说话人识别中的因子分析以及空间拼接

郭 武 [1] 李轶杰 [1] 戴礼荣 [1] 王仁华 [1]

摘 要 联合因子分析可以有效拟合混合高斯模型中的说话人和信道差异, 在说话人识别中得到广泛应用. 一般情况下, 该算法在对说话人和信道两个载荷矩阵进行联合估计时, 说话人残差矩阵无法发挥作用, 信道载荷矩阵的因子数不能提高. 本文提出说话人载荷矩阵、说话人残差载荷矩阵采用串行的训练模式, 在信道载荷矩阵训练中采用矩阵拼接的方法, 能够有效提高识别率; 在 NIST SRE 2008 年核心测试数据库的五个部分分别达到等错误率 3.3 %, 5.1 %, 5.0 %, 5.3 % 和 5.0 %.

关键词 说话人识别, 联合因子分析, 本征音因子, 说话人确认, 期望最大化
中图分类号 TN912.34

Factor Analysis and Space Assembling in Speaker Recognition

GUO Wu[1] LI Yi-Jie[1] DAI Li-Rong[1] WANG Ren-Hua[1]

Abstract Factor analysis is a model of the speaker and session variability in Gaussian mixture models and is widely used in text-independent speaker recognition. There exist two issues when the loading matrices of the eigenvoice and eigenchannel are estimated jointly. First, the speaker diagonal matrix (residual) will not take effect; second, the channel factors can not be very large. In this paper, the loading matrices of eigenvoice and the diagonal are calculated serially and different eigenchannel matrices are assembled to form a large channel loading matrix. The performance can be improved by the proposed algorithm. In the NIST speaker recognition evaluation (SRE) 2008 core test corpus, the equal error rates (EERs) of the five sub sessions were 3.3 %, 5.1 %, 5.0 %, 5.3 %, and 5.0 %.

Key words Speaker recognition, joint factor analysis, eigenvoice, speaker verification, expectation maximization

按照混合高斯−通用背景模型[1](Gaussian mixture model−universal background model, GMM-UBM) 的思想, 所有的说话人信息都包含在混合高斯函数所形成的均值超矢量[2] 中, 具体均值超矢量形成的过程如图 1, 也就是把所有的 C 个 GMM 的均值按照顺序排列起来, 形成一个大的超矢量. 假设高斯数目为 C, 特征矢量的维数为 F, 那么超矢量的维数就是 FC.

因子分析[3−5](Factor analysis) 就是建立在均值超矢量上的一种算法, 由于其优异的性能得到研究者广泛认可的一种算法, 因子分析的优势在于: 1) 与 GMM-UBM 的说话人识别主流算法结合紧密, 理论完备; 2) 与最大后验概率 (Maximum a posterior, MAP) 的说话人建模的方法能够有效地结合.

按照因子分析, 对每句话 s 形成的均值超矢量 $\boldsymbol{m}(s)$, 用式 (1) 表示.

$$\boldsymbol{m}(s) = \boldsymbol{m}_0 + V\boldsymbol{y} + D\boldsymbol{z} + U\boldsymbol{x} \tag{1}$$

其中, $\boldsymbol{m}_0$ 代表的是背景模型 (UBM) 训练得到的均值超矢量, $\boldsymbol{m}_0$ 和 $\boldsymbol{m}(s)$ 的维数都是 $FC \times 1$. V 是说话人本征音 (Eigenvoice) 载荷矩阵, 是一个 $FC \times R_v$ 的矩阵, R_v 是说话人因子数; D 是残差载荷矩阵, 是用 V 空间无法拟合的每次说话形成的与个人特性相关的一个空间, 是一个 $FC \times FC$ 的对角阵; U 是信道载荷矩阵, 是一个 $FC \times R_u$ 的矩阵, R_u 是信道因子数; $\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}$ 分别是对应的因子. 式 (1) 在联合空间估计及最终的实验效果中发现存在以下问题: $D\boldsymbol{z}$ 矩阵在数值上非常小, 对实际的识别影响不大; 另外, 信道因子的数目 R_u 一般比较低, 在 30 ∼ 50 范围, 因子数增加对性能几乎没有任何提高; 本文中把这两种现象定义为饱和. 正是由于这种原

