高斯 PLDA 在说话人确认中的应用及其联合估计

许云飞 [1] 杨 海 [1] 周若华 [1] 颜永红 [1]

摘 要 近年来, 基于总变化因子的说话人识别方法成为说话人识别领域的主流方法. 其中, 概率线性鉴别分析 (Probabilistic linear discriminant analysis, PLDA) 因其优异的性能而得到学者们的广泛关注. 然而, 在估计 PLDA 模型时, 传统的因子分析方法只更新模型空间, 因此, 模型均值不能很好地与更新后的模型空间耦合. 提出联合估计法对模型均值和模型空间同时估计, 得到更为严格的期望最大化更新公式, 在美国国家标准与技术局说话人识别评测 2010 扩展测试数据库以及 2012 核心测试数据库上, 等错率得到一定提升.

关键词 因子分析, 总变化因子, 概率线性鉴别分析, 联合估计, 期望最大化

引用格式 许云飞, 杨海, 周若华, 颜永红. 高斯 PLDA 在说话人确认中的应用及其联合估计. 自动化学报, 2014, **40**(6): 1068−1074

DOI 10.3724/SP.J.1004.2014.01068

Gaussian PLDA for Speaker Verification and Joint Estimation

XU Yun-Fei[1] YANG Hai[1] ZHOU Ruo-Hua[1] YAN Yong-Hong[1]

Abstract Recently the approaches based on i-vector have become very popular in the speaker recognition domain. Among these methods, the probabilistic linear discriminant analysis (PLDA) has attracted much attention due to its promising performance. However, the traditional factor analysis method only updates model space, thus making model mean couple with the model space unsuitably. This paper propose an approach of joint estimation for both model mean and model space, resulting in more strict expectation maximization (EM) formula. The equal error rate has been improved on the NIST SRE 2010 extended test corpus and NIST SRE 2012 core test corpus.

Key words Factor analysis, i-vector, probabilistic linear discriminant analysis (PLDA), joint estimation, expectation-maximization (EM)

Citation Xu Yun-Fei, Yang Hai, Zhou Ruo-Hua, Yan Yong-Hong. Gaussian PLDA for speaker verification and joint estimation. *Acta Automatica Sinica*, 2014, **40**(6): 1068−1074

在过去 10 年里, 说话人识别领域中有多种信道补偿技术被相继提出. 其中, 因子分析方法由于理论完备, 与基于混合高斯-通用背景模型 (Gaussian mixture model-universal background model, GMM-UBM)[1] 的说话人识别主流算法结合紧密, 并且与最大后验概率的说话人建模方法能够有效结合[2], 因此, 得到研究者的广泛认可.

根据 GMM-UBM 思想, 所有的信息都包含在混合高斯函数所形成的均值超矢量中, 假设高斯数为 C, 特征的维数为 F, 将所有的均值拼接, 最终形成一个 CF 维的超矢量. 联合因子分析 (Joint factor analysis, JFA)[2−5] 认为, 该超矢量可以被分解为说话人与信道两部分之和. 然而, 文献 [6] 中的实验证明, 在 JFA 进行信道补偿时, 不可避免地包含了说话人信息. 不仅如此, JFA 还存在空间掩盖和空间重叠问题[5].

针对 JFA 不能分别对说话人和信道进行准确建模和区分的缺点, Dehak 等提出总变化因子技术. 该技术源于 JFA, 但是在 GMM 超矢量空间, 它不能区分说话人和信道, 而是将两部分当成一个整体建模. 通过该技术, 帧数不定的语音被映射为固定长度的低维矢量, 与此同时, 尽可能保留了说话人信息[7]. 该技术的优势在于: 1) 只需在这个低维的空间, 而不是高维的 GMM 超矢量空间进行信道补偿; 2) 它为探索说话人识别的本质打开了新的思路. 至

收稿日期 2013-01-06 录用日期 2013-08-12
Manuscript received January 6, 2013; accepted August 12, 2013
国家高技术研究发展计划 (863 计划) (2012AA012503), 国家自然科学基金 (10925419, 90920302, 61072124, 11074275, 11161140319, 91120001, 61271426), 中国科学院战略性先导科技专项 (XDA06030100, XDA06030500), 中科院重点部署项目 (KGZD-EW-103-2) 资助.
Supported by National High Technology Research and Development Program of China (863 Program) (2012AA012503), National Natural Science Foundation of China (10925419, 90920302, 61072124, 11074275, 11161140319, 91120001, 61271426), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA06030100, XDA06030500), and Priority Deployment Project of Chinese Academy of Sciences (KGZD-EW-103-2)
本文责任编委 吴玺宏
Recommended by Associate Editor WU Xi-Hong
1. 中国科学院语言声学与内容理解重点实验室 北京 100190
1. Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190

此, 可以从经典模式识别的角度, 来处理说话人识别问题. 最初, Dehak 等采用线性鉴别分析 (Linear discriminant analysis, LDA) 及类内协方差规整技术. 来构建总变化因子识别系统, 用余弦 (Cosine) 打分进行分类[7].

后来, McLaren 等提出归一化带权重的 LDA, 妥当地解决了各种信道下, 训练数据不均衡的问题[8].

