联合因子分析中的本征信道空间拼接方法

何 亮 [1] 史永哲 [1] 刘 加 [1]

摘 要 为了使联合因子分析适用于多种信道条件下的文本无关说话人识别, 提出了一种本征信道空间的正交拼接法. 在多信道条件下, 可以通过混合数据法或简单拼接法估计本征信道空间, 但前者存在空间掩盖, 后者虽解决了空间掩盖但引入了空间重叠. 本文首先证明说话人建模和测试的核心运算是斜投影, 基于上述证明, 通过将待拼接空间正交的方法移除了空间重叠. 在 NIST SRE 2008 核心评测数据库上的实验表明, 本文所提算法优于混合数据法和简单拼接法.

关键词 说话人识别, 因子分析, 空间拼接, 投影分析

DOI 10.3724/SP.J.1004.2011.00849

Eigenchannel Space Combination Method of Joint Factor Analysis

HE Liang[1] SHI Yong-Zhe[1] LIU Jia[1]

Abstract For application of joint factor analysis on the condition of multiple channels in text-independent speaker recognition, this paper proposes an eigenchannel space orthogonal combination method. The eigenchannel space can be estimated by a mix data method or a simple combination method on the condition of multiple channels. However, the former has space masking effects while the latter introduces space overlapping effects. This paper proves that the core computation of the speaker enrollment and test is an oblique projection. Space overlapping effects can be removed subsequently by an orthogonal method based on the above proof. On the NIST SRE 2008 core tasks corpus, the proposed method has a better performance than the mix data method and the simple combination method.

Key words Speaker recognition, factor analysis, space combination, projection analysis

说话人识别隶属于语音识别, 其基本任务是判断两段语音是否属于同一个说话人, 在信息安全领域和人机交互领域有广泛的应用[1].

高斯混合模型 — 通用背景模型 (Gaussian mixture models universal background model, GMM-UBM) 系统是解决文本无关的说话人识别问题的主流系统[2]. 实验结果证明, GMM-UBM 系统在实验室环境下录制的、纯净的、长语音数据库上有优异的性能, 但在实际应用中的性能远达不到要求. 由于实际采集的语音数据与录音环境、录音设备、被录音对象是否合作和录音长度等多种因素相关, 多方面综合的结果使得 GMM-UBM 系统性能急剧下降. 其中最重要的两个因素是: 1) 训练语音和识别语音的信道条件不匹配, 即信道失配问题; 2) 训练数据相对较少与说话人建模过程中需要估计较多参数之间的矛盾. 联合因子分析 (Joint factor analysis, JFA) 很好地解决了上述两个问题[3−5]. JFA 的基本假设是将说话人高斯混合模型的均值超矢量所在的空间划分为三个组成部分: 本征信道空间、本征音空间和残差空间. 通过移除说话人均值超矢量在本征信道空间的影响, JFA 有很好的抗信道失配能力. 此外, JFA 在建立说话人模型过程中需要估计的参数明显减少 (本征音空间维数), 更适用于训练数据相对不足的情况. 在近年来美国国家标准技术署举行的说话人评测 (NIST SRE) 中[6], JFA 一直是诸多参赛单位的主要子系统.

尽管信道条件难以使用数学公式明确界定, 但根据语音的采集方式、采集环境和采集设备, 可以对信道条件进行划分. 例如 NIST SRE 定义的电话信道、麦克风信道和采访信道. 各种信道下的语音的听觉感受有较为明显的不同. 在应用 JFA 技术时, 使本征信道矩阵适应多种信道条件具有重要的理论意义和应用价值. 目前有两种解决策略: 1) 混合各种信道条件的数据进行训练; 2) 简单的空间拼接[7].

第一种策略是将各种信道条件下的数据混合起来, 也就是说不区分信道情况, 将所有数据训练一个统计的信道载荷矩阵, 在各种信道条件的训练数据较为均衡的情况下, 可以达到比较好的效果. 然而,

收稿日期 2010-09-26 录用日期 2011-02-01
Manuscript received September 26, 2010; accepted February 1, 2011
国家高技术研究发展计划 (863 计划) (2008AA02Z414, 2008AA040201), 国家自然科学基金 (90920302), 国家自然科学基金委员会与微软亚洲研究院项目 (60776800) 资助
Supported by National High Technology Research and Development Program of China (863 Program) (2008AA02Z414, 2008AA040201), National Natural Science Foundation of China (90920302), and National Natural Science Foundation of China and Microsoft Research Asia (60776800)
1. 清华大学电子工程系, 清华信息科学与技术国家实验室 北京 100084
1. Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084

实际常会遇见如下情况: 某一信道的语音数据较为丰富, 其他信道的语音数据则相对缺乏. 例如在 NIST SRE 2008 评测时, 电话信道语音数据丰富, 麦克风信道语音数据较少. 如果将两者数据混合起来训练, 则估计的本征信道载荷矩阵偏向电话信道, 而不能完全反映麦克风信道的情况. 本文将这种现象定义为 "空间掩盖".

第二种策略是为每种信道条件的数据训练一个信道矩阵, 再将各矩阵拼接起来. 这种策略虽然解决了空间掩盖, 但带来了空间重叠. 简单的空间拼接方法是利用不同信道的数据训练不同的本征信道载荷矩阵, 并将这些本征信道载荷矩阵拼接起来. 然而, 实验证明各本征信道载荷矩阵对应的空间有重叠部分, 简单地将各本征信道载荷矩阵拼接起来, 使重叠部分在建立说话人模型和测试过程中被重复计算、多次移除, 降低了系统的识别性能. 本文将这种现象定义为 "空间重叠".