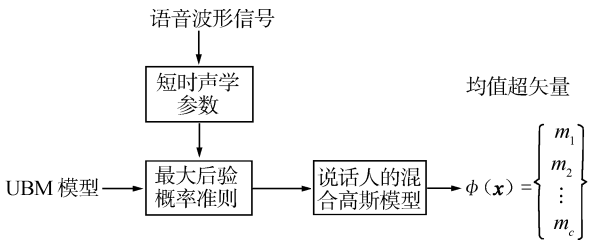

图 1 均值超矢量的形成过程

Fig. 1 The GMM mean supervector

收稿日期 2008-06-11 收修改稿日期 2009-01-03
Received June 11, 2008; in revised form January 3, 2009
国家自然科学基金 (60970161) 和多媒体计算与通信教育部-微软重点实验室科研基金资助 (07122803)
Supported by National Nature Science Foundation of China (60970161) and the Science Research Fund of Ministry of Education-Microsoft Key Laboratory of Multimedia Computing and Communication (07122803)
1. 中国科学技术大学多媒体计算与通信教育部 - 微软重点实验室 合肥 230027
1. Ministry of Education-Microsoft Key Laboratory of Multimedia Computing and Communication, University of Science and Technology of China, Hefei 230027
DOI: 10.3724/SP.J.1004.2009.01193

因, 大部分的因子分析系统[4−5] 都简化成式 (2):

$$\boldsymbol{m}(s) = \boldsymbol{m}_0 + \Delta\boldsymbol{y} + U\boldsymbol{x} \tag{2}$$

式 (2) 中将因子分析与 MAP 结合起来, 不去估计说话人因子部分, 而仅仅估计信道因子部分, 将影响说话人识别最大的信道因素去除. 在实际识别效果上, 式 (1) 和式 (2) 基本上差不多.

在文献 [6] 中, Kenny 试图解决式 (1) 中的饱和问题, 采用了单独估计 V、D 和 U 的方法, 但是在实现中发现仍然存在着饱和问题, 主要是无法将信道因子数目提升到比较大的数目. 另一个问题是, 采用文献 [6] 的算法在 U 空间估计时运算量非常繁杂. 在本文中, 采用完全独立的方法去估计三个空间, 对于信道空间采用多种信道单独估计最后拼接的方法, 可以解决空间饱和的问题, 另外由于采用相对简单的模型估计 U 矩阵, 运算量大大减少.

本文的安排如下: 第 1 节介绍 V、D 和 U 的估计, 第 2 节介绍模型训练及测试, 第 3 节是在 NIST SRE 2008 核心测试数据库上的实验结果, 最后一部分是分析和总结.

1 空间估计

1.1 说话人因子载荷矩阵估计

由于是采用串行的训练方式, 训练说话人载荷矩阵估计就是一个纯粹的本征音因子[7](Eigenvoice) 的训练过程, 采用式 (3) 来训练 V 空间, 不考虑 U、D 的影响, 是一个典型的 EM (Expectation maximization) 过程, 具体理论可以参见文献 [7].

$$\boldsymbol{m}(s) = \boldsymbol{m}_0 + V\boldsymbol{y} \tag{3}$$

对每个人 s 的声学特征参数 $\boldsymbol{x}_t$ 而言, 首先计算出相对 UBM 的均值超矢量 $\boldsymbol{m}_0$ 的统计量, $\boldsymbol{m}_c$ 代表 $\boldsymbol{m}_0$ 中的第 c 个高斯的均值超矢量, $\gamma_t(c)$ 是 $\boldsymbol{x}_t$ 相对每个高斯函数 c 的状态占有率, $\mathrm{diag}\{\cdot\}$ 表示取对角运算. 在以后的公式中, 如果变量有 c 下标, 就表示 GMM 第 c 个高斯函数对应的统计量, 如果没有 c 下标, 表示所有 GMM 的统计量; 有变量 s 就表示第 s 个说话人的统计量, 其他与此类同.

$$\begin{aligned} N_c(s) &= \sum_t \gamma_t(c) \\ \boldsymbol{S}_{X,c}(s) &= \sum_t \gamma_t(c)(\boldsymbol{x}_t - \boldsymbol{m}_c) \\ S_{XX^{\mathrm{T}},c}(s) &= \mathrm{diag}\{\sum_{\mathrm{t}} \gamma_t(c)(\boldsymbol{x}_t - \boldsymbol{m}_c)(\boldsymbol{x}_t - \boldsymbol{m}_c)^{\mathrm{T}}\} \end{aligned} \tag{4}$$

超向量 $\boldsymbol{S}_X(s)$ 是 $FC\times 1$ 维, $S_{XX^{\mathrm{T}}}(s)$ 是 $FC\times FC$ 维的矩阵, 由 $N_c(s)$ 为主对角线元素可以构成 $FC\times FC$ 维的方阵 $N(s)$, 可以分别看作语音相对于 UBM 模型的一阶、二阶和零阶统计量.