近两年, 学者们通常采用生成式模型对总变化因子进行信道补偿, 忽略总变化因子的提取机制, 认为它们是由概率生成模型得到的观察量. 在生成式模型中, 语音 $\boldsymbol{D}$ 被分解为说话人分量 $\boldsymbol{S}$ 和信道分量 $\boldsymbol{C}$ 之和, 即 $\boldsymbol{D}=\boldsymbol{S}+\boldsymbol{C}$. 其中 $\boldsymbol{S}$ 与 $\boldsymbol{C}$ 统计独立, $\boldsymbol{S}$ 和 $\boldsymbol{C}$ 服从高斯分布. 一个典型的生成式模型例子为概率线性鉴别分析 (Probabilistic linear discriminant analysis, PLDA), 其最初用于人脸识别[9], 后来被用于说话人识别[10−12]. PLDA 可以看成 JFA 在单高斯下的一种特例. 另一个经典的生成式模型例子为 2 Cov 模型[13], 本质上, 它是一个简单的线性高斯生成模型[14], 又可以看成 PLDA 的一种特例. 考虑到 2 Cov 估计参数的不确定性, 文献 [15] 从全贝叶斯的角度进行推导, 将这种不确定引入到打分过程, 从而获得一定的性能提升. 然而, 上面两个模型的基本假设都不够准确, 原因有两个: 1) 信道影响并不是与说话人无关的 (比如性别相关的本征信道建模比性别无关的要好), 也就是独立性假设不成立; 2) 高斯先验这一假设，要求同一个说话人的总变化因子不能过大地偏离均值, 但在实际情况中，这是很常见的 (比如非母语交流下的录音或严重的信道失真). 为了解决问题 2), 寻找合适的先验概率就显得尤为重要. 文献 [16] 通过将高斯先验分布替换成 Student′s t 分布 (Heavy tailed PLDA, HT-PLDA), 并采用变分贝叶斯方法计算对数似然比, 很好地解决了这个问题; 并且, 相对高斯形式的 PLDA (Gussian PLDA, G-PLDA), 有较大优势. 同样, 采用拉普拉斯分布取代高斯分布 (Sparse PLDA, SP-PLDA), 也获得了性能提升[17].

尽管 HT-PLDA 和 SP-PLDA 大大提升了系统性能, 但是, 其计算量也非常大, 尤其当总变化因子维数较高时, 这一问题更为突出. 采用 LDA 技术对总变化因子降维, 然后送入 HT-PLDA, 可以减少计算量[11, 16], 但这会带来另外一个问题: 由于总变化因子本身包含了丰富的信息, LDA 过程中不可避免地会丢失一些有用信息. 文献 [11] 提出长度规整技术 (Length normalization, LNorm) 先对总变化因子进行预处理, 再送入 G-PLDA, 不但计算量大大降低, 而且性能与 HT-PLDA 相当. 究其原因, 是规整后的总变化因子, 其分布与 G-PLDA 基本假设相对吻合. 基于这个思路, 文献 [12] 提出了方差谱规整技术 (Variance-spectra based normalization, SphNorm), 该技术基于总变化因子的均值和类内协方差矩阵, 在循环迭代过程中, 将总变化因子映射到合适的空间. 最终得到的总变化因子, 不但模为单位长度, 而且, 其分布与模型假设更加吻合.

通过观察发现, 传统的因子分析方法都是预先计算好 G-PLDA 的均值[10−12], 期望最大化 (Expectation-maximization, EM) 过程中, 不进行更新, 因此, 均值不能很好地与更新后的模型空间耦合. 同样, 与 JFA 类似, 采用本征音和本征信道两个载荷矩阵建模并不好. 本文关注 G-PLDA, 去掉本征信道, 只保留本征音, 提出模型参数的联合估计法, 并给出相应的推导. 结果表明, 尽管 G-PLDA 在规整技术的基础上已经可以达到很好的性能, 联合估计法仍有相对 5 %∼6 % 的等错率 (Equal error rate, EER) 提升.

本文结构如下: 第 1 节简要介绍总变化因子系统, 第 2 节介绍 G-PLDA 及其联合估计法, 第 3 节为实验及结果, 第 4 节给出结论.

1 总变化因子系统

1.1 总变化因子

总变化因子技术旨在对说话人和信道同时建模, 在 GMM 超向量空间不区分说话人和信道. 给定一段语音, 与说话人及信道相关的 GMM 超向量 $\boldsymbol{s}$ 由下式表示:

$$\boldsymbol{s}=\boldsymbol{m}+T\boldsymbol{w}+\boldsymbol{\epsilon} \tag{1}$$

$\boldsymbol{m}$ 为通用背景模型的超矢量, T 矩阵表示总变化空间, $\boldsymbol{w}$ 矢量是与说话人及信道相关的总变化因子, 也就是最终得到的低维矢量总变化因子, $\boldsymbol{\epsilon}$ 表示残差, 和 $\boldsymbol{w}$ 均服从高斯分布.

$$\begin{aligned}\boldsymbol{w}&\sim \mathrm{N}(\boldsymbol{0},I)\\ \boldsymbol{\epsilon}&\sim \mathrm{N}(\boldsymbol{0},\Sigma)\end{aligned} \tag{2}$$

其中, Σ 为对角协方差阵. T 矩阵训练及 $\boldsymbol{w}$ 的计算参见文献 [3, 7].