为了解决空间重叠, 本文通过适当的近似, 证明了 JFA 在说话人建模和测试过程中的核心运算是斜投影过程[8], 并通过强制待拼接矩阵彼此正交化的方法, 解决了空间重叠的现象, 并给出所提算法与数据混合法、简单拼接法在 NIST SRE 2008 核心测试上的对比实验结果.

本文安排如下: 第 1 节介绍联合因子分析的一种实现算法; 第 2 节提出正交拼接法; 第 3 节是所提算法在 NIST SRE 2008 核心测试上的实验结果及结果分析; 第 4 节总结全文.

1 联合因子分析

JFA 的基本假设是

$$\boldsymbol{m}_{s,h} = \boldsymbol{m}_u + U\boldsymbol{x}_{s,h} + V\boldsymbol{y}_s + D\boldsymbol{z}_s \tag{1}$$

其中, $\boldsymbol{m}$ 是高斯混合模型的均值超矢量, U 是 $CF \times R_u$ 维本征信道空间载荷矩阵, V 是 $CF \times R_v$ 维本征音空间载荷矩阵, D 是 $CF \times CF$ 维对角残差空间载荷矩阵, F 是特征序列的维数, C 是高斯混合模型的混合分量数, R_u 是本征信道空间因子数, R_v 是本征音空间因子数, 一般而言, $10 < R_u < 200$, $100 < R_v < 400$, 下标 u 代表 UBM, 下标 s 代表说话人 s, 下标 s, h 代表说话人 s 的第 h 段语音. 有下标 s 说明该变量仅有说话人相关, 有下标 s, h 说明该变量不仅与说话人相关, 也与该段语音对应的信道条件相关.

JFA 的应用主要有三个步骤: 1) 估计空间的载荷矩阵; 2) 建立说话人模型; 3) 测试. 每个步骤都有多种估计方法. 估计空间矩阵时有: 1) 联合估计 UV 法; 2) 先估计 U, 移除信道空间影响, 再估计 V, D 的顺序估计方法; 3) 分别估计 U 和 V, 再估计 D 的方法. 建立说话人模型时有: 1) UVD 联合估计法; 2) UV 联合估计法; 3) 类似于 Gauss-Seidel 的迭代估计法. 测试时有[9]: 1) 逐帧打分法; 2) 点估计法; 3) 积分法.

根据每个步骤使用的方法及其组合的不同, JFA 有诸多变种. 其中, 性能较好的变种有文献 [3, 7] 提供的方法.

本文采用参考文献 [3−4, 7, 10] 所提技术相综合的方法: 在估计载荷矩阵时采用顺序估计法; 在建立说话人模型时采用类似于 Gauss-Seidel 的迭代法; 在测试时采用积分法. 本文的主要工作是空间拼接, 空间拼接推导基于说话人建模公式和测试公式, 而它们又依赖于空间载荷矩阵的训练, 因而本文首先介绍载荷矩阵的训练方法.

1.1 载荷矩阵顺序估计法

JFA 的基本假设中含有四个部分: 1) UBM 均值超矢量; 2) 本征信道空间; 3) 本征音空间; 4) 残差空间. UBM 均值超矢量的作用比较简单, 在计算时相当于将统计的原点统计量转化为中心统计量. 在估计本征信道载荷矩阵、本征音载荷矩阵和残差载荷矩阵的过程中, 需要将式 (1) 适当简化或变形, 再进行估计.

1.1.1 本征信道载荷矩阵估计法

在估计本征信道空间载荷矩阵时, V 和 D 是未知的, 需要将式 (1) 简化为

$$\boldsymbol{m}_{s,h} - \boldsymbol{m}_u - \triangle\boldsymbol{y}_s = U\boldsymbol{x}_{s,h} \tag{2}$$

其中, $\triangle\boldsymbol{y}$ 代表与信道无关的说话人均值超矢量. 此时 JFA 的模型假设退化为本征信道的模型假设.

估计 U 的数据要求是一个说话人有多段语音 $X_{s,h}$ (最好能覆盖多种信道条件), 设共有 H_s 段. 使用 EM 算法估计 U 时有 4 个步骤:

步骤 1. 积累零阶、一阶、二阶中心统计量:

$$\begin{aligned} N_{s,h,c} &= \sum_t \gamma_{c,t} \\ F_{s,h,c} &= \sum_t \gamma_{c,t}\left(\boldsymbol{x}_{s,h,t} - \boldsymbol{m}_{u,c}\right) \\ S_{s,h,c} &= \mathrm{diag}\Big\{ \sum_t \gamma_{c,t}\left(\boldsymbol{x}_{s,h,t} - \boldsymbol{m}_{u,c}\right) \times \\ &\qquad \left(\boldsymbol{x}_{s,h,t} - \boldsymbol{m}_{u,c}\right)^{\mathrm{T}} \Big\} \end{aligned} \tag{3}$$

其中, t 是时间索引, $\gamma_{c,t}$ 代表 $\boldsymbol{x}_{s,h,t}$ 在 UBM 第 c 个高斯混合分量上的占有率, $\mathrm{diag}\{\cdot\}$ 代表取对角运算.

