

正交拉普拉斯语种识别方法

杨绪魁¹ 屈丹¹ 张文林¹

摘 要 提出了一种正交拉普拉斯语种识别方法,即在提取语音的 i-vector 后,采用正交局部保持投影进行子空间映射,将信号整体空间映射到语言信息加信道信息子空间,然后对映射后的矢量进行信道补偿处理,最后用支持向量机进行识别. 尽管 i-vector 最大限度地保留了语音的声学信息,但是并没有发现这些信息之间的内在结构. 利用正交局部保持投影在去除声学无关信息的基础上,进一步发现声学特征的内在结构,能够有效地提高特征的可区分性. 在对 NIST LRE 2003 测试数据库实验后,发现新方法相较于基线系统来说,平均代价降低了 28.91%.

关键词 因子分析, 辨识矢量, 流形学习, 正交局部保持投影, 语种识别

引用格式 杨绪魁, 屈丹, 张文林. 正交拉普拉斯语种识别方法. 自动化学报, 2014, 40(8): 1812–1818

DOI 10.3724/SP.J.1004.2014.01812

An Orthogonal Laplacian Language Recognition Approach

YANG Xu-Kui¹ QU Dan¹ ZHANG Wen-Lin¹

Abstract An orthogonal Laplacian language recognition approach is proposed. In this approach, the i-vector of an utterance, after being extracted, is mapped into a subspace by an orthogonal locality preserving projection. Then, channel compensation is done for the mapped vector. At last, recognition is done with a support vector machine. Though the i-vector preserves the acoustics information as much as possible, it cannot find the inner structure among this information. Whereas the intrinsic structure of acoustics feature can be found by the orthogonal locality preserving projection algorithm on the basis of removing the irrelevant information. Experiments on the NIST LRE 2003 evaluation corpus show that this new approach can reduce a 28.91% average detection cost compared to the baseline.

Key words Factor analysis, identifying vector, manifold learning, orthogonal locality preserving projection, language recognition

Citation Yang Xu-Kui, Qu Dan, Zhang Wen-Lin. An orthogonal Laplacian language recognition approach. *Acta Automatica Sinica*, 2014, 40(8): 1812–1818

基于并行音素识别器结合语言模型 (Parallel phone recognition language model, PPRLM) 和基于统计模型是当前语种识别的主要方法^[1]. 高斯混合模型 (Gaussian mixture model, GMM) 利用最底层的声学信息, 根据特征向量空间的概率统计分布构建与语种相关的模型, 由于其良好的统计表征能力及优越的性能, 且不需要专门的语言学知识, 使得高斯超矢量支持向量机 (GMM-supervector support vector machine, GSV-SVM)^[2] 方法成为统计模型的典型代表得到广泛应用. 模型中 GMM 自适

应算法将低维的声学特征映射到高维空间构造超矢量, 带来了性能提升的同时也增加了后端识别的运算复杂度. 但更重要一点, 由于语种识别性能常常受说话人、信道环境、噪声等影响, 而 GMM 在映射的过程中, 并没有将信道、说话人、环境等与语种识别无关的信息进行有效抑制, 因此在复杂环境下, GSV-SVM 模型的性能受到较大影响.

联合因子分析 (Joint factor analysis, JFA)^[3–4] 作为一种信道补偿及降维技术在说话人识别中得到广泛应用. 但是由于 JFA 模型结构复杂, 为了发挥它的优势, 需要大量的带标注的训练数据, 同时算法的运算量也非常大. 辨识矢量 (identifying-vector, i-vector) 方法^[5–6] 作为 JFA 的改进算法, 降低了算法复杂度, 并且可以用无监督的方法训练 i-vector 模型, 在语种识别相关实验中证明, 该算法具有优良的性能.

i-vector 算法实质上是概率主分量分析 (Probabilistic principle component analysis, PPCA)^[7] 的典型应用. 作为主分量分析 (PCA)^[8] 的概率化模型, PPCA 在计算数据主分量的过程中, 不需要计

收稿日期 2013-03-12 录用日期 2013-08-28
Manuscript received March 12, 2013; accepted August 28, 2013
国家高技术研究发展计划 (863 计划) (2012AA011603), 国家自然科学基金 (61175017), 全军军事学研究生课题 (2010JY0258-144) 资助
Supported by National High Technology Research and Development Program of China (863 Program) (2012AA011603), National Natural Science Foundation of China (61175017), Research of The Military Science Graduate of PLA (2010JY0258-144)
本文责任编辑 吴玺宏
Recommended by Associate Editor WU Xi-Hong
1. 解放军信息工程大学信息工程学院 郑州 450002
1. Institute of Information Engineering, PLA Information Engineering University, Zhengzhou 450002

算数据协方差矩阵的特征向量, 因此对数据的维数没有限制. 但同 PCA 一样, PPCA 属于流形学习 (Manifold learning)^[9] 中的线性方法. 这类方法能够有效地发现高维观测空间中的全局线性特性. 但是当观测数据中的低维流形是非线性时, 这类方法难以发现高维数据的内在结构. 而语音中的语种相关的信息不是一个简单的线性结构可以描述的, 因此 i-vector 最大限度的保留了语音的声学信息, 但并没有发现这些信息之间的内在结构, 因此系统的识别性能必然受到影响.

