')}=\boldsymbol{\mu}_m+\left(\sum_{k=1}^{K}\boldsymbol{x}_k^{(s)}\widetilde{\boldsymbol{V}}_k\right)\tilde{\boldsymbol{y}}_m=\boldsymbol{\mu}_m+\sum_{k=1}^{K}\boldsymbol{x}_k^{(s)}\boldsymbol{p}_{k,m},\quad m=1,2,\cdots,M \tag{9}$$

其中，$\boldsymbol{p}_{k,m}=\widetilde{\boldsymbol{V}}_k\tilde{\boldsymbol{y}}_m$ 为一个 D 维的矢量(矩阵 $\widetilde{\boldsymbol{V}}_k$ 维数为 $D\times(N+1)$，矢量 $\boldsymbol{y}_m$ 维数为 $N+1$)。

进一步，定义矩阵 $\boldsymbol{P}_m=[\boldsymbol{p}_{1,m}\ \boldsymbol{p}_{2,m}\ \cdots\ \boldsymbol{p}_{K,m}]$，则式(9)等价为

$$\boldsymbol{\mu}_m^{(s')}=\boldsymbol{\mu}_m+\boldsymbol{P}_m\boldsymbol{x}^{(s')} \tag{10}$$

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，采用期望最大

(Expectation Maximization, EM)算法，由式(10)，说话人 s' 坐标矢量 $\boldsymbol{x}^{(s')}$ 的最大似然估计目标函数可以写为

$$Q\left(\boldsymbol{x}^{(s')}\right)=-\frac{1}{2}\sum_{t}\sum_{m}\left(\boldsymbol{o}_t-\boldsymbol{\mu}_m-\boldsymbol{P}_m\boldsymbol{x}^{(s')}\right)^{\mathrm{T}}\cdot\boldsymbol{\Sigma}_m^{-1}\left(\boldsymbol{o}_t-\boldsymbol{\mu}_m-\boldsymbol{P}_m\boldsymbol{x}^{(s')}\right)\tag{11}$$

令 $\dfrac{\partial Q\left(\boldsymbol{x}^{(s')}\right)}{\partial \boldsymbol{x}^{(s')}}=0$，整理可得

$$\boldsymbol{x}^{(s')}=\boldsymbol{A}^{-1}\boldsymbol{b}\tag{12}$$

其中，矩阵 $\boldsymbol{A}$ 和矢量 $\boldsymbol{b}$ 的定义分别为

$$\boldsymbol{A}=\sum_{m}\gamma_m\boldsymbol{P}_m^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1}\boldsymbol{P}_m\tag{13}$$

$$\boldsymbol{b}=\sum_{m}\boldsymbol{P}_m^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1}\left(\boldsymbol{s}_m-\gamma_m\boldsymbol{\mu}_m\right)\tag{14}$$

其中，γ_m 与 $\boldsymbol{s}_m$ 分别为属于第 m 个高斯混元的特征矢量的零阶与一阶统计量。

3.3 与说话人子空间自适应方法的比较

在基于说话人子空间的自适应方法中，其基本假设是说话人超矢量 $\boldsymbol{\mu}^{(s)}$ 位于一个低维线性子空间 Γ^K 中(K 为子空间维数，$K<S$)。设 Γ^K 的一组基矢量为 $\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_K\}$，其中第 k 个基矢量 $\boldsymbol{e}_k$ 中第 m 个高斯混元对应的均值矢量为 $\boldsymbol{e}_{k,m}$。设 $\boldsymbol{\mu}^{(s)}$ 在这组基下的坐标矢量为 $\boldsymbol{x}_K^{(s)}$，称 $\boldsymbol{x}_K^{(s)}$ 为说话人因子；令 $\boldsymbol{E}_{K,m}=[\boldsymbol{e}_{1,m}\ \ \boldsymbol{e}_{2,m}\ \ \cdots\ \ \boldsymbol{e}_{K,m}]$，则 $\boldsymbol{\mu}_m^{(s)}$ 可以分解为

$$\boldsymbol{\mu}_m^{(s)}=\boldsymbol{\mu}_m+\boldsymbol{E}_{K,m}\boldsymbol{x}_K^{(s)}\tag{15}$$

