基于通用背景–联合估计 (UB-JE) 的说话人识别方法

汪海彬 [1] 郭剑毅 [1,2] 毛存礼 [1,2] 余正涛 [1,2]

摘 要 在说话人识别中, 有效的识别方法是核心. 近年来, 基于总变化因子分析 (i-vector) 方法成为了说话人识别领域的主流, 其中总变化因子空间的估计是整个算法的关键. 本文结合常规的因子分析方法提出一种新的总变化因子空间估计算法, 即通用背景–联合估计 (Universal background-joint estimation algorithm, UB-JE) 算法. 首先, 根据高斯混合–通用背景模型 (Gaussian mixture model-universal background model, GMM-UBM) 思想提出总变化矩阵通用背景 (UB) 算法; 其次, 根据因子分析理论结合相关文献提出了一种总变化矩阵联合估计 (JE) 算法; 最后, 将两种算法相结合得到通用背景–联合估计 (UB-JE) 算法. 采用 TIMIT 和 MDSVC 语音数据库, 结合 i-vector 方法将所提的算法与传统算法进行对比实验. 结果显示, 等错误率 (Equal error rate, EER) 和最小检测代价函数 (Minimum detection cost function, MinDCF) 分别提升了 8.3 % 与 6.9 %, 所提方法能够提升 i-vector 方法的性能.

关键词 总变化因子分析, 总变化因子空间, 通用背景–联合估计算法, 说话人识别

引用格式 汪海彬, 郭剑毅, 毛存礼, 余正涛. 基于通用背景–联合估计 (UB-JE) 的说话人识别方法. 自动化学报, 2018, **44**(10): 1888−1895

DOI 10.16383/j.aas.2017.c170051

Speaker Recognition Based on Universal Background-Joint Estimation (UB-JE)

WANG Hai-Bin[1] GUO Jian-Yi[1,2] MAO Cun-Li[1,2] YU Zheng-Tao[1,2]

Abstract In the speaker recognition, the effective identification method is the core. In recent years, i-vector method has become the mainstream in the field of speaker recognition, and estimation of the total variation factor space is the key of whole algorithm. In this paper, we propose a new algorithm for total variation factor space estimation named UB-JE, which is combined with conventional factor analysis method. Firstly, the universal background algorithm of total variation matrix is proposed according to Gaussian mixture model-universal background model (GMM-UBM). Secondly, the joint estimation algorithm of total variation matrix is proposed according to the factor analysis theory and related works. Finally, the two algorithms are combined to get the universal background-joint estimation algorithm (UB-JE). TIMIT and MDSVC corpus are adopted in the experiment to compare the proposed algorithm with the traditional algorithm. Experimental results show that the equal error rate (EER) and the minimum detection cost function (MinDCF) are improved by 8.3 % and 6.9 %, respectively. The proposed method can improve the performance of i-vector method.

Key words I-vector, total variation factor space, universal background-joint estimation algorithm (UB-JE), speaker recognition

Citation Wang Hai-Bin, Guo Jian-Yi, Mao Cun-Li, Yu Zheng-Tao. Speaker recognition based on universal background-joint estimation (UB-JE). *Acta Automatica Sinica*, 2018, **44**(10): 1888−1895

语音是人们用来交流和沟通的最自然、最直接的方式之一, 因此, 语音是一种重要的生物特征. 作为一种重要的身份鉴定技术, 目前说话人识别[1−2]已广泛运用于国家安全、司法鉴定、电话银行及门禁安全等领域. 与此同时, 说话人识别仍有许多问题需要解决, 例如信道多样化的识别、噪声对识别性能的影响等, 这就涉及到对说话人识别算法的研究.

2000 年左右, Reynolds 等[3] 提出的高斯混合模型–通用背景模型 (Gaussian mixture model-universal background model, GMM-UBM), 以其特有的良好性能和灵活的模型结构, 降低了说话人模型对训练集的依赖, 迅速成为当时说话人识别领域的主流方法之一, 推动了整个领域的发展[4−5]. 由 GMM-UBM 的思想可知, 在高斯混合函数的均值超向量 (Gaussian mixture model supervector, GSV) 中包含有说话人语句的所有信息. 根据该思想, Kenny 等[6−7] 提出了联合因子分析方法 (Joint factor analysis, JFA), 认为说话人语句中包含说话人信息和信道信息两部分, 因此, GSV 又可被分解

收稿日期 2017-01-20 录用日期 2017-08-08
Manuscript received January 20, 2017; accepted August 8, 2017
国家自然科学基金 (61262041, 61472168, 61562052) 资助
Supported by National Natural Science Foundation of China (61262041, 61472168, 61562052)
本文责任编委 吴玺宏
Recommended by Associate Editor WU Xi-Hong
1. 昆明理工大学信息工程与自动化学院 昆明 650500 2. 昆明理工大学智能信息处理重点实验室 昆明 650500
1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500 2. Intelligent Information Processing Key Laboratory, Kunming University of Science and Technology, Kunming 650500