文献 [10] 在 i-vector 算法基础上提出了一种新的因子分析方法. 该方法在提取语音信号的 i-vector 后, 用局部保持投影 (Locality preserving projection, LPP)^[11-12] 算法对 i-vector 进行降维. 实验证明, 该方法较 i-vector 算法而言, 性能有所提升, 尤其是在女性说话人识别中, 性能提升明显.

LPP 和正交局部保持投影 (Orthogonal locality preserving projection, OLPP)^[13-14] 作为全局线性化的局部保持的流形学习算法, 利用数据样本点的邻接图对流形结构进行建模, 明确地考虑了高维空间中的局部结构, 具有优良的局部保持能力. 基于 LPP 和 OLPP, 何小飞等^[12] 和蔡登等^[14] 等在人脸识别中分别提出了拉普拉斯脸 (Laplacianface) 和正交拉普拉斯脸 (Orthogonal laplacianface). 这类方法首先用主分量分析 (Principle component analysis, PCA) 对高维数据进行投影、降维去噪的同时, 确保了数据协方差矩阵非奇异; 然后利用 LPP 或 OLPP 发现数据之间的内在局部结构. 在分类中, 这种局部结构具有良好的区分性, 因此这类方法具有优良的性能. 同时, 由于 OLPP 保证了基函数的正交性, 提升了算法的局部保持能力, 也使得算法性能对低维流形的维数不敏感.

本文将 OLPP 引入基于 i-vector 的语种识别中, 提出了正交拉普拉斯语种识别方法. 一方面, 这种方法在利用 i-vector 算法去除声学无关信息的基础上, 进一步应用 OLPP 算法发现声学特征的内在结构, 提高了数据的可分性, 最后使用信道补偿技术去除语种无关信息的影响. 另外一方面, 本文将 i-vector 替代正交拉普拉斯脸中的 PCA, 使得该方法在 GSV 的维数十分高的情况下仍能适用. 在 NIST LRE 2003 上的测试表明, 与基于 i-vector 的基线系统性能相比, 新系统的性能较基线系统有了一定的提高.

本文的组织如下: 第 1 节对 i-vector 的模型假设及参数估计方法进行了介绍; 第 2 节对正交拉普拉斯语种识别方法进行了介绍; 第 3 节介绍实验设置及实验结果; 最后为结论部分.

1 语种识别的 i-vector 算法

1.1 i-vector 简介

i-vector 是说话人确认领域中最先进的技术, 同时在语种识别中得到了成功的应用^[5-6]. i-vector 将输入的高维特征向量映射到一个低维的特征空间中, 同时保留输入特征绝大部分的相关信息. 它的主要思想是, 一个与语种、信道都有关的高斯混合模型均值超矢量 $\mathbf{M}$ 可以表示成:

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

其中, $\mathbf{m}$ 是一个与语种、信道都无关的均值超矢量, T 是由张成保留超矢量空间主成分的子空间的一组基构成的矩阵, $\mathbf{w}$ 是一个服从正态分布的隐变量, 而 i-vector 就是 $\mathbf{w}$ 的点估计.

1.2 模型参数估计

对于 GMM 模型中任意一个混元 c , 令 $\mathbf{M}_c$ 为 $\mathbf{M}$ 中对应混元 c 的均值向量, Σ_c 为协方差矩阵, ω_c 为混元权重. 令 Σ 为以 $\Sigma_1, \Sigma_2, \dots, \Sigma_C$ 为对角块的 $CF \times CF$ 维块对角阵, 权重向量 $\Omega = [\omega_1, \omega_2, \dots, \omega_C]$, 其中 C 为 GMM 模型混元数, F 为声学特征维数. 令 $\chi(h) = \{x_t^h\}_{t=1,2,\dots,T}$ 为训练集中任意一段训练语音 h 的特征矢量, $P(\chi(h)|\Omega, \mathbf{M}, \Sigma)$ 为 $\chi(h)$ 的似然概率, 表示语音 h 在 i-vector 模型下的概率, 其中 h 为训练集中语音的标号, t 为语音特征的帧标号, T 为语音段长度. 因此, i-vector 的目标函数为所有语音段的整体概率最大化, 即:

$$\prod_h \int_{\mathbf{w}(h)} P(\chi(h)|\Omega, \mathbf{m} + T\mathbf{w}(h), \Sigma) P(\mathbf{w}(h)) d\mathbf{w}(h) \quad (2)$$