步骤 2. 将 $N_{s,h,c}$ 构成 $CF \times CF$ 对角矩阵 $N_{s,h}$; 将 $F_{s,h,c}$ 和 $S_{s,h,c}$ 拼接成 CF 维超矢量 $F_{s,h}$ 和 $S_{s,h}$. 估计本征信道空间因子 $\boldsymbol{x}$ 的一阶和二阶期

望值

$$\begin{aligned} L_{s,h} &= I + U^{\mathrm{T}}\Sigma^{-1}N_{s,h}U \\ \mathrm{E}[\boldsymbol{x}_{s,h,c}] &= L_{s,h}^{-1}U^{\mathrm{T}}\Sigma^{-1}F_{s,h,c} \\ \mathrm{E}[\boldsymbol{x}_{s,h,c}\boldsymbol{x}_{s,h,c}^{\mathrm{T}}] &= \mathrm{E}[\boldsymbol{x}_{s,h,c}]\mathrm{E}[\boldsymbol{x}_{s,h,c}^{\mathrm{T}}] + L_{s,h}^{-1} \end{aligned} \tag{4}$$

其中, $L_{s,h}$ 是临时变量, Σ 是 UBM 的协方差矩阵. 在说话人识别中, Σ 常被简化为对角阵.

步骤 3. 将同一说话人各段语音累加, 并使用 MAP 自适应, 所得统计量减去信道因素得到 $\triangle\boldsymbol{y}_s$. 对于第 c 个高斯分量对应的 $\triangle\boldsymbol{y}_{s,c}$ 计算公式如下:

$$\triangle\boldsymbol{y}_{s,c} = \frac{1}{N_{s,c}+\tau}\left[\sum_{h=1}^{H_s}F_{s,h,c} - \sum_{h=1}^{H_s}N_{s,h,c}U_c\mathrm{E}[\boldsymbol{x}_{s,h,c}]\right] \tag{5}$$

步骤 4. 更新本征信道载荷矩阵 U 和协方差矩阵 Σ. 第 c 个高斯分量的 U_c 和 Σ_c 的更新公式如下

$$\begin{aligned} &\sum_s\sum_{h=1}^{H_s}N_{s,h}U_c\mathrm{E}[\boldsymbol{x}_{s,h,c}\boldsymbol{x}_{s,h,c}^{\mathrm{T}}] = \\ &\qquad\sum_s\sum_{h=1}^{H_s}\left[F_{s,h,c} - N_{s,h,c}\triangle\boldsymbol{y}_{s,c}\right]\mathrm{E}[\boldsymbol{x}_{s,h,c}^{\mathrm{T}}] \end{aligned} \tag{6}$$

$$\begin{aligned} \mathrm{diag}\{\Sigma_c\} = &\left[\sum_s\sum_{h=1}^{H_s}N_{s,h,c}\right]^{-1}\left\{\sum_s\sum_{h=1}^{H_s}S_{s,h,c} - \right. \\ &\left.\mathrm{diag}\left\{\left(\sum_s\sum_{h=1}^{H_s}\left[F_{s,h,c} - N_{s,h,c}\triangle\boldsymbol{y}_{s,c}\right]\mathrm{E}[\boldsymbol{x}_{s,h,c}^{\mathrm{T}}]\right)U^{\mathrm{T}}\right\}\right\} \end{aligned} \tag{7}$$

由于要对本征信道载荷矩阵进行拼接, 为了保持拼接矩阵对应协方差的一致性, 不更新 Σ. 经过 6 次迭代, 可以近似认为 U 收敛.

1.1.2　本征音载荷矩阵估计法

在已知 U 的条件下, 估计 V. 式 (1) 简化成

$$\boldsymbol{m}_{s,h} - \boldsymbol{m}_u - U\boldsymbol{x}_{s,h} = V\boldsymbol{y}_s \tag{8}$$

估计 V 的方法与步骤与估计 U 的基本类似, 主要区别在于使用的统计量不同.

训练本征音载荷矩阵的数据也要求一个说话人对应多段语音, 然而, 与本征信道载荷矩阵对数据的要求相比, 本征音载荷矩阵对数据的要求相对宽松, 一个说话人对应的语音段数可以相对较少, 甚至一段都可以.

在使用 EM 算法估计 V 有以下 3 个步骤:

步骤 1. 积累零阶、一阶和二阶中心统计量:

$$\begin{aligned} N_{s,c} &= \sum_t\sum_h^{H_s}\gamma_{c,t} \\ F_{s,c} &= \sum_t\sum_h^{H_s}\gamma_{c,t}\left(\boldsymbol{x}_{s,h,t} - \boldsymbol{m}_{u,c} - U_c\boldsymbol{x}_{s,h}\right) \\ S_{s,c} &= \mathrm{diag}\left\{\sum_t\sum_h^{H_s}\gamma_{c,t}\left(\boldsymbol{x}_{s,h,t} - \boldsymbol{m}_{u,c} - U_c\boldsymbol{x}_{s,h}\right)\times \right. \\ &\qquad\left.\left(\boldsymbol{x}_{s,h,t} - \boldsymbol{m}_{u,c} - U_c\boldsymbol{x}_{s,h}\right)^{\mathrm{T}}\right\} \end{aligned} \tag{9}$$

注意, 下标带时间索引 t 的 $\boldsymbol{x}$ 代表第 t 个时刻的频谱特征及其衍生特征, 不带时间下标索引的代表本征信道空间的因子. 本文用 $\boldsymbol{x}$ 代表两种不同类型的变量, 是为了与诸多参考文献的符号标记保持一致.

