一种新的基于子空间的说话人自适应方法

张文林 [1] 张卫强 [2] 刘 加 [2] 李弼程 [1] 屈 丹 [1]

摘 要 提出了一种新的基于子空间的快速说话人自适应方法. 该方法在本征音 (Eigen-voice, EV) 自适应方法基础上, 进一步在音子空间寻找低维子空间, 得到更为紧凑的 "说话人 – 音子" 联合子空间. 该子空间不仅包含了说话人间的模型参数相关性信息, 而且对音子间的模型参数相关性信息也进行了显式建模, 在大大降低模型存储量的同时更为全面地反映模型参数的先验信息. 在基于连续语音识别的无监督自适应实验中, 在少量的自适应数据条件下, 新方法取得了比最大似然线性回归和聚类最大似然线性基方法更好的效果.

关键词 连续语音识别, 说话人自适应, 本征音, 本征音子

DOI 10.3724/SP.J.1004.2011.01495

A New Subspace Based Speaker Adaptation Method

ZHANG Wen-Lin[1] ZHANG Wei-Qiang[2] LIU Jia[2] LI Bi-Cheng[1] QU Dan[1]

Abstract A new speaker adaptation method based on subspace modeling is proposed. After performing eigen-voice (EV) analysis and finding the speaker subspace, another low dimensional subspace is found in the phone space. The new subspace can capture the inter-speaker variability as well as intra-speaker variability of the hidden Markov model (HMM) model parameters. This joint speaker-phone subspace is both robust and compact. In large vocabulary continuous speech recognition experiments, the new method showed better unsupervised adaptation than the baseline maximum likelihood linear regression and clustered maximum-likelihood linear basis adaptation method, especially when the adaptation data were less than 30 s.

Key words Continuous speech recognition, speaker adaptation, eigen-voice (EV), eigen-phone (EP)

对于连续语音识别和关键词检出系统, 说话人自适应是一项关键的实用化技术. 通常为了提高系统的泛化性能, 训练声学模型所用的语料库包含多个说话人和多种环境下的语音, 由此得到说话人无关 (Speaker independent, SI) 声学模型. 然而当训练语料充分时, 使用特定人语料训练得到的说话人相关 (Speaker dependent, SD) 声学模型比 SI 模型具有更高的识别率[1]. 因此, 如何在少量自适应语料下, 由 SI 模型快速自适应得到 SD 模型, 一直是语音识别技术研究的热点.

根据自适应数据有无人工标注, 说话人自适应技术可以分为有监督的说话人自适应和无监督的说话人自适应[1]. 前者要求提供自适应数据的人工标注, 而对于后者, 自适应数据的标注是利用 SI 模型进行解码识别得到的 1-best 或 lattice 形式的结果. 在实际应用中, 由于人工标注难以得到, 无监督自适应技术更加实用, 但通过解码器得到的自动标注通常会有识别错误, 大大影响了无监督自适应的效果.

目前, 主流的说话人自适应技术可以分为三大类[2]: 1) 基于最大后验概率 (Maximum a posteriori, MAP) 准则的自适应, 其基本思想是根据贝叶斯原理, 将 SI 模型参数与由自适应数据得到的说话人相关信息进行线性组合, 从而得到 SD 模型. 2) 基于变换的自适应方法, 它通常首先对 SI 模型的参数进行聚类, 在自适应阶段对同一类参数在某种准则下估计一个变换, 进而得到 SD 模型. 其中最具代表性的就是基于最大似然线性回归 (Maximum likelihood linear regression, MLLR)[3] 的说话人自适应方法, 它在最大似然准则下估计一组线性变换. 3) 基于说话人聚类的自适应方法, 其基本假设是认为所有可能的 SD 模型参数构成说话人空间, 其维数是有限的, 或者存在某种聚类结构, 通过寻找说话人空间中的一个低维子空间近似 (称为说话人子空间) 或聚类结构来对说话人先验信息进行显式建模; 在自适应阶段, 利用子空间的基矢量 (称为本征音) 或聚类结构对未知说话人的模型参数进行近似. 这类方法的典型代表是基于本征音 (Eigen-voice, EV)[4] 和基于参考说话人加权 (Reference speaker weight-

收稿日期 2011-01-13 录用日期 2011-07-07
Manuscript received January 13, 2011; accepted July 7, 2011
国家自然科学基金 (60872142, 61005019, 61175017) 资助
Supported by National Natural Science Foundation of China (60872142, 61005019, 61175017)
1. 中国人民解放军信息工程大学信息工程学院 郑州 450002 2. 清华大学电子工程系 北京 100084
1. Information Engineering Institute, PLA Information Engineering University, Zhengzhou 450002 2. Department of Electronic Engineering, Tsinghua University, Beijing 100084

ing, RSW)[5] 的说话人自适应算法. 前者通过主成分分析 (Principle component analysis, PCA)[4] 或最大似然估计[6] 寻找说话人子空间中的一组基, 后者利用说话人聚类或最大似然准则[7], 在训练说话人中挑选与待识说话人相近的 SD 模型进行加权组合得到新的 SD 模型.