1.2 余弦打分

文献 [18−19] 中的实验表明, 用支持向量机 (Support vector machine, SVM) 在因子空间进行分类, 余弦核可以达到很好的性能. 本文基线实验采用类似的方法. $\boldsymbol{w}_1$ 和 $\boldsymbol{w}_2$ 两个总变化因子之间的距离为

$$k(\boldsymbol{w}_1,\boldsymbol{w}_2)=\frac{\langle \boldsymbol{w}_1,\boldsymbol{w}_2\rangle}{\|\boldsymbol{w}_1\|\cdot\|\boldsymbol{w}_2\|} \tag{3}$$

这个值和门限 θ 进行比较, 然后做出最终决定. 这种方法的好处是, 不像 SVM 或者经典的因子分析系统, 它不需要说话人的注册数据. 在降低复杂度的同时, 大大提高了计算速度.

1.3 方差谱规整

方差谱规整最初由 Bousquet 等[12] 提出, 用于对总变化因子预处理. 在该方法下, PLDA 的性能获得显著提升. 本文实验显示, 规整后总变化因子更加具备鉴别能力, 同样显著提升了基线系统性能. 其训练过程如下:

1) For $i = 1$ to 总迭代次数;

2) 计算训练集里所有总变化因子的均值 $\boldsymbol{\mu}_i$ 和类内协方差矩阵 W_i;

3) 对于集内的每一个总变化因子做变换

$$\boldsymbol{w} := \frac{W_i^{-\frac{1}{2}}(\boldsymbol{w} - \boldsymbol{\mu}_i)}{\|W_i^{-\frac{1}{2}}(\boldsymbol{w} - \boldsymbol{\mu}_i)\|} \tag{4}$$

测试时, 对每一个总变化因子使用式 (4), 其中 $\boldsymbol{\mu}_i$ 和 W_i 对应于训练时候的值.

2 高斯 PLDA

2.1 原始模型

G-PLDA 与 JFA 类似, 前者用于总变化因子, 后者用于 GMM 超向量. 假定第 i 个人的第 j 个总变化因子表示为 $\boldsymbol{m}_{ij}$, 那么 G-PLDA 模型可表示成:

$$\boldsymbol{m}_{ij} = \boldsymbol{\mu} + V\boldsymbol{y}_i + U\boldsymbol{x}_{ij} + \boldsymbol{z}_{ij} \tag{5}$$

其中, $\boldsymbol{\mu}$ 为所有训练数据的均值, 矩阵 V 表示说话人子空间 (本征音矩阵), 矩阵 U 表示信道子空间 (本征信道矩阵), 矢量 $\boldsymbol{y}_i$ 和 $\boldsymbol{x}_{ij}$ 为对应的子空间因子, 它们服从标准高斯分布. $\boldsymbol{z}_{ij}$ 表示残差, 由一个对角矩阵 D 表示.

$$\begin{aligned} &\boldsymbol{y}_i \sim \mathrm{N}(\boldsymbol{0}, I) \\ &\boldsymbol{x}_{ij} \sim \mathrm{N}(\boldsymbol{0}, I) \\ &\boldsymbol{z}_{ij} \sim \mathrm{N}(\boldsymbol{0}, D^{-1}) \end{aligned} \tag{6}$$

为了应用 G-PLDA, 必须在已标注数据集上通过最大似然方法估计参数 $\lambda = (\boldsymbol{\mu}, V, U, D)$, 其训练方法参照文献 [9−10].

2.2 G-PLDA 在说话人识别中的应用及联合估计

虽然, G-PLDA 模型可以成功地用于人脸识别[9], 然而, 完全不变地将之用于说话人识别, 其性能并不好. 前文已经提到, G-PLDA 为 JFA 的一个特例, 既然 JFA 无法准确地将说话人和信道信息分开, 那么 G-PLDA 也存在这个问题. 仿照文献 [11−12], 本文移除本征信道部分, 即模型为

$$\boldsymbol{m}_{ij} = \boldsymbol{\mu} + V\boldsymbol{y}_i + \boldsymbol{z}_{ij} \tag{7}$$

同时, 假设 D^{-1} 为全角协方差矩阵. 因此, 残差可以包含更多的信息, 以此弥补去掉本征信道所带来的损失.

2.2.1 模型训练

传统的因子分析方法在一开始就估算好均值 $\boldsymbol{\mu}$, 在 EM 更新过程中只更新 V 和 D. 但是, 在更新后的 V 空间, 其均值也理应做出相应的调整. 否则, 在不断更新的过程中, 会导致模型参数之间无法很好地耦合. 本文提出联合估计法, 通过拼接均值 $\boldsymbol{\mu}$ 和载荷矩阵 V, 进行 EM 推导, 给出更为严格的更新公式. 令 M_i 表示第 i 个说话人的总变化因子集合, $M_i = \{\boldsymbol{m}_{ij}, j = 1, 2, \cdots, n_i\}$, n_i 代表第 i 个说话人的语音数目, 总变化因子的维数为 p, 因子 $\boldsymbol{y}_i$ 的维数为 q, 所有训练数据的数目总和为 N.