为说话人和信道两部分. Dehak[8] 研究发现, 在对 JFA 进行信道补偿时, 信道空间存在掩盖和重叠问题, 信道空间中不可避免地包含了一部分说话人的信息, 即不能准确地对说话人与信道分别建模. 在此基础上, Dehak 等[9−11] 提出了 i-vector 方法, 该方法认为对 GSV 进行处理时不应该区分说话人和信道, 而应该把它们看成一个整体, 即总变化空间. 但是, 在总变化空间中存在信道失配问题, Dehak 等[9] 又提出了一些信道补偿技术: 线性鉴别分析 (Linear discriminant analysis, LDA) 和类内协方差规整 (Within class covariance normalization, WCCN) 等. 近几年来, 基于 i-vector 方法的说话人识别模型 (图 1) 明显提升了说话人识别系统的性能, 是目前说话人识别领域中最热门的建模方法之一[12−13]. 在美国国家标准技术局组织的说话人评测 (The National Institute of Standards and Technology speaker recognition evaluation, NIST SRE) 中, 该方法的性能明显优于 GMM-UBM[3] 和 GSV-SVM (Gaussian mixture model supervector-support vector machine)[14−15] 等方法, 是处于国际研究前沿的一种说话人识别方法.

i-vector 是一种有效的因子分析方法, 其中总变化因子空间的估计是基础和关键. 为了得到性能更好的 i-vector 方法, 本文结合常规的因子分析方法提出了一种新的总变化因子空间估计算法, 即通用背景−联合估计 (Universal background-joint estimation algorithm, UB-JE) 算法. 首先, 针对说话人识别任务中正负样本分布不平衡问题, 本文借鉴 GMM-UBM 的思想, 结合 i-vector 方法, 通过大量的非训练数据来训练形成一个包含大量说话人的通用背景初始总变化空间, 从而提出了总变化矩阵通用背景 (Universal background, UB) 算法; 其次, 在 i-vector 模型中由于均值不能很好地与更新后的总变化因子空间耦合, 我们根据因子分析理论结合文献 [16−17] 提出了一种总变化矩阵联合估计 (Joint estimation, JE) 算法; 最后, 将两种算法相结合得到通用背景−联合估计 (UB-JE) 算法.

本文结构如下: 第 1 节介绍了因子分析方法的理论, 主要是高斯混合模型超向量、联合因子分析方法和总变化因子分析方法; 第 2 节提出通用背景−联合估计总变化矩阵估计算法, 包含两种总变化因子空间估计算法, 即通用背景算法和联合估计算法; 第 3 节是针对提出的三种总变化因子空间估计算法的实验与结果分析; 第 4 节是结论.

1　因子分析方法理论

1.1　高斯混合模型超向量

由于 GMM-UBM[3] 是先用一些无关数据训练一个通用背景模型 (Universal background model, UBM), 然后利用训练数据对该 UBM 进行数据更新, 得到代表单个说话人的高斯混合模型 (Gaussian mixture model, GMM). 按照 GMM-UBM 模型的原理, 说话人所有的语音信息都包含在由说话人 GSV[14−15] 中 (GSV 形成过程见图 2). 一般情况下, 在说话人识别领域里常用的超向量是均值超向量, 因此, 下文中的超向量如果没有特别指明, 均默认为均值超向量.

1.2　联合因子分析方法

随着科技的发展, 语音可以通过多种渠道获取, 在说话人识别任务中相应地产生了信道失配问题. Kenny 等[6−7] 认为一段语音信号中应包含说话人信息和信道信息两部分, 因此对说话人进行识别时, GSV 应该被分解为说话人和信道两部分, 分别对它们建立模型, 然后去除无关信息 (信道模型), 留下有用信息 (说话人模型), 然后再进行估计, 这就是 JFA 的思想.

根据 Kenny 提出的 JFA 思路, 假设有一个混合度为 C 的 GMM-UBM 模型, 训练时的语音特征参数为 F 维, 则形成一个 FC 维的均值超向量, 则

