

基于总体变化子空间自适应的 i-vector 说话人识别系统研究

栗志意¹ 张卫强¹ 何亮¹ 刘加¹

摘要 在说话人识别研究中, 基于身份认证矢量 (identity vector, i-vector) 的子空间建模被证明是目前最前沿最有效的说话人建模技术, 其中如何有效地准确地估计总体变化子空间矩阵 T 成为影响系统性能好坏的关键问题. 本文针对 i-vector 技术如何新的应用环境下进行总体变化子空间矩阵 T 的自适应估计问题进行了研究, 并提出了两种行之有效的自适应估计算法. 在由美国国家标准技术局 (American National Institute of Standard and Technology, NIST) 组织的 2008 年说话人识别核心评测数据库以及自行采集的测试数据库上的实验结果显示, 不论采用测试集数据本身还是与测试集较匹配的开发集数据, 通过本文所提的自适应算法来更新总体变化子空间矩阵均可以使更新后的子空间更有利于新测试数据下的低维子空间描述, 在新的测试环境下都更有利于说话人分类. 此外实验结果还表明基于多子空间拼接的子空间自适应方法性能明显优于迭代自适应方法, 而且两者的结合可达到最优的识别性能, 且此时利用开发集数据进行自适应可以接近其利用测试集数据进行自适应得到的最优性能.

关键词 身份认证矢量, 总体变化子空间, 自适应, 说话人识别

引用格式 栗志意, 张卫强, 何亮, 刘加. 基于总体变化子空间自适应的 i-vector 说话人识别系统研究. 自动化学报, 2014, 40(8): 1836–1840
DOI 10.3724/SP.J.1004.2014.01836

Total Variability Subspace Adaptation Based Speaker Recognition

LI Zhi-Yi¹ ZHANG Wei-Qiang¹ HE Liang¹
LIU Jia¹

Abstract In text-independent speaker recognition, the identity vector (i-vector) based modeling method has recently been proved to be the most popular and efficient method. It is a key problem to estimate the total variability subspace T efficiently and accurately. In this paper, two adaptation algorithms are proposed in order to improve the performance of the i-vector base system in practical environments. Experiments on the 2008 core speaker recognition evaluation dataset of American NIST and Technology and the self-collected speaker recognition evaluation dataset demonstrate that using the proposed adaptation algorithms to adapt to the total variability subspace T from either the test dataset or the developing dataset is effective for improving the performance. In addition, the combination of the two adaptation algorithms can achieve almost the best performance using the developing dataset rather than the test dataset.

Key words i-vector, total variability subspace, adaptation, speaker recognition

Citation Li Zhi-Yi, Zhang Wei-Qiang, He Liang, Liu Jia. Total variability subspace adaptation based speaker recognition. *Acta Automatica Sinica*, 2014, 40(8): 1836–1840

收稿日期 2013-11-13 录用日期 2013-11-23
Manuscript received November 13, 2013; accepted November 23, 2013

本文责任编辑 吴玺宏
Recommended by Associate Editor WU Xi-Hong
国家自然科学基金 (61370034, 61273268, 61005019, 90920302), 北京市自然科学基金项目 (KZ201110005005) 资助

说话人识别技术是指利用从说话人语音信号中提取出的声纹特征进行辨识或确认说话人身份的一项技术. 作为一项重要的生物特征身份鉴定技术, 该技术可广泛应用于国家安全、司法鉴定、语音拨号、电话银行等诸多领域^[1].

近几年来, 以身份认证矢量 i-vector 为基础的说话人建模技术取得了非常大的成功, 使得说话人识别系统的性能有了非常显著的提升^[2–3], 在由美国国家标准技术局 (American National Institute of Standards and Technology, NIST) 组织的国际说话人评测中, 基于该技术的说话人识别系统的性能明显优于之前广泛采用的高斯混合模型超矢量-支持向量机 (Gaussian mixture model super vector-support vector machine, GSV-SVM)^[4]、联合因子分析 (Joint factor analysis, JFA)^[5–6], 成为目前占主导地位的话人识别系统.