其中，SI 模型的均值矢量 $\boldsymbol{\mu}_m$ 可视为第 m 个高斯混元所在说话人子空间的原点。根据训练数据得到说话人子空间的基矢量，则在自适应阶段，只需要根据自适应数据估计未知说话人 s' 的说话人因子 $\boldsymbol{x}_K^{(s')}$，然后根据式(15)即可得到自适应后各高斯混元的均值矢量 $\boldsymbol{\mu}_m^{(s')}$。

不难发现，本征音子说话人子空间的自适应方法与说话人子空间的自适应方法非常类似。对比式(10)和式(15)可见，$\boldsymbol{P}_m$ 相当于第 m 个高斯混元对应的本征音矩阵 $\boldsymbol{E}_{K,m}$。在说话人子空间自适应方法中，说话人子空间的基由一组说话人超矢量构成，其中每一个超矢量的维数为 $M\times D$；而基于本征音子说话人子空间的自适应方法中，说话人子空间的基是由若干个扩展本征音子矩阵构成，其中每一个矩阵的维数为 $(N+1)\times D$。由于 $N\ll M$，因此本文方法所需要的存储空间要小得多。对于一个实际的大词汇量连续语音识别系统，M 通常高达十万级，而 N 往往只需数百左右，因此存储空间的节省是非常可观的。

4　实验结果及分析

为了验证本文算法的性能，采用微软中文语料库[16]针对 HMM-GMM 框架下的连续语音识别系统说话人自适应实验。训练集中包括 100 个男性说话人，每人大约 200 句话，每句话时长大约 5 s，共有 19688 句话，总时长为 33 h。测试集中共有 25 个说话人，每人 20 句话，每句话时长也是大约 5 s。

声学特征矢量采用 13 维的 MFCC 参数及其一阶、二阶差分，总的特征维数为 39 维。帧长和帧移分别为 25 ms 和 10 ms。实验中，借助语音开源工具箱 HTK(Hidden Markov Toolkit)(版本 3.4.1)[15]训练得到 SI 基线系统。首先训练单音子声学模型，其中每个单音子对应一个汉语有调音节。根据发音字典，对单音子进行上下文扩展，得到 295180 个跨词的三音子有调音节，其中 95534 个三音子在训练语料中得到覆盖。每一个三音子用一个包含 3 个发射状态的、自左向右无跨越的隐马尔可夫模型进行建模。采用基于决策树的三音子状态聚类后，系统中共有 2392 个不同的上下文相关状态。最终训练得到的说话人无关(SI)声学模型中每个状态含有 8 个高斯混元，因此声学模型中的总的高斯混元数为 19136 个。

在测试阶段，使用 HTK 自带的 HVite 工具作为解码器，使用音节全连接的解码网络，不采用任何语法模型。采用这种解码网络的语音识别系统对声学模型的要求最高，可以充分展示声学模型的识别性能。在原始测试集上，SI 基线系统的平均有调音节正确识别率为 53.04%(文献[16]中结果为 51.21%)。

4.1 扩展本征音子超矢量的说话人子空间存在性实验

本节通过对训练说话人的扩展本征音子超矢量进行主分量分析来验证其说话人子空间的存在性。根据训练说话人的初始 SD 声学模型，首先得到各高斯混元对应的音子变化超矢量 $\boldsymbol{u}_m$ (式(2))，每个音子变化超矢量的维数为 $S\times D=100\times 39=3900$，对 $\{\boldsymbol{u}_m\}_{m=1}^{19136}$ 进行主分量分析，保留前 100 个最大的特征值对应的特征矢量作为基矢量矩阵 $\boldsymbol{V}$ (式(2))的列；根据基矢量矩阵 $\boldsymbol{V}$ 及音子变化超矢量的均值矢量 $\boldsymbol{u}_0$ (式(2))，得到 100 个训练说话人的扩展本征音子超矢量 $\tilde{\boldsymbol{v}}^{(s)}$ (式(7))，每个扩展本征音子超矢量的维数为 $D\times(N+1)=39\times(N+1)$；最后，对 $\{\tilde{\boldsymbol{v}}^{(s)}\}_{s=1}^{100}$ 再次进行主分量分析，将其协方差矩阵的特征值从大到小排序，并计算各特征值的累积贡献率。将音子变化子空间的维数 N 从 25 调整到 250，各种参数设置下的特征值累积贡献率变化曲线如图 1 所示。