步骤 2. 将搜集的中心统计量拼接为超矢量, 并估计本征音因子 $\boldsymbol{y}$ 的一阶和二阶期望值:

$$\begin{aligned} L_s &= I + V^{\mathrm{T}}\Sigma^{-1}N_sV \\ \mathrm{E}[\boldsymbol{y}_{s,c}] &= L_s^{-1}V^{\mathrm{T}}\Sigma^{-1}F_{s,c} \\ \mathrm{E}[\boldsymbol{y}_{s,c}\boldsymbol{y}_{s,c}^{\mathrm{T}}] &= \mathrm{E}[\boldsymbol{y}_{s,c}]\mathrm{E}[\boldsymbol{y}_{s,c}^{\mathrm{T}}] + L_s^{-1} \end{aligned} \tag{10}$$

其中, L_s 是临时变量, Σ 是 UBM 的协方差矩阵.

步骤 3. 更新本征音载荷矩阵 V 和协方差矩阵 Σ. 第 c 个高斯分量的 V_c 和 Σ_c 的更新公式如下:

$$\sum_s N_{s,h}V_c\mathrm{E}[\boldsymbol{y}_{s,c}\boldsymbol{y}_{s,c}^{\mathrm{T}}] = \sum_s F_{s,c}\mathrm{E}[\boldsymbol{y}_{s,c}^{\mathrm{T}}] \tag{11}$$

$$\begin{aligned} \mathrm{diag}\{\Sigma_c\} = &\left[\sum_s N_{s,c}\right]^{-1}\times \\ &\left\{\sum_s S_{s,c} - \mathrm{diag}\left\{\left(\sum_s F_{s,c}\mathrm{E}[\boldsymbol{y}_{s,c}^{\mathrm{T}}]\right)V^{\mathrm{T}}\right\}\right\} \end{aligned} \tag{12}$$

经过 6 次迭代, 可以近似认为 V 收敛.

1.1.3　残差载荷矩阵估计法

在已知 U 和 V 的条件下, 估计 D. 此时式 (1) 为

$$\boldsymbol{m}_{s,h} - \boldsymbol{m}_u - U\boldsymbol{x}_{s,h} - V\boldsymbol{y}_s = D\boldsymbol{z}_s \tag{13}$$

D 是对角阵, 运算比较简单. 首先, 累积统计量

$$\begin{aligned} N_{s,c} &= \sum_t\sum_h^{H_s}\gamma_{c,t} \\ F_{s,c}^{'} &= \sum_t\sum_h^{H_s}\gamma_{c,t}\left(\boldsymbol{x}_{s,h,t} - \boldsymbol{m}_{u,c} - U_c\boldsymbol{x}_{s,h} - V_c\boldsymbol{y}_s\right) \end{aligned}$$

$$S_{s,c}^{'}=\mathrm{diag}\Big\{\sum_{t}\sum_{h}^{H_s}\gamma_{c,t}\big(\boldsymbol{x}_{s,h,t}-\boldsymbol{m}_{u,c}-U_c\boldsymbol{x}_{s,h}-V_c\boldsymbol{y}_s\big)\big(\boldsymbol{x}_{s,h,t}-\boldsymbol{m}_{u,c}-U_c\boldsymbol{x}_{s,h}-V_c\boldsymbol{y}_s\big)^{\mathrm{T}}\Big\} \tag{14}$$

再次, 拼接统计量并求解隐含变量

$$\begin{aligned}G_s&=I+D^{\mathrm{T}}\Sigma^{-1}N_sD\\ \mathrm{E}[\boldsymbol{z}_{s,c}]&=G_s^{-1}D^{\mathrm{T}}\Sigma^{-1}F_{s,c}^{'}\\ \mathrm{E}[\boldsymbol{z}_{s,c}\boldsymbol{z}_{s,c}^{\mathrm{T}}]&=\mathrm{E}[\boldsymbol{z}_{s,c}]\mathrm{E}[\boldsymbol{z}_{s,c}^{\mathrm{T}}]+G_s^{-1}\end{aligned} \tag{15}$$

最后, 更新 D 和 Σ.

$$\sum_s N_sD_c\mathrm{E}[\boldsymbol{z}_{s,c}\boldsymbol{x}_{s,c}^{\mathrm{T}}]=\sum_s F_{s,c}^{'}\mathrm{E}[\boldsymbol{z}_{s,c}^{\mathrm{T}}] \tag{16}$$

$$\mathrm{diag}\{\Sigma_c\}=\left[\sum_s N_{s,c}\right]^{-1}\times\left\{\sum_s S_{s,c}^{'}-\mathrm{diag}\Big\{\sum_s F^{'}\mathrm{E}[\boldsymbol{z}_{s,c}^{\mathrm{T}}]D^{\mathrm{T}}\Big\}\right\} \tag{17}$$

经过 3 次迭代, 可以近似认为 D 收敛. 至此, 空间矩阵 U, V 和 D 都估计完毕.

1.2 采用类似 Gauss-Seidel 的迭代估计法建立说话人模型

采用类似 Gauss-Seidel 的迭代估计法建立说话人模型的过程与估计载荷矩阵的过程类似[11], 可以使用相同的程序实现, 区别是不使用更新步骤. 首先根据式 (3), (4), (9), (10), (14) 和 (15) 估计本征信道因子 $\boldsymbol{x}_{s,h}$、本征音因子 $\boldsymbol{y}_s$ 和残差因子 $\boldsymbol{z}_s$, 再利用式 (18) 建立说话人模型

$$\boldsymbol{m}_s=\boldsymbol{m}_u+V\boldsymbol{y}_s+D\boldsymbol{z}_s \tag{18}$$

上述过程中, 迭代 1 次即可, 多次迭代反而会降低系统性能[12].