基于身份认证矢量 i-vector 的说话人建模方法与先前的 GSV-SVM、JFA 建模方法一样, 都是基于高斯混合模型-通用背景模型 (Gaussian mixture model-universal background model, GMM-UBM)^[7], 其基本思想是假设说话人信息以及信道信息同时处于高斯混合模型高维均值超矢量空间中的一个低维线性子空间流形结构中, 如式 (1) 所示.

$$\mathbf{M} = \mathbf{m} + T\mathbf{w} \quad (1)$$

其中, \mathbf{M} 表示高斯混合模型均值超矢量, \mathbf{m} 表示一个与特定说话人和信道都无关的超矢量. 而总体变化子空间矩阵 T 完成从高维空间到低维空间的映射, 从而使得降维后的矢量更有利于进一步地分类和识别. 因此该建模过程中, 首先通过因子分析的方法, 训练得到矩阵 T , 然后再将高维的高斯混合模型均值超矢量在该子空间上进行投影, 得到低维的总体变化因子矢量, 也称之为身份认证矢量 i-vector, 最后将得到的低维身份认证矢量 i-vector 进行线性鉴别性分析 (Linear discriminate analysis, LDA) 降维和类内协方差归一化 (Within class covariance normalization, WCCN). 前者线性鉴别性分析技术 LDA 的目的在于在满足最小化类内说话人距离和最大化类间说话人距离的鉴别性优化准则下进一步降低 i-vector 的维数, 后者类内协方差归一化技术 WCCN 的目的是通过白化协方差矩阵, 使得变换后的子空间的基尽可能正交. 最终将经过 LDA 和 WCCN 变换后的 i-vector 矢量, 作为输入特征送入后续的分类器进行分类判决. 经典的 i-vector 分类器包括余弦距离打分 (Cosine distance scoring, CDS) 分类器和 SVM 分类器^[8] 等, 本文采用与文献 [2] 中一致的余弦距离打分 CDS 分类器.

从 i-vector 建模方法的基本假设中可以看出, 如何准确地估计总体变化子空间 T 是一个非常基础性且关键性的环节. 准确的估计子空间 T 意味着投影后的低维 i-vector 矢量对于说话人和信道信息描述的更具有区分性, 更有利于进一步的分类. 在近年来由 NIST 组织的说话人评测中, 我们发现 i-vector 系统的性能一致性优于 GSV-SVM 系统, 其中一个因素在于 NIST 评测数据库的数据资源充足, 数据质量较好, 数据来源比较一致, 因此会对训练矩阵带来很多利好条件. 而在实际应用部署 i-vector 系统的过程中发现,

Supported by National Natural Science Foundation of China (61370034, 61273268, 61005019, 90920302) and Beijing Natural Science Foundation (KZ201110005005)

1. 清华大学电子工程系 清华信息与科学技术国家实验室 北京 100084

1. Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084

由于实际测试场景复杂, 数据资源短缺, 数据质量较差, 常导致子空间估计的不甚稳健^[9], 使得 i-vector 说话人系统在实际部署测试时的识别性能在很多情况下反而没有 GSV-SVM 系统稳健.

由于 i-vector 建模方法的核心和基础是对总体变化子空间 T 矩阵的估计, 因此本文针对 i-vector 说话人识别系统在子空间自适应估计问题进行了深入研究, 在此基础上提出了两种行之有效的子空间自适应算法, 并通过实验对比给出了最优的自适应策略.

本文安排如下: 第 1 节介绍了 i-vector 说话人系统中总体变化子空间 T 矩阵的估计过程; 第 2 节介绍了 i-vector 说话人系统中模型的训练和测试过程; 第 3 节介绍了本文提出的两种总体变化子空间矩阵 T 的自适应算法, 并给出所对应的系统框图; 第 4 节给出本文所提算法在测试数据集上的实验结果和分析; 最后在第 5 节给出总结和结论.

1 总体变化子空间 T 矩阵估计

1.1 统计量估计

在 i-vector 系统总体变化子空间 T 的估计过程中, 由于高斯混合模型均值超矢量是通过计算声学特征相对于通用背景模型 UBM 均值超矢量的零阶、一阶和二阶统计量得到的, 因此本小节将首先给出声学特征各阶统计量的估计过程. 为了估计各阶统计量, 需要首先利用一些训练数据通过期望最大化 (Expectation maximum, EM) 算法训练得到通用背景模型 UBM, 该模型提供了一个统一的参考坐标空间, 并且可以在一定程度上解决由于说话人训练数据较少导致的小样本问题. 而高斯混合模型则可通过训练数据在该 UBM 上面进行最大后验概率 (Maximum a posterior, MAP) 自适应得到.