上述三种主流技术各有其优缺点: 首先, 基于 MAP 的自适应技术具有很好的渐近性能, 即随着自适应数据量的增加, 自适应得到的模型会越来越接近真实的 SD 模型, 其缺点是对于自适应数据中没有出现的声学模型无法进行自适应; 而基于变换的自适应方法, 为了使变换矩阵得到稳健的估计, 对于每一个变换类也需要一定量的自适应数据, 而且通常为了数学上的易处理性, 其线性变换的假设是不精确的, 不具有 MAP 自适应技术的渐近性能; 基于说话人聚类的自适应方法, 特别是基于本征音的说话人自适应技术, 由于自适应时待估参数的数量较少 (取决于说话人子空间维数), 且对自适应数据标注中的错误不太敏感, 特别适用于少量自适应数据下的快速无监督自适应. 但它需要估计并存储多个本征音模型, 对于大词汇量连续语音识别来说, 模型参数数量及其需要的存储空间是很大的. 为了解决数据稀疏问题, 并降低模型存储量, Tang 等[8] 在本征音自适应技术基础上提出了一种基于聚类最大似然线性基 (Clustered maximum likelihood linear basis, CMLLB) 的自适应方法, 通过对模型中的高斯混元进行聚类, 将模型参数数量降至原来的十分之一左右.

近年来, Jeong[9] 提出了一种基于张量分解的说话人自适应方法, 其基本思想是将训练说话人模型看作一个三维的张量, 对其进行 Tucker 分解 (类似于矩阵的 PCA 分解), 在对分解形式进行适当变换后, 在特征矢量的每一维估计一个说话人因子, 从而进行说话人自适应. 这种基于张量的方法, 需要估计的参数数量大于 MLLR 方法, 在自适应数据足够时, 可以达到比 MLLR 方法更好的自适应效果, 然而在少量自适应数据条件下, 易于出现过训练的问题, 性能反而不如经典的本征音方法.

本文研究无监督的快速说话人技术, 提出了一种新的基于说话人 – 音子联合子空间的说话人自适应方法. 新方法在训练阶段, 进一步在本征音空间对高斯混元 (对应音子模型) 之间的相关性进行分析, 得到说话人 – 音子联合子空间; 在识别阶段, 利用最大似然准则, 估计待识说话人的说话人因子系数, 从而达到自适应的目的. 新方法充分利用了不同 SD 模型在 "说话人" 和 "音子" 两个维度上变化的相关性, 既考虑了声学模型参数在说话人间 (Inter-speaker) 的相关性信息, 又考虑了其在单个说话人内 (Intra-speaker) (即音子间的) 的相关性信息. 与基于张量的方法[9] 不同, 新方法没有增加原始本征音方法的待估参数数量, 适用于少量自适应数据条件下的快速无监督自适应. 相关实验表明, 新方法在无监督说话人自适应中, 在少量的自适应数据条件下, 明显优于 MLLR 自适应方法和基于 CMLLB 的自适应方法.

1 基于本征音的说话人自适应

基于本征音的说话人自适应方法[4], 借鉴了人脸识别中 "本征人脸" 技术的思想, 将每个 SD 模型参数用低维子空间中的一组基的线性加权来表示. 本节简要介绍基于本征音的说话人自适应原理, 及其改进算法 — 基于聚类最大似然线性基的说话人自适应方法, 同时引入本文使用的数学符号.

1.1 基于本征音的说话人自适应方法原理

令语音特征维数为 D, 高斯混元数为 M, 训练数据中说话人数量为 S, 设第 s 个 SD 模型的第 m 个高斯混元对应的均值矢量为 $\boldsymbol{\mu}_m^s$, SI 模型的第 m 个高斯混元均值矢量为 $\boldsymbol{\mu}_m$, 其中 $\boldsymbol{\mu}_m^s \in \mathbf{R}^D$, $\boldsymbol{\mu}_m \in \mathbf{R}^D$.

对于第 s 个说话人, 将其 SD 模型的所有高斯混元均值矢量组合为一个 $M \cdot D$ 维的超矢量, 记为 $\boldsymbol{\mu}^s = [\ (\boldsymbol{\mu}_1^s)^{\mathrm{T}} \quad (\boldsymbol{\mu}_2^s)^{\mathrm{T}} \quad \cdots \quad (\boldsymbol{\mu}_M^s)^{\mathrm{T}}\ ]^{\mathrm{T}}$, 称之为说话人 s 的高斯超矢量. 相应地, SI 模型的高斯超矢量为 $\boldsymbol{\mu} = \left[(\boldsymbol{\mu}_1)^{\mathrm{T}} \quad (\boldsymbol{\mu}_2)^{\mathrm{T}} \quad \cdots \quad (\boldsymbol{\mu}_M)^{\mathrm{T}}\right]^{\mathrm{T}}$.

则在本征音自适应[4] 的基本假设中有:

$$\boldsymbol{\mu}^s = \boldsymbol{\mu} + \sum_{k=1}^{K} w_k^s \cdot \boldsymbol{e}^k \tag{1}$$

上式中, 若将 SI 模型的高斯超矢量 $\boldsymbol{\mu}$ 视为说话人空间中的原点, 则 $\{\boldsymbol{e}^k\}_{k=1}^K$ 是说话人高斯超矢量在 K 维说话人子空间中的一组基, 每一个基矢量 $\boldsymbol{e}^k$ 称为一个本征音 (EV). $\boldsymbol{w}^s = \left[w_1^s \quad w_2^s \quad \cdots \quad w_K^s\right]$ 是 SD 模型的投影系数向量, 它反映了 SD 模型在说话人子空间中的位置信息.