对所有说话人数据进行处理, 估计出每一段语音的说话人因子 $\boldsymbol{y}(s)$ 的一阶和二阶统计量的期望值,

$$\begin{aligned} L(s) &= I + V^{\mathrm{T}}\Sigma^{-1}N(s)V \\ \mathrm{E}[\boldsymbol{y}(s)] &= L^{-1}(s)V^{\mathrm{T}}\Sigma^{-1}\boldsymbol{S}_X(s) \\ \mathrm{E}[\boldsymbol{y}(s)\boldsymbol{y}^{\mathrm{T}}(s)] &= \mathrm{E}[\boldsymbol{y}(s)]\mathrm{E}[\boldsymbol{y}^{\mathrm{T}}(s)] + L^{-1}(s) \end{aligned} \tag{5}$$

其中 $L(s)$ 是临时变量, Σ 是 UBM 模型的协方差矩阵.

V 矩阵的更新公式如式 (6), 可利用文献 [7] 中的快速算法求出 V.

$$\sum_s N(s)V\mathrm{E}[\boldsymbol{y}(s)\boldsymbol{y}^{\mathrm{T}}(s)] = \sum_s \boldsymbol{S}_X(s)\mathrm{E}[\boldsymbol{y}^{\mathrm{T}}(s)] \tag{6}$$

在此基础上, 实现对原来 UBM 模型的协方差矩阵 Σ 的调整, 更新均值 $\boldsymbol{m}_0$ 不符合因子分析的理论, 因此只更新协方差矩阵 Σ:

$$\begin{aligned} \Sigma \;=\; & N^{-1}\sum_s S_{XX^{\mathrm{T}}}(s) - \\ & N^{-1}\mathrm{diag}\{\sum_s \boldsymbol{S}_X(s)\mathrm{E}[\boldsymbol{y}^{\mathrm{T}}(s)]V^{\mathrm{T}}\} \end{aligned} \tag{7}$$

矩阵 N 代表所有人 s 的零阶统计量之和.

采用式 (4)∼(7) 反复迭代 5∼6 遍, V 和 Σ 收敛.

1.2 说话人残差载荷矩阵估计

在估计完说话人载荷矩阵之后, 接下来的工作是估计说话人残差载荷矩阵, 采用式 (8) 估计 D, 其中 V, $\boldsymbol{m}_0$, 以及对应的协方差矩阵 Σ 已在前面得到, D 矩阵是一个 $FC\times FC$ 的对角阵, 相对而言, 运算量小很多.

$$\boldsymbol{m}(s) = \boldsymbol{m}_0 + V\boldsymbol{y} + D\boldsymbol{z} \tag{8}$$

同样也需要由式 (4) 得到相对于 UBM 的 $0, 1, 2$ 阶统计量 $N(s)$, $\boldsymbol{S}_X(s)$, $S_{XX^{\mathrm{T}}}(s)$, 并采用式 (5) 计算出说话人因子的一阶期望 $\mathrm{E}[\boldsymbol{y}(s)]$ 和二阶的期望 $\mathrm{E}[\boldsymbol{y}(s)\boldsymbol{y}^{\mathrm{T}}(s)]$, 在此基础上, 通过式 (9) 得到残差的一阶和二阶的统计量, $\boldsymbol{v}(s)$ 是一个中间变量.

$$\begin{aligned}\boldsymbol{v}(s) &= V\mathrm{E}[\boldsymbol{y}(s)]\\ \boldsymbol{d}_{X,c}(s) &= \boldsymbol{S}_{X,c}(s) - N_c(s)\boldsymbol{v}_c(s)\\ d_{XX^{\mathrm{T}},c}(s) &= \mathrm{diag}\{S_{XX^{\mathrm{T}},c}(s) - 2\boldsymbol{S}_{X,c}(s)\boldsymbol{v}_c^{\mathrm{T}}(s) + \\ &\quad N_c(s)\boldsymbol{v}_c(s)\boldsymbol{v}_c^{\mathrm{T}}(s)\}\end{aligned} \tag{9}$$