各阶统计量的估计过程如下所示, 假设说话人 s 的声学特征表示为 $\mathbf{x}_{s,t}$, 则其相对于 UBM 均值超矢量 \mathbf{m} 的零阶统计量 $N_{c,s}$, 一阶统计量 $\mathbf{F}_{c,s}$ 以及二阶统计量 $S_{c,s}$ 可如式 (2) 所示.

$$\begin{aligned} N_{c,s} &= \sum_t \gamma_{c,s,t} \\ \mathbf{F}_{c,s} &= \sum_t \gamma_{c,s,t} (\mathbf{x}_{s,t} - \mathbf{m}_c) \\ S_{c,s} &= \text{diag} \left\{ \sum_t \gamma_{c,s,t} (\mathbf{x}_{s,t} - \mathbf{m}_c) (\mathbf{x}_{s,t} - \mathbf{m}_c)^T \right\} \end{aligned} \quad (2)$$

式中 \mathbf{m}_c 代表 UBM 均值超矢量 \mathbf{m} 中的第 c 个高斯均值分量. t 表示时间帧索引. $\gamma_{c,s,t}$ 表示 UBM 第 c 个高斯分量的后验概率. $\text{diag}\{\cdot\}$ 表示取对角运算. 假设单高斯模型的维数为 F , 则将所有 C 个高斯模型的均值矢量拼接成的高维均值超矢量维数为 FC .

1.2 子空间 T 估计

对于上述得到的各阶统计量, 子空间 T 的估计可以采用如下的期望最大化 (Expectation maximum, EM) 算法得到, 首先随机初始化子空间矩阵 T , 然后固定 T , 在最大似然准则下估计隐变量 \mathbf{w} 的一阶和二阶统计量, 估计过程如式 (3) 所示. 其中超矢量 \mathbf{F}_s 是由 $\mathbf{F}_{c,s}$ 矢量拼接成的 $FC \times 1$ 维的矢量. N_s 是由 $N_{c,s}$ 作为主对角元拼接成的 $FC \times FC$ 维的矩阵.

$$L_s = I + T^T \Sigma^{-1} N_s T$$

$$\begin{aligned} E[\mathbf{w}_s] &= L_s^{-1} T^T \Sigma^{-1} \mathbf{F}_s \\ E[\mathbf{w}_s \mathbf{w}_s^T] &= E[\mathbf{w}_s] E[\mathbf{w}_s^T] + L_s^{-1} \end{aligned} \quad (3)$$

式中 L_s 是临时变量, Σ 是 UBM 的协方差矩阵.

接着更新 T 矩阵和协方差矩阵 Σ . T 矩阵的更新过程可利用式 (4) 来实现, 也可根据文献 [10] 中的快速算法来实现.

$$\sum_s N_s T E[\mathbf{w}_s \mathbf{w}_s^T] = \sum_s \mathbf{F}_s E[\mathbf{w}_s] \quad (4)$$

对 UBM 协方差矩阵 Σ 的更新过程如式 (5) 所示.

$$\Sigma = N^{-1} \sum_s S_s - N^{-1} \text{diag} \left\{ \sum_s \mathbf{F}_s E[\mathbf{w}_s^T] T^T \right\} \quad (5)$$

式中 S_s 是由 $S_{c,s}$ 进行矩阵对角拼接成的 $FC \times FC$ 维的矩阵, $N = \sum N_s$ 为所有说话人的零阶统计量之和.

对于上述步骤反复进行迭代 6~8 次后, 可近似认为 T 和 Σ 收敛.

2 i-vector 模型训练和测试

本文对于 i-vector 模型的训练和测试采用与文献 [2] 一致的过程, 即首先通过 LDA 和 WCCN 对于上述子空间投影后的 i-vector 矢量进行进一步的鉴别性降维和白化, 然后利用余弦距离打分对处理后的 i-vector 进行最终打分和判决.

2.1 线性鉴别性分析

线性鉴别性分析^[11] (Linear discriminant analysis, LDA) 是模式识别领域广泛采用的一种鉴别性降维技术. 在基于身份认证矢量 i-vector 的说话人系统中, 由于前述基于因子分析的方法没有采用鉴别性的准则, 因此通常首先采用 LDA 对因子分析后的 i-vector 矢量进行鉴别性降维. 训练 LDA 矩阵的过程主要通过优化如式 (6) 所示的目标函数, 即在最小化类内说话人距离和最大化类间说话人距离的鉴别性准则下求该目标函数的最优解.