设第 s 个说话人的高斯超矢量与 SI 模型高斯超矢量之差为 $\Delta\ \boldsymbol{\mu}^s = \boldsymbol{\mu}^s - \boldsymbol{\mu}$, 其中第 m 个高斯分量的均值与对应 SI 模型高斯分量均值之差为 $\Delta\ \boldsymbol{\mu}_m^s = \boldsymbol{\mu}_m^s - \boldsymbol{\mu}$. 定义总的训练说话人模型参数矩阵为

$$\triangle \Xi = \begin{bmatrix} (\triangle \boldsymbol{\mu}^1)^{\mathrm{T}} \\ (\triangle \boldsymbol{\mu}^2)^{\mathrm{T}} \\ \vdots \\ (\triangle \boldsymbol{\mu}^S)^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} (\triangle \boldsymbol{\mu}_1^1)^{\mathrm{T}} & (\triangle \boldsymbol{\mu}_2^1)^{\mathrm{T}} & \cdots & (\triangle \boldsymbol{\mu}_M^1)^{\mathrm{T}} \\ (\triangle \boldsymbol{\mu}_1^2)^{\mathrm{T}} & (\triangle \boldsymbol{\mu}_2^2)^{\mathrm{T}} & \cdots & (\triangle \boldsymbol{\mu}_M^2)^{\mathrm{T}} \\ \vdots & \vdots & \ddots & \vdots \\ (\triangle \boldsymbol{\mu}_1^S)^{\mathrm{T}} & (\triangle \boldsymbol{\mu}_2^S)^{\mathrm{T}} & \cdots & (\triangle \boldsymbol{\mu}_M^S)^{\mathrm{T}} \end{bmatrix} \tag{2}$$

这里 $\triangle \Xi$ 的每一行都是一个 $M \cdot D$ 维矢量, 分别对应一个训练说话人模型, 其行空间构成了训练说话人的 SD 模型参数空间. 则根据本征音自适应的基本假设 (1), 训练说话人 SD 模型参数的本征音分解为

$$\triangle \Xi = \begin{bmatrix} w_1^1 & w_2^1 & \cdots & w_K^1 \\ w_1^2 & w_2^2 & \cdots & w_K^2 \\ \vdots & \vdots & \ddots & \vdots \\ w_1^S & w_2^S & \cdots & w_K^S \end{bmatrix}_{S \times K} \times \begin{bmatrix} (\boldsymbol{e}^1)^{\mathrm{T}} \\ (\boldsymbol{e}^2)^{\mathrm{T}} \\ \vdots \\ (\boldsymbol{e}^K)^{\mathrm{T}} \end{bmatrix}_{K \times (M \cdot D)} \tag{3}$$

由式 (3) 可见, 本征音自适应实际上是在 $\triangle \Xi$ 的 S 个行矢量构成的行空间中寻找一个 K 维的子空间, 它通常是对 $\triangle \Xi$ 在说话人维度上 ($\triangle \Xi$ 的行) 进行主成分分析[4] 或最大似然估计[6] 得到. 在自适应阶段, 以最大似然为准则, 利用自适应数据及其标注, 找到未知说话人模型参数在这 K 维子空间中的投影矢量 $\boldsymbol{w}^s$, 从而得到新的 SD 模型. 该分解过程称为最大似然本征分解 (Maximum likelihood eigen decomposition, MLED)[4].

令第 k 个本征音 $\boldsymbol{e}^k$ 中对应第 m 个高斯分量的子矢量为 $\boldsymbol{e}_m^k$, 记 K 维说话人子空间的 K 个基矢量组成的本征音矩阵为

$$E = \begin{bmatrix} (\boldsymbol{e}^1)^{\mathrm{T}} \\ (\boldsymbol{e}^2)^{\mathrm{T}} \\ \vdots \\ (\boldsymbol{e}^K)^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{e}_1^1)^{\mathrm{T}} & (\boldsymbol{e}_2^1)^{\mathrm{T}} & \cdots & (\boldsymbol{e}_M^1)^{\mathrm{T}} \\ (\boldsymbol{e}_1^2)^{\mathrm{T}} & (\boldsymbol{e}_2^2)^{\mathrm{T}} & \cdots & (\boldsymbol{e}_M^2)^{\mathrm{T}} \\ \vdots & \vdots & \ddots & \vdots \\ (\boldsymbol{e}_1^K)^{\mathrm{T}} & (\boldsymbol{e}_2^K)^{\mathrm{T}} & \cdots & (\boldsymbol{e}_M^K)^{\mathrm{T}} \end{bmatrix} \tag{4}$$

则新的 SD 模型的高斯超矢量可以表示为

$$\boldsymbol{\mu}^s = \boldsymbol{\mu} + \sum_{k=1}^{K} w_k^s \cdot \boldsymbol{e}^k = \boldsymbol{\mu} + E^{\mathrm{T}} \cdot \boldsymbol{w}^s \tag{5}$$

