doi:10.11959/j.issn.1000-436x.2017096

基于 LDOF 准则的自适应高斯后端语种识别方法

叶中付[1,2,3]，戚婷[1,2]，李赛峰[1,2]，宋彦[1,2]

（1. 中国科学技术大学信息科学技术学院，安徽 合肥 230027；
2. 中国科学技术大学语音及语言信息处理国家工程实验室，安徽 合肥 230027；
3. 数学工程与先进计算国家重点实验室，江苏 无锡 214125）

摘 要：针对由语种类内多样性引起的测试样本和训练模型不匹配的问题，提出一种基于局部距离离群因子准则（LDOF, local distance-based outlier factor）的自适应高斯后端语种识别方法。定义 LDOF 准则，实现有效的参数寻优过程并动态地在多类语种训练集上挑选出与测试样本特性相近的训练样本，调整原高斯后端，进而得到改进的语种识别方法。在 NIST LRE 2009 的 6 个易混淆语种任务集上的实验结果表明，所提方法的等错误概率（EER, equal error rate）和平均检测代价有显著提升。

关键词：语种识别；类内多样性；自适应高斯后端；LDOF
中图分类号：TN912.34 **文献标识码**：A

Adaptive Gaussian back-end based on LDOF criterion for language recognition

YE Zhong-fu[1,2,3], QI Ting[1,2], LI Sai-feng[1,2], SONG Yan[1,2]

(1. School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China;
2. National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei 230027, China;
3. State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi 214125, China)

Abstract: In order to alleviate the mismatch in model between training and testing samples caused by inter-language variations, adaptive Gaussian back-end based on LDOF criterion was proposed for language recognition. The local distance-based outlier factor (LDOF) criterion was defined to find the appropriate model parameters and dynamically select the training data subset similar to the testing samples from multiple class training sets. Then original back-end was adjusted to obtain a more matched recognition model. Experimental results on NIST LRE 2009 easily-confused language data set show that proposed method achieves an obvious performance improvement on both the equal error rate (ERR) and average decision cost function.

Key words: language recognition, inter-language variations, adaptive Gaussian back-end, LDOF

1 引言

语种识别(LID, language recognition)是一个典型的模式识别问题，利用计算机自动判断出语音段所属的语言种类。LID 是多语言智能语音技术中关键的前端处理技术[1]，在民用以及国家安全方面都有重要的应用。近年来，基于全差异(TV, total variability)空间建模的语种识别方法在 NIST 语种识别任务中取得优越性能[2]，该方法用一个低维的矢量（称为 i-vector）表示语音段，在 i-vector 空间下的语种识别方法也越来越受到研究者的关注[3~5]。相对一些较复杂的算法[6,7]，高斯后端(GB, Gaussian back-end)[8]是一种基于 i-vector 先验假设的方法，复杂度低，同时可实现较好的识别性能，已成为主流

收稿日期：2016-08-15；**修回日期**：2017-02-09
基金项目：数学工程与先进计算国家重点实验室开放基金资助项目（No.2015A15）
Foundation Item: The Open Project Program of the State Key Laboratory of Mathematical Engineering and Advanced Computing (No.2015A15)

的 LID 方法之一。

然而，随着语种识别的实用需求日益增强，语种数据呈现出明显的多样性。例如，即使属于同一语种的语音数据，也可能来自电话、广播、卫星电话等不同信道或性别、年龄、情绪有差异的说话人，以及具有不同的时长。这些因素给语种数据的 i-vector 带来丰富的类内变化，使同一类语种的 i-vector 会因不同的特性形成多样的“小群体”。但利用这些多样 i-vector 训练的目标语种模型将出现问题。由于在测试过程中，一段未知语种类别的 i-vector 只可能和某一类语种数据部分训练样本具有相同的特性，而这种特性在训练模型中被模糊，使测试样本和已训练好的模型不匹配，造成识别性能下降。针对这种不匹配，如何得到适用于不同测试样本的语种识别方法是本文的研究重点。

文献[9]考虑语音段之间不同的信道信息，提出在训练语种分类器之前，先训练一个多信道的分类器。文献[10, 11]分析了训练语音段时长的不同对系统性能的影响，提出根据时长不同划分训练数据，训练出各自的全差异空间，获得时长相关的 i-vector。类似地，文献[12]根据性别不同划分训练数据，训练出性别相关的模型。这些方法虽然对性能有所提升，但是都只针对特定的某一种引起类内差异的因素，且系统的计算量成倍增加。文献[13]则提出先对所有训练时语音段的 i-vector 聚类，再采用矢量量化的方式，进行后续分类。这种方法对参数的设置敏感，最优的聚类数无法确定，且系统很难根据测试语音段的不同调整训练模型。