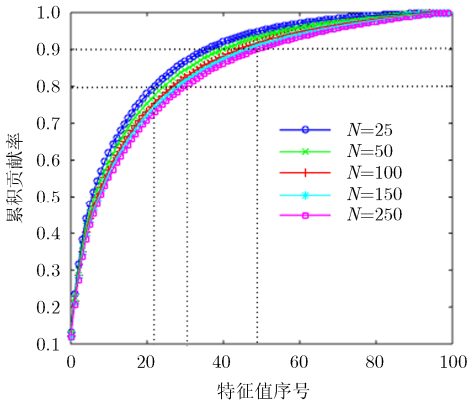

图 1 扩展本征音子超矢量协方差矩阵的特征值累积贡献率变化曲线

由图 1 可见，在各种音子变化子空间维数(N)下，对训练说话人扩的展本征音子超矢量进行主分量分析(Principal Component Analysis, PCA)后，前 20 至 50 个特征值即具有 80%~90%的累积贡献率，这表明扩展本征音子超矢量空间中的确存在一个低维的说话人子空间。此外从图 1 还可看出，音子变化子空间维数(N)越小，说话人子空间越明显：当 $N=25$ 时，前 22 个特征值具有 80%的累积贡献率；而当 $N=250$ 时，前 31 个特征值才能达到 80%的累积贡献率。

4.2 基于本征音子说话人子空间的自适应实验

为了比较本文算法的性能，实验中，本文针对以下 3 种说话人自适应算法进行对比实验：

(1)本征音(Eigen Voice, EV)：基于主分量分析的本征音说话人自适应算法，本征音的个数 K 从 20 调整到 100。

(2)最大似然本征音子(EigenPhone based on Maximum Likelihood, EP-ML)：基于最大似然估计的本征音子说话人自适应算法，在训练阶段采用 2.2 节给出的主分量分析方法得到各高斯混元的坐标矢量$\{\boldsymbol{y}_m\}_{m=1}^{M}$，然后在测试阶段采用 3.3 节最大似然估计准则计算每个说话人的扩展本征音子矩阵；

(3)本征音子说话人子空间(EigenPhone based on Speaker Subspace, EP-SS)：本文提出的基于本征音子说话人子空间自适应算法，其中说话人子空间维数 K 从 20 调整到 100。

其中，(1)为经典的说话人子空间自适应算法，(2)为原始的本征音子自适应算法，(3)为本文提出的基于本征音子说话人子空间的自适应算法。在所有的本征音子自适应算法实验中，本征音子的个数 N 均取为 100。

在训练阶段，对每一个训练说话人，利用其训练语料，采用 MLLR+MAP 自适应方法得到其对应的 SD 声学模型及其对应的说话人超矢量。利用这 100 个训练说话人超矢量，采用经典的主分量分析方法得到 100 个本征音超矢量。其中，在 MLLR+MAP 自适应方法中，将回归树中的回归类数分别设置为 16, 32 和 64, MLLR 变换矩阵分别设置为对角矩阵、分块对角矩阵和满阵，将 MAP 自适应的先验权重从 10 调整到 40。最终发现在所有自适应数据量条件下，当回归类数为 32、线性变换矩阵为分块对角矩阵(每个子矩阵均为 13×13 维，分别对应原始的美尔频率倒谱系数(Mel Frequency Cepstral Coefficients, MFCC)及其一阶和二阶差分参数)、先验权重为 10 时，得到最佳的平均正确识别率。

在测试阶段，为了测试各方法在不同数据量下的自适应性能，对于每一个说话人，从其 20 句话中随机选取 1 句话、2 句话、4 句话作为自适应语料，从剩下的语料中随机选取 10 句话作为测试语料。为了保证实验结果的可靠性，每种自适应语料条件下，使用交叉验证的方法对每一个说话人重复 8 次实验，统计所有 8 词实验测试语料上的平均结果作为系统性能指标，表 1 给出了各种说话人自适应算法的实验结果(为简洁起见，对于本征音自适应算法，表中仅给出了其最佳结果)。其中黑体字所示为每种自适应数据量条件下的最好实验结果，斜体字所示为相比基线 SI 系统平均正确识别率(53.04%)下降的实验结果。

由表 1 中结果可见，随着自适应语料的增加，为了获得最佳的自适应性能，本征音自适应算法中说话人子空间的维数(K)也要相应地增大。

本文实验中，由于自适应语料相对较少，本征音子的个数($N=100$)相对较大，因此原始的本征音子自适应方法(EP-ML)出现严重的过拟合现象，在 1 句话自适应语料条件下其平均正识率(19.45%)甚至远低于自适应前 SI 声学模型的实验结果(53.04%)。