该优化问题可以通过 EM 算法迭代求解, 具体步骤如下^[3-4]:

步骤 1 (E-step). 对于训练集中任意一段语音 h , 给定 T 和 Σ 的当前估计值, 获取 $\mathbf{w}(h)$ 的后验概率分布, 即

$$P(\chi(h)|\Omega, \mathbf{m} + T\mathbf{w}(h), \Sigma, \chi(h))$$

步骤 2 (M-step). 通过最大化式 (3), 更新 T 和 Σ .

$$\sum_h \int_{\mathbf{w}(h)} \{P(\chi(h)|\Omega, \mathbf{m} + T\mathbf{w}(h), \Sigma, \chi(h)) \log P(\chi(h)|\Omega, \mathbf{m} + T\mathbf{w}(h), \Sigma)\} d\mathbf{w}(h) \quad (3)$$

估计出矩阵 T 后, 对于 i-vector 的计算, 即为计算使得 $\chi(h)$ 似然概率最大的 $\mathbf{w}(h)$, 即,

$$\mathbf{w}(h) = \arg \max P(\chi(h)|\Omega, \mathbf{m} + T\mathbf{w}, \Sigma)$$

2 正交拉普拉斯语种识别方法

本文提出了正交拉普拉斯语种识别方法, 其系统框图如图 1 所示. 在建立通用背景模型 (Universal background model, UBM) 基础上, 训练阶段, 首先采用 EM 算法进行 i-vector 的训练, 其次采用 OLPP 流形学习方法进行子空间映射, 然后将映射后的矢量采用信道补偿进行处理, 最后送入支持向量机用于训练; 识别阶段, 将特征矢量经过同样的处理用于识别.

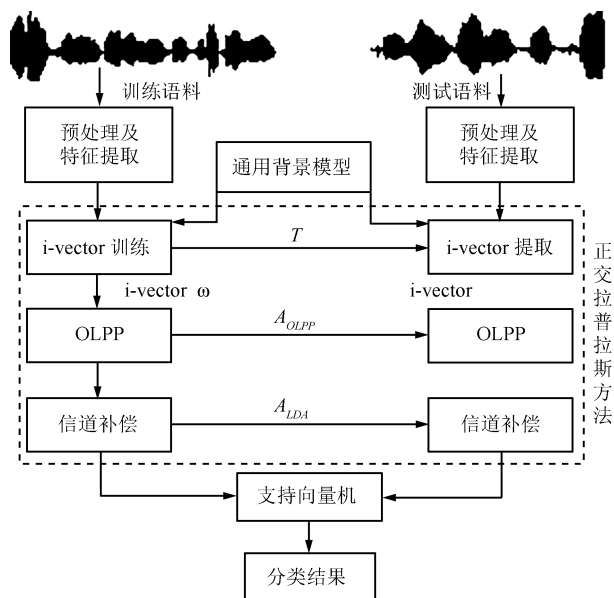


图 1 正交拉普拉斯语种识别方法
Fig. 1 Orthogonal Laplacian language recognition method

如图 1 所示, 本文所提出的正交拉普拉斯语种识别方法包括三个部分: i-vector 矢量获取、OLPP 映射以及信道补偿算法. 同正交拉普拉斯脸算法^[14]相比, 本文提出的方法用 i-vector 代替了正交拉普拉斯脸算法中的 PCA 映射. i-vector 的理论基础是 PPCA, PPCA 在 PCA 的基础上引入了概率模型, 使得观测数据的主轴可以由一个与因子分析 (Factor analysis, FA) 类似的隐变量最大似然估计方法得到. 因此, 主分量的计算是通过 EM 算法迭代计算. 由于不需要计算观测数据协方差矩阵的特征向量, 因此观测数据的维数不必受限.

PCA 等传统的线性降维技术在高维观测数据具有线性结构时, 能够在高维空间和低维空间之间建立显式的线性变换, 但对于数据集中的非线性几何结构不能很好的反映. 拉普拉斯特征映射 (Laplacian eigenmaps, LE) 等局部保持的流形学习方法依照局部到整体的思想, 保持高维空间和内在嵌入空间的局部几何共性, 发现高维空间中的低

维流形. 但这些算法建立的映射都是隐式的非线性映射, 只定义在训练数据上, 新的观测数据不能快速明确地映射到低维空间中, 因此限制了其应用. LPP 作为 LE 的线性化算法, 其原理是限制高维空间到低维空间的变换为线性投影变换, 然后通过与 LE 类似的方法得到最优的投影方向. 因此, LPP 不仅保持了高维空间的局部几何特性, 而且具有线性子空间投影的优点. 虽然 LPP 得到的基函数 (Basis function) 是低维流形中拉普拉斯-贝尔特拉米算子 (Laplace Beltrami operator) 特征函数的线性近似, 但是这些基函数是非正交的. OLPP 算法保证了基函数的正交性, 保持了流形空间的度量结构, 因此, 局部保持能力优于 LPP 算法. 由于算法的局部保持能力与算法的区分性能直接相关, 因此 OLPP 算法的区分性能优于 LPP 算法.