为了联合估计均值 $\boldsymbol{\mu}$ 和因子载荷矩阵 V, 首先, 定义因子矢量 $\tilde{\boldsymbol{y}}_i = [\boldsymbol{y}_i^{\mathrm{T}} \quad 1]^{\mathrm{T}}$ 以及因子载荷矩阵 $\tilde{V} = [V \quad \boldsymbol{\mu}]$, 则 G-PLDA 模型可表示为

$$\boldsymbol{m}_{ij} = \tilde{V}\tilde{\boldsymbol{y}}_i + \boldsymbol{z}_{ij} \tag{8}$$

给定模型及说话人的隐藏因子, 其条件概率为

$$\begin{aligned} p(M_i|\tilde{\boldsymbol{y}}_i, \lambda) = &\\ &\prod_{j=1}^{n_i} \frac{1}{(2\pi)^{p/2}|D^{-1}|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{m}_{ij} - \tilde{V}\tilde{\boldsymbol{y}}_i)^{\mathrm{T}} \times \\ &D(\boldsymbol{m}_{ij} - \tilde{V}\tilde{\boldsymbol{y}}_i)\} = \\ &(2\pi)^{-\frac{n_i p}{2}}|D|^{\frac{n_i}{2}} \exp\{-\frac{n_i}{2}\tilde{\boldsymbol{y}}_i^{\mathrm{T}}\tilde{V}^{\mathrm{T}}D\tilde{V}\tilde{\boldsymbol{y}}_i - \\ &\frac{1}{2}\sum_{j=1}^{n_i}\boldsymbol{m}_{ij}^{\mathrm{T}}D\boldsymbol{m}_{ij} + \sum_{j=1}^{n_i}\boldsymbol{m}_{ij}^{\mathrm{T}}D\tilde{V}\tilde{\boldsymbol{y}}_i\} \end{aligned} \tag{9}$$

隐藏因子 $\tilde{\boldsymbol{y}}_i$ 的似然为

$$\begin{aligned} p(\tilde{\boldsymbol{y}}_i|\lambda) = p(\boldsymbol{y}_i|\lambda) = &\\ &(2\pi)^{-\frac{q}{2}} \exp\{-\frac{1}{2}\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{y}_i\} = \\ &(2\pi)^{-\frac{q}{2}} \exp\{-\frac{1}{2}\tilde{\boldsymbol{y}}_i^{\mathrm{T}}\tilde{\boldsymbol{y}}_i - \frac{1}{2}\} \end{aligned} \tag{10}$$

给定第 i 个说话人的语音, 其隐藏因子的后验为

$$\begin{aligned} p(\tilde{\boldsymbol{y}}_i|M_i, \lambda) \propto p(M_i, \tilde{\boldsymbol{y}}_i|\lambda) = &\\ p(M_i|\tilde{\boldsymbol{y}}_i, \lambda)p(\tilde{\boldsymbol{y}}_i|\lambda) & \end{aligned} \tag{11}$$

将式 (9) 和式 (10) 代入式 (11), 并写成 $\tilde{\boldsymbol{y}}_i$ 的一次项和二次项的形式, 有:

$$p(\tilde{\boldsymbol{y}}_i|M_i,\lambda)= C\exp\{-\frac{1}{2}\tilde{\boldsymbol{y}}_i^{\mathrm{T}}(n_i\widetilde{V}^{\mathrm{T}}D\widetilde{V}+I)\tilde{\boldsymbol{y}}_i+ \sum_{j=1}^{n_i}\boldsymbol{m}_{ij}^{\mathrm{T}}D\widetilde{V}\tilde{\boldsymbol{y}}_i\}+const \tag{12}$$

C、$const$ 为常数, 在 EM 算法中的 E-Step 需要对隐藏因子求期望, 由于 $\tilde{\boldsymbol{y}}_i$ 为二次型, 可知服从高斯分布, 于是有:

$$\begin{aligned}&\mathrm{E}[\tilde{\boldsymbol{y}}_i]=J_i^{-1}K_i^{\mathrm{T}}\\&\mathrm{E}[\tilde{\boldsymbol{y}}_i\tilde{\boldsymbol{y}}_i^{\mathrm{T}}]=J_i^{-1}+\mathrm{E}[\tilde{\boldsymbol{y}}_i]\mathrm{E}[\tilde{\boldsymbol{y}}_i]^{\mathrm{T}}\end{aligned} \tag{13}$$

其中,

$$\begin{aligned}J_i&=n_i\widetilde{V}^{\mathrm{T}}D\widetilde{V}+I\\K_i&=\sum_{j=1}^{n_i}\boldsymbol{m}_{ij}^{\mathrm{T}}D\widetilde{V}\end{aligned} \tag{14}$$