本文从 GB 模型出发，充分考虑由各种因素带来的语种类内多样性，提出一种与测试样本相关的自适应高斯后端(AGB, adaptive Gaussian back-end)语种识别方法，并给出一种有效的基于局部距离离群因子(LDOF, local distance-based outlier factor)准则的参数寻优解决方案。该方法利用定义的 LDOF 准则，在每次测试中，动态地从多类语种训练样本集上挑选出与测试样本（i-vector）特性相近的训练样本，调整原高斯后端，进而得到改进语种识别方法。这种调整不受限于某一种变化因素，根据测试样本的特性自动调整处理，灵活、顽健性强且复杂度低。

2 语音段 i-vector 的获取

2.1 DBN-UBM- DBF 系统

i-vector 的基本思想是将各语音段高斯均值超矢量之间的差异用一个低维的矢量表示，数学形式可以表示为

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{v} \tag{1}$$

其中，$\boldsymbol{M}$ 为对应每段语音的高斯混合模型(GMM, Gaussian mixture model)均值超矢量，$\boldsymbol{m}$ 为通用背景模型(UBM, universal background model)的 GMM 均值超矢量，$\boldsymbol{T}$ 为描述全差异空间的载荷矩阵，$\boldsymbol{v}$ 为每段语音在载荷矩阵 $\boldsymbol{T}$ 定义下的低维矢量表示 i-vector。在提取 i-vector 过程中，假设 i-vector 服从高斯分布。

文献[14,15]提出一种 DBN-UBM-DBF 系统，利用深度瓶颈神经网络(DBN, deep bottleneck network)同时完成了前端特征提取和 TV 建模，不仅实现音素状态对齐下的差异信息提取，同时充分利用了语种数据在音素状态下的先验信息，是目前性能最为优越的语音段 i-vector 的获取系统之一。

2.2 对 i-vector 的信道补偿

信道补偿的目的是找到 i-vector 空间中利于分类的投影方向，从而得到最终的语音段区分性表示。线性判别分析(LDA, linear discriminant analysis)结合类内方差归一化(WCCN, within-class covariance normalization)的策略可以达到较好的补偿性能[2,16]。LDA 和 WCCN 都是一种线性变换，并不改变信道补偿前后的 i-vector 分布形式，即信道补偿后的 i-vector 仍服从高斯先验假设分布。在本文中，所有 i-vector 均信道补偿，补偿方法是通过将 LDA 与 WCCN 相结合来实现的。

3 改进的高斯后端

3.1 高斯后端(GB)

2.1 节中，i-vector 服从高斯分布的先验假设，故用多维高斯分布对每类语种进行建模，每个语种的高斯分布共享同一协方差矩阵[7]，即语种 l 的 i-vector 服从均值为 $\boldsymbol{u}_l$、方差为矩阵 $\boldsymbol{\Sigma}$ 的高斯分布，记为

$$\boldsymbol{w}_l \sim N(\boldsymbol{u}_l, \boldsymbol{\Sigma}) \tag{2}$$

其中，$\boldsymbol{w}_l$ 表示语种 l 的 i-vector。

3.2 自适应高斯后端

由于语种数据类内的丰富变化，同一类语种的 i-vector 也会因不同的特性形成多样的“小群体”，而通过直接对训练样本依据某种特性分类或聚类划分，从本质上都只考虑了有限的变化因素，得到

的语种识别系统并不适用于所有测试样本，仍然存在测试样本和训练模型不匹配的问题。本文提出的自适应高斯后端目的在于综合考虑各种变化因素，可根据测试样本动态选择与之特性相近的训练样本，进而调整原高斯模型。新的模型是根据测试样本的特性自动调整处理的，这也正是“自适应高斯后端”的意义。图 1 给出了自适应过程的示意。