得到了残差的统计量之后, 可以采用与式 (5) 相似的式 (10) 求得残差因子的一阶和二阶期望值, $G(s)$ 是临时变量.

$$\begin{aligned}G(s) &= I + D^{\mathrm{T}}\Sigma^{-1}N(s)D\\ \mathrm{E}[\boldsymbol{z}(s)] &= G^{-1}(s)D^{\mathrm{T}}\Sigma^{-1}\boldsymbol{d}_X(s)\\ \mathrm{E}[\boldsymbol{z}(s)\boldsymbol{z}^{\mathrm{T}}(s)] &= \mathrm{E}[\boldsymbol{z}(s)]\mathrm{E}[\boldsymbol{z}^{\mathrm{T}}(s)] + G^{-1}(s)\end{aligned} \tag{10}$$

D 矩阵的更新公式如式 (11), 由于是对角阵, 运算非常简单.

$$\sum_s N(s)D\mathrm{E}[\boldsymbol{z}(s)\boldsymbol{z}^{\mathrm{T}}(s)] = \sum_s \boldsymbol{d}_X(s)\mathrm{E}[\boldsymbol{z}^{\mathrm{T}}(s)] \tag{11}$$

同样也只更新 UBM 模型的协方差矩阵 Σ:

$$\begin{aligned}\Sigma &= N^{-1}\sum_s d_{XX^{\mathrm{T}}}(s) -\\ &\quad N^{-1}\mathrm{diag}\{\sum_s \boldsymbol{d}_X(s)\mathrm{E}[\boldsymbol{z}^{\mathrm{T}}(s)]D^{\mathrm{T}}\}\end{aligned} \tag{12}$$

采用式 (9) ∼ (12) 反复迭带 5 ∼ 6 遍, D 和 Σ 收敛.

1.3　信道空间载荷矩阵估计

信道载荷矩阵的估计与说话人和残差载荷矩阵的估计采用相同的步骤, 但是在数据的组织上存在显著不同. 对于说话人和残差载荷矩阵估计而言, 理论上一个人只需要一句话就可以估计出两个空间, 而信道载荷矩阵需要一个人在各种信道情况下的多段话来估计, 在后面的公式中, 对于每个说话人 s, 不妨假设其有 $H(s)$ 段语音, 采用 h 下标来表示说话人的第 h 段语音. 另外一个问题是估计信道空间采用式 (1) 的标准公式计算量很大, 为了简化运算, 采用式 (2) 进行估计.

对于说话人 s 的第 h 段语音 X, 由式 (4) 得到相对于 UBM 的统计量 $N_h(s)$, $\boldsymbol{S}_{X,h}(s)$, $S_{XX^{\mathrm{T}},h}(s)$, 对所有说话人数据进行处理, 根据模型参数的初始值和训练数据, 利用式 (13) 估计出每一段语音的信道因子的一阶和二阶统计量, $J(s)$ 是一个临时变量.

$$\begin{aligned}J(s) &= I + U^{\mathrm{T}}\Sigma^{-1}N_h(s)U\\ \mathrm{E}[\boldsymbol{x}_h(s)] &= J^{-1}(s)U^{\mathrm{T}}\Sigma^{-1}\boldsymbol{S}_{X,h}(s)\\ \mathrm{E}[\boldsymbol{x}_h(s)\boldsymbol{x}_h^{\mathrm{T}}(s)] &= \mathrm{E}[\boldsymbol{x}_h(s)]\mathrm{E}[\boldsymbol{x}_h^{\mathrm{T}}(s)] + J^{-1}(s)\end{aligned} \tag{13}$$

利用得到的信道因子, 对第 s 人的所有 $H(s)$ 段语音, 利用类似 MAP 的方法估计说话人模型与 UBM 模型的变化量 $\Delta\boldsymbol{y}(s)$, 如式 (14), 其中 τ 因子就是 MAP 中的相关因子, 一般取 $8\sim 20$[1], 把所有 $\Delta\boldsymbol{y}_c(s)$ 串起来就是 $\Delta\boldsymbol{y}(s)$.