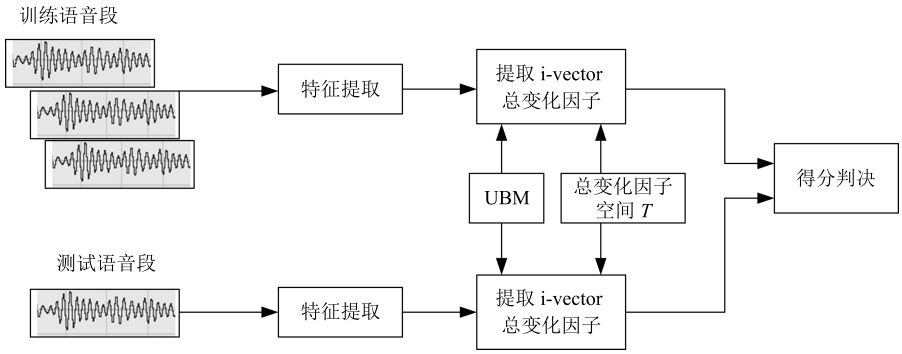


图 1　i-vector 说话人识别系统

Fig. 1　i-vector speaker recognition system

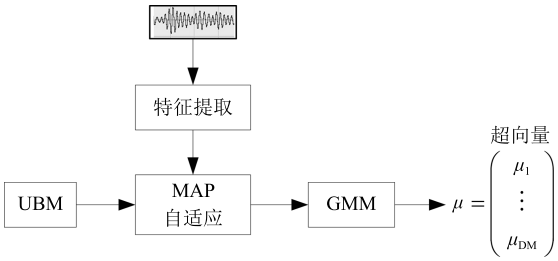


图 2 GMM 均值超向量的形成过程

Fig. 2 The formation process of GMM mean super vector

该超向量可表示为

$$m_{s,h} = m_u + Ux_{s,h} + Vy_s + Dz_s \tag{1}$$

其中, $m_{s,h}$ 为特定说话人 s 的第 h 段语音所形成的 $FC \times 1$ 维的 GSV, m_u 是 $FC \times 1$ 维的 UBM 超矢量, U 表示信道空间, 是一 $FC \times R_u$ 维矩阵, 其中 R_u 为信道因子数, V 表示说话人空间, 是一 $FC \times R_v$ 维矩阵, R_v 是说话人因子数, D 是残差空间, 是一个 $FC \times FC$ 维的对角矩阵, $x_{s,h}$ 表示信道因子, y_s 表示说话人 s 的因子, z_s 是残差因子. 一般来说, $10 < R_u < 200$, $100 < R_v < 400$.

由式 (1) 可知, 在 JFA 中需要对 m_u, U, V 以及 D 进行预先估计, 由于 m_u 已预先得到, 因此只需要估计其他三个矩阵, 即 $\lambda = (U, V, D)$. 关于这三个矩阵的估计参见文献 [16].

1.3 总变化因子分析方法

由于 JFA 存在空间掩盖和空间重叠问题[8], 不能很好地区分说话人和信道. Dehak 等提出了 i-vector[9−10], i-vector 把说话人和信道看成一个整体, 根据 JFA 可把 i-vector 表示为

$$s = m + Tw + \varepsilon \tag{2}$$

其中, s 为特定说话人 GSV, m 为 UBM 超矢量, T 表示总变化因子空间, 迭代更新时, 先随机初始化, w 为总变化因子, 即 i-vector, ε 为残差.

$$\begin{aligned} w &\sim \mathrm{N}(\mathbf{0}, I) \\ \varepsilon &\sim \mathrm{N}(\mathbf{0}, \Sigma) \end{aligned} \tag{3}$$

其中, Σ 为对角协方差矩阵, 可用 UBM 协方差代替.

由式 (2) 可知, i-vector 的建模可简化为对模型参数 $\lambda = (s, m, T, \Sigma)$ 的估计, 由上述理论可知, 训练数据的 s, m 很容易得到, 因此, 可简化为对 $\lambda = (T, \Sigma)$ 的估计. 其中最关键的是对总变化因子空间 T 的估计, T 的估计类似于 JFA 中说话人空间估计, 可以采用最大期望 (EM) 算法得到, 参见文献 [9]. 步骤如下:

步骤 1. 估计统计量. 一段语音特征参数为 $x_{s,t}$, UBM 超矢量为 m, 则

$$N_{c,s} = \sum_t \gamma_{c,s,t} \tag{4}$$

$$F_{c,s} = \sum_t \gamma_{c,s,t}(x_{s,t} - m_c) \tag{5}$$

$$S_{c,s} = \mathrm{diag}\left\{\sum_t \gamma_{c,s,t}(x_{s,t} - m_c)(x_{s,t} - m_c)^{\mathrm{T}}\right\} \tag{6}$$

其中, $N_{c,s}$ 为零阶统计量, $F_{c,s}$ 为一阶统计量, $S_{c,s}$ 为二阶统计量, m_c 为 m 中的第 c 个分量, $\gamma_{c,s,t}$ 为第 c 个高斯密度函数后验概率.

步骤 2. (E 步) 计算总变化因子 w 的一阶统计量和二阶统计量.

$$L_s = I + T^{\mathrm{T}}\Sigma^{-1}N_sT \tag{7}$$

$$\mathrm{E}[w_s] = L_s^{-1}T^{\mathrm{T}}\Sigma^{-1}F_s \tag{8}$$

$$\mathrm{E}\left[w_s w_s^{\mathrm{T}}\right] = \mathrm{E}[w_s]\,\mathrm{E}\left[w_s^{\mathrm{T}}\right] + L_s^{-1} \tag{9}$$

其中, L_s 为临时中间变量, $\mathrm{E}[w_s]$, $\mathrm{E}[w_s w_s^{\mathrm{T}}]$ 为 w 的一阶统计量 (需要的结果) 和二阶统计量, N_s 为 $N_{c,s}$ 的对角拼接 $FC \times FC$ 维矩阵, F_s 为 $F_{c,s}$ 拼接的 FC 维矢量, Σ 为 UBM 协方差.

步骤 3. (M 步) 更新 T 和 Σ.