基于以上原因, 因此本文采用 i-vector 方法与 OLPP 进行结合, 考虑到信道的进一步影响, 还需要采用 LDA 等信道补偿技术.

2.1 局部保持投影

令 $W = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_C]$ 为 i-vector 数据集, 其中 $\mathbf{w}_i$ ($i = 1, 2, \dots, n$) 代表语音段的 i-vector 矢量. 则 LPP 的实现步骤如下^[11]:

步骤 1. 近邻选择, 构造邻接图 G ;

步骤 2. 计算近邻相似度矩阵 H ;

步骤 3. 通过最小化近邻保持的目标函数, 计算投影向量, 即

$$\min \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\mathbf{y}_i - \mathbf{y}_j)^2 H_{ij}$$

$$\text{s.t. } \mathbf{Y}\mathbf{Y}^T = \mathbf{I}$$

可以证明, 投影向量可以通过如下特征方程^[11]

$$\mathbf{W}\mathbf{L}\mathbf{W}^T\boldsymbol{\alpha} = \lambda\mathbf{W}\mathbf{D}\mathbf{W}^T\boldsymbol{\alpha}$$

的 d 个最小特征值对应的特征向量得到. 其中, $\mathbf{Y}$ 为数据集的低维全局坐标, $\mathbf{L} = \mathbf{D} - \mathbf{H}$, $\mathbf{D}$ 为对角阵, 其对角线元素为

$$D_{ii} = \sum_j H_{ij}$$

步骤 4. 设特征方程的 d 个最小特征值对应的特征向量为 $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_d$, 则构成保持近邻相似度特性的线性变换矩阵为

$$\mathbf{A}_{\text{LPP}} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_d]$$

因此, $\mathbf{Y} = \mathbf{A}_{\text{LPP}}^T \mathbf{W}$.

2.2 正交局部保持投影

令投影映射 $\mathbf{a} \in \mathbf{R}^{CF}$, 则局部保持函数定义如下^[13]:

$$f(\mathbf{a}) = \frac{\mathbf{a}^T \mathbf{W} \mathbf{L} \mathbf{W}^T \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{D} \mathbf{W}^T \mathbf{a}} \quad (5)$$

若直接最小化式 (5) 可以导出 LPP 算法. OLPP 算法的目标是找到一组正交基使得局部保持函数最小. 因此, OLPP 算法的目标函数可以表示为:

$$\mathbf{a}_1 = \arg \min_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{W} \mathbf{L} \mathbf{W}^T \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{D} \mathbf{W}^T \mathbf{a}} \quad (6)$$

和

$$\begin{aligned} \mathbf{a}_k &= \arg \min_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{W} \mathbf{L} \mathbf{W}^T \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{D} \mathbf{W}^T \mathbf{a}} \\ \text{s.t. } \mathbf{a}_k^T \mathbf{a}_1 &= \mathbf{a}_k^T \mathbf{a}_2 = \cdots = \mathbf{a}_k^T \mathbf{a}_{k-1} = 0, \\ k &= 2, 3, \cdots, d \end{aligned} \quad (7)$$

容易证明, $\mathbf{a}_1$ 是式 (8) 所示特征方程最小特征值对应的特征向量^[13].

$$\mathbf{W} \mathbf{L} \mathbf{W}^T \mathbf{a} = \lambda \mathbf{W} \mathbf{D} \mathbf{W}^T \mathbf{a} \quad (8)$$

如果 $\mathbf{W} \mathbf{D} \mathbf{W}^T$ 非奇异, 则 $\mathbf{a}_1$ 是矩阵 $(\mathbf{W} \mathbf{D} \mathbf{W}^T)^{-1} \times \mathbf{W} \mathbf{L} \mathbf{W}^T$ 最小特征值对应的特征向量.