由式 (5) 可见, 在训练阶段, 不仅要得到 SI 模型的参数, 还要得到维说话人子空间基矢量组成的矩阵 E. 对于一个典型的大词汇量连续语音识别系统, 采用上下文相关三音子模型, 假设状态聚类后共有 4 000 个不同状态, 每个状态的高斯混元数为 16, 则总的高斯混元数为 $M = 4\,000 \times 16 = 64\,000$, 说话人子空间维数通常取为 $10 < K < 100$, 特征维数为 $D = 36$, 此时本征音矩阵 E 的参数个数至少为 $K \times M \times D = 10 \times 64\,000 \times 36 = 23\,040\,000$, 实际中通常没有足够的数据来对 E 进行稳健的估计, 而且模型存储的数据量太大. 因此, 需要某种方法对 E 进行降维.

1.2 基于聚类最大似然基的说话人自适应方法

为了对上述 E 矩阵进行降维, 同时为了解决训练数据的稀疏性问题, Tang 等[8] 提出一种基于聚类最大似然线性基 (CMLLB) 的说话人自适应方法. 其基本思想是, 首先对 SI 模型的所有 M 个高斯混元进行聚类, 认为属于同一类的高斯混元, 其本征音基矢量完全相同. 即:

$$\boldsymbol{e}_m^k = \boldsymbol{e}_{\phi(m)}^k \tag{6}$$

其中, $\phi(m)$ 是高斯混元标号 m 到类别号的一个映射. Tang 等[8] 还给出了基于最大似然准则训练上述聚类后的基矢量的方法.

设高斯混元聚类结果得到 N 类, 则 CMLLB 方法将本征音矩阵 E 的参数个数从 $K \times M \times D$ 个降为 $K \times N \times D$ 个. 由于聚类基矢量的个数远远少于原始本征音基矢量的个数, 因此在少量训练数据情况下模型也能得到较充分的估计.

2 基于说话人 – 音子联合子空间的说话人自适应

原始的本征音自适应方法利用 PCA 在说话人空间中寻找低维子空间, 所得到的子空间反映了说话人间 (Inter-speaker) 模型参数的相关性信息[6]. 事实上, 进一步的考查可以发现, 对于同一个说话人, 其不同音子之间的参数变化也具有很强的相关性. 上述 CMLLB 方法的成功实际上就是利用了这一点, 它认为属于同一类的高斯混元的参数变化是完全相同的, 然而这种对于音子空间参数变化信息的建模是很粗糙的. 因此, 本文考虑在得到说话人子

空间后, 在音子空间也寻找一个子空间. 与本征音相对应, 这里我们可以将音子子空间的基矢量称为 “本征音子 (Eigen-phone, EP)”[10]. 下面给出说话人 – 音子联合子空间构造及相应的自适应方法.

2.1 说话人 – 音子联合子空间的构造

对本征音矩阵 E 进一步分析可以发现, 其列矢量分别对应声学模型的 M 个高斯混元; 对每个高斯元 m, 可以将其对应的本征音均值矢量 $\{\boldsymbol{e}_m^k\}_{k=1}^K$ 排列为一个 $K\cdot D$ 维的列矢量, 这种列矢量所构成的空间可视为 “音子空间”, 其维数为混元数 M. 与本征音类似, 在音子空间进行主成分分析, 也可以得到一个 N ($N << M$) 维子空间, 其基矢量称之为 “本征音子 (EP)”[10]; N 个本征音子构成了一个 $N\times K\cdot D$ 维新的基矢量矩阵, 该矩阵同时反映了说话人间与说话人内的模型参数相关性信息. 由于新的音子子空间的构造过程是在 K 维说话人子空间中进行的, 这里我们将其称为说话人 – 音子联合子空间, 其构造过程如图 1 所示.

设所得到的第 n 个本征音子为 $\boldsymbol{u}_n = \left[(\boldsymbol{u}_n^1)^{\mathrm{T}} \quad (\boldsymbol{u}_n^2)^{\mathrm{T}} \quad \cdots \quad (\boldsymbol{u}_n^K)^{\mathrm{T}}\right]^{\mathrm{T}}$, 第 m 个高斯混元对应的坐标矢量为 $\boldsymbol{l}_m = \left[l_m^1 \quad l_m^2 \quad \cdots \quad l_m^N\right]^{\mathrm{T}}$. 则本征音矩阵 E 在该 N 维子空间上的分解为