在 M-step 中, 根据上面 E-step 求得的期望来更新模型参数, 为此, 需要先求出 Q 函数[14]:

$$\begin{aligned}Q=&\left\langle\sum_i\ln p(M_i|\tilde{\boldsymbol{y}}_i,\lambda)\right\rangle=\\&\left\langle-\frac{Np}{2}\ln(2\pi)+\frac{N}{2}\ln|D|-\frac{1}{2}\sum_{ij}\mathrm{tr}(\boldsymbol{m}_{ij}\boldsymbol{m}_{ij}^{\mathrm{T}}D)-\right.\\&\left.\frac{1}{2}\sum_i\mathrm{tr}(n_i\tilde{\boldsymbol{y}}_i\tilde{\boldsymbol{y}}_i^{\mathrm{T}}\widetilde{V}^{\mathrm{T}}D\widetilde{V})+\sum_{ij}\mathrm{tr}(\tilde{\boldsymbol{y}}_i\boldsymbol{m}_{ij}^{\mathrm{T}}D\widetilde{V})\right\rangle=\\&\frac{N}{2}\ln|D|-\frac{1}{2}\mathrm{tr}(SD)+\mathrm{tr}(TD\widetilde{V})-\\&\frac{1}{2}\mathrm{tr}(R\widetilde{V}^{\mathrm{T}}D\widetilde{V})+const\end{aligned} \tag{15}$$

其中,

$$\begin{aligned}S&=\sum_{ij}\boldsymbol{m}_{ij}\boldsymbol{m}_{ij}^{\mathrm{T}}\\T&=\sum_{ij}\mathrm{E}[\tilde{\boldsymbol{y}}_i]\boldsymbol{m}_{ij}^{\mathrm{T}}\\R&=\sum_{ij}\mathrm{E}[\tilde{\boldsymbol{y}}_i\tilde{\boldsymbol{y}}_i^{\mathrm{T}}]\end{aligned} \tag{16}$$

为了使 Q 函数最大, 分别使 Q 函数对 $\widetilde{V}$ 和 D 求导. 对 $\widetilde{V}$ 求导:

$$\frac{\partial Q}{\partial\widetilde{V}}=D^{\mathrm{T}}T^{\mathrm{T}}-\frac{1}{2}(D\widetilde{V}R+D^{\mathrm{T}}\widetilde{V}R^{\mathrm{T}}) \tag{17}$$

令其等于 0, 得到 $\boldsymbol{\mu}$ 和 V 的联合更新公式:

$$\widetilde{V}=T^{\mathrm{T}}R^{-1} \tag{18}$$

对 D 求导:

$$\frac{\partial Q}{\partial D}=\frac{N}{2}D^{-1}-\frac{1}{2}S^{\mathrm{T}}+T^{\mathrm{T}}\widetilde{V}^{\mathrm{T}}-\frac{1}{2}\widetilde{V}R^{\mathrm{T}}\widetilde{V}^{\mathrm{T}} \tag{19}$$

令其等于 0, 得到残差的更新公式:

$$\begin{aligned}D^{-1}=&\frac{1}{N}(S-2T^{\mathrm{T}}\widetilde{V}^{\mathrm{T}}+\widetilde{V}R^{\mathrm{T}}\widetilde{V}^{\mathrm{T}})=\\&\frac{1}{N}(S-T^{\mathrm{T}}\widetilde{V}^{\mathrm{T}})\end{aligned} \tag{20}$$

至此, EM 中 M-step 所对应的更新式 (18) 和式 (20) 已经全部得到. 可以看出, 更新 $\widetilde{V}$ 意味着同时更新了均值 $\boldsymbol{\mu}$ 和载荷矩阵 V.

2.2.2　打分

估计好模型参数后, 给定两个总变化因子 $\boldsymbol{w}_1$ 和 $\boldsymbol{w}_2$, 其对数似然比由式 (21) 计算, 其中，假设 θ_{tar} 表示 $\boldsymbol{w}_1$ 和 $\boldsymbol{w}_2$ 来自同一个说话人, θ_{non} 表示来自不同说话人.

$$\begin{aligned}score=&\ln\frac{p(\boldsymbol{w}_1,\boldsymbol{w}_2|\theta_{tar})}{p(\boldsymbol{w}_1,\boldsymbol{w}_2|\theta_{non})}=\\&\ln\mathrm{N}\left(\begin{bmatrix}\boldsymbol{w}_1\\\boldsymbol{w}_2\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}\\\boldsymbol{\mu}\end{bmatrix},\begin{bmatrix}\Sigma_{tot}&\Sigma_{ac}\\\Sigma_{ac}&\Sigma_{tot}\end{bmatrix}\right)-\\&\ln\mathrm{N}\left(\begin{bmatrix}\boldsymbol{w}_1\\\boldsymbol{w}_2\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}\\\boldsymbol{\mu}\end{bmatrix},\begin{bmatrix}\Sigma_{tot}&0\\0&\Sigma_{tot}\end{bmatrix}\right)\end{aligned} \tag{21}$$

其中, $\Sigma_{tot}=VV^{\mathrm{T}}+D^{-1}$, $\Sigma_{ac}=VV^{\mathrm{T}}$. 因为 $\boldsymbol{\mu}$ 代表所有总变化因子的均值, 在测试时, 可以事先移除, 那么令上式的 $\boldsymbol{\mu}=\boldsymbol{0}$, 通过 2×2 分块矩阵求逆[14], 上式可化简为:

$$score=\boldsymbol{w}_1^{\mathrm{T}}Q\boldsymbol{w}_1+\boldsymbol{w}_2^{\mathrm{T}}Q\boldsymbol{w}_2+2\boldsymbol{w}_1^{\mathrm{T}}P\boldsymbol{w}_2+const \tag{22}$$