对于第 k 个基向量 $\mathbf{a}_k$ 是矩阵 $M^{(k)}$ 最小特征值对应的特征向量, 其中

$$\begin{aligned} M^{(k)} &= \{I - (\mathbf{W} \mathbf{D} \mathbf{W}^T)^{-1} A^{(k-1)} [B^{(k-1)}]^{-1} \times \\ &\quad [A^{(k-1)}]^T\} (\mathbf{W} \mathbf{D} \mathbf{W}^T)^{-1} \mathbf{W} \mathbf{L} \mathbf{W}^T \\ A^{(k-1)} &= [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{k-1}] \\ B^{(k-1)} &= [B_{ij}^{(k-1)}] \\ B_{ij}^{(k-1)} &= \mathbf{a}_i^T (\mathbf{W} \mathbf{D} \mathbf{W}^T)^{-1} \mathbf{a}_j \end{aligned}$$

将得到的正交基 $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_d$ 组合构成 OLPP 投影矩阵 A_{OLPP} , 即

$$A_{OLPP} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_d]$$

因此, $Y = A_{OLPP}^T W$.

2.3 信道补偿技术

由于 i-vector 算法投影得到的子空间不仅包含了语种的信息, 也包含了信道、说话人的信息. 这些信息的存在对系统的性能有一定的影响, 因此, 必须用信道补偿技术将这些信息去除, 常用的信道补偿方法有线性鉴别分析 (Linear discriminant analysis, LDA)^[5-6] 和类内方差规整 (Within class covariance normalization, WCCN)^[15]. 本文选用 LDA 作为信道补偿算法.

LDA 的目标是寻找一组基, 使得类间方差最大的同时类内方差最小. 定义 LDA 的投影矩阵为 A_{LDA} , 则该投影矩阵的求解可以通过如下特征方程得到:

$$\begin{aligned} \Sigma_b \mathbf{v} &= \Lambda \Sigma_w \mathbf{v} \\ \Sigma_b &= \sum_{l=1}^L (\bar{\mathbf{y}}_l - \bar{\mathbf{y}})(\bar{\mathbf{y}}_l - \bar{\mathbf{y}})^T \\ \Sigma_w &= \sum_{l=1}^L \sum_{i=1}^{n_l} (\mathbf{y}_l^i - \bar{\mathbf{y}}_l)(\mathbf{y}_l^i - \bar{\mathbf{y}}_l)^T \\ \bar{\mathbf{y}}_l &= \frac{1}{n_l} \sum_{i=1}^{n_l} \mathbf{y}_l^i \end{aligned}$$

其中, Λ 是对角阵, 其对角线元素为特征值. Σ_b 和 Σ_w 分别表示类间方差矩阵和类内方差矩阵, L 为语种种类数, n_l 为第 l 种语言的样本点数, $\bar{\mathbf{y}}_l$ 为第 l 种语言特征超矢量的均值, $\bar{\mathbf{y}}$ 为所有特征超矢量的均值. 将求解得到的特征向量 $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_L$ 组合构成信道补偿矩阵 A_{LDA} , 即

$$A_{LDA} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_L]$$

3 实验

3.1 实验数据

美国国家标准与技术研究院 (National Institute of Standards and Technology, NIST) 自 1996 年举行语种识别大赛以来, 其测试数据库逐步贴近实际的应用情况, 甚至超过实际应用中的复杂程度. 本文实验选用 NIST LRE 2003 作为实验测试集. 对于 NIST LRE 2003 数据库, 其包含了 12 种目标语言, 分别是: 英语、阿拉伯语、波斯语、法语、汉语、德语、日语、西班牙语、韩语、北印度语、泰米尔语、越南语; 包含了一种非目标语种 (俄语). 本文选用时长 30 秒的数据集作为测试数据集.

选用 Call Friend 和 Call Home 数据库作为实验训练集和开发集. Call Friend 数据库包括 15 个语种, 分别是: 美式英语 (南方口音)、美式英语 (非南方口音)、阿拉伯语、法语、波斯语、汉语 (台湾口音)、汉语 (大陆口音)、德语、日语、西班牙语 (加勒比口音)、西班牙语 (非加勒比口音)、韩语、北印度语、泰米尔语、越南语. 每个语种包括训练、开发、测试三个数据集, 每个数据集包含 20 段长度约为 30 分钟的对话语音. Call Home 数据库包括 6 个语种, 分别是: 美式英语、阿拉伯语、德语、汉语、日语、西班牙语. 每个语种包括训练、开发、测试三个数据集, 每个数据集包含 20 段长度约为 30 分钟的对话语音.

3.2 评测指标

语种识别是一种典型的分类问题, 一般的研究都集中用两个量来表示, 即虚警概率 P_{Fa} 和漏检概率 P_{Miss} :

$$P_{Fa} = \frac{N_{Fa}}{N_{Non-target}}$$

$$P_{Miss} = \frac{N_{Miss}}{N_{Target}}$$

其中, N_{Fa} 为非目标语种语音中检测为目标语种的段数, $N_{Non-target}$ 为非目标语种语音段数, N_{Miss} 为目标语种语音中检测为非目标语种的段数, N_{Target} 为目标语种语音段数. 在虚警概率和漏检概率的基础上, 又有等错误率 (Equal error rate)、检测代价 (Detection cost)、平均代价函数 (Average performance)、平衡图 (Detection error trade-off) 等评测指标. 本文选用平均代价函数作为系统性能评价指标.