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \quad (6)$$

式 (6) 中类间协方差矩阵 S_B 和类内协方差矩阵 S_W 的计算过程如式 (7) 和式 (8) 所示.

$$S_B = \sum_{s=1}^S (\mathbf{w}_s - \bar{\mathbf{w}}) (\mathbf{w}_s - \bar{\mathbf{w}})^T \quad (7)$$

$$S_W = \sum_{s=1}^S \frac{1}{n_s} \sum_{i=1}^{n_s} (\mathbf{w}_i^s - \bar{\mathbf{w}}_s) (\mathbf{w}_i^s - \bar{\mathbf{w}}_s)^T \quad (8)$$

其中 $\bar{\mathbf{w}}_s = (1/n_s) \sum_{i=1}^{n_s} \mathbf{w}_i^s$ 代表第 s 个说话人的 i-vector 均值矢量. S 表示说话人的个数, n_s 表示第 s 个说话人的 i-vector 段数. 最终求解式 (6) 中最优化目标函数的过程可转换为求解如式 (9) 所示广义特征值的过程.

$$S_B \mathbf{w} = \lambda S_W \mathbf{w} \quad (9)$$

2.2 类内方差归一化

类内方差归一化^[12] (Within class covariance normalization, WCCN) 通过白化说话人因子使得变换后说话人子空间的基尽可能正交. WCCN 矩阵可由式 (10) 估计.

$$W = \frac{1}{S} \sum_{s=1}^S \frac{1}{n_s} \sum_{i=1}^{n_s} (\mathbf{w}_i^s - \bar{\mathbf{w}}_s) (\mathbf{w}_i^s - \bar{\mathbf{w}}_s)^T \quad (10)$$

其中 $\bar{\mathbf{w}}_s = (1/n_s) \sum_{i=1}^{n_s} \mathbf{w}_i^s$ 代表第 s 个说话人的 i-vector 均值矢量. S 表示说话人个数, n_s 表示第 s 个说话人的 i-vector 段数.

2.3 余弦距离打分

余弦距离打分^[2] 是一种对称式的核函数分类器, 即说话人模型 i-vector 矢量与测试段 i-vector 矢量交换后打分结果不变. 在对 i-vector 分类时, 该分类器将说话人矢量 \mathbf{w}_{tar} 和测试段矢量 \mathbf{w}_{tst} 的余弦距离分数直接作为判决分数, 并与阈值 θ 进行比较, 给出判决结果, 如式 (11) 所示.

$$score(\mathbf{w}_{tar}, \mathbf{w}_{tst}) = \frac{\langle \mathbf{w}_{tar}, \mathbf{w}_{tst} \rangle}{\|\mathbf{w}_{tar}\| \cdot \|\mathbf{w}_{tst}\|} \stackrel{\geq}{\leq} \theta \quad (11)$$

本质上讲, 该分类器通过归一化矢量的模去除了矢量幅度的影响, 将两个矢量的余弦角度作为分类的依据, 计算简单快捷, 而且该分类器可以通过合并计算, 加快后续分数归一化的过程.

3 总体变化子空间 T 的自适应算法

与联合因子分析 JFA 子空间建模方法不同, i-vector 子空间建模过程中 T 矩阵的估计不需要任何标签信息, 属于无监督训练过程, 这为在实际应用中通过大量的无标注数据来自适应估计 T 矩阵来提高子空间估计的性能提供了可能. 在实际应用中, 为了在新测试环境下对子空间 T 矩阵进行稳健地估计, 本文源于不同的出发点, 提出了两种子空间自适应算法, 并提出了两者结合的自适应算法.

3.1 迭代自适应算法

该算法的基本思想类似于高斯混合模型-通用背景模型的思想, 首先离线利用已有的训练数据来按照 1.2 中的算法流程估计一个与测试条件不相关的子空间矩阵 T_o , 称之为通用子空间矩阵, 然后从通用子空间矩阵 T_o 上进行子空间自适应, 完成子空间的迁移变换, 如图 1 所示.