同样也只更新 UBM 模型的协方差矩阵 Σ.

$$\begin{aligned}\Delta\boldsymbol{y}_c(s) &= \frac{1}{N_c(s)+\tau}\{\sum_{h=1}^{H(s)}\boldsymbol{S}_{hc}(s,\boldsymbol{m}_c) -\\ &\quad \sum_{h=1}^{H(s)}N_{hc}(s)U_c(s)\mathrm{E}[\boldsymbol{x}_h(s)]\}\end{aligned} \tag{14}$$

$$\begin{aligned}\Sigma &= N^{-1}\{\sum_s\sum_{h=1}^{H(s)}\boldsymbol{S}_{X,h}(s) -\\ &\quad \mathrm{diag}\{\sum_s\sum_{h=1}^{H(s)}\boldsymbol{S}_{X,h}(s)\{\Delta\boldsymbol{y}^{\mathrm{T}}(s) + \mathrm{E}[\boldsymbol{x}_h^{\mathrm{T}}(s)]U^{\mathrm{T}}\}\}\end{aligned} \tag{15}$$

采用公式 (13) ∼ (15) 反复迭带 5 ∼ 6 遍, U 和 Σ 收敛.

对于复杂情况的信道空间的训练, 由于各种信道情况的数据会出现互相干扰影响, 在实际应用中如果明确知道信道的情况, 可以对于每种信道情况都训练一个信道空间 U_i, 最后把 K 个 U_i 拼起来形成一个大的信道矩阵 U. $U = [U_1\ U_2\ \cdots\ U_K]$. 由于是单独训练的 U_i, 每种情况下的协方差矩阵肯定不同, 为了维持大的信道因子矩阵 U 对应的协方差矩阵的一致性, 在这种情况下, EM 迭代中不采用式 (15) 更新 Σ.

2　说话人模型训练及测试

2.1　说话人模型训练

在 V, D, U 三个空间训练得到之后, 就可以用来训练说话人模型了. 对于任意一段语音 X, 最简单的方法就是采用空间训练中的方法用式 (5), (10), (13) 分别得到 $\boldsymbol{y}(s)$, $\boldsymbol{z}(s)$ 和 $\boldsymbol{x}(s)$, 然后用 $\boldsymbol{y}(s)$ 和 $\boldsymbol{z}(s)$ 采用式 (8) 拼接得到说话人 s 的 GMM 均值超矢量, $\boldsymbol{x}(s)$ 是需要去除的信道部分, 不在说话人 GMM 中, 但是在实验中发现识别的效果非常差. 在说话人模型的训练中, 需要采用文献 [8] 相同的算法, 也就是联合估计 $\boldsymbol{y}(s)$, $\boldsymbol{z}(s)$ 和 $\boldsymbol{x}(s)$ 的方法.

首先把 V, D, U 拼接成一个矩阵 W, 三个需要估计的因子 $\boldsymbol{y}(s)$, $\boldsymbol{z}(s)$ 和 $\boldsymbol{x}(s)$ 也拼接成一个大

矩阵 $\boldsymbol{f}(s)$, 如式 (16). 如果把式 (5) 中的 V 替换成 W, 把 $\boldsymbol{y}(s)$ 替换成 $\boldsymbol{f}(s)$, 那么采用式 (5) 就能够求出因子 $\boldsymbol{f}(s)$, 从而完成说话人的建模. 在实际运算过程中, W 矩阵是一个非常大的矩阵, 维度为 $FC \times (R_v + R_u + FC)$. 相应地, 式 (5) 中的 $L^{-1}(s)$ 计算是非常繁杂的, 由于 D 矩阵是对角阵, 可以简化, 具体见文献 [8] 的附录.

$$W = [V \quad D \quad U]$$
$$\boldsymbol{f}(s) = [\boldsymbol{y}^{\mathrm{T}}(s) \quad \boldsymbol{z}^{\mathrm{T}}(s) \quad \boldsymbol{x}^{\mathrm{T}}(s)]^{\mathrm{T}} \tag{16}$$

2.2 测试

在模型训练好以后, 测试过程就相对简单多了. 对一个已经训练好的模型 Λ, 需要测试的语句为 X, 不妨假设 X 就是 Λ, 仅仅是信道不同罢了, 在这个假设前提下求对数似然度函数 $\log(X|\Lambda)$. 由于仅仅是信道的不同, 只需要采用说话人模型训练中同样的步骤求出 $\boldsymbol{x}(s)$, 再采用文献 [8] 中的式 (19) 就可以很容易地求得对数似然度函数得分, 这种运算方法与文献 [1] 中的 TOP-N 快速算法不矛盾, 如果一句话需要对多个模型进行测试, 那么也可以采用文献 [1] 中的 TOP-N 快速算法.