其中,

$$\begin{aligned}Q&=\Sigma_{tot}^{-1}-(\Sigma_{tot}-\Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1}\\P&=\Sigma_{tot}^{-1}\Sigma_{ac}(\Sigma_{tot}-\Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1}\end{aligned} \tag{23}$$

化简后, 打分复杂度降低, 大大提高了测试速度. 本文实验在美国国家标准与技术局（National Institute of Standards and Technology, NIST) 说话人识别评测（Speaker recognition evaluation, SRE）2010 扩展测试集上, 完成全部测试只需要几分钟. 通过对 P 求本征值和对角化, 速度还可以进一步提高[11].

2.2.3 长度规整

通过对总变化因子进行长度规整, 可以有效提升 G-PLDA 性能. 具体做法是: 对总变化因子每一维同等缩放, 使其模为单位长度.

$$\boldsymbol{w} := \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \tag{24}$$

3 实验及结果

本文采用等错率[20] 评价说话人识别系统.

3.1 前端处理

实验中所使用的特征为美尔倒谱系数 (Mel frequency cepstral coefficient, MFCC) 60 维特征, 其基本特征由 19 维的基本倒谱系数和一维能量构成; 然后, 对基本特征做一阶差分和二阶差分. 在提取特征前, 采用布尔诺科大学 (BUT) Hungarian phoneme recognizer[21] 对数据进行语音/静音切割, 对于麦克风数据, 采用 NIST 提供的语音识别结果处理. 首先, 抛弃小于 10 s 的语音. 然后, 按窗长 25 ms 和窗移 10 ms 提取 60 维的 MFCC 特征; 最后, 使用倒谱均值减及倒谱方差规整技术对特征进行规整.

3.2 训练数据选取

本文使用 NIST SRE 2004、2005、2006 电话语音训练两个性别相关的 GMM-UBM, 高斯数为 2 048, 其中, 男性 (Male) 8 316 句话, 女性 (Female) 11 445 句话. 使用 NIST SRE 2004、2005、2006、Switchboard I、Switchboard II Phase 1&2&3、Switchboard Cellular Parts 1&2、Fisher 的电话语音数据, 来训练总变化矩阵 T, 其中, Male 31 170 句话, Female 38 969 句话, 采用随机初始矩阵, 迭代 7 次, 最终得到 600 维的总变化因子. 方差谱规整训练所用数据和 PLDA 训练数据相同, 采用性别相关数据训练模型, 数据来自 NIST SRE 2004、2005、2006、2008、Switchboard II Phase 2. 其中, Male 18 871 句话, Female 17 308 句话. PLDA 训练采用随机矩阵, 在本文实验中, 迭代 5 次即可达到最优.

3.3 实验结果

重估均值势必会带来计算性能的降低, 但是, 在某种程度上可以忽略. 在本文实验中, 在 Male 上, G-PLDA 的总训练时间为 107 s, 在联合高斯概率线性鉴别分析 (Joint Gussian PLDA, JG-PLDA) 上为 126 s, 降低 17.8 %; 在 Female 上, G-PLDA 的总训练时间为 128 s, 在 JG-PLDA 上为 200 s, 降低 56.3 %. 虽然下降许多, 但是, 总训练时间本身并不长.

表 1 给出了 G-PLDA 和 JG-PLDA 在 NIST SRE 2010 扩展测试集上的性能 (括号中的数值表示性能相对提升或下降的百分值). 从表 1 可以看出, 除了在 3 个条件下, JG-PLDA 相对 G-PLDA 性能有所下降外, 在其他 15 个条件下均有提升. 其中, 有 10 个测试条件达到了 10 % 左右的提升.

表 2 列出了基于总变化因子的各个系统在 NIST SRE 2010 "电话-电话" 下的实验结果. 方差谱规整通过学习类内与类间信息, 这一过程加入了鉴别信息, 即便使用简单的余弦打分, 等错率也能得到 50 % 左右的提升. 不但如此, 方差谱规整后, 总变化因子分布与 G-PLDA 模型假设更加吻合, 这在文献 [12] 和表 2 的实验中, 都得到了验证. 方差谱规整后, 虽然 G-PLDA 可以达到较好性能, 但本文提出的 JG-PLDA 仍有 5.19 % 的提升.

表 3 给出了 G-PLDA 和 JG-PLDA 在 NIST SRE 2010 扩展测试集所有条件下, 采用 SphNorm 规整技术的对比结果. 与 G-PLDA 相比, JG-PLDA 除了在 Male 条件 6 下性能有所下降和在 Female 条件 8 和条件 9 下不变外, 在其他条件下, 均有一定的提升. 将所有相对变化求平均, JG-PLDA 相对 G-PLDA 有 5.85 % 的提升.