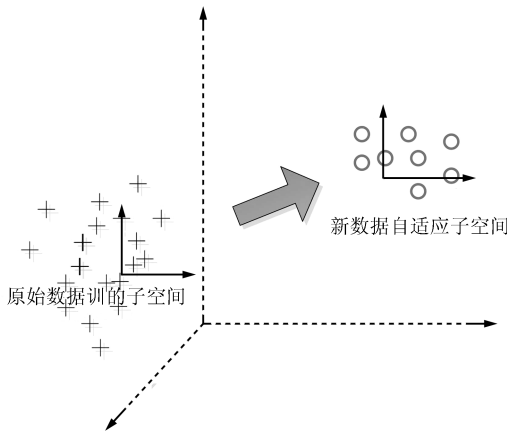


图 1 总体变化子空间 T 的迭代自适应算法示意图
Fig. 1 Total variability diagram of total variability subspace T iteration adaptation algorithm

具体实施过程中, 首先通过已有的训练数据得到通用子空间矩阵 T_o , 然后将该矩阵作为初始化种子, 利用 EM 算法在新的测试数据集上进行迭代自适应. 该迭代过程与 1 中的迭代过程类似, 其自适应过程如算法 1 所示.

算法 1. 总体变化子空间 T 的迭代自适应算法

步骤 1. 用已有数据训练得到通用子空间变化矩阵 T_o 和 UBM 协方差矩阵 Σ , 作为初始化种子;

步骤 2. 根据 T , 计算临时变量 L , 并估计总体变化矢量 \mathbf{w} 的一阶统计量 $E[\mathbf{w}_s]$ 和二阶统计量 $E[\mathbf{w}_s \mathbf{w}_s^T]$;

步骤 3. 根据步骤 2 中的统计量对 T 矩阵进行更新, 如式 (4) 所示;

步骤 4. 根据步骤 2 和步骤 3 中结果更新 UBM 协方差矩阵 Σ , 更新过程如式 (5) 所示;

步骤 5. 未达到迭代次数则返回步骤 2 中继续; 否则结束退出.

3.2 拼接自适应算法

该算法的基本思想源于文献中对联合因子分析建模技术中信道空间拼接的思想^[13-14], 在联合因子分析建模过程中, 通过拼接信道子空间, 从而可以更加有效地去掉信道因子分量. 而在 i-vector 建模过程中, 我们认为在利用原始训练数据和新环境下的数据集可以分别训练得到两个反映不同角度的总体变化子空间, 意味着通过在这两个总体变化子空间构成的联合子空间中对高维矢量进行投影, 可以从不同角度来反映低维的总体变化因子.

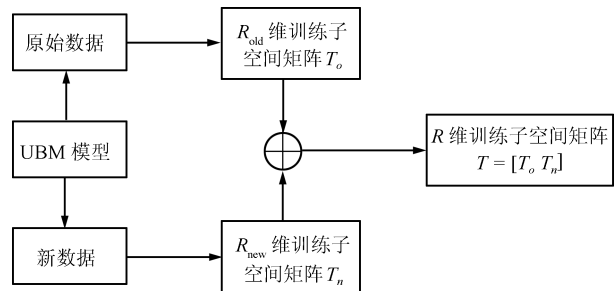


图 2 总体变化子空间 T 拼接自适应算法示意图
Fig. 2 Diagram of total variability subspace T combination adaptation algorithm

该算法如图 2 所示, 首先通过原始训练数据和新数据集利用第 1.2 中的估计算法分别训练得到两个不同的总体变化子空间矩阵 T_o 和 T_n , 然后将两个子空间进行拼接, 得到自适应后的子空间矩阵. 该算法流程如下所示:

算法 2. 总体变化子空间 T 拼接自适应算法

步骤 1. 利用原始数据训练得到子空间变化矩阵 T_o ;

步骤 2. 利用新的测试数据训练得到子空间变化矩阵 T_n ;

步骤 3. 拼接 T_o 和 T_n 得到最终的自适应子空间 T .

3.3 两者结合的自适应算法

针对上述提出的两种子空间自适应算法, 我们进一步提出了两者相结合的自适应算法, 该结合算法可以实现之前两种算法的有效互补, 更有利于总体变化因子在低维子空间中的表示. 图 3 给出了两者结合的自适应算法流程结构图.