1.3 采用积分法进行测试

在计算对数似然比的时候, 积分法通过积分将本征信道空间的影响消除掉

$$p\left(X|\lambda_s\right)=\int p\left(X|\lambda_s,\boldsymbol{x}\right)\mathcal{N}(\boldsymbol{x};0,I)\mathrm{d}\boldsymbol{x} \tag{19}$$

其中, X 是测试语音提取出的频谱特征及衍生特征序列, 上述积分有闭式解

$$p\left(X|\lambda_s\right)=\sum_{c=1}^{C}N_c\frac{1}{(2\pi)^{\frac{F}{2}}|\Sigma_c|^{\frac{1}{2}}}-\frac{1}{2}\mathrm{tr}\{\Sigma^{-1}S_s^{''}\}-\frac{1}{2}\log|L|+\frac{1}{2}\|L^{-\frac{1}{2}}U^{\mathrm{T}}\Sigma^{-1}F_s^{''}\|^2 \tag{20}$$

其中, $F_s^{''}$ 和 $S_s^{''}$ 是以说话人均值 $\boldsymbol{m}_s$ 为中心的中心统计量, L 是临时变量. 计算公式如下:

$$\begin{aligned}N_{s,c}&=\sum_t\gamma_{c,t}\\ F_{s,c}^{''}&=\sum_t\gamma_{c,t}\left(\boldsymbol{x}_{s,t}-\boldsymbol{m}_s\right)\\ S_{s,c}^{''}&=\mathrm{diag}\Big\{\sum_t\gamma_{c,t}\left(\boldsymbol{x}_{s,t}-\boldsymbol{m}_s\right)\left(\boldsymbol{x}_{s,t}-\boldsymbol{m}_s\right)^{\mathrm{T}}\Big\}\end{aligned} \tag{21}$$

$$L=I+U^{\mathrm{T}}\Sigma^{-1}N_s \tag{22}$$

在计算对数似然比的过程中, 式 (20) 中的第 1 项和第 3 项都被消除掉, 仅计算第 2 项和第 4 项即可.

2 空间拼接

尽管 JFA 理论体系框架完备, 但是 JFA 的建模是针对单一信道情况. 对于多种复杂的信道条件, 简单拼接法是目前国际主流的解决方案. 然而, 简单拼接法会引入空间重叠的问题. 举例说明, 令 U_{tel} 和 U_{mic} 分别代表电话信道和麦克风信道语音数据训练所得的本征信道载荷矩阵. 简单拼接的本征信道载荷矩阵为 $U_{\mathrm{com}}=[U_{\mathrm{tel}},U_{\mathrm{mic}}]$. 在建立说话人模型过程中, 需要移除信道部分的影响, 其计算过程如下:

$$U\mathrm{E}[\boldsymbol{x}_{s,h}]=U\left[I+U^{\mathrm{T}}\Sigma^{-1}N_{s,h}U\right]^{-1}U^{\mathrm{T}}\Sigma^{-1}F_{s,h}\approx\Sigma^{\frac{1}{2}}Q\left[Q^{\mathrm{T}}Q\right]^{-1}Q^{\mathrm{T}}\left[\Sigma^{-\frac{1}{2}}\boldsymbol{m}_{s,h}\right] \tag{23}$$

其中, $Q=\Sigma^{-\frac{1}{2}}U$. 上述推导使用了两次近似: 第 1 个近似是去除了单位矩阵, 这是由于当有效训练语音长度大于 2 分钟时, $N_{s,h}$ 对角值远大于 1, 起主导作用; 第 2 个近似是 $\boldsymbol{m}_{s,h}\approx N_{s,h}^{-1}F_{s,h}$. 由于在 MAP 自适应过程中, 可以通过参数进行调节 UBM 和说话人统计量在说话人模型所占的比例. 近似号成立时是 UBM 所占比例为 0 的一种特殊情况. 令

$$P=Q\left[Q^{\mathrm{T}}Q\right]^{-1}Q^{\mathrm{T}} \tag{24}$$

则 P 是斜投影算子, 式 (23) 证明说话人建模过程的核心运算是 $\Sigma^{-\frac{1}{2}}\boldsymbol{m}_{s,h}$ 向 P 的投影[11].

将简单拼接矩阵 $U_{\rm com}=[U_{\rm tel},U_{\rm mic}]$ 代入,

$$\begin{aligned}&U_{\rm com}{\rm E}[\boldsymbol{x}_{{\rm com},s,h}]=\\&\quad[U_{\rm tel},U_{\rm mic}]\left[{\rm E}[\boldsymbol{x}_{{\rm tel},s,h}],{\rm E}[\boldsymbol{x}_{{\rm mic},s,h}]\right]=\\&\quad\Sigma^{\frac{1}{2}}\left[Q_{\rm tel},Q_{\rm mic}\right]\left[Q_{\rm com}^{\rm T}Q_{\rm com}\right]^{-1}Q_{\rm com}^{\rm T}\Sigma^{-\frac{1}{2}}\boldsymbol{m}_{s,h}\end{aligned}\tag{25}$$