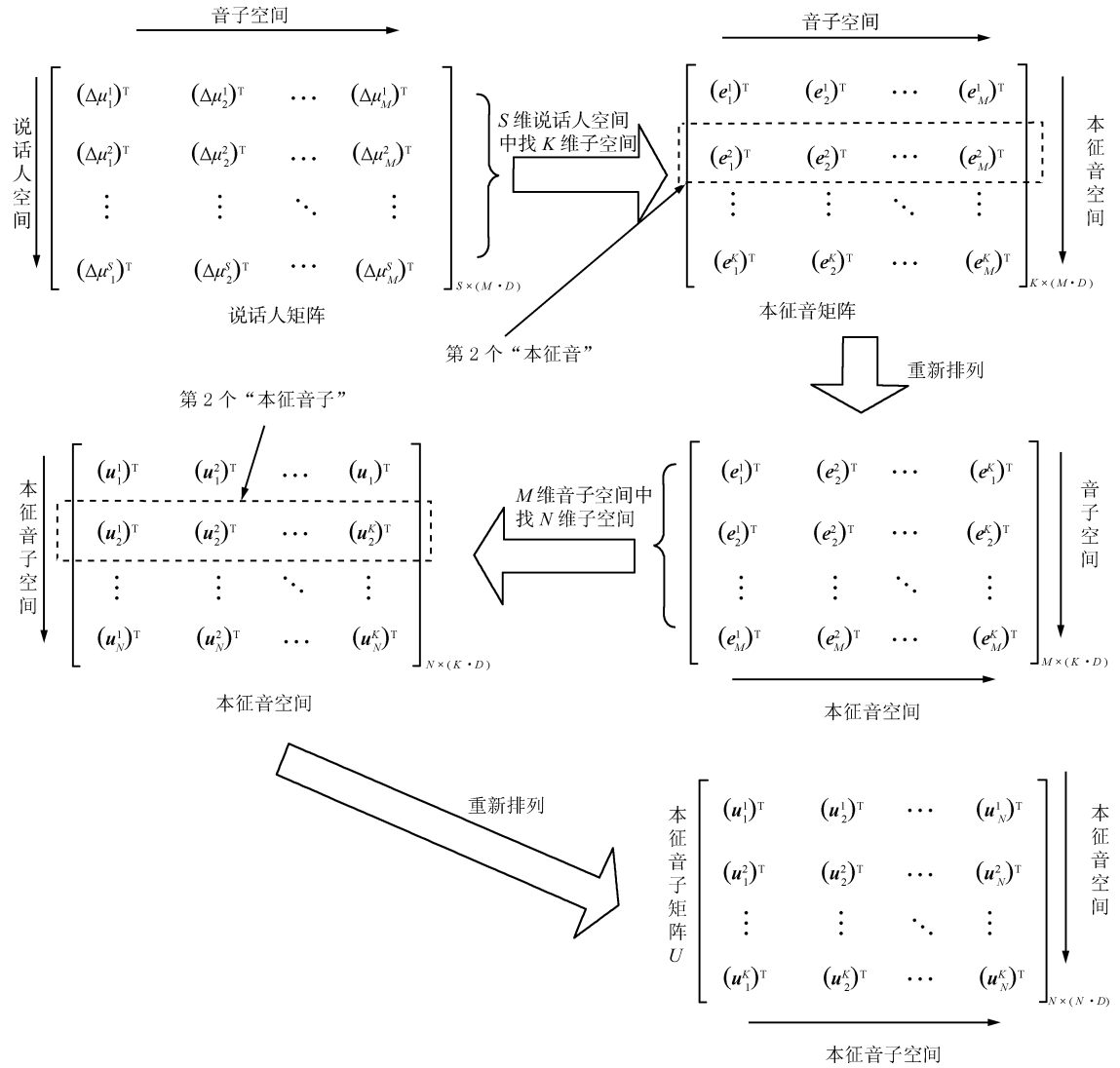

图 1 “说话人 – 音子” 联合子空间的构造过程

Fig. 1 The construction of the joint “speaker-phone” subspace

$$E=\begin{bmatrix}(\boldsymbol{u}_1^1)^{\mathrm{T}} & (\boldsymbol{u}_2^1)^{\mathrm{T}} & \cdots & (\boldsymbol{u}_N^1)^{\mathrm{T}}\\ (\boldsymbol{u}_1^2)^{\mathrm{T}} & (\boldsymbol{u}_2^2)^{\mathrm{T}} & \cdots & (\boldsymbol{u}_N^2)^{\mathrm{T}}\\ \vdots & \vdots & \ddots & \vdots\\ (\boldsymbol{u}_1^K)^{\mathrm{T}} & (\boldsymbol{u}_2^K)^{\mathrm{T}} & \cdots & (\boldsymbol{u}_N^K)^{\mathrm{T}}\end{bmatrix}_{K\times(N\cdot D)}\times\begin{bmatrix}l_1^1 I & l_2^1 I & \cdots & l_M^1 I\\ l_1^2 I & l_2^2 I & \cdots & l_M^2 I\\ \vdots & \vdots & \ddots & \vdots\\ l_1^N I & l_2^N I & \cdots & l_M^N I\end{bmatrix}_{(N\cdot D)\times(M\cdot D)} \tag{7}$$

其中, I 为一个 $D\times D$ 维的单位矩阵; 式 (7) 左边矩阵称为 “本征音子 (Eigen-phone) 矩阵”, 记为 U (如图 1 中第 3 行所示); 右边矩阵称为 “本征音子系数矩阵”, 记为了 L. 第 k 个本征音 (Eigen-voice) 的第 m 个高斯分量的均值矢量在本征音子子空间中的分解为

$$\boldsymbol{e}_m^k=\sum_{j=1}^{N} l_m^j\cdot\boldsymbol{u}_j^k \tag{8}$$

若令

$$l_m^j=\begin{cases}1, & \phi(m)=j\\ 0, & \phi(m)\neq j\end{cases}$$

其中 $\phi(m)$ 是 CMLLB 方法中的聚类函数, 则上述方法与 CMLLB 方法是完全等价的. 由此可见 CMLLB 自适应方法可以视为说话人 – 音子联合子空间自适应方法的一个特例.