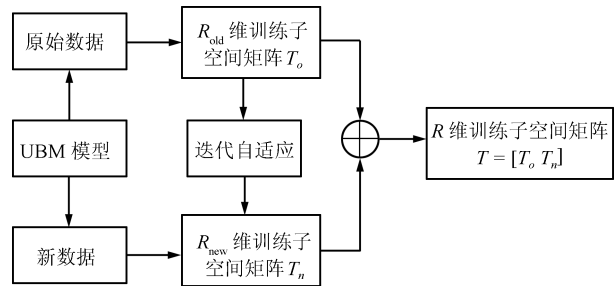


图 3 总体变化子空间 T 的迭代自适应和拼接自适应结合的自适应算法示意图
Fig. 3 Diagram of total variability subspace T integration algorithm of iteration adaptation and subspace combination adaptation

该算法如图 3 所示, 首先通过已有的训练数据利用第 1.2 节中的估计算法训练得到总体变化子空间矩阵 T_0 , 然后再根据算法 1 在该子空间上对新数据集进行迭代得到 T_n , 最后对两个子空间进行拼接, 得到最终的子空间矩阵. 该算法流程如下所示.

算法 3. 总体变化子空间 T 拼接自适应算法

步骤 1. 利用原始数据训练得到子空间变化矩阵 T_0 ;

步骤 2. 利用迭代自适应算法 1, 对上述矩阵在新数据集上迭代得到子空间变化矩阵 T_n ;

步骤 3. 拼接 T_0 和 T_n 得到最终的自适应子空间 T .

3.4 算法复杂度分析

通过分析以上三种自适应算法的流程可以看出, 与不进行自适应相比, 每种自适应算法都需要增加额外的自适应时间, 而具体的时间与自适应所用的数据量大小呈线性关系. 此外, 三种子空间自适应算法本身的时间复杂度是一样的, 而不同点在于基于子空间拼接的自适应算法在后续低维投影上的时间和空间是第一种自适应算法的一倍左右, 因此在实际应用中也需要根据系统实时率和响应需求来采用最合适的自适应算法.

4 实验配置与实验结果

为了验证本文所提算法的性能, 本文将针对所提子空间自适应 i-vector 说话人系统性能进行实验验证和结果对比, 并分析得出结论. 在实际部署中, 不同的应用环境使得我们是否可以将新数据用于自适应也有所不同. 比如离线测试环境或者说人检索情况下, 我们可以获得所有的测试数据, 因此可以直接利用测试数据. 而对于某些应用环境下, 比如在线测试情况下, 我们不能得到所有的测试集数据, 但我们可以退而求其次, 提前获得一些接近该测试集数据的开发集数据.

4.1 实验配置

本文实验中采用两个数据集, 一个是 NIST 发布的 SRE 2008 的核心数据集, 原始训练子空间数据来自 Switchboard I 和 II 约 20 000 条, 这些数据同样用于训练 UBM 模型, ZTnorm 伙伴集以及训练 LDA 和 WCCN 矩阵, 新数据集包括新的说话人训练数据集以及 12 922 条的测试数据集, 以及 3 000 条左右的开发集. 另一个是自行采集的数据集, 原始训练子空间数据和新数据来源于不同的地域和采集卡. 原始数据采集了约 20 000 条, 这些数据用于训练 UBM 模型, ZTnorm 伙伴集以及训练 LDA 和 WCCN 矩阵, 新数据集包括新的说话人训练数据集以及 8 000 条的测试数据集, 以及 2 000 条左右的开发集.

实验中采用 Mel 频率倒谱特征 (Mel-frequency cepstral coefficients, MFCC) 作为声学层特征. 在预处理阶段采用 G.723.1 进行有效语音端点检测 (Voice activity detection, VAD) 以及采用倒谱均值减 (Cepstral mean subtraction, CMS) 技术来去除或抑制信道的卷积噪声, 并设置了 3s 窗长用于特征弯折 (Feature warping), 进行了 25% 低能量删减以及预加重 (因子为 0.95). 在上述预处理基础上, 首先提取 13 维基本特征, 并与一阶、二阶差分特征一起构成最终的 39 维 MFCC 声学层特征. 实验中使用 UBM 的混合数设置为 1 024, 高斯概率密度的方差采用对角阵. i-vector 中总体变化子空间矩阵 T 的子空间维数也即列数均设置为 400, 训练时迭代次数取 6 次. LDA 矩阵降维后的维数设置为 200.

4.2 实验结果

本文中衡量系统性能指标采用等错误率 (Equal error rate, EER) 和最小检测代价函数 (Minimum detection cost function, MinDCF).