T 更新:

$$\sum_s N_s T \mathrm{E}\left[w_s w_s^{\mathrm{T}}\right] = \sum_s F_s \mathrm{E}[w_s] \tag{10}$$

Σ 更新:

$$\Sigma = N^{-1}\sum_s S_s - N^{-1}\mathrm{diag}\left\{\sum_s F_s \mathrm{E}\left[w_s^{\mathrm{T}}\right] T^{\mathrm{T}}\right\} \tag{11}$$

其中, S_s 为 $S_{c,s}$ 拼接的 $FC \times FC$ 维矩阵, $N = \sum_c N_s$ 为所有说话人零阶统计量之和. (当反复迭代几次后, 就可得到收敛的 T 和 Σ).

2 总变化因子空间通用背景–联合估计算法

在 i-vector 中 T 的估计是关键和基础, 由上一节可知, T 是通过随机初始化然后通过迭代产生的, 但并没有考虑到通用背景的情况. 本文根据 GMM-UBM 的思想, 先通过背景无关数据产生一个初始化的 T_{ubm}, 然后再进行迭代更新, 提出了一种总变化因子空间通用背景算法. 在常规的 i-vector 算法中, 用均值最大化算法 (Expectation maximum, EM) 对数据更新时, 仅仅考虑到 T 和 Σ, 而没有对 m 进

行更新. 为了使得 i-vector 能够有更好的结合性, 本文又提出了一种同时更新 m 和 T 的联合估计算法. 在本节最后, 我们把上述两种算法相结合, 提出通用背景–联合估计算法.

2.1　通用背景算法 (UB)

首先, 利用大量的无关数据训练一个 UBM 超向量, 并根据 i-vector 中 T 估计方法估计一通用背景变化空间 $T_{\rm ubm}$. 然后, 将 $T_{\rm ubm}$ 作为 EM 算法中对 T 估计的初始矩阵, 进行自适应计算 (如图 3 所示). 具体如下:

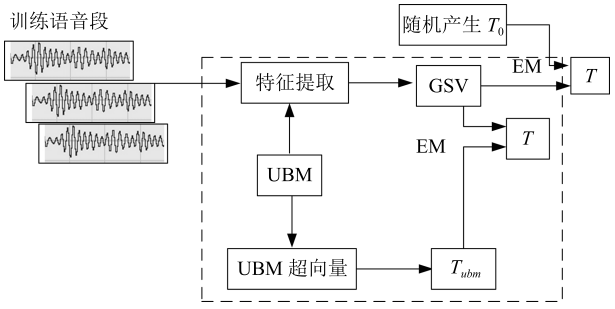


图 3　总变化因子的常规估计算法和 UB 算法 (虚线框) 比较

Fig. 3　Comparison of conventional estimation algorithm of total variation factor with UB (dashed frame)

步骤 1. 通过前端处理将大量的无关数据训练成一个 UBM 超向量, 并结合第 1.3 节中 T 的估计算法, 生成一个通用背景下的总变化空间, 记为 $T_{\rm ubm}$.

步骤 2. 将 $T_{\rm ubm}$ 代入式 (7), 生成 L_s.

步骤 3. 将 L_s 代入式 (8) 和式 (9) 生成 $\mathrm{E}[w_s]$ 和 $\mathrm{E}[w_s w_s^{\rm T}]$.

步骤 4. 结合 $T_{\rm ubm}$, $\mathrm{E}[w_s]$ 以及 $\mathrm{E}[w_s w_s^{\rm T}]$, 根据第 1.3 节中的 EM 算法, 依次对 T 和 Σ 进行更新.

步骤 5. 观察 T 和 Σ 是否收敛或者达到迭代次数. 如果没有, 则返回步骤 2 继续; 否则退出.

2.2　联合估计算法 (JE)

由第 1.3 节可知, i-vector 的建模可转化为对 $\lambda = (s, m, T, \Sigma)$ 的估计. 其中, s 与 m 已预先估计好了, 即 GSV 和 UBM 超矢量, 因此, 只更新 $\lambda = (T, \Sigma)$. 事实上, 在更新 T 和的 Σ 的同时, 也应该更新 m, 即 $\lambda = (m, T, \Sigma)$, 只有这样, 不断更新的参数模型才会更加耦合.

本文提出一种 m, T 联合估计算法, 即

$$T_1 = [T\ m] \tag{12}$$

把 T 和 m 看成一个整体, 在更新 T 的同时, 也更新 m. 此时, i-vector 模型表示为 $\lambda_1 = (T_1, \Sigma)$, 根据式 (2) 可写为

$$s = T_1 w_1 + \varepsilon \tag{13}$$

其中, s 为 GSV, $w_1 = [w^{\rm T}\ \ 1]^{\rm T}$, 称为联合变化因子, T 称为联合变化空间, ε 为残差.

$$\begin{aligned} w_1 &\sim \mathrm{N}(\mathbf{0}, I) \\ \varepsilon &\sim \mathrm{N}(\mathbf{0}, \Sigma) \end{aligned} \tag{14}$$

根据第 1.3 节中 i-vector 中的 EM 更新算法可得:

步骤 1. (E 步)

$$J_s = I + T_1^{\rm T}\Sigma^{-1}N_s T \tag{15}$$

$$\mathrm{E}[w_{1s}] = J_s^{-1}T_1^{\rm T}\Sigma^{-1}F_s \tag{16}$$

$$\mathrm{E}\left[w_{1s}w_{1s}^{\rm T}\right] = \mathrm{E}[w_{1s}]\,\mathrm{E}\left[w_{1s}^{\rm T}\right] + J_s^{-1} \tag{17}$$

步骤 2. (M 步)

对 T_1 更新:

$$\sum_s N_s T_1 \mathrm{E}\left[w_{1s}w_{1s}^{\rm T}\right] = \sum_s F_s \mathrm{E}[w_{1s}] \tag{18}$$

对 Σ 更新:

$$\Sigma = N^{-1}\sum_s S_s - N^{-1}\mathrm{diag}\left\{\sum_s F_s \mathrm{E}\left[w_{1s}^{\rm T}\right] T_1^{\rm T}\right\} \tag{19}$$

对 T_1 的更新就是对 T 和 m 同时更新.

2.3　通用背景–联合估计算法 (UB-JE)

基于上述两种算法, 本文将它们相结合形成互补, 提出了一种新的算法, 即 UB-JE (如图 4 所示). 具体如下:

步骤 1. 通过大量无关数据得到 UBM 超向量, 集合 JE 算法中 T_1 的估计算法, 得到通用背景–联合总变化空间 $T_{\rm 1ubm}$.

步骤 2. 将 $T_{\rm 1ubm}$ 代入式 (15), 生成 J_s.

步骤 3. 将 J_s 代入式 (16) 和式 (17), 生成 $\mathrm{E}[w_{1s}]$ 和 $\mathrm{E}[w_{1s}w_{1s}^{\rm T}]$.

步骤 4. 结合 $T_{\rm 1ubm}$, $\mathrm{E}[w_{1s}]$ 以及 $\mathrm{E}[w_{1s}w_{1s}^{\rm T}]$, 根据 JE 算法中的 EM 算法, 依次对 T_1 和 Σ 进行更新.

步骤 5. 观察 T_1 和 Σ 是否收敛或者达到迭代次数. 如果没有, 则返回步骤 2 继续; 否则退出.

3　实验与分析

3.1　实验设置

实验的测试数据采用 TIMIT 语音库[18]、MDSVC 语音库[19] 以及一组由 MDSVC 语音库组成的长时语音数据. 实验在预处理阶段包括: 有效语音端

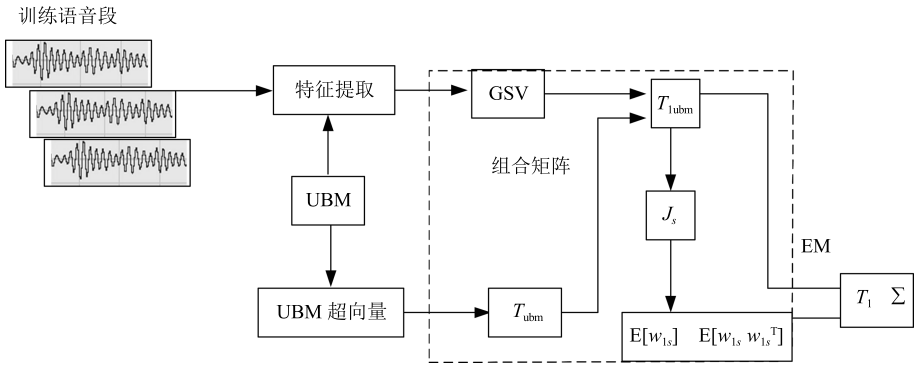


图 4 通用背景−联合估计算法 (虚线框)

Fig. 4 Diagram of universal background-joint estimation algorithm (dashed frame)

点检测 (短时能量与平均过零率相结合的方法)、预加重 (因子为 0.95)、分帧 (帧长 25 ms, 帧移 12.5 ms) 和加窗. 实验采用 39 维美尔倒谱系数 (Mel frequency cepstral coefficients, MFCC) 特征参数 (基本特征包括 1 维能量和 12 维倒谱、13 维一阶差分特征以及 13 维二阶差分特征). 实验中, UBM 混合数为 512, 密度函数方差采用对角矩阵. 在 i-vector 训练中, 总变化因子空间维数设置为 400, 训练时迭代次数取 6 次.

运用 5 300 句 TIMIT 语音 (女性 1 620 句语音, 男性 3 680 句语音)、MDSVC 中 Enroll_Session1 + Enroll_Session2 以及用 HTK 工具[20] 对 MDSVC 语音数据组合长句 (48Enroll_Session1 + 48Enroll_Session2 + 40Imposter 共 136 长句) 分别训练 UBM 模型和 T. 实验训练数据为 100 人 (30 个女性, 70 个男性, 每人 9 句语音) TIMIT 语音、MDSVC 中 Imposter (23 个文件的男性和 17 个文件的女性中各 50 句) 以及 MDSVC 中部分数据 (30 个文件的男性和 30 个文件的女性中各 50 句); 测试数据为 TIMIT 中 100 人 (30 个女性, 70 个男性, 每人 1 句语音)、MDSVC Imposter (23 个文件的男性和 17 个文件的女性中的剩余 4 句) 以及 MDSVC 中部分数据 (30 个文件的男性和 30 个文件的女性中各 4 句) (详见表 1). 本文设置了一个基线实验 (GMM-UBM)[3] 来验证因子分析方法 (i-vector) 的有效性.