2.2　基于说话人 – 音子联合子空间的说话人自适应

基于上述说话人 – 音子联合子空间, 第 s 个说话人的第 m 个高斯混元的均值矢量可以分解为

$$(\boldsymbol{\mu}_m^s)^{\mathrm{T}}=(\boldsymbol{\mu}_m)^{\mathrm{T}}+(\boldsymbol{w}^s)^{\mathrm{T}}\cdot U\cdot(\boldsymbol{l}_m\otimes I) \tag{9}$$

其中, $\otimes$ 表示矩阵之间的 Kronecker 积. 式 (9) 的分解过程可以由图 2 清晰地看出.

与本征音自适应方法类似, 对于一个新的说话人 s, 只需要利用某种准则重新估计说话人权重因子 $\boldsymbol{w}^s=\begin{bmatrix}w_1^s & w_2^s & \cdots & w_K^s\end{bmatrix}$ 即可. 这里采用最大似然准则, 其自适应方法与本征音自适应中的 MLED 算法[4] 类似, 本文不再赘述.

3　实验结果及分析

3.1　实验语料库及实验设置

为了验证本文自适应方法的有效性, 我们针对一个典型的连续语音识别系统进行了实验. 实验语料采用微软语料库[11], 其中训练语料包含 100 个男性说话人, 每个人 200 句话, 约 33 小时的话音数据, 测试语料包含另外 20 个男性说话人, 每人 20 句话, 每句话大约 5 s 的话音. 实验中, 特征参数采用原始的 13 维 Mel 频率倒谱系数 (Mel frequency cepstrum coefficient, MFCC) 特征及其一阶差分和二阶差分, 总的特征矢量维数为 39 维. 基线系统中的 SI 模型利用开源 HTK 工具箱 (3.4.1 版本)[12] 训练得到, 采用上下文相关的三音子有调音节作为声学模型单元, 每个隐马尔科夫模型 (Hidden Markov

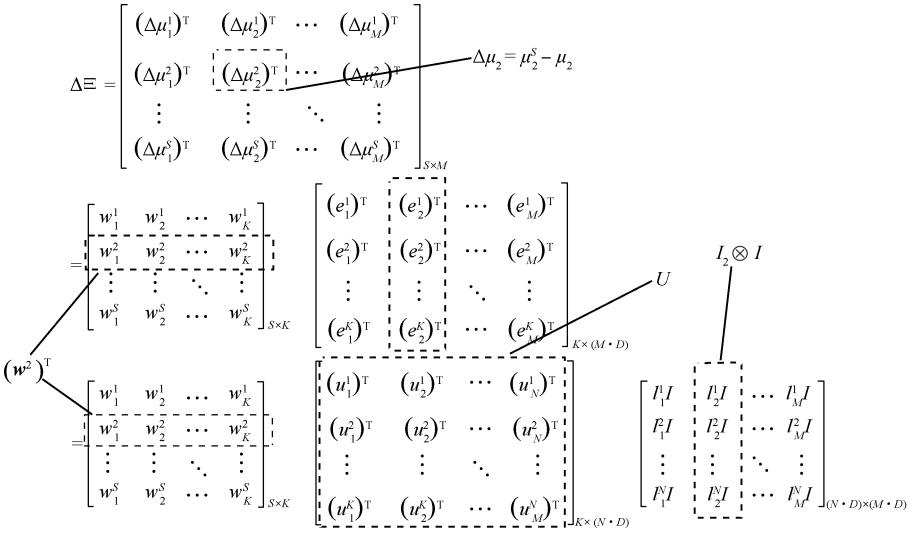

图 2　第 2 个说话人第 2 个高斯分量的分解过程

Fig. 2　The decomposition of the second Gaussian component for the second speaker

model, HMM) 模型共 3 个输出状态, 每个状态 8 个高斯混元, 进行三音子聚类后共 19 136 个高斯混元. 解码器采用 HTK 自带的一遍解码器 HVite, 不采用语法模型.

3.2 音子子空间存在性实验

首先, 为了验证音子子空间的存在性, 我们对音子空间进行了主分量分析. 具体做法如下: 对每个训练说话人, 利用基于回归树 (32 个回归类) 的 MLLR 自适应方法得到其说话人相关模型, 进而利用 PCA 得到本征音矩阵 $E_{K\times(M\cdot D)}$. 对每个高斯元 m, 将其对应的本征音均值矢量 $\{\boldsymbol{e}_m^k\}_{k=1}^K$ 排列为一个 $K\cdot D$ 维的列矢量, 对所有的 M 个这样的列矢量进行主分量分析, 得到与主分量相对应的前 150 个特征值的分布如图 3 所示 ($M = 19\,136$).