表 1 和表 2 给出了在两个数据集上由原始数据集训练得到子空间 T 和迭代自适应算法训练 T 在新开发集数据以及新测试集数据上的性能比较结果. 可以看出, 不管利用新开发集还是新测试集的数据进行自适应训练, 均可以较之有性能提升. 从表中数据还可以看出, 在实际应用中, 虽然开发集数据较之原始数据更接近测试集数据, 但是由于开发集数据的获取有限, 所以采用开发集数据进行迭代自适应获得的性能提升也有限. 而由于测试集数据本身的匹配程度最高, 因此可以得到最好的自适应性能. 因此在实际应用中应该首先选择利用测试集数据本身来进行子空间 T 的自适应. 在某些在线测试应用下, 若无法利用测试集数据, 也可以考虑采用开发集数据来做自适应.

表 1 原始数据训练 T 与本文所提迭代自适应 T 在 NIST SRE 2008 核心数据集上的性能比较

Table 1 Performance comparison of baseline training T algorithm and the proposed iteration adaptation T algorithm on NIST SRE 2008 core dataset

算法	EER	MinDCF
原始数据训练 T	5.41	0.029
新开发集数据自适应 T	4.92	0.026
新测试集数据自适应 T	4.67	0.023

表 2 原始数据训练 T 与本文所提迭代自适应 T 在自行采集数据集上的性能比较

Table 2 Performance comparison of baseline training T algorithm and the proposed iteration adaptation T algorithm on actual application dataset

算法	EER	MinDCF
原始数据训练 T	3.00	0.014
新开发集数据自适应 T	2.99	0.013
新测试集数据自适应 T	2.00	0.011

表 3 原始数据训练 T 与本文所提迭代自适应和子空间拼接自适应相结合的自适应算法在 NIST SRE 2008 核心数据集上的性能比较

Table 3 Performance comparison of baseline training T algorithm and the proposed integration algorithm of iteration adaptation and subspace combination adaptation on NIST SRE 2008 core dataset

算法	EER	MinDCF
原始数据训练 T	5.41	0.029
新开发集数据自适应 T	4.01	0.021
新测试集数据自适应 T	3.89	0.020

从表 1 和表 2 中的实验结果可以看出, 迭代自适应在两个数据集上均可以一致性地提高系统的性能. 因此接下来直接对第 3.3 节中所提出的迭代自适应与拼接自适应相结合的自适应算法上进行实验比较. 如表 3 和表 4 所示实验结果所示, 通过与空间拼接自适应相结合, 识别性能有更进一步的改善. 且此时利用开发集数据进行自适应可以接近其利用测试集数据进行自适应得到的最优性能. 因此实际应用中, 如

果在可

表 4 原始数据训练 T 与本文所提迭代自适应和子空间拼接自适应相结合的自适应算法在自行采集数据集上性能比较

Table 4 Performance comparison of baseline training T algorithm and the proposed integration algorithm of iteration adaptation and subspace combination adaptation on actual application dataset

算法	EER	MinDCF
原始数据训练 T	3.00	0.014
新开发集数据自适应 T	1.99	0.012
新测试集数据自适应 T	1.99	0.010

以获得测试集数据情况下, 利用测试集数据进行自适应可以取得最优的自适应效果. 当测试集数据不可用于训练子空间的情况下, 可以退而求其次, 利用与测试集较为匹配的开发集, 可以取得同样不错的性能. 这样也为我们在实际应用环境中, 如何有效地通过自适应来提高 i-vector 说话人识别系统的性能提供了参考依据.

5 讨论与结论

从基于身份认证矢量 i-vector 建模的说话人识别的原理假设来看, 有效准确的估计子空间总体变化矩阵 T 是一个基本性和关键性的问题, 会直接影响系统识别性能的好坏, 同时也是影响该建模技术在实际应用中稳健性的关键问题. 本文针对 i-vector 技术如何在实际应用中根据新数据来自适应子空间矩阵 T 进行了深入研究, 提出几种切实可行的自适应估计算法, 并针对不同的测试条件下给出了最优的自适应策略. 本文所提算法在 NIST SRE 2008 核心测试数据集和自行采集的测试数据库上的实验结果均显示, 不论采用测试集本身还是与测试集较匹配的开发集数据, 通过本文所提的自适应算法均可以使更新后的子空间更有利于新测试数据下的低维子空间描述, 从而更有利于说话人识别. 此外实验结果还表明基于多子空间拼接的自适应算法的性能明显优于迭代自适应算法, 而且两者的结合可或得到最优的系统性能, 且此时利用开发集数据进行自适应可以接近其利用测试集数据进行自适应得到的性能提升, 这样为在实际应用环境下如何有效地通过子空间自适应来提高 i-vector 说话人识别系统的性能提供了重要的参考依据.