3 实验配置与实验结果

3.1 测试数据库描述

本文使用 NIST SRE 2008 中的核心测试 (Core test) 作为实验数据库[9], 该部分数据的训练集叫做 Short 2, 主要包括两部分数据, 第 1 部分是电话信道录制的语音 (Telephone), 第 2 部分是从一个复杂面试场景中的麦克风阵列中录制的语音 (Interview); 测试的语音叫做 Short 3, 除了与 Short 2 相同的两种情况之外, 还包括与电话并行连接的麦克风录制的语音 (Microphone), 这部分语音没有经过电话线. 训练和测试语音的组合包括以下 5 种: 1) 面试训练−面试测试; 2) 面试训练−电话测试; 3) 电话训练−电话测试; 4) 电话训练−面试测试; 5) 电话训练−麦克风测试. 这 5 部分测试内容包含总计 98 766 次测试, 3 263 个说话人模型, 可以说不管是模型的人数、测试数目还是信道的复杂度都相对以前的比赛复杂得多. 另外面试数据是以前从来没有出现的数据, NIST 给出了男女各 3 人的 Mix5 数据库作为开发集.

3.2 开发集的选择

本实验选用的数据库包括 SwitchBoard phase II, III, SwitchBoard Cellular, NIST SRE 2004, NIST SRE 2005, NIST SRE 2006, NIST 给的 6 个人的 Mix5 开发集的数据库. 其中采用 NIST SRE 2004 年的 1side 数据库训练 UBM 模型, 采用 SwitchBoard 数据库训练说话人空间 V, 采用 NIST SRE 2004 年数据训练残差空间 D, 采用 NIST SRE 2004, 2005, 2006, Mix5 数据库来训练信道空间 U.

需要注意的是 Mix5 男女数据都只有 3 个人, 每个人在 9 种信道的情况下各有 180 分钟的录音, 我们把 180 分钟的语音按照 5 分钟一段分成 36 段, 每个人总计有 $9 \times 36 = 324$ 段语音, 再采取随机挑选的方法把这 324 段语音映射到 18 个假定的说话人, 每个人有 18 段语音, 通过映射 Mix5 男女声各有 54 个说话人. 这样可以很方便地应用于信道空间的训练.

3.3 声学特征参数

本实验采用感知线性预测参数 (Perceptual linear predictive, PLP). 对于 PLP 参数提取, 首先预加重 (因子为 0.97), 经过帧宽 20 ms, 帧移是 10 ms 的汉明窗. 采用基于能量的寂静帧检测算法去除静音帧. 抽取 $0 \sim 12$ 维 PLP, 总计为 13 维, 特征参数通过去均值 (Cepstrum mean subtraction, CMS) 去除信道卷积噪声, 通过一阶差分、二阶差分总计构成 39 维, 最后通过高斯化以提高识别率.

3.4 系统描述

本实验是男女性别分开单独进行 (Gender-dependent), 首先用 NISTSRE 2004 中 1side 数据分别训练一个 1 024 的男声和女声 UBM 模型, 以后所有的训练测试过程都是男女性别分开进行的.

采用 SwitchBoard phase II, III, SwitchBoard Cellular 部分数据训练说话人载荷矩阵 V, 女声选取了 846 人, 男声选取了 742 个说话人, 每人平均有 11 句话左右; 说话人因子 R_v 取 300.

残差载荷矩阵 D 的训练采用 NISTSRE 2004 的所有说话人, 由于 2004 年的数据总计只有 310 人 (女生 186 人, 男生 124 人), 因此所有这些人都挑选了进行训练, 但是对每个人的语句进行了限制, 一个人最多不超过 18 句话.

由于 NISTSRE 2008 数据分为多种情况, 因此信道载荷矩阵 U 采用拼接的方法训练得到. 首先采用 NIST SRE 2004、2005、2006 的所有电话语音数据训练一个因子数为 100 的载荷矩阵 U_1, 该部分空间对应的训练或测试数据的麦克风情况. 采用 NIST SRE 2005 和 NIST SRE 2006 的麦克风数据训练一个因子数为 50 的载荷矩阵 U_2, 该部分空间对应的训练或测试数据的麦克风情况. 采用 Mix5 的数据训练一个因子数为 50 的载荷矩阵 U_3, 该部分空间对应的训练或测试数据的面试情况. 最后把 U_1, U_2, U_3 拼接形成一个信道因子数为 200 的大载荷矩阵 u 用于模型的训练和测试.