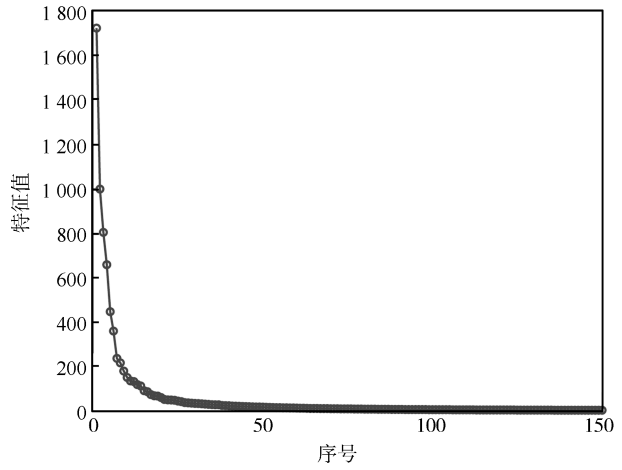

图 3 对音子空间进行主成分分析, 前 150 个特征值的分布图

Fig. 3 The distribution of the first 150 eigenvalues in the principal component analysis of the phone space

由图 3 可见, 特征值集中分布在前面 50 ~ 100 维. 因此, 在音子空间中确实存在一个低维的子空间.

3.3 无监督说话人自适应实验

为了比较本文方法的自适应效果, 在测试集上进行无监督的说话人自适应实验. 实验中, 我们构造了三套基线系统和一套基于本文方法的新系统:

1) SI: 直接利用说话人无关声学模型进行识别, 不采用说话人自适应技术;

2) MLLR: 采用基于回归树的 MLLR 自适应技术, 回归类个数设为典型值 32;

3) CMLLB: 采用聚类最大似然线性基的本征音自适应技术, 本征音 (Eigen-voice) 子空间维数设为 20, 音子聚类类数取为 100;

4) JSPS: 基于本文提出的说话人 – 音子联合子空间技术, 本征音 (Eigen-voice) 子空间维数设为 20, 本征音子 (Eigen-phone) 子空间的维数取为 50 (记为 JSPS-50) 或 100 (记为 JSPS-100).

实验中解码器均采用 HTK 工具箱中的一遍解码器 HVite, 不采用语言模型. 实验步骤分为两步: 首先将每个测试说话人 20 句话分为两组, 其中 10 句话作为自适应数据, 用于对 SI 声学模型进行无监督自适应得到 SD 模型, 剩下的 10 句话作为测试数据.

为了比较各方法在不同长度的自适应数据下的自适应效果, 分别对 1 句话、2 句话、4 句话、6 句话、8 句话、10 句话的自适应数据进行了实验. 最终, 有调音节及无调音节的平均识别率分别如表 1 和表 2 所示 (SI 模型的有调音节平均识别率为 53.04 %, 无调音节识别率为 76.21 %).

由结果可见, 无论是有调音节还是无调音节识别, 本文的方法在 1 ~ 4 句话 (相当于 5 ~ 20 秒左右) 的自适应语料下, 明显优于经典的 MLLR 自适应效果; 特别是在 1 ~ 2 句话 (5 ~ 10 秒) 自适应数据下, 本文方法即可快速达到饱和, 相比 SI 模型识别率提高了两个百分点, 而 MLLR 方法却几乎没有任何提高. 这是由于 MLLR 自适应方法需要估计一

表 1 无监督自适应实验结果 (有调音节识别率)

Table 1 Unsupervised speaker adaptation results (recognition rates of tonal syllables)

自适应方法	参数数量	自适应句数					
		1	2	4	6	8	10
MLLR	$\geq 39^2$	53.04	53.17	55.53	**55.71**	**56.29**	**56.76**
CMLLB	20	53.90	54.05	54.20	54.18	54.45	54.22
JSPS-50	20	54.74	**55.48**	**55.60**	**55.71**	55.83	55.83
JSPS-100	20	**54.87**	55.27	55.31	55.48	55.62	55.71

表 2 无监督自适应实验结果 (无调音节识别率)

Table 2 Unsupervised speaker adaptation results (recognition rates of toneless syllables)

自适应方法	参数数量	自适应句数					
		1	2	4	6	8	10
MLLR	$\geq 39^2$	76.21	76.67	78.30	**78.91**	**79.02**	**79.82**
CMLLB	20	76.94	77.40	77.59	77.53	77.59	77.63
JSPS-50	20	77.79	78.09	78.39	78.58	78.66	78.66
JSPS-100	20	**78.01**	**78.39**	**78.54**	78.60	78.75	78.81

个或多个变换矩阵, 所需要估计的参数数量较多 ($\geq 39^2$), 在少量自适应数据量下无法得到可靠的估计. 与聚类最大似然线性基 (CMLLB) 自适应方法相比, 尽管所需要估计的参数数量相同 (= 20), 但不论在哪种自适应数据量条件下, 本文方法对于系统平均识别率均有明显提高. 这是由于本文方法通过对音子空间进行主分量分析, 更精确地得到了各个音子之间的相关性信息, 对说话人先验信息有着更强的表征能力, 因此具有更好的快速自适应效果.

此外, 由上面的实验结果可见, 当本征音子 (Eigen-phone) 子空间取为 50 时即可取得很好的自适应效果, 相比原始的音子空间, 维数从 19 136 降为 50, 所需要估计的参数数量降低了到原来的 1/383, 大大提高了模型参数的稳健性, 节省了存储空间.