对本征音子算法引入说话人子空间后，EP-SS 算法的自适应性能得到明显提升。在 1 句话与 2 句

表 1 各种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 句	2 句	4 句
EV	K=10~100	55.90 (K=30)	57.10 (K=60)	57.64 (K=70)
EP-ML		*19.45*	*41.46*	54.45
EP-SS	K=20	55.32	55.45	56.12
	K=40	**55.80**	56.12	56.68
	K=60	55.12	**56.92**	57.12
	K=80	54.86	56.76	**57.57**
	K=100	54.78	56.48	57.32

话，与 EP-ML 算法相比，其最佳平均正识率相对提高了 187%与 37%。同时，随着自适应数据量的增加，为达到最佳平均正确识别率，说话人子空间的维数也要相应地增大，这一点与本征音自适应算法的变化趋势是一致的。实际应用中应根据实验确定最佳的说话人子空间维数，或利用数据拟合的方法得到说话人子空间维数随着自适应语料数据量变化的经验公式。

将本征音子说话人子空间自适应算法(EP-SS)与本征音自适应算法(EV)进行比较，可以看出，在所有自适应数据量下前者的平均正确识别率略低于后者，但已十分接近。这是由于前者的说话人子空间是针对本征音子超矢量进行构建的，它只能得到说话人相关高斯混元均值矢量的一个近似表达；而后者的说话人子空间是针对说话人超矢量构建的，它是说话人相关高斯混元均值的原始表达；因此，在训练本征音子超矢量时，都会对原始高斯混元均值矢量的表示造成一定误差。

为了更好地比较两种算法的性能时，采用 NIST 公布的开源工具包 SCTK[1)]进行显著性水平测试(Significance test)以检验识别结果之间的差异在统计上是否显著。3 种显著性测试(MP 测试、SI 测试及 WI 测试)结果均表明在 5%的显著性水平之下，在 1 句话与 4 句话自适应语料时，两种方法的的最佳实验结果之间差异是不显著的；而在 2 句话自适应语料时，本征音自适应算法(EV)的 MP 测试相对更优一些，而其它两种测试显示其差异也是不显著的。这就说明两者的性能从统计上讲几乎是相同的。

下面讨论本文提出的 EP-SS 算法的时间复杂度和空间复杂度。首先分析一下时间复杂度，根据式(10)和式(15)，在 3.3 节讨论了本征音子说话人子空间自适应方法和本征音子自适应算法的相似性。并且由式(12)可以看出，两种方法的时间复杂度只与说话人音子 $\boldsymbol{x}^{(s')}$ 的维数 K 有关，即对于相同的说话人音子维数，两种方法的时间复杂度完全相同。即使两种方法最佳的说话人音子 $\boldsymbol{x}^{(s')}$ 的维数 K 不同，从实验可知，二者相差不大，因此时间复杂度也相差不大。例如 EP-SS 方法当 $K = 40$ 时与 EV 算法当 K =30 进行比较，从式(12)可知，由于只是一个 K 维的矩阵求逆和向量相乘，二者的时间复杂度差别可忽略。

然而就空间复杂度而言，正如 3.3 节中的分析所指出，与原始的说话人子空间自适应方法相比，在基于本征音子说话人子空间方法中，说话人子空间的基矢量维数大大压缩(从 $M \times D$ 维压缩为 $N \times D$ 维)，使得在实际应用中针对大词汇量连续语音识别的实现变得更为简单与现实。例如，在原始说话人子空间自适应方法中，当训练语料达到百小时数量级时，高斯混元数量(M)会达到十万级，存储 200 个说话人超矢量将耗费约几 G 内存(高斯混元数×特征维数×浮点数精度字节数×说话人个数)，耗费了大量的内存资源；而在基于本征音子的说话人子空间方法中，由于 N 可以取为 100 左右即可，存储 200 个说话人超矢量只需几 M 内存，这种对内存资源的节约是非常可观的。因此，基于本征音子说话人子空间的自适应方法在牺牲少许性能的代价下，换来了说话人子空间自适应方法实用性的大幅提高。