表 1 实验所用语音库

Table 1 The corpus used in the experiment

类型	TIMIT		MDSVC		MDSVC 长句	
	male	female	male	female		
UBM	3 860	1 620	2 808	2376	136	
T	3 860	1 620	2 808	2 376	136	
训练 GSV	630	270	1 150	850	1 500	1 500
测试	70	30	92	68	120	120

3.2 评价指标

本文采用等错误率 (Equal error rate, EER) 和 2010 年的 NIST SRE 中的最小检测代价函数 (Minimum detection cost function 2010, MinDCF10)[21] 作为性能评测指标. MinDCF10 与 EER 越小说明系统的性能越好.

检测代价函数计算公式为

$$C_{\rm det} = C_{\rm Miss} \times P_{\rm Miss|Target} \times P_{\rm Target} + C_{\rm FalseAlarm} \times P_{\rm FalseAlarm|NonTarget} \times (1 - P_{\rm Target}) \tag{20}$$

其中, $C_{\rm Miss}$ 和 $C_{\rm FalseAlarm}$ 分别为漏警和虚警的代价; $P_{\rm Miss|Target}$ 和 $P_{\rm FalseAlarm|NonTarget}$ 分别为给定门限 θ 情况下的漏警率和虚警率. $P_{\rm Target}$ 为目标实验的先验概率, 参数设定见表 2. 当 $P_{\rm Miss|Target} = P_{\rm FalseAlarm|NonTarget}$ 时, $EER = C_{\rm det}$.

表 2 MinDCF10 参数设定

Table 2 MinDCF10 parameter setting

$C_{\rm Miss}$	$C_{\rm FalseAlarm}$	$P_{\rm Target}$
1	1	0.001

3.3 实验设计与结果分析

为了验证所提算法的有效性, 本文基于两种不同的语音库设置了四个实验: 实验 1 基于 TIMIT 语音库; 实验 2 基于 MDSVC 语音库; 实验 3 是在两者综合的语音库中完成的; 实验 4 基于 MDSVC 语音库组成的长时语音数据. 每一个实验做 6 次比较试验, 即基线实验 (GMM-UBM)[3]、本文提出的三个新算法、总变化矩阵传统算法以及文献 [22] 中 i-vector 矢量规整 PLDA 技术. 由于不同语音库的录音条件、方式等不同, 四个实验分别代表了四种不同的实验环境. 表 3 ∼ 6 分别给出在不同语音库上

各算法训练 T 后的系统性能比较. 表中括号里的是性能提升值.

表 3　GMM-UBM、传统算法估计 T、本文所提出算法估计 T 以及 PLDA 在 TIMIT 语音库上的性能对比
Table 3　Performance comparison of GMM-UBM, the traditional algorithm to estimate T, the proposed algorithms to estimate T, and the PLDA on TIMIT corpora

算法	EER (%)	MinDCF10
GMM-UBM	6.26	0.076
传统算法估计 T	4.76	0.025
通用背景估计 T	4.28	0.021
联合估计 T	4.01	0.020
通用背景−联合估计 T	**3.76 (21 %)**	**0.019 (24 %)**
PLDA	3.94	0.022

表 4　GMM-UBM、传统算法估计 T、本文所提出算法估计 T 以及 PLDA 在 MDSVC 语音库上的性能对比
Table 4　Performance comparison of GMM-UBM, the traditional algorithm to estimate T, the proposed algorithms to estimate T, and the PLDA on MDSVC corpora

算法	EER (%)	MinDCF10
GMM-UBM	7.57	0.072
传统算法估计 T	4.96	0.027
通用背景估计 T	4.92	0.026
联合估计 T	4.71	0.024
通用背景−联合估计 T	**4.67 (5.8 %)**	**0.023 (14.8 %)**
PLDA	4.67	0.024

表 5　GMM-UBM、传统算法估计 T、本文所提出算法估计 T 以及 PLDA 在 TIMIT + MDSVC 语音库上的性能对比
Table 5　Performance comparison of GMM-UBM, the traditional algorithm to estimate T, the proposed algorithms to estimate T, and the PLDA on TIMIT mixed MDSVC corpora

算法	EER (%)	MinDCF10
GMM-UBM	8.33	0.071
传统算法估计 T	5.41	0.029
通用背景估计 T	5.19	0.028
联合估计 T	5.11	0.028
通用背景−联合估计 T	**4.96 (8.3 %)**	0.027 (6.9 %)
PLDA	5.01	**0.025**