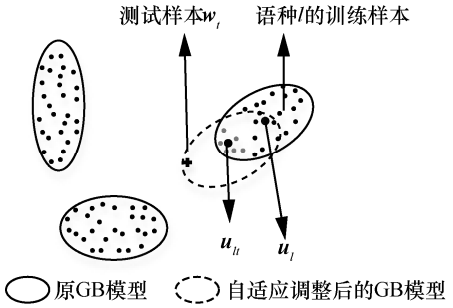


图 1　高斯后端自适应过程示意

在原 GB 模型中，每类语种的训练样本都服从高斯分布，在自适应 GB 模型中，根据测试样本 $\boldsymbol{w}_t$ 的特性，从原 GB 训练模型中的训练样本挑选出与测试样本特性相近的训练样本子集（图 1 中灰色点），从而得到自适应调整后的 GB 模型。此时，语种 l 的 i-vector 在已知测试样本 $\boldsymbol{w}_t$ 的条件下，服从的高斯分布记为

$$\boldsymbol{w}_l \left| \boldsymbol{w}_t \right. \sim N(\boldsymbol{u}_{lt}, \boldsymbol{\Sigma}) \tag{3}$$

其中，$\boldsymbol{u}_{lt}$ 是根据测试样本 $\boldsymbol{w}_t$ 动态地从训练样本中选取若干样本计算出的新均值向量，即自适应调整后的高斯模型的均值。

3.3　基于 k 近邻的自适应高斯后端

为了得到改进的 AGB 模型中的均值向量 $\boldsymbol{u}_{lt}$，需选出与测试样本 $\boldsymbol{w}_t$ 特性相近的训练样本子集，一种自然的方法是采用 k 近邻，挑选前 k 个与测试样本距离最小的训练样本。对于所有长度归一化后 i-vector，平方欧式距离是一种合适的度量方式，且可通过计算测试样本和训练样本的内积得到

$$\left\| \boldsymbol{w}_t - \boldsymbol{w}_j \right\|^2 = 2(1 - \boldsymbol{w}_t^{\mathrm{T}} \boldsymbol{w}_j) \tag{4}$$

其中，$\boldsymbol{w}_j$ 表示某一训练样本，符号 $\|\cdot\|^2$ 表示计算向量的二范数的平方。

记训练样本集 ξ，按已知语种标签划分为 L 类子集 $\xi_1, \xi_2, \cdots, \xi_L$，$N_1, N_2, \cdots, N_L$ 分别为每类子集样本数，则测试样本 $\boldsymbol{w}_t$ 在 ξ_l 中选出的训练样本个数，即 k 值，可通过在开发集上遍历所有 k 在各类样本上可能的取值，取性能最好情况下的 k，则需要遍历的次数为 $N_1, N_2, \cdots, N_L$。显然，这种遍历计算量大，对于数据量庞大的语种识别任务来说是不实际的。故基于 k 近邻的自适应高斯后端，测试样本 $\boldsymbol{w}_t$ 在每类语种的训练样本上选择一致的样本数，即 k 值对于不同类别的语种训练样本是固定的。记 $\boldsymbol{w}_t$ 在 ξ_l 中选出的 k 个训练样本构成的集合为 $\left\{ A_t^k \right\}_l$，则在本文中，基于 k 近邻的自适应高斯后端算法如下。

算法 1　基于 k 近邻的自适应高斯后端算法

S_1：给定 k 值;

S_2：For $l = 1, 2, \cdots, L$

1) 计算 $\boldsymbol{w}_t$ 与 ξ_l 中所有训练样本的平方欧式距离，并升序排序；
2) 选出前 k 个训练样本构成的 $\left\{ A_t^k \right\}_l$；
3) 计算 $\left\{ A_t^k \right\}_l$ 的均值向量 $\boldsymbol{u}_{lt}^k$；

S_3：更新均值向量为 $\boldsymbol{u}_{lt}^k$，得到 L 类语种 AGB 模型。

4　基于 LDOF 准则的 AGB 语种识别方法

4.1　LDOF 准则

在基于 k 近邻的 AGB 中，为了减少计算量，测试样本在每类语种的训练样本上选择样本数 k 是固定的，同时，一旦在开发集上确定了 k 值，在测试过程中，每个测试样本也只能共用同一 k 值。LDOF 准则的提出是为了解决如何在较短的时间内，找到基于 k 近邻的 AGB 中适用于不同测试样本和多类语种训练集的 k 值。基于 k 近邻的 AGB 中，将与测试样本的平方欧式距离最小的前 k 个训练样本视作和测试样本具有类似属性，但并未衡量这一结论的可信度。本文采用局部距离离群因子(LDOF, local distance-based outlier factor)来判断所选出的训练样本子集是否和测试样本归属同一特性，并定义 LDOF 准则，为每个测试样本在多类语种训练集上设计 k 值寻优方法。