References

- Kinnunen T, Li H Z. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, 2010, **52**(1): 12–40
- Dehak N, Kenny P, Ouellet P, Dumouchel P. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2011, **19**(4): 788–798
- Li Zhi-Yi, He Liang, Zhang Wei-Qiang, Liu Jia. Speaker recognition based on discriminant i-vector local distance preserving projection. *Journal of Tsinghua University (Science and Technology)*, 2012, **52**(5): 598–601
(栗志意, 何亮, 张卫强, 刘加. 基于鉴别性 i-vector 局部距离保持映射的说话人识别. 清华大学学报 (自然科学版), 2012, **52**(5): 598–601)
- Campbell W M, Campbell J P, Reynolds D A, Singer E, Torres-Carrasquillo P A. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 2006, **20**(2–3): 210–229

- Kenny P, Boulianne G, Ouellet P, Dumouchel P. Speaker and session variability in GMM-based speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2007, **15**(4): 1448–1460
- Kenny P, Boulianne G, Ouellet P, Dumouchel P. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2007, **15**(4): 1435–1447
- Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 2000, **10**(1–3): 19–41
- Cortes C, Vapnik V. Support vector networks. *Machine Learning*, 1995, **20**(3): 273–297
- Zhang Wen-Lin, Zhang Wei-Qiang, Liu Jia, Li Bi-Cheng, Qu Dan. A new subspace based speaker adaptation method. *Acta Automatica Sinica*, 2011, **37**(12): 1495–1502
(张文林, 张卫强, 刘加, 李弼程, 屈丹. 一种新的基于子空间的说话人自适应方法. 自动化学报, 2011, **37**(12): 1495–1502)
- Kenny P, Boulianne G, Dumouchel P. Eigenvoice modeling with sparse training data. *IEEE Transactions on Audio, Speech, and Language Processing*, 2005, **13**(3): 345–354
- Bishop C M. *Pattern Recognition and Machine Learning*. Berlin: Springer, 2008
- Hatch A O, Kajarekar S, Stolcke A. Within-class covariance normalization for SVM-based speaker recognition. In: *Proceedings of the International Conference on Spoken Language Processing*. Pittsburgh, PA, 2006. 1471–1474
- He Liang, Shi Yong-Zhe, Liu Jia. Eigenchannel space combination method of joint factor analysis *Acta Automatica Sinica*, 2011, **37**(7): 849–856
(何亮, 史永哲, 刘加. 联合因子分析中的本征信道空间拼接方法. 自动化学报, 2011, **37**(7): 849–856)
- Guo Wu, Li Yi-Jie, Dai Li-Rong, Wang Ren-Hua. Factor analysis and space assembling in speaker recognition. *Acta Automatica Sinica*, 2009, **35**(9): 1193–1198
(郭武, 李轶杰, 戴礼荣, 王仁华. 说话人识别中的因子分析以及空间拼接. 自动化学报, 2009, **35**(9): 1193–1198)

栗志意 清华大学电子工程系博士研究生. 主要研究方向为说话人识别与语种识别. 本文通信作者.

E-mail: lizhiyi06@mails.tsinghua.edu.cn

(LI Zhi-Yi Ph. D. candidate in the Department of Electronic Engineering, Tsinghua University. His research interest covers speaker recognition and language recognition. Corresponding author of this paper.)

张卫强 清华大学电子工程系助理研究员. 主要研究方向为说话人识别与语种识别. E-mail: wqzhang@tsinghua.edu.cn

(ZHANG Wei-Qiang Assistant professor in the Department of Electronic Engineering, Tsinghua University. His research interest covers speaker recognition and language recognition.)

何亮 清华大学电子工程系助理研究员. 主要研究方向为说话人识别与语种识别. E-mail: heliang@tsinghua.edu.cn

(HE Liang Assistant professor in the Department of Electronic Engineering, Tsinghua University. His research interest covers 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和信号处理. E-mail: liuj@tsinghua.edu.cn

(LIU Jia Professor in the Department of Electronic Engineering, Tsinghua University. His research interest covers speech recognition and signal processing.)