表 6　GMM-UBM、传统算法估计 T、本文所提出算法估计 T 以及 PLDA 在 MDSVC 长句语音库上的性能对比
Table 6　Performance comparison of GMM-UBM, the traditional algorithm to estimate T, the proposed algorithms to estimate T, and the PLDA on MDSVC long sentence corpora

算法	EER (%)	MinDCF10
GMM-UBM	6.58	0.067
传统算法估计 T	4.45	0.022
通用背景估计 T	3.96	0.021
联合估计 T	3.73	0.021
通用背景−联合估计 T	**3.72 (16.40 %)**	**0.020 (9.09 %)**
PLDA	3.88	0.021

为了更加直观地观察实验结果, 本文分别作了图 5 和图 6. 图 5 是在不同语音库中各个算法的性能比较 (表 3～6 的内容), 图 6 是不同算法在四种语音库中的性能比较 (包含表 7).

表 7　通用背景−联合估计算法在不同语音库中的性能对比
Table 7　Performance comparison of universal background-joint estimation algorithm on different speech corpus

语音库	EER (%)	MinDCF10
TIMIT	3.76	**0.019**
MDSVC	4.67	0.023
TIMIT + MDSVC	4.96	0.027
MDSVC 长句	**3.72**	0.020

从以上图表可以看出, 在 TIMIT 数据集、MDSVC 数据集、TIMIT + MDSVC 综合集或 MDSVC 长句集上, 1) 因子分析方法相较于基线系统 (GMM-UBM) 性能都有显著提升. 2) 新提出算法的性能都有一定提升, 特别是通用背景−联合估计算法的性能在 TIMIT 中提升了较为明显, EER 和 MinDCF10 分别提升 21 % 和 24 %. 同时, 在实验环境更加复杂的综合集 TIMIT + MDSVC 上性能也有一定提升, EER 和 MinDCF10 分别提升了 8.3 % 和 6.9 %. 通过每单个实验的对比发现, 联合估计算法的性能要一致性的优于通用背景算法, 两者结合的算法 (通用背景−联合估计) 可以得到更优的系统. 3) 相比于文献 [22] 中 i-vector 矢量规整 PLDA 技术, 本文提出的算法 (UB-JE) 在不同语音数据库中, 性能有一定的提升.

但是对比表 3～6 的数据可知, 表 6 中算法的性能最好, 其次是表 3, 然后是表 4, 最后才到表 5 (从

表 7 和图 6 可以看出), 这是由于长时语音相对短时语音更能准确地代表说话人信息以及随着语音数据复杂程度的提高, 系统的性能受到一定的影响. 现阶段说话人识别领域一个热门方向就是针对语音数据复杂程度展开的, 即多信道下的说话人识别, 这是说话人识别发展的一个趋势.

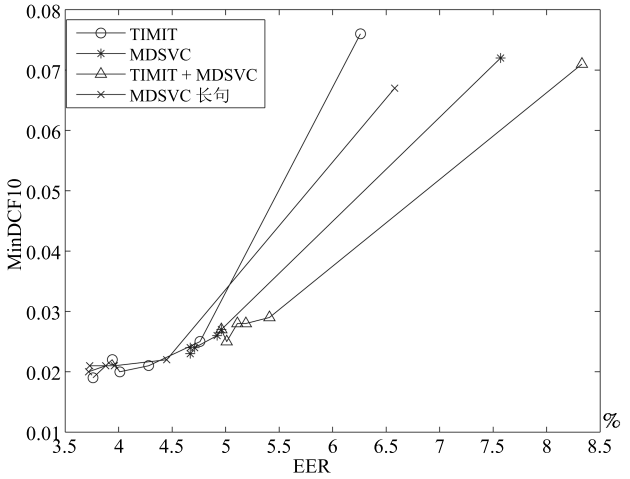


图 5 不同语音库中各算法性能对比

Fig. 5 Performance comparison of algorithms on different speech corpus

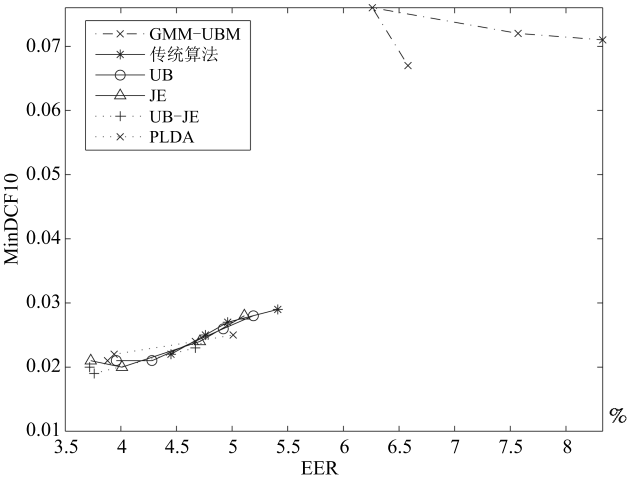


图 6 不同算法在四种语音库中的性能对比

Fig. 6 Performance comparison of different algorithms on four speech corpus

4 结论

本文主要研究了说话人识别算法 i-vector 中总变化因子空间 T 的估计, 提出了四种 T 估计算法. 实验结果显示, 在三种语音库中, 新提出的三种算法对系统的性能都有一定的提升 (如图 5), 且不同语音库对每一种算法的性能都有一定的影响 (如图 6). 实验结果证明有效估计 T 对整个 i-vector 模型起着至关重要的作用, 验证了前面 i-vector 理论分析, T 的估计引领着整个模型. 语音库的选择对整个系统的性能有一定影响, 下一步将在更加复杂的语音库 (如 NIST SRE 语音库) 上进行评测实验.