LDOF 的提出是为了检测出散点簇中的离群点[17]，在说话人识别任务中，也被用于为概率线性判别分析建模挑选合适样本[18]。在本文中，利用 LDOF 为测试样本从训练样本中选择合适 k 值的近邻训练样本。给定 k，测试样本 $\boldsymbol{w}_t$ 在第 l 类语种的 LDOF 定义为

$$LDOF_{lt}^k = \frac{d_{lt}^k}{D_{lt}^k} \tag{5}$$

其中，d_{lt}^k 为测试样本 $\boldsymbol{w}_t$ 和第 l 类语种训练样本集上 k 个近邻的距离平均，D_{lt}^k 是这 k 个近邻两两之间距离的平均，分别定义为

$$\begin{aligned} d_{lt}^k &= \frac{1}{k}\sum_{\boldsymbol{w}_i\in\left\{A_t^k\right\}_l}\left\|\boldsymbol{w}_t-\boldsymbol{w}_i\right\|^2 \\ &= \left\|\boldsymbol{w}_t-\boldsymbol{u}_{lt}^k\right\|^2+\frac{1}{k}\sum_{\boldsymbol{w}_i\in\left\{A_t^k\right\}_l}\left\|\boldsymbol{w}_i-\boldsymbol{u}_{lt}^k\right\|^2 \end{aligned} \tag{6}$$

$$\begin{aligned} D_{lt}^k &= \frac{1}{k(k-1)}\sum_{\boldsymbol{w}_i,\boldsymbol{w}_j\in\left\{A_t^k\right\}_l,i\neq j}\left\|\boldsymbol{w}_i-\boldsymbol{w}_j\right\|^2 \\ &= \frac{2}{k-1}\sum_{\boldsymbol{w}_i\in\left\{A_t^k\right\}_l}\left\|\boldsymbol{w}_i-\boldsymbol{u}_{lt}^k\right\|^2 \end{aligned} \tag{7}$$

其中，$\left\{A_t^k\right\}_l$ 表示 $\boldsymbol{w}_t$ 从第 l 类语种训练样本集中选出的 k 个样本构成的集合。文献[18]指出，当 $LDOF_{lt}^k>\theta$ 时，判定对于集合 $\left\{A_t^k\right\}_l$，样本 $\boldsymbol{w}_t$ 是一个离群点，反之，则判定样本 $\boldsymbol{w}_t$ 和集合 $\left\{A_t^k\right\}_l$ 具有类似属性。θ 为常数，对于一般散点分布，取 $\theta=1$。

将式(6)和式(7)代入式(5)，定义本文中合适的 k 值需要满足的 LDOF 准则

$$\left\|\boldsymbol{w}_t-\boldsymbol{u}_{lt}^k\right\|^2\leqslant\frac{(2\theta-1)k+1}{k(k-1)}\sum_{\boldsymbol{w}_i\in\left\{A_t^k\right\}_l}\left\|\boldsymbol{w}_i-\boldsymbol{u}_{lt}^k\right\|^2 \tag{8}$$

图 2 是 LDOF 准则如何指导测试样本在训练样本集上选择合适的 k 近邻训练样本，其中，叉号表示测试样本 $\boldsymbol{w}_t$，黑色圆点表示训练集 ξ_l，黑色星号表示训练集 ξ_l 的均值 $\boldsymbol{u}_l$，灰色圆点表示选中的 k 个近邻训练样本 $\left\{A_t^k\right\}_l$，黑色三角符号表示 $\left\{A_t^k\right\}_l$ 的均值 $\boldsymbol{u}_{lt}^k$，虚线表示以 $\boldsymbol{u}_{lt}^k$ 为中心、D_{lt}^k 为半径的圆，实线表示以 $\boldsymbol{u}_{lt}^k$ 为中心、d_{lt}^k 为半径的圆。当 $k=1$ 时，式(7)没有意义，所以，利用 LDOF 准则寻找合适的 k 值时，k 值变化从 2 开始。可以看出，随着 k 值从 2 到 5 的递增，所选的样本点也逐渐变化，直至图 2 (d)中 $LDOF$ 值满足条件，至此，利用 LDOF 准则，指导测试样本动态地从训练样本中选取的样本子集计算出 AGB 模型的均值过程结束。

如文献[17]论述，当样本点服从高斯分布时，若 $\boldsymbol{w}_t$ 对于集合是离群点，则 $LDOF_{lt}^k$ 值存在下界，$LDOF_{lt}^k>0.5$。在本文中，采用高斯后端进行语种识别，训练模型均为高斯模型，故式(8)中取 $\theta=0.5$ 更适用于自适应调整的高斯后端。