平均代价函数是在检测代价基础上给出的评测指标. 检测代价定义如下^[16]:

$$C(L_T, L_N) = C_{Miss}P_{Target}P_{Miss}(L_T) + C_{Fa}(1 - P_{Target})P_{Fa}(L_T, L_N)$$

其中, L_T 和 L_N 分别表示目标语种和非目标语种, C_{Miss} 、 C_{Fa} 和 P_{Target} 与具体的应用有关, 在 NIST LRE 2003 中可以设为:

$$C_{Miss} = C_{Fa} = 1$$

$$P_{Target} = 0.5$$

从定义中可以看出, 检测代价是针对目标-非目标语种对而定义的指标, 因此平均代价函数^[16] 定义如下:

$$C_{arg} = \frac{1}{N_L} \sum_{L_T} \{C_{Miss}P_{Target}P_{Miss}(L_T) + C_{Fa}P_{Non-target}P_{Fa}(L_T, L_N) + C_{Fa}P_{Out-of-set}P_{Fa}(L_T, L_O)\}$$

其中, N_L 是闭集测试中目标语种的数量, L_O 表示集外语种. 在闭集测试时, 上式中第三个累加项可以省略. 在开集测试时,

$$P_{Out-of-set} = 0.2$$

$$P_{Non-target} = \frac{(1 - P_{Target} - P_{Out-of-set})}{N_L - 1}$$

3.3 实验设置

选用 7 维标准 MFCC 特征加上参数为 (7, 1, 3, 7) 移位差分倒谱 (Shifted delta cepstra, SDC)^[17] 特征, 共 56 维特征, 使用说话人归一化技术 (倒谱均值规整、RASTA 滤波、高斯化) 规整后, 作为声学特征. 在开发集中, 选用男女各约 5 小时数据, 训练混元数为 1024, 性别相关通用背景模型 (Universal background model, UBM), 然后将这两个 UBM 融合为一个混元数为 2048 的 UBM. 选用约 110 小时的数据 (各语种数据大致均衡) 训练 i-vector 中矩阵, i-vector 维数为 400 维, 迭代 20 次后认为其收敛.

对于不同 OLPP 映射子空间维数, 没有直接指定, 而是根据 $M^{(k)}$ 最小特征值的大小来确定. 当 $M^{(k)}$ 最小特征值小于某一阈值时, 此时 OLPP 迭代结束, 可将该最小特征值近似为 0, 则映射维数即为当前值减 1. 在本文中, 阈值选取为 0.001, 根据阈值确定 OLPP 子空间维数为 128. 由于 LDA 算法中类间方差矩阵 Σ_b 的秩不超过类别数, 因此本文将 LDA 映射子空间的维数定为 11 维.

对于后端分类, 使用 LIBSVM^[18] 实现 SVM 模型. 在训练集中, 为各语种选取 1000 段语音数据作为 SVM 训练数据, 为各语种选取 100 段语音数据作为得分规整训练数据. 选用 NIST LRE 2003 时长为 30 秒的数据集作为实验测试集.

3.4 实验结果

为了选取 LPP 算法和 OLPP 算法的最优维数, 选择维数范围为 20 到 200, 步长为 5, 分别进行实验. 图 2 给出了拉普拉斯语种识别方法和正交拉普拉斯语种识别方法在不同维数情况下的平均代价函数曲线. 从图中结果可以看出, 正交拉普拉斯语种识别方法的性能优于拉普拉斯语种识别方法. 这是由于 OLPP 算法保证了基函数的正交性, 保持了流型空间的度量结构, 因此局部保持能力优于 LPP 算法, 最终得到的向量也更具区分性. 图 2 还表明拉普拉斯语种识别方法在不同维数下, 系统性能曲线起伏波动较大, 而正交拉普拉斯语种识别方法系统性能曲线较为平坦, 这也验证了 OLPP 算法由于基函数的正交性, 使得算法对低维流形的维数不敏感. 同时, 图 2 中 OLPP 算法最优维数为 125, 这与根据阈值 0.001 来自动确定 OLPP 子空间维数 128 十分接近, 验证了文本 OLPP 子空间维数选取的正确性.

对 NIST LRE 2003 数据库中时长为 30 秒测试集进行测试. 分别对 i-vector 基线系统、i-vector 加上信道补偿、拉普拉斯语种识别方法、正交拉普拉斯语种识别方法, 4 种系统分别进行闭集和开集的测试. 由第三部分可知, 本文拉普拉斯语种识别方法

即为 i-vector+LPP+LDA, 正交拉普拉斯语种识别方法为 i-vector+OLPP+LDA. 4 种方法的相应开集测试的 DET 曲线如图 3 所示. 实验结果见表 1.