利用分块矩阵求逆法可得

$$\begin{aligned}&\left[Q_{\rm com}^{\rm T}Q_{\rm com}\right]^{-1}=\\&\quad\left[\begin{array}{c}\left(A_{tt}-A_{mt}A_{mm}^{-1}A_{tm}\right)^{-1}\\-\left(A_{mm}-A_{tm}A_{tt}^{-1}A_{mt}\right)^{-1}A_{tm}A_{tt}^{-1}\end{array}\right.\\&\qquad\left.\begin{array}{c}-\left(A_{tt}-A_{mt}A_{mm}^{-1}A_{tm}\right)^{-1}A_{mt}A_{mm}^{-1}\\\left(A_{mm}-A_{tm}A_{tt}^{-1}A_{mt}\right)^{-1}\end{array}\right]\end{aligned}\tag{26}$$

其中,

$$\begin{aligned}&A_{tt}=Q_{\rm tel}^{\rm T}Q_{\rm tel},\quad A_{tm}=Q_{\rm tel}^{\rm T}Q_{\rm mic}\\&A_{mt}=Q_{\rm mic}^{\rm T}Q_{\rm tel},\quad A_{mm}=Q_{\rm mic}^{\rm T}Q_{\rm mic}\end{aligned}\tag{27}$$

由于

$$A_{tm}\neq 0,\quad A_{mt}\neq 0\tag{28}$$

则在计算 ${\rm E}[\boldsymbol{x}_{{\rm tel},s,h}]$ 会混入麦克风信道载荷矩阵的影响; 同理在计算 ${\rm E}[\boldsymbol{x}_{{\rm mic},s,h}]$ 会混入电话信道载荷矩阵的影响. 可见, 简单拼接法在说话人建模过程存在空间重叠现象.

测试需要计算式 (20) 中第 2 项和第 4 项. 其中第 4 项涉及本征信道载荷矩阵, 将其单独展开如下

$$\begin{aligned}&\frac{1}{2}\|\boldsymbol{l}^{-\frac{1}{2}}U^{\rm T}\Sigma^{-1}F_s''\|^2=\left(F_s''\right)^{\rm T}\Sigma^{-\frac{1}{2}}\times\\&\qquad\frac{1}{2}Q_{\rm com}\left[Q_{\rm com}^{\rm T}Q_{\rm com}\right]^{-1}Q_{\rm com}^{\rm T}\Sigma^{-\frac{1}{2}}F_s''\end{aligned}\tag{29}$$

上式表明测试的核心运算也是斜投影运算. 简单拼接法同样存在空间重叠现象.

尽管各本征信道载荷矩阵的训练数据是彼此独立的, 但是训练所得的本征信道载荷矩阵并不是彼此无交连的. 其中所有本征信道载荷矩阵的交连是所有本征信道载荷矩阵的共有部分, 这部分代表所有本征信道载荷矩阵的共性部分; 而每个本征信道载荷矩阵除去共有部分是每种本征信道载荷矩阵的特有部分, 这部分代表某个信道条件对应的载荷矩阵的特性部分. 简单拼接法使共性部分被重复计算, 多次移除, 进而影响系统的识别性能. 空间拼接时, 应将各本征信道载荷矩阵特性部分进行拼接, 而仅保留一份共性部分. 换言之, 空间拼接的 U 所对应的 Q 矩阵应满足如下条件

$$Q_i^{\rm T}Q_j=0,\quad i\neq j\tag{30}$$

基于上述思想, 本文提出如下空间拼接的方法. 假设待拼接的本征信道载荷矩阵为 $U_0,U_1,U_2,\cdots,U_M$, 拼接步骤如下:

1) 初始化: 计算

$$U_0,U_1,U_2,\cdots,U_M$$

对应的

$$Q_0,Q_1,Q_2,\cdots,Q_M$$

并令 $Q_{\rm com}=Q_0$.

2) 循环拼接: 对 $i=1,2,\cdots,M$ 进行循环:

a) 使用格拉姆 - 施密特正交化的方法寻找 $Q_{\rm com}$ 对应的标准正交基

$$\{\boldsymbol{\eta}_{{\rm com},1},\boldsymbol{\eta}_{{\rm com},2},\cdots,\boldsymbol{\eta}_{{\rm com},R_{u,{\rm com}}}\}$$

b) 移除待拼接矩阵 Q_i 各列向量在

$${\rm span}\left\{\boldsymbol{\eta}_{{\rm com},1},\boldsymbol{\eta}_{{\rm com},2},\cdots,\boldsymbol{\eta}_{{\rm com},R_{u,{\rm com}}}\right\}$$

中的部分, 得到 Q_i';

c) 拼接 $[Q_{\rm com},Q_i']$, 并令新拼接的矩阵为 $Q_{\rm com}$.

3) 通过 $Q_{\rm com}$ 反向计算 $U_{\rm com}$.

图 1 和图 2 是本文使用的本征信道矩阵的简单拼接法和正交拼接法的 $Q_{\rm com}^{\rm T}Q_{\rm com}$. 对比图 1 和图 2 可以看出, 空间拼接法移除了空间重叠.