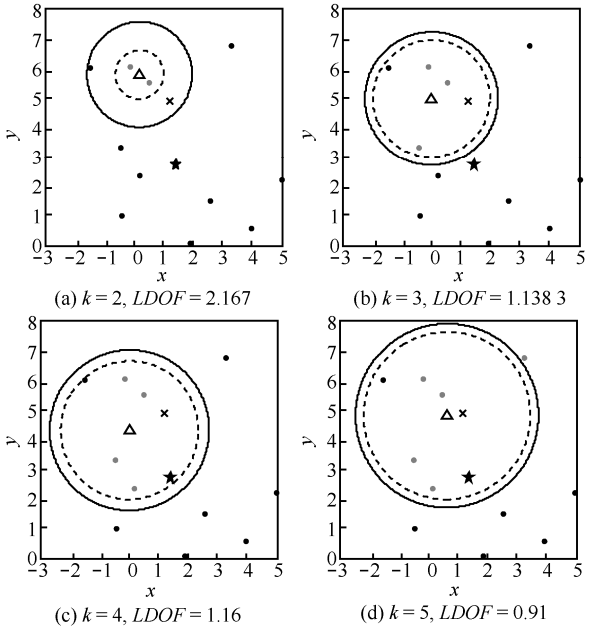


图 2 基于 LDOF 的 k 值选择过程

当 k 值较小时，$LDOF_{lt}^k$ 值不稳定的情况，在利用 LDOF 准则寻找合适的 k 值时，定义

$$\Delta LDOF_{lt}^k=\frac{\left|LDOF_{lt}^k-LDOF_{lt}^{k-1}\right|}{LDOF_{lt}^k} \tag{9}$$

其中，符号 $|\cdot|$ 表示取绝对值运算。当 $\Delta LDOF_{lt}^k>\gamma$ 时，k 值的选择不合适，γ 定义为收敛阈值。

4.2 AGB 语种识别方法

在语种识别阶段，采用最大对数似然函数作为判决依据，需计算测试样本 $\boldsymbol{w}_t$ 在各个语种训练模型上的对数似然函数。在语种识别任务中，该对数似然函数也被称作测试样本 $\boldsymbol{w}_t$ 在语种模型上的得分，即测试样本 $\boldsymbol{w}_t$ 在语种 l 上的 GB 得分为

$$\begin{aligned} s_{lt} = &-\frac{1}{2}\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}_t+\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}_l- \\ &\frac{1}{2}\boldsymbol{u}_l^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}_l+\mathrm{const} \end{aligned} \tag{10}$$

其中，$\boldsymbol{u}_l$ 是语种 l 的 GB 模型的均值。const 是常数项，二次项 $\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}_t$ 和语种类别无关，可以忽略。当对 i-vector 进行 WCCN 信道补偿后，所有语段 i-vector 都对同一类内方差归一化，则测试样本 $\boldsymbol{w}_t$ 在语种 l 上的 GB 得分简化为

$$s_{lt}=\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{u}_l-\frac{1}{2}\boldsymbol{u}_l^{\mathrm{T}}\boldsymbol{u}_l \tag{11}$$

利用定义的 LDOF 准则，可以使测试样本在较短时间内，在多类训练样本集上选择合适的训练样

本子集，进而调整得到均值为 $\boldsymbol{u}_{lt}$ 的语种 l 的 AGB 模型。此时，测试样本 $\boldsymbol{w}_t$ 在语种 l 上的 AGB 得分为

$$s_{lt} = \boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{u}_{lt} - \frac{1}{2}\boldsymbol{u}_{lt}^{\mathrm{T}}\boldsymbol{u}_{lt} \tag{12}$$

其中，$\boldsymbol{u}_{lt}$ 是测试样本 $\boldsymbol{w}_t$ 基于 LDOF 准则，动态地从训练样本中选取的样本子集计算出 AGB 模型的均值。