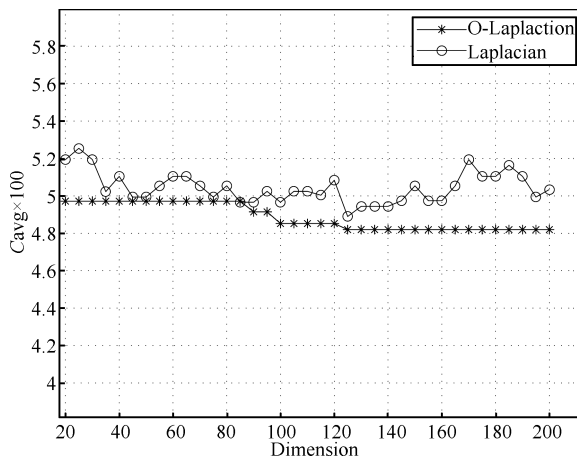


图 2 不同维数对系统性能的影响

Fig. 2 The effects of different dimensions to system performance

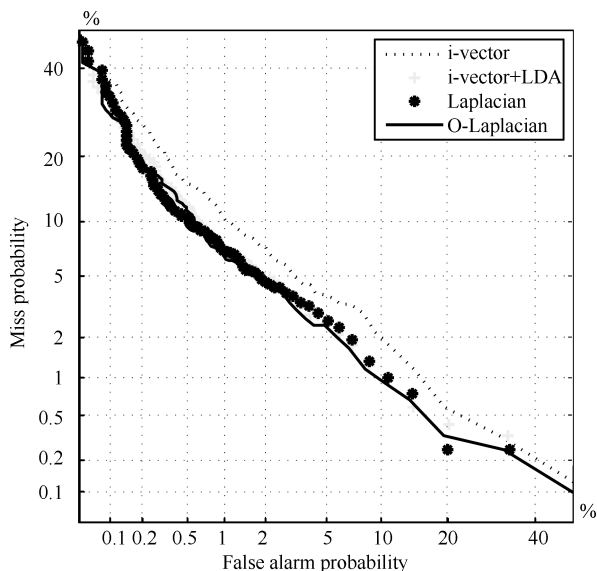


图 3 语种识别系统的 DET 曲线

Fig. 3 The DET curves of different language recognition systems

表 1 语种识别系统性能 C_{avg} (30 s)

Table 1 The system performance of language recognition C_{avg} (30 s)

系统	闭集	开集
i-vector	6.78	7.04
i-vector + LDA	5.09	5.13
Laplacian (i-vector + LPP + LDA)	4.89	4.90
O-Laplacian (i-vector + OLPP + LDA)	4.82	4.82

从图 3 中不难发现, 相比于基线系统, 其它三种语种识别系统的性能都有所提升. 具体分析表 1 中的实验结果, 可以看出, 相较于基线系统, i-vector 加上信道补偿系统的性能有了明显提升, 这是 i-vector 算法在投影的过程中并没有将类内和类间方差分开, 而是将它们一起投影到总变量空间中. 这样必然导致估计得到的 i-vector 特征中不仅存在语种相关的信息, 还存在信道和说话人相关的信息, 这些与语种无关信息对系统性能的提升存在一定的影响. LDA 算法能够最大化类间方差, 同时最小化类内方差. i-vector 经过 LDA 变换后, 区分性得到增强, 因此系统的性能有了明显的提升. 对比 i-vector 加上信道补偿、拉普拉斯语种识别方法、正交拉普拉斯语种识别方法的系统性能, 后两者的性能都有了一定的提升. 这是因为 i-vector 的理论基础是 PPCA, 属于流形学习中的线性方法. 这类方法能够有效地发现高维数据集的全局特性, 但是难以精确地描述数据之间的局部结构, 然而在分类问题中这种局部结构更具区分性. LPP 和 OLPP 算法属于流形学习中的全局线性化算法, 这类算法不仅保持了高维空间的局部几何特性, 而且具有线性子空间投影的优点. 拉普拉斯语种识别方法和正交拉普拉斯语种识别方法通过 i-vector 算法, 发现语音声学空间的整体特性, 去除声学无关信息的影响, 然后通过 LPP 或 OLPP 发现声学空间中的内在结构, 最后用 LDA 与语种无关信息. 因此这两种方法的性能优于 i-vector 加上信道补偿系统的性能.

4 结论

本文提出了正交拉普拉斯语种识别方法. 为了解决经典的 i-vector 因子分析中整体空间与语种无关信息 (例如说话人信息、信道信息的影响), 本文采用了 OLPP 流形学习算法将整体空间映射到语言信息加信道空间, 然后采用 LDA 信道补偿技术去除信道的影响, 这样所获得的向量突出语言相关的信息, 因此将其用于训练与识别, 可以达到语种识别的目的. 相关实验结果表明, 引入这两种技术, 在相同的条件下, 语种识别系统的识别率得以改善.