表 4 给出了 G-PLDA 和 JG-PLDA 在 NIST SRE 2012 核心测试集所有条件下的对比结果. 为了进一步验证本文联合估计方法相对传统方法有显著提升, 本文作者整理两个系统的结果, 分析识别结果的差异性, 并给出 McNemar 统计量[22]. 结果表明, 在该测试集上, JG-PLDA 均优于 G-PLDA, 平均提升 4.72 %. 虽然性能提升幅度不大, 但从 McNemar 统计量结果看, 本文方法相对传统方法有显著提升.

表 1 G-PLDA 和 JG-PLDA 在 NIST SRE 2010 扩展测试集上的性能

Table 1 Performance of G-PLDA and JG-PLDA on NIST SRE 2010 extend test corpus

模型估计方法	条件 1	条件 2	条件 3	条件 4	条件 5	条件 6	条件 7	条件 8	条件 9
G-PLDA/Male	2.12	4.66	4.11	3.39	2.83	3.88	3.35	1.25	1.81
JG-PLDA/Male	1.92(↑9.43)	4.27(↑8.37)	3.65(↑11.2)	3.23(↑4.72)	2.39(↑15.6)	3.48(↑10.3)	3.78(↓12.8)	1.33(↓6.40)	1.74(↑3.87)
G-PLDA/Female	4.56	8.78	5.07	7.33	2.90	4.86	11.1	1.82	2.31
JG-PLDA/Female	3.77(↑17.3)	7.94(↑9.57)	4.81(↑5.13)	6.27(↑14.5)	2.62(↑9.66)	4.84(↑0.41)	10.53(↑5.14)	1.84(↓1.10)	1.74(↑24.7)

表 2 余弦打分、G-PLDA、JG-PLDA 系统性能比较

Table 2 Comparison of performance among Cosine, G-PLDA and JG-PLDA

模型估计方法	Cosine	Cosine+SphNorm	G-PLDA	G-PLDA+LNorm	G-PLDA+SphNorm	JG-PLDA+SphNorm
Male EER (%)	5.57	2.76	2.83	1.47	1.35	1.28
Female EER (%)	7.83	3.93	2.90	2.48	2.31	2.19

表 3 方差谱规整后 G-PLDA 和 JG-PLDA 在 NIST SRE 2010 扩展测试集上的性能

Table 3 Performance of G-PLDA and JG-PLDA after SphNorm on NIST SRE 2010 extend test corpus

模型估计方法	条件 1	条件 2	条件 3	条件 4	条件 5	条件 6	条件 7	条件 8	条件 9
G-PLDA/Male	2.33	5.20	1.84	3.69	1.35	2.62	3.91	1.05	1.86
JG-PLDA/Male	2.14 (↑8.15)	4.99 (↑4.04)	1.67 (↑9.24)	3.29 (↑10.8)	1.28 (↑5.19)	2.70 (↓3.05)	3.37 (↑13.8)	0.90 (↑14.3)	1.70 (↑8.60)
G-PLDA/Female	5.16	9.54	4.14	7.75	2.31	4.49	11.68	1.64	2.31
JG-PLDA/Female	4.64 (↑10.1)	8.88 (↑6.92)	4.13 (↑0.24)	7.52 (↑2.97)	2.19 (↑5.19)	4.31 (↑4.01)	11.1 (↑4.79)	1.64 (↑0.00)	2.31 (↑0.00)

表 4 G-PLDA 和 JG-PLDA 在 NIST SRE 2012 核心测试集上的性能

Table 4 Performance of G-PLDA and JG-PLDA on NIST SRE 2012 core test

模型估计方法	条件 1	条件 2	条件 3	条件 4	条件 5
G-PLDA	7.04	2.92	7.79	5.39	3.53
JG-PLDA	6.77(↑3.84)	2.78(↑4.79)	7.45(↑4.36)	4.97(↑7.79)	3.43(↑2.83)
McNemar 统计量	87.41	337.33	413.00	915.00	162.60

4 结论

本文介绍了总变化因子技术、总变化因子规整技术以及总变化因子基线系统的构建.

本文提出联合估计法, 对 G-PLDA 的均值和载荷空间同时更新, 使得两者可以更好地耦合. 虽然, 联合估计法导致了计算性能下降, 但是其总的训练时间很短, 可以忽略. 在 NIST SRE 2010 扩展测试集上, 不对总变化因子做任何规整技术, JG-PLDA 相对 G-PLDA, 平均提升 7.20 %; 采用 SphNorm 后, JG-PLDA 相对 G-PLDA, 平均提升 5.85 %. 在 NIST SRE 2012 核心测试集上, 等错率及 McNemar 统计量同时表明, JG-PLDA 相对 G-PLDA, 有显著提升.