以属于第 l 类语种的测试样本为例，计算在第 l 类语种模型上的得分时，利用本文定义的 LDOF 准则，可以合理地在第 l 类语种训练样本集上挑选出和测试样本特性相近的训练样本子集，计算 AGB 模型中的 $\boldsymbol{u}_{lt}$，得到和测试样本更匹配的训练模型，从而测试样本在第 l 类语种模型上比在原 GB 模型上更高的得分，利于最终判决；计算在其他类语种模型上的得分时，由于测试样本并不属于这些类别，则可能出现在某类语种训练样本集上只挑选出少量训练样本的情况，此时 AGB 模型中，计算得到的训练样本子集的均值不稳定，从而影响得分计算。针对这一问题，本文在 AGB 模型上得分计算过程中保留原 GB 模型部分信息，即语种 l 的全部训练样本的均值 $\boldsymbol{u}_l$，改进后的得分计算方法为

$$s_{lt} = \boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{u}_{lt} - \frac{1}{2}\boldsymbol{u}_l^{\mathrm{T}}\boldsymbol{u}_{lt} \tag{13}$$

当对所有 i-vector 长度归一化后，式(11)变换为

$$s_{lt} = \boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{u}_l - \frac{1}{2} \tag{14}$$

对比式(14)，式(13)中保留了一项与语种 l 全部训练样本的均值有关的二次项 $\boldsymbol{u}_l^{\mathrm{T}}\boldsymbol{u}_{lt}$，该二次项计算调整前后 2 个模型的均值向量的余弦距离，即衡量调整前后模型的差异大小。当模型未调整时，$\boldsymbol{u}_{lt} = \boldsymbol{u}_l$，$\boldsymbol{u}_l^{\mathrm{T}}\boldsymbol{u}_{lt} = 1$，式(13)和式(14)等价；当模型调整时，$\boldsymbol{u}_l^{\mathrm{T}}\boldsymbol{u}_{lt} \leqslant 1$，且调整前后 2 个模型差异越大，$\boldsymbol{u}_l^{\mathrm{T}}\boldsymbol{u}_{lt}$ 值越小，式(13)中计算的得分越高。调整前后 2 个模型差异大说明了测试样本选出了只和自身特性相近的部分训练样本，相比于模糊了这种特性的原 GB 模型，调整后的 AGB 模型和测试样本更加匹配。故式(13)中保留的与语种 l 全部训练样本的均值有关的二次项是对 AGB 模型的合理约束。

基于以上分析，在识别阶段，得到测试样本 $\boldsymbol{w}_t$ 在各类语种上的 AGB 得分后，依据最大对数似然，做出最终判决。基于 LDOF 准则的 AGB 语种识别方法步骤如下。

算法 2　基于 LDOF 准则的 AGB 语种识别方法

S$_1$：给定 θ 和收敛阈值 γ；

S$_2$：For $l = 1, 2, \cdots, L$

1) 设定初始值：$k = 2$，$LDOF_{lt}^1 = 0$；

2) 基于 k 近邻的 AGB，计算 $\left\{A_t^k\right\}_l$ 的均值向量 $\boldsymbol{u}_{lt}^k$；

3) 计算 $\Delta LDOF_{lt}^k$；

4) 如果式(8)不满足或者 $\Delta LDOF_{lt}^k > \gamma$，则 $k = k+1$；

跳到步骤 2)；

5) 根据式(13)计算得分 s_{lt}；

S$_3$：若 $s_{lt} = \max\left\{s_{it}, 1 \leqslant i \leqslant L\right\}$，则测试样本属于第 l 类语种。

5　实验与结果分析

5.1　实验数据配置

为了验证自适应高斯后端和 LDOF 准则的有效性，本文利用 Matlab 编写了算法的仿真程序。所有实验在美国国家标准技术署的标准测试集 NIST LRE 2009 的 6 个易混淆的 NIST 语种任务集上进行，仿真实验分为 3 种语音段时长测试环境，分别对长时 30 s 语音段、10 s 语音段以及短时 3 s 语音段进行测试[15]。实验系统性能评价采用 NIST LRE 标准的 2 种指标：等错误率（*EER*）[20]和平均检测代价（*Cavg*）[21]。

EER 定义为系统通过调整判决门限，使虚警概率与漏警概率一致时的错误率。*EER* 的值越小，则系统的性能表现越优越。

平均检测代价（*Cavg*）的计算式为

$$Cavg = \frac{1}{N_L}\sum_{L_T}\left\{C_{\mathrm{Miss}}P_{\mathrm{Target}}P_{\mathrm{Miss}}(L_T) + \sum_{L_N}C_{\mathrm{FA}}P_{\mathrm{Non\text{-}Target}}P_{\mathrm{FA}}(L_T, L_N)\right\} \tag{15}$$