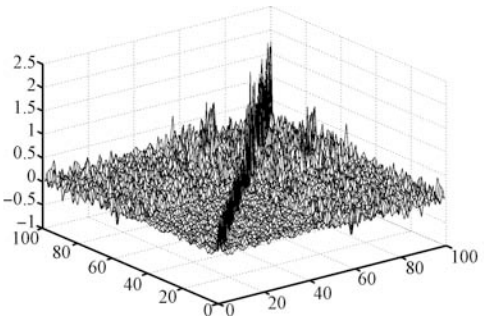


图 1　简单拼接法

Fig. 1　Simple combination method

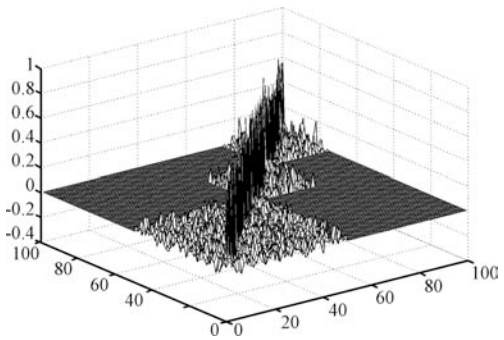


图 2　正交拼接法

Fig. 2　Orthogonal combination method

3 实验配置与实验结果

3.1 测试数据库

本文使用 NIST SRE 2008 规定的核心测试集作为测试数据库. 核心测试集包含 8 组测试任务, 共 98 766 个测试任务、3 263 个说话人模型. 其中最有代表性的是第 1、4、5、6 组测试任务. 第 2、3 组测试任务是第 1 组测试任务的子集, 第 7、8 组测试任务是第 6 组测试任务的子集. 根据语音方式、采集设备等不同, NIST 将语音数据分成三种信道条件: 采访信道 (int)、电话信道 (tel) 和麦克风信道 (mic). 其中, 第 1 组测试任务的训练语音、测试语音均是采访信道; 第 4 组测试任务的训练语音是采访信道, 测试语音是电话信道; 第 5 组测试任务的训练语音是电话信道, 测试语音是麦克风信道; 第 6 组测试任务的训练语音、测试语音均是电话信道.

3.2 开发数据集

开发集数据库包括 NIST SRE 1996-2006, NIST SRE 2008 开发集以及 NIST SRE 2008 Followup 测试集. 其中 NIST SRE 1996-2004 数据用于训练 UBM 模型; NIST SRE 1996-2004, SRE 2008 Followup 测试集数据用于训练本征音载荷矩阵; NIST 2004-2006, NIST SRE 2008 开发集和 NIST SRE 2008 Followup 测试集数据用于训练本征信道载荷矩阵; NIST SRE 2004-2006 数据用于训练残差载荷矩阵.

3.3 前端处理

使用 G.723.1 进行活动语音端点检测 (VAD); 对检测出的语音删除前 25 % 低能量; 预加重, 预加重因子 0.95; 提取 13 维 Mel 频率倒谱系数 (MFCC) 以及 1、2 阶差分特征构成 39 维特征; 对上述 39 维特征使用频率弯曲 (Feature warping)[13], 得到最终的特征序列.

3.4 JFA 系统

实验是基于性别相关的 UBM, 混合模型数是 1 024, 本征音因子数是 300.

为了验证正交拼接的有效性, 分成 3 组实验:

实验 1. 将训练数据混合, 估计一个可以覆盖各种信道条件的本征信道载荷矩阵. 本征信道因子数为 100.

实验 2. 使用简单拼接法获得本征信道空间矩阵. 本征信道载荷矩阵的训练数据如下: 训练 Tel 本征信道载荷矩阵的数据来源于 NIST SRE 2004-2006 的 Tel 语音; 训练 Mic 本征信道载荷矩阵的数据来源于 NIST SRE 2005-2006 的 Mic 语音; 训练 Int 本征信道载荷矩阵的数据来源于 NIST SRE 2008 的开发集和 Followup 测试集. Tel 本征信道因子数为 50, Mic 本征信道因子数为 25, Int 本征信道因子数为 25. 采用简单拼接法获得最终的本征信道载荷矩阵.

实验 3. 训练本征信道载荷矩阵的数据, 各本征信道载荷矩阵因子数以及拼接后的信道载荷矩阵因子数同实验 2, 但采用本文所提出的方法进行空间拼接.

3.5 实验结果

本文仅关注第 1、4、5、6 组测试任务. 衡量指标采用等错点 (EER) 以及 NIST SRE 2008 规定的最小检测代价 (MinDCF). 测试结果如下:

第 1 组测试条件下 JFA 的性能非常好, 并且简单拼接法和正交拼接法都不如混合数据法, 这与训练载荷矩阵时, 使用了 NIST SRE 08 Followup 数据有关. NIST SRE 2008 Followup 数据是采访数据, 与第 1 组测试的训练、测试条件都匹配. 此外, 它们所含的说话人有部分重叠 (注意, 测试语音不重叠). 混合数据法所得的载荷矩阵已能很好描述重叠的说话人所对应的本征信道空间, 因而简单拼接法和正交拼接法都失效. 对于第 4、5、6 组测试条件, 首先关注没有使用分数归一化这部分结果. 综合而言, 无论是同信道的第 6 组测试, 还是跨信道的第 4、5 组测试条件, 正交拼接法均优于混合数据法和简单拼接法. 由于空间掩盖, 混合数据法的性能差于正交拼接法; 由于空间重叠, 简单拼接法的性能不仅差于正交拼接法, 还差于混合数据法. 其次, 关注使用 ztnorm 这部分结果, 归一化使得正交拼接的优势有所降低. 在 JFA 应用中, ztnorm 是非常有效的归一化策略. 然而, 归一化需要大量的冒充者数据, 并要求数据源于不同的说话人. 由于 NIST 提供的数据规模所限, 在进行 ztnorm 的过程中不可能完全按照理论要求进行, 实际采用的折中方案是, 选取的 ztnorm 集存在大量 1 个人对应多段语音数据的现象. 这种折中方案可以近似达到 ztnorm 的效果, 但会由于冒充者数据主要源于几十个人, 使得 ztnorm 的结果具有一定不稳定性. 这种不稳定性使正交拼接法的优势弱化, 但综合而言, 正交拼接法的性能还是最优.