4 结论与进一步研究方向

本文提出了一种新的基于子空间的说话人自适应方法. 该方法在 "说话人" 和 "音子" 两个方向上寻找联合子空间, 不仅考虑了说话人间的模型参数相关性, 还考虑了音子间的模型参数相关性, 从而能够更好地对说话人模型参数变化的先验信息进行建模. 实验表明, 在少量自适应数据条件下, 新方法在大大降低模型的存储空间的同时具有良好的自适应效果. 本文的联合子空间的获得是利用 PCA 得到的, 作为进一步的研究方向, 可以考虑如何在最大似然或最大区分性准则下训练得到联合子空间的基矢量, 以及如何将新的自适应算法与说话人自适应训练框架相结合.

References

1 Li Hu-Sheng, Liu Jia, Liu Run-Sheng. Technology of speaker adaptation in speech recognition and its development trend. *Acta Electronica Sinica*, 2003, **31**(1): 103−108
(李虎生, 刘加, 刘润生. 语音识别说话人自适应研究现状及发展趋势. 电子学报, 2003, **31**(1): 103−108)

2 Woodland P C. Speaker adaptation: techniques and challenges. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, USA: IEEE, 1999. 85−90

3 Mak B K W, Lai T C, Tsang I W, Kwok J T Y. Maximum penalized likelihood kernel regression for fast adaptation. *IEEE Transactions on Audio, Speech and Language Processing*, 2009, **17**(7): 1372−1381

4 Kuhn R, Junqua J C, Nguyen P, Niedzielski N. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 2000, **8**(6): 695−707

5 Teng W X, Gravier G, Bimbot F, Soufflet F. Speaker adaptation by variable reference model subspace and application to large vocabulary speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Taipei, China: IEEE, 2009. 4381−4384

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

7 Mak B, Lai T C, Hsiao R. Improving reference speaker weighting adaptation by the use of maximum-likelihood reference speakers. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Toulouse, France: IEEE, 2006. 229−232

8 Tang Y, Rose R. Rapid speaker adaptation using clustered maximum-likelihood linear basis with sparse training data. *IEEE Transactions on Audio, Speech, and Language Processing*, 2008, **16**(3): 607−616

9 Jeong Y. Speaker adaptation based on the multilinear decomposition of training speaker models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Dallas, USA: IEEE, 2010. 4870−4873

10 Kenny P, Boulianne G, Ouellet P, Dumouchel P. Speaker adaptation using an eigenphone basis. *IEEE Transactions on Speech and Audio Processing*, 2004, **12**(6): 579−589

11 Chang E, Shi Y , Zhou J L, Huang C. Speech lab in a box: a mandarin speech toolbox to jumpstart speech related research. In: Proceedings of the 7th European Conference on Speech Communication and Technology. Aalborg, Denmark: ISCA, 2001. 2799−2802

12 Young S, Evermann G, Gales M, Gales M, Hain T, Kershaw D, Liu X. The HTK book (for HTK Version 3.4) [Online], available: http://htk.eng.cam.ac.uk/docs/docs.shtml, October 28, 2011

张文林 中国人民解放军信息工程大学信息工程学院博士研究生, 主要研究方向为语种识别, 连续语音识别, 机器学习. 本文通信作者.
E-mail: zwlin_2004@163.com
(**ZHANG Wen-Lin** Ph. D. candidate at the Institute of Information Engineering, PLA Information Engineering University. His research interest covers language identification, continuous speech recognition, and machine learning. Corresponding author of this paper.)

张卫强 清华大学电子工程系助理研究员. 主要研究方向为时频分析, 高阶统计量, 音频检索, 说话人识别, 语种识别.
E-mail: wqzhang@tsinghua.edu.cn
(**ZHANG Wei-Qiang** Research assistant in the Department of Electronic Engineering, Tsinghua University. His research interest covers time-frequency analysis, higher-order statistics, audio retrieval, speaker recognition, and language recognition.)

刘 加 清华大学电子工程系教授. 主要研究方向为语音识别, 说话人识别, 语种识别, 语音合成, 语音编码以及语言理解.
E-mail: liuj@tsinghua.edu.cn
(**LIU Jia** Professor in the Department of Electronic Engineering, Tsinghua University. His research interest covers speech recognition, speaker recognition, language recognition, expressive speech synthesis, speech coding, and spoken language understanding.)

李弼程 中国人民解放军信息工程大学信息工程学院教授. 主要研究方向为文本分析与理解, 语音处理与识别, 图像/视频处理与识别, 信息融合.
E-mail: lbclm@163.com
(**LI Bi-Cheng** Professor at the Institute of Information Engineering, PLA Information Engineering University. His research interest covers text analysis and understanding, speech/image/video processing and recognition, and information fusing.)

屈 丹 中国人民解放军信息工程大学信息工程学院副教授. 2005 年获解放军信息工程大学博士学位. 主要研究方向为语音信号处理与模式识别.
E-mail: qudanqudan@sina.com
(**QU Dan** Associate professor at the Institute of Information Engineering, PLA Information Engineering University. Her research interest covers speech signal processing and pattern recognition.)