其中，N_L 表示所有待识别语种数目，L_T 和 L_N 分别表示目标语种和非目标语种，C_{Miss} 和 C_{FA} 表示漏判决一个和错判决一个的检测代价，P_{Target} 和 $P_{\mathrm{Non\text{-}Target}}$ 表示目标语种和非目标语种的先验概率。在计算指标时取 C_{Miss}=C_{FA}，P_{Target}=$P_{\mathrm{Non\text{-}Target}}$=0.5。

所采用的DBN-UBM-DBF系统中的DBN网络输出节点数为 3 020，获取的 i-vector 为 400 维，其他参数配置同文献[14]。所有语音段的 i-vector 进行 LDA 结合 WCCN 的信道补偿技术并对所有的补偿后的

i-vector 长度归一化。图 3 利用 t-SNE 方法[19]，可视化每类语种的训练语音段 i-vectors。如图 3 所示，每类语种的样本整体服从高斯的先验假设分布，而每类语种的类内差异也会形成各自多样的“小群体”。

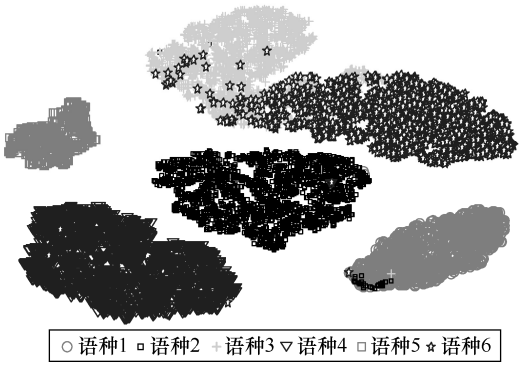


图 3 NIST LRE 2009 的 6 个易混淆语种训练数据 i-vector 分布

5.2 相关方法符号定义

方法 1（GB）如 3.1 节中介绍的高斯后端，计算对数似然比得分的语种识别方法，即调整前的原高斯后端，利用训练集所有样本训练得到。

方法 2（CDS）通过计算测试样本与每类语种上训练数据均值向量的余弦距离，作为测试样本的得分(CDS, cosine distance scoring)，是文献[14]中所用的语种识别方法，也是本文的基线方法。

方法 3（KNN-AGB）如 3.3 节中介绍的基于 k 近邻的自适应调整高斯后端语种识别方法。

方法 4（LDOF-AGB）取 $\theta=1$ 时的基于 LDOF 准则的自适应调整高斯后端语种识别方法。

方法 5（G-LDOF-AGB）考虑采用高斯后端进行语种识别，训练模型均为高斯模型的情况，取 $\theta=0.5$ 时的基于 LDOF 准则的自适应调整高斯后端语种识别方法。

实验中，LDOF-AGB 和 G-LDOF-AGB 方法的收敛阈值均取 $\gamma=0.000\,1$。

5.3 实验结果与分析

图 4 是 KNN-AGB 方法在不同测试语段时长下，性能指标 *EER* 随 k 值的变化情况，当 $k=600$ 时，系统平均性能最优。结合表 1 和表 2 中给出的各种方法的 2 种性能评测指标可以看出，当 $k=600$ 时，KNN-AGB 在 30 s、10 s 和 3 s 这 3 种测试时长下，*EER* 和 *Cavg* 均有优于自适应调整前的 GB，且对比基线 CDS 方法，也有一定性能提升，实验结果验证了 AGB 的有效性。但从图 4 可以看出，KNN-AGB 方法性能受 k 值变化的影响明显，当 k 值较小时（小于 300），性能不稳定且出现低于基线甚至低于调整前 GB 的情况，只有当 k 值大于一定值时，才能取得较好性能。较大的 k 值使 AGB 模型在各类语种训练样本集上挑选出过多的训练样本，带来冗余，不能保证挑选出的训练样本子集完全和测试样本具有类似属性。故 KNN-AGB 方法中固定 k 值，不适用于不同的测试样本，影响了系统性能的进一步提升。