对于最终的得分采用 ZTnorm 的方法进行规

整, 也就是对得分先做 Znorm, 然后做 Tnorm. 数据选择是从 NIST SRE 2005、NIST SRE 2006 的电话和麦克风数据中各随机挑选 1 000 句话, 总计是 4 000 句话 (有些语句是同一个人所说, 我们不去考虑); 男女性别的语句大约各半, 由于 NIST 是男女性别分开测试的, 因此我们的得分规整也是采用性别相关的规整策略. 具体策略如下:

1) NIST SRE 2008 年测试有 3 263 句话来训练说话人模型, 另外我们选用了 NIST SRE 2006 的电话和麦克风的 2 000 句话做 Tnorm, 我们首先将这 5 263 句话按照性别不同采用第 2.1 节所述的说话人模型训练过程来映射得到说话人模型.

2) 采用 NIST SRE 2005 的 2 000 句话来分别对这 5 263 个说话人模型测试得到 Znorm 的参数, 如果模型是电话的情况, 采用 NIST SRE 2005 的电话数据计算, 否则就采用 NIST SRE 2 005 的麦克风数据进行计算, 这样可以得到 5 263 个模型的 Znorm 的规整参数 (均值, 标准差).

3) 测试时, 如果测试的语句是电话的情况, 选用上面的 NIST SRE 2006 的电话数据得到的说话人模型作为 Tnorm 的冒认者模型, 对于测试的语句是麦克风和面试的情况, 选用上面的 NIST SRE 2006 的麦克风数据得到的说话人模型作为 Tnorm 的冒认者模型, 每次计算的得分都是先采用 Znorm 的规整参数规整后, 然后再做 Tnorm.

另外, 我们建立了两套系统与本文算法进行对比.

第一套系统是 GMM-UBM 系统, 该系统作为基线系统, 高斯数为 1 024, 不采用因子分析, 采用 MAP 由 UBM 模型训练得到说话人模型, 测试中采用标准的对数似然度算法求得分, 得分采用 NIST 2006 的男女各 1 000 句话做 Tnorm 规整.

第二套系统采用式 (2) 建立的一套因子分析系统, 高斯数仍为 1 024; 信道空间 U 的因子数为 100, 信道空间训练的数据采用上述三个信道空间训练的所有数据; 采用 ZTnorm, 所有数据和规整策略也与上面的实验一致.

3.5 实验结果

本文采用等错误率 (Equal error rate, EER) 和规整后 (Cnorm) 的最小检测错误代价 (Minimum detect cost function, MDCF) 对系统进行评价[9], 表 1 列出了采用常规的 GMM-UBM 系统, 采用式 (2) 的因子分析系统, 以及本文推荐的因子分析的结果.

对比表 1 中的三个系统, 可以发现采用因子分析的系统性能明显优于传统的 GMM-UBM 系统. 而从采用因子分析的两个系统性能比较来看, 不管是 EER, 还是 MDCF, 采用本文中推荐的因子分析的系统性能都要优越一些, 五个部分 EER 减少的幅度在 10 % ∼ 26 % 之间. 面试−面试部分性能提高最大, 采用本文提出的算法可以更好地拟合面试部分的信道特征. 而在采用式 (2) 的信道空间训练中, 由于 Mix5 数据比较少 (只有几百句话), 被掩盖在电话语音的数据中, 因此性能相对较差.