References

- 1 Zissman M A. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions Speech and Audio Process*, 1996, **4**(3): 31–44
- 2 Campbell W M, Sturim D E, Reynolds D A. Support vector machine using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 2006, **13**(5): 308–311
- 3 Kenny P. Factor Analysis of Speaker and Session Variability: Theory and Algorithms, Technical Report CRIM-06/08–13. Montreal, CRIM, 2005

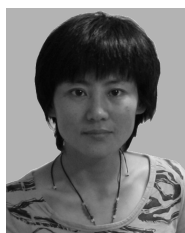
- 4 Kenny P, Boulianne G, Oullet P, Dumouchel P. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**(4): 1435–1447
- 5 Martinez D, Plchot O, Burget L, Glembek O, Matejka P. Language Recognition in iVectors Space. In: INTER-SPEECH. Florence, Italy: ISCA, 2011. 861–864
- 6 Dehak N, Torres P A, Reynolds D, Dehak R. Language recognition via iVectors and dimensionality reduction. In: INTERSPEECH. Florence, Italy: ISCA, 2011. 857–860
- 7 Tipping M E, Bishop C M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1999, **61**(3): 611–622
- 8 Turk M, Pentland A P. Face recognition using eigenfaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Maui, Hawaii: IEEE, 1991. 586–591
- 9 Zeng Xian-Hua. Researches on Related Issues of Spectral Method for Manifold Learning [Ph. D. dissertation], Beijing Jiaotong University, China, 2009
(曾宪华. 流形学习的谱方法相关问题研究 [博士学位论文], 北京交通大学, 中国, 2009)
- 10 Yang J C, Liang C Y, Yang L, Suo H B, Wang J J, Yan Y H. Factor analysis of Laplacian approach for speaker recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Kyoto, Japan: IEEE, 2012. 4221–4224
- 11 He X F, Niyogi P. Locality preserving projections. In: Proceedings of the Neural Information Processing Systems 16 (NIPS). Vancouver, Canada: The MIT Press, 2003. 153–160
- 12 He X F, Yan S C, Hu Y X, Niyogi P, Zhang H J. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2005, **27**(3): 328–340
- 13 Cai D, He X F. Locality preserving projections. In: Proceedings of the 28th Annual International ACM SIGIR Conference (SIGIR'05). Salvador, Brazil: ACM, 2005
- 14 Cai D, He X F, Han J W, Zhang H J. Orthogonal Laplacianfaces for face recognition. *IEEE Transactions on Image Processing*, 2006, **15**(11): 3608–3614
- 15 Hatch A O, Kajarekar S, Stolcke A. Within-class covariance normalization for SVM-based speaker recognition. In: INTERSPEECH. Pittsburgh, PA, USA, 2006. 1471–1474
- 16 The NIST 2003 Language Recognition Evaluation Plan [Online], available: <http://www.itl.nist.gov/iad/mig//test/lre/2003/LRE03EvalPlan-v1.pdf>, September 3, 2011
- 17 Torres-Carrasquillo P A, Singer E, Kohler M A, Greene R J, Reynolds D A, John R, Deller J R Jr. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In: Proceedings of the International Conferences on Spoken Language Processing (ICSLP). Denver, 2002. 89–92
- 18 Chang C C, Lin C J. LIBSVM: a library for support vector machines [Online], available: <http://www.csie.ntu.edu.tw/~cjlin/>, October 10, 2011



杨绪魁 中国人民解放军信息工程大学信息工程学院硕士研究生. 主要研究方向为语种识别, 连续语音识别和机器学习. 本文通信作者.

E-mail: gzyangxk@163.com

(**YANG Xu-Kui** Master student at the Institute of Information Engineering, PLA Information Engineering University. His research interest covers language identification, continuous speech recognition and machine learning. Corresponding author of this paper.)



屈丹 中国人民解放军信息工程大学信息工程学院副教授, 2005 年获解放军信息工程大学博士学位. 主要研究方向为语音信号处理与模式识别.

E-mail: qudanqudan@sina.com

(**QU Dan** Associate professor at the Institute of Information Engineering, PLA Information Engineering University. She received her Ph.D degree from PLA Information Engineering University in 2005. Her research interest covers speech signal processing and pattern recognition.)



张文林 解放军信息工程大学信息工程学院博士生. 主要研究方向为语种识别, 连续语音识别和机器学习.

E-mail: zwlin_2004@163.com

(**ZHANG Wen-Lin** Ph.D. candidate at the Institute of Information Engineering, PLA Information Engineering University. His research interest covers language identification, continuous speech recognition and machine learning.)