4 结论

JFA 是目前性能最为优异的说话人识别系统. 对于单一信道条件, JFA 理论完备. 对于复杂信道条件, 目前有混合数据法和简单拼接法获得本征信道载荷矩阵. 混合数据法所得的载荷矩阵在训练数据不均衡时存在空间掩盖, 简单拼接法解决了空间掩盖, 但引入空间重叠. 空间掩盖和空间重叠在不同

表 1 空间拼接法在 NIST SRE 2008 年核心测试第 1、4、5、6 组测试条件上的对比实验结果

Table 1 Comparison of space combination methods on the 1, 4, 5, 6 trial conditions of the NIST SRE 2008 core task

	拼接方法	归一化方法	EER, 男	MinDCF, 男	EER, 女	MinDCF, 女
第 1 组测试条件	混合数据	—	1.19	0.0421	0.73	0.0407
第 1 组测试条件	混合数据	ztnorm	0.83	0.0215	0.36	0.0169
第 1 组测试条件	简单拼接	—	1.29	0.0569	0.98	0.0476
第 1 组测试条件	简单拼接	ztnorm	0.83	0.0215	0.36	0.0169
第 1 组测试条件	正交拼接	—	1.60	0.0789	1.12	0.0560
第 1 组测试条件	正交拼接	ztnorm	0.98	0.0214	0.37	0.0194
第 4 组测试条件	混合数据	—	5.77	0.3170	9.67	0.4457
第 4 组测试条件	混合数据	ztnorm	3.64	0.1104	3.03	0.1578
第 4 组测试条件	简单拼接	—	7.29	0.3390	10.51	0.5099
第 4 组测试条件	简单拼接	ztnorm	3.63	0.0996	3.50	0.1549
第 4 组测试条件	正交拼接	—	6.11	0.2817	8.68	0.4424
第 4 组测试条件	正交拼接	ztnorm	3.43	0.1011	3.59	0.1582
第 5 组测试条件	混合数据	—	6.43	0.2537	7.61	0.3786
第 5 组测试条件	混合数据	ztnorm	4.84	0.1585	5.69	0.2345
第 5 组测试条件	简单拼接	—	6.57	0.2745	7.98	0.3555
第 5 组测试条件	简单拼接	ztnorm	5.30	0.1733	5.43	0.2175
第 5 组测试条件	正交拼接	—	5.86	0.2787	8.03	0.3703
第 5 组测试条件	正交拼接	ztnorm	4.83	0.1724	5.20	0.2073
第 6 组测试条件	混合数据	—	6.12	0.2732	8.77	0.3883
第 6 组测试条件	混合数据	ztnorm	4.75	0.2255	6.79	0.3254
第 6 组测试条件	简单拼接	—	6.29	0.2825	8.67	0.3964
第 6 组测试条件	简单拼接	ztnorm	4.68	0.2243	6.76	0.3294
第 6 组测试条件	正交拼接	—	5.98	0.2793	8.56	0.3858
第 6 组测试条件	正交拼接	ztnorm	4.56	0.2202	6.71	0.3203

程度上会引起系统性能的降低. 本文提出的正交拼接法, 解决了简单拼接法存在的空间重叠. 在 NIST SRE 2008 年核心测试数据库上的实验结果证明了正交拼接法优于混合数据法和简单拼接法.

References

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

2 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

3 Kenny P, Ouellet P, Dehak N, Gupta V, Dumouchel P. A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 2008, **16**(5): 980−988

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 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

6 NIST speaker recognition evaluation [Online], available: http://www.itl.nist.gov/iad/mig/tests/spk/2008/index.html,

December 19, 2010

7 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)

8 Zhang Xian-Da. *Matrix Analysis and Application*. Beijing: Tsinghua University, 2004. 687−698
(张贤达. 矩阵分析与应用. 北京: 清华大学出版社, 2004. 687−698)

9 Glembek O, Burget L, Dehak N, Brummer N, Kenny P. Comparison of scoring methods used in speaker recognition with joint factor analysis. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Taipei, China: IEEE, 2009. 4057−4060

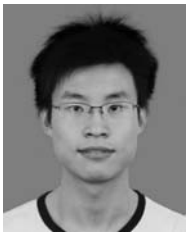
10 Vogt R, Sridharan S. Experiments in session variability modelling for speaker verification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Toulouse, France: IEEE, 2006. 897−900

11 Campbell W, Karam Z, Sturim D. Speaker comparison with inner product discriminant functions. *Advances in Neural Information Processing Systems*. Cambridge: The MIT Press, 2009. 207−215

12 Vogt R J, Baker B J, Sridharan S. Modelling session variability in text independent speaker verification. In: Proceedings of the 9th European Conference on Speech Communication and Technology. Lisbon, Portugal: ISCA, 2005. 3117−3120

13 Xiang B, Chaudhari U V, Navratil J, Ramaswamy G N, Gopinath R A. Short-time Gaussianization for robust speaker verification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Minneapolis, USA: IEEE, 2002. 681−684

何 亮 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和语种识别. 本文通信作者.
E-mail: heliang06@mails.thu.edu.cn
(**HE Liang** 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.)



史永哲 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和语音识别.
E-mail: shiyz09@mails.tsinghua.edu.cn
(**SHI Yong-Zhe** Ph. D. candidate in the Department of Electronic Engineering, Tsinghua University. His research interest covers speaker recognition and speech 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.)