5　结束语

本文提出了一种基于本征音子说话人子空间的说话人自适应方法。本文在分析了本征音子说话人自适应算法基本原理的基础上，利用了本征音子矩阵的说话人相关特性定义了本征音子的说话人子空间，并且通过对训练说话人的扩展本征音子超矢量进行主分量分析来验证其说话人子空间的存在性。然后详细推导了本征音子说话人子空间自适应的具体算法，并且将该方法与已有的相关自适应算法进行比较。由于对本征音子的说话人相关性建模，因此与本征音子自适应算法相比，当自适应数据量较少(小于 4 句)时，本征音子说话人子空间的自适应算法能够大幅提高系统的识别性能，较好解决了本征音子自适应算法由于自适应数据不足带来的过拟合问题。与本征音方法比较可以发现，二者算法非常相似，但前者的说话人子空间是针对本征音子超矢量构建的，而后者说话人子空间是针对说话人超矢量构建的，前者在牺牲少许性能的代价下，节省了大量的存储空间，具有较小的空间复杂度而更具实用性。

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] Solomonoff A, Campbell W M, and Boardman I. Advances in channel compensation for SVM speaker recognition[C]. Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, United States, 2005: 629–632.

[3] Kumar D S P, Prasad N V, Joshi V, *et al*.. Modified splice

1) ftp://jaguar.ncsl.nist.gov/pub/sctk-2.4.0-20091110-0958.tar.bz2

and its extension to non-stereo data for noise robust speech recognition[C]. Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop(ASRU), Olomouc, Czech Republic, 2013: 174–179.

[4] Ghalehjegh S H and Rose R C. Two-stage speaker adaptation in subspace Gaussian mixture models[C]. Proceedings of International Conference on Audio, Speech and Signal Processing(ICASSP), Florence, Italy, 2014: 6374–6378.

[5] Wang Y Q and Gale M J F. Tandem system adaptation using multiple linear feature transforms[C]. Proceedings of International Conference on Audio, Speech and Signal Processing(ICASSP), Vancouver, Canada, 2013: 7932–7936.

[6] Kenny P, Boulianne G, and Dumouchel P. Eigenvoice modeling with sparse training data[J]. *IEEE Transactions on Speech and Audio Processing*, 2005, 13(3): 345–354.

[7] Kenny P, Boulianne G, Dumouchel P, *et al.*. Speaker adaptation using an eigenphone basis[J]. *IEEE Transaction on Speech and Audio Processing*, 2004, 12(6): 579–589.

[8] 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(ASRU), Hawaii, United States, 2011: 48–52.

[9] 张文林, 张连海, 陈琦, 等. 语音识别中基于低秩约束的本征音子说话人自适应方法[J]. 电子与信息学报, 2014, 36(4): 981–987.

Zhang Wen-lin, Zhang Lian-hai, Chen Qi, *et al.*. Low-rank constraint eigenphone speaker adaptation method for speech recognition[J]. *Journal of Electronics & Information Technology*, 2014, 36(4): 981–987.

[10] Zhang Wen-lin, Qu Dan, and Zhang Wei-qiang. Speaker adaptation based on sparse and low-rank eigenphone matrix estimation[C]. Proceedings of Annual Conference on International Speech Communication Association (INTERSPEECH), Singapore, 2014: 2972–2976.

[11] Wang N, Lee S, Seide F, *et al.*. Rapid speaker adaptation using a priori knowledge by eigenspace analysis of MLLR parameters[C]. Proceedings of International Conference on Audio, Speech and Signal Processing(ICASSP), Salt Lake City, United States, 2001: 345–348.

[12] Povey D and Yao K. A basis representation of constrained MLLR transforms for Robust adaptation[J]. *Computer Speech and Language*, 2012, 26(1): 35–51.

[13] Miao Y, Metze F, and Waibel A. Learning discriminative basis coefficients for eigenspace MLLR unsupervised adaptation[C]. Proceedings of International Conference on Audio, Speech and Signal Processing(ICASSP), Vancouver, Canada, 2013: 7927–7931.

[14] Saz O and Hain T. Using contextual information in joint factor eigenspace MLLR for speech recognition in diverse scenarios[C]. Proceedings of International Conference on Audio, Speech and Signal Processing(ICASSP), Florence, Italy, 2014: 6364–6368.

[15] 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.

[16] 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 7th European Conference on Speech Communication and Technology(Eurospeech), Aalborg, Denmark, 2001: 2799–2802.

屈 丹： 女，1974 年生，博士，副教授，研究方向为语音处理与识别、机器学习、自然语言处理.

张文林： 男，1982年生，博士，讲师，研究方向为语音处理与识别、机器学习、自然语言处理.