表 1　NIST SRE 2008 不同任务上性能比较

Table 1　The performance of different tasks on NIST SRE 2008

测试情况	GMM-UBM		传统的因子分析		本文算法	
训练−测试	EER	MDCF	EER	MDCF	EER	MDCF
面试−面试	8.2 %	0.321	4.5 %	0.172	3.3 %	0.118
面试−电话	8.9 %	0.383	5.7 %	0.263	5.1 %	0.219
电话−电话	9.2 %	0.450	5.7 %	0.308	5.0 %	0.227
电话−麦克风	9.4 %	0.377	6.0 %	0.237	5.3 %	0.219
电话−面试	8.8 %	0.411	5.9 %	0.278	5.0 %	0.247

4 讨论和结论

从因子分析的基本理念来看[10], 在式 (1) 的联合估计中存在一个逻辑上的问题, D 虽然是对角阵, 但是是一个满秩的空间, 必然会与 U、V 两个空间出现重叠. 只是在具体运算每句话的时候把对角阵 D 隐含地作为了 U、V 两个空间描述剩下的残差部分, 因此在估计的时候即使有些部分无法由 U、V 很好描述, 由于采用的是最大似然度概率准则, 会尽量地把本来是每个人个性的一些东西往 U、V 两个方向上引导, 因此会使得 D 代表的部分尽量得小. 另外因子分析中最大似然的原则求出的 U、V 两个空间不是正交的, 也就是每个空间找到的本征矢量并不代表本身空间最大的那些本征值. 在串行的估计中, 式 (2) 和式 (3) 的估计, 都把自身空间无法估计的部分放到残差部分, 因此 U、V 求出的每个子空间中的那些本征矢量都对应着 (旋转后的) 最大的本征值. 通过式 (8) 把对角阵 D 作为一个空间来估计, 拉大了 U、V 两个空间不能描述的部分, 达到在说话人建立说话人模型的时候更能突出说话人个性的东西的目标.

按照因子分析在信道处理中的应用, 不要求每个说话人有所有信道情况下的数据, 但是要求任意选择几个人的数据 (人数不能太大, 比如随机选择的 4 ∼ 5 个人) 能够覆盖所有的信道情况, 这样才能把信道的矩阵通过最大似然准则估计出来. 而 NIST SRE 2008 主要分为三种不同类型的信道, 分布在三种不同情况下的人完全不同, 如果采用把所有人的录音数据聚集在一起, 在最大似然准则的情况下, 训练出来的信道空间应该是三种信道平均的情况 (或

者往数据多的那部分空间倾斜), 空间描述相对也不准确, 信道因子数目上也无法达到本实验中的 200 个, 如果我们采用已有的先验知识, 人为地把信道空间的数据分开单独训练不同矩阵进行拼接, 这样可以把信道因子数目增大, 相对空间描述也准确一些. 当然, 这三个空间会出现不正交或者重叠的可能性, 但是在模型注册和测试的时候采用的是最大似然度的策略来求因子的, 而不是像 PCA 那样采用投影的方法来去除信道的影响, 因此这种影响不会很大. 从实验的结果以及与国际同行的比较中发现, 这种算法还是有一定的优势.

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 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

3 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

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

5 Castaldo F, Colibro D, Dalmasso E, Laface P, Vair C. Compensation of nuisance factors for speaker and language recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**(7): 1969−1978

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

7 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

8 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

9 NIST. The NIST Year 2008 Speaker Recognition Evaluation Plan [Online], available: http://www.nist.gov/speech/tests/sre/2008/index.html, March 20, 2008

10 Bishop C M. *Pattern Recognition and Machine Learning*. Berlin: Springer, 2008. 583−586

郭 武 中国科学技术大学电子工程与信息科学系讲师. 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: guowu@mail.ustc.edu.cn
(**GUO Wu** Lecturer in the Department of Electronic Engineering and Information Science, University of Science and Technology of China. His research interest covers speaker recognition and language identification. Corresponding author of this paper.)

李轶杰 中国科学技术大学硕士研究生. 主要研究方向为说话人识别中的信道处理. E-mail: andylyj@mail.ustc.edu.cn
(**LI Yi-Jie** Master student in the Department of Electronic Engineering and Information Science, University of Science and Technology of China. His research interest covers session variability in speaker recognition.)

戴礼荣 中国科学技术大学电子工程与信息科学系教授. 主要研究方向为语音识别和信号处理.
E-mail: lrdai@ustc.edu.cn
(**DAI Li-Rong** Professor in the Department of Electronic Engineering and Information Science, University of Science and Technology of China. His research interest covers speech recognition and signal processing.)

王仁华 中国科学技术大学电子工程与信息科学系教授. 主要研究方向为文语合成和语音识别.
E-mail: rhw@ustc.edu.cn
(**WANG Ren-Hua** Professor in the Department of Electronic Engineering and Information Science, University of Science and Technology of China. His research interest covers speech synthesis and speech recognition.)