References

1 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

2 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)

3 Kenny P, Boulianne G, Dumouchel P. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech Audio Processing*, 2005, **13**(3): 345−359

4 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

5 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)

6 Dehak N. Discriminative and generative approches for long- and short-term speaker characteristics modeling: Application to speaker verification [Ph. D. dissertation], *École* de Technologie Supérieure, Montreal, QC, Canada, 2009

7 Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2011, **19**(4): 788−798

8 McLaren M, Leeuwen D A V. Sourcenormalised and weighted lda for robust speaker recognition using i-vectors. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Prague, Czech Republic: IEEE, 2011. 5456−5459

9 Simon J D P, James H E. Probabilistic linear discriminant analysis for inferences about identity. In: Proceedings of International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE, 2007. 1−8

10 Dehak N, Karam Z, Reynolds D, Dehak R, Campbell W, Glass J. A channel-blind system for speaker verification. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Prague, Czech Republic: IEEE, 2011. 4536−4539

11 Garcia Romero D, Espy Wilson C. Analysis of i-vector length normalization in speaker recognition systems. In:

Proceedings of International Conference on Speech Communication and Technology. Florence, Italy: IEEE, 2011. 249−252

12 Bousquet P M, Larcher A, Matrouf D, Bonastre J F, O Plchot. Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis. In: Proceedings of Odyssey Speaker and Language Recognition Workshop. Biopolis, Singapore: 2012. 157−164

13 Brummer N, de Villiers E. The speaker partitioning problem. In: Proceedings of Odyssey Speaker and Language Recognition Workshop. Brno, Czech Republic: 2010. 194−201

14 Christopher M Bishop. *Pattern Recognition and Machine Learning*. Singapore: Springer, 2006. 439−441

15 Villalba J, Brümmer N. Towards fully Bayesian speaker recognition: integrating out the between speaker covariance. In: Proceedings of International Conference on Speech Communication and Technology. Florence, Italy: IEEE, 2011. 505−508

16 Kenny P. Bayesian speaker verification with heavy-tailed priors. In: Proceedings of Odyssey Speaker and Language Recognition Workshop. Brno, Czech Republic: 2010.

17 Yang Hai, Liang Chun-Yan, Xu Yun-Fei, Yang Lin, Yan Yong-Hong. Sparse probabilistic linear disciminant analysis for speaker verification. In: Proceedings of International Conference on Speech Communication and Technology. Portland, Oregon: IEEE, 2012.

18 Dehak N, Dehak R, Kenny P, Brummer N, Ouellet P, Dumouchel P. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Proceedings of International Conference on Speech Communication and Technology. Brighton, UK: IEEE, 2009. 1559−1562

19 Dehak N, Kenny P, Dehak R, Glembek O, Dumouchel P, Burget L, Hubeika V. Support vector machines and joint factor analysis for speaker verification. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Taipei, China: IEEE, 2009. 4237−4240

20 NIST speaker recognition evaluation [Online], available: http://www.itl.nist.gov/iad/mig/tests/spk/2010/index.html, April 21, 2010. 325−328

21 Schwarz P, Pavel M, Cernocky J. Hierarchical structures of neural networks for phoneme recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Toulouse, France: IEEE, 2006.

22 McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 1947, **12**(2): 153−157

许云飞 中国科学院声学研究所博士研究生. 2010 年获南开大学电子科学与技术学士学位. 主要研究方向为语音信号处理, 说话人识别及机器学习.
E-mail: xuyunfei@hccl.ioa.ac.cn
(**XU Yun-Fei** Ph. D. candidate at the Institute of Acoustics, Chinese Academy of Sciences. He received his bachelor degree from Nankai University in 2010. His research interest covers speech signal processing, speaker recognition and machine learning.)

杨 海 2008 年获西安交通大学计算机学士学位. 2013 年获中国科学院信号与信息处理博士学位. 主要研究方向为语音信号处理及鲁棒性说话人识别.
E-mail: yanghai@hccl.ioa.ac.cn
(**YANG Hai** Received his bachelor degree from Xi′an Jiaotong University in 2008 and Ph. D. degree from Chinese Academy of Sciences in 2013. His research interest covers speech signal processing, noise robust speaker recognition.)

周若华 中国科学院声学研究所语言声学与内容理解重点实验室研究员. 1994 年获北京理工大学学士学位, 1997 年获中国科学院微电子研究所硕士学位, 2006 年获瑞士联邦理工学院博士学位. 主要研究方向为音乐信号处理, 语音信号处理. 本文通信作者.
E-mail: zhouruohua@hccl.ioa.ac.cn
(**ZHOU Ruo-Hua** Professor at the Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences. He received his bachelor degree from Beijing Institute of Technology in 1994 and the master degree of engineering in microelectronics and semiconductor devices from Microelectronics R&D Center, Chinese Academy of Sciences in 1997, and the Ph. D. degree from Swiss Federal Institute of Technology, Switzerland in 2006. His research interest covers music signal processing, and speech signal processing. Corresponding author of this paper.)

颜永红 中科院声学所语言声学与内容理解重点实验室教授. 1990 年获清华大学学士学位, 1995 年获美国俄勒冈研究院获计算机科学和工程博士学位. 主要研究方向为语音信号处理及识别, 语种/说话人识别, 人机交互.
E-mail: yanyonghong@hccl.ioa.ac.cn
(**YAN Yong-Hong** Professor of Key Laboratory of Speech Acoustics and Content Understanding at IOACAS. He received his bachelor degree from Tsinghua University in 1990 and Ph. D. degree from Oregon Graduate Institute in 1995. His research interest covers speech signal processing and recognition, and language/speaker recognition and human computer interface.)