References

1 Reynolds D A. An overview of automatic speaker recognition technology. In: Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Orlando, FL, USA: IEEE, 2002. IV-4072−IV-4075

2 Kinnunen T, Li H Z. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, 2010, **52**(1): 12−40

3 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

4 Cumani S, Laface P. Large-scale training of pairwise support vector machines for speaker recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, **22**(11): 1590−1600

5 Yessad D, Amrouche A. SVM based GMM supervector speaker recognition using LP residual signal. In: Proceedings of the 2012 International Conference on Image and Signal Processing. Agadir, Morocco: Springer, 2012. 579−586

6 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

7 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

8 Dehak N. Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification [Ph. D. dissertation], École de Technologie Supérieure, Montreal, QC, Canada, 2009.

9 Dehak N, Kenny P J, 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

10 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 the 10th Annual Conference of the International Speech Communication Association. Brighton, UK: DBLP, 2009. 1559−1562

11 Cumani S, Laface P. I-vector transformation and scaling for PLDA based speaker recognition. In: Proceedings of the 2016 Odyssey Speaker and Language Recognition Workshop. Bilbao, Spain: IEEE, 2016. 39−46

12 Rouvier M, Bousquet P M, Ajili M, Kheder W B, Matrouf D, Bonastre J F. LIA system description for NIST SRE 2016. In: Proceedings of the 2016 International Speech Communication Association. San Francisco, USA: Elsevier, 2016.

13 Xu Y, McLoughlin I, Song Y, Wu K. Improved i-vector representation for speaker diarization. *Circuits, Systems, and Signal Processing*, 2016, **35**(9): 3393−3404

14 Fine S, Navratil J, Gopinath R A. Enhancing GMM scores using SVM "hints". In: Proceedings of the 7th European Conference on Speech Communication and Technology. Aalborg, Denmark: DBLP, 2001. 1757−1760

15 Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 2006, **13**(5): 308−311

16 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)

17 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)

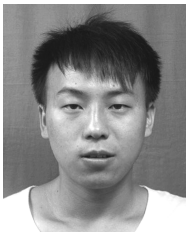
18 Jankowski C, Kalyanswamy A, Basson S, Spitz J. NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. In: Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Albuquerque, NM, USA: IEEE, 1990, **1**: 109−122

19 Woo R H, Park A, Hazen T J. The MIT mobile device speaker verification corpus: data collection and preliminary experiments. In: Proceedings of the 2016 IEEE Odyssey: the Speaker and Language Recognition Workshop. San Juan, Puerto Rico: IEEE, 2006. 1−6

20 Young S, Evermann G, Gales M, Hain T, Liu X Y, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland P. *The HTK Book (for HTK Version 3.4)*. Cambridge: Cambridge University Engineering Department, 2006.

21 NIST Speaker Recognition Evaluation [Online], available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/index.html, April 21, 2010

22 Chen L P, Lee K A, Ma B, Li H Z, Dai L R. Adaptation of PLDA for multi-source text-independent speaker verification. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). New Orleans, USA: IEEE, 2017. 5380−5384



汪海彬 昆明理工大学硕士研究生. 主要研究方向为语音信号处理, 语音识别.
E-mail: thankswhb@163.com
(**WANG Hai-Bin** Master student at Kunming University of Science and Technology. His research interest covers speech signal process and speech recognition.)

郭剑毅 昆明理工大学教授. 1990 年获得西安交通大学硕士学位. 主要研究方向为自然语言处理, 信息抽取, 知识获取. 本文通信作者.
E-mail: gjade86@hotmail.com
(**GUO Jian-Yi** Professor at Kunming University of Science and Technology. She received her master degree from Xi′an Jiaotong University in 1990. Her research interest covers natural language process, information extraction, and knowledge acquisition. Corresponding author of this paper.)

毛存礼 昆明理工大学副教授. 2014 年获得昆明理工大学博士学位. 主要研究方向为自然语言处理, 信息检索.
E-mail: maocunli@163.com
(**MAO Cun-Li** Associate professor at Kunming University of Science and Technology. He received his Ph. D. degree from Kunming University of Science and Technology in 2014. His research interest covers natural language process and information retrieval.)

余正涛 昆明理工大学教授. 2005 年获得北京理工大学博士学位. 主要研究方向为自然语言处理, 机器翻译, 信息检索.
E-mail: ztyu@hotmail.com
(**YU Zheng-Tao** Professor at Kunming University of Science and Technology. He received his Ph. D. degree from Beijing Institute of Technology in 2005. His research interest covers natural language process, machine translation, and information retrieval.)