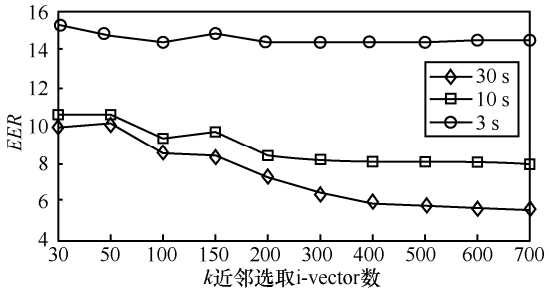


图 4 KNN-AGB 在不同测试语段时长下，随 k 值变化的 *EER*

基于 LDOF 准则的 AGB 语种识别方法可以根据不同测试样本的特性，动态地确定适合各类训练样本的 k 值。表 1 和表 2 分别给出了各类方法的性能评测指标。实验结果验证了当 LDOF 准则中 $\theta=0.5$ 时，相比于 $\theta=1$，更符合语种识别任务，即 G-LDOF-AGB 性能较为优越。对比于自适应调整前的 GB，在 3 种测试时长下，*EER* 平均有 12.4%的相对提升，*Cavg* 平均有 10.2%的相对提升。

表 1 在 6 个混淆语种数据集上不同后端方法的 *EER* 性能对比

方法	*EER* 性能		
	30 s	10 s	3 s
GB	6.07%	8.08%	15.23%
CDS	5.84%	8.02%	15.27%
KNN-AGB($k=600$)	5.61%	8.01%	14.46%
LDOF-AGB	5.01%	7.19%	14.42%
G-LDOF-AGB	**4.87%**	**7.12%**	**14.38%**

表 2 在 6 个混淆语种数据集上不同后端方法的 *Cavg* 性能对比

方法	*Cavg* 性能		
	30 s	10 s	3 s
GB	183%	248%	475%
CDS	180%	244%	477%
KNN-AGB($k=600$)	176%	240%	471%
LDOF-AGB	157%	**215%**	476%
G-LDOF-AGB	**154%**	217%	**468%**

以实验中随机选取的属于语种 1 的一个测试样本为例，图 5 给出了在基于 LDOF 准则的 AGB 语种识别方法实施时，该测试样本在 6 种易混淆语种的 AGB 模型上的得分随选取的样本数变化曲线。采用 G-LDOF-AGB 方法，该测试样本在 6 种语种上选出的样本数 k 分别为 201、96、78、40、43、147。从图 5 可以看出，G-LDOF-AGB 方法实施时，测试样本在非所属语种 AGB 模型上得分随着选取样本数增加均呈明显的下降趋势，且与所属语种 AGB 模型上得分之间的区分性逐渐增大，更利于最终语种判决，这也进一步验证了 G-LDOF-AGB 方法的有效性。

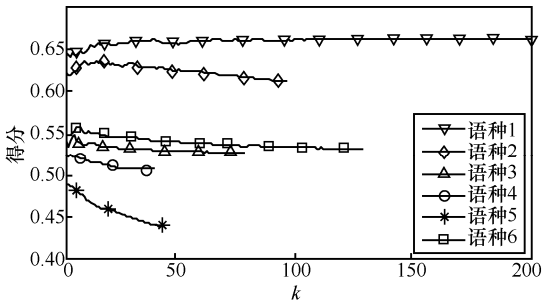


图 5　某测试样本在 AGB 模型上的得分随选取的样本数变化曲线

LDOF 准则保证了从训练样本中选出的样本和测试样本具有类似属性，对原 GB 语种识别方法做了更匹配于测试样本的调整。图 6 给出了测试样本集在原 GB 和改进后的 G-LDOF-AGB 上的得分情况。为了便于观察和比较，在图 6 中用 M_1~M_6 标记 6 种目标语种训练模型，将测试样本集分成 $test_1$~$test_6$ 的 6 种测试样本子集，按所属语种类别大小依次排列，且将测试样本集在 6 种易混淆语种的 GB 和 AGB 模型上的得分均做[0,1]规整。

由图 6 可知，图 6(b)中非主对角线上图像灰度相较于图 6(a)明显变浅，即测试样本在非所属语种 AGB 模型上具有更低的得分。这说明改进后的 AGB 模型可以大大降低各语种间的混淆程度，从而带来语种识别系统性能的提升。

对比于基线，G-LDOF-AGB 语种识别方法在 30 s、10 s 和 3 s 测试时长下，*EER* 分别相对提升 16.6%、11.2%、5.8%，*Cavg* 分别相对提升 14.4%、11.9%、1.9%，说明了提出的新方法在短时语种识别任务中仍能保持性能的优越，顽健性强。同时，LDOF 准则（式(8)）的计算只涉及测试样本和训练样本之间平方欧式距离的计算，当 i-vector 长度归一化后，平方欧式距离可通过直接计算测试样本和训练样本的内积得到，这和基线系统的得分计算方式是一致的，故提出的新方法和基线系统具有相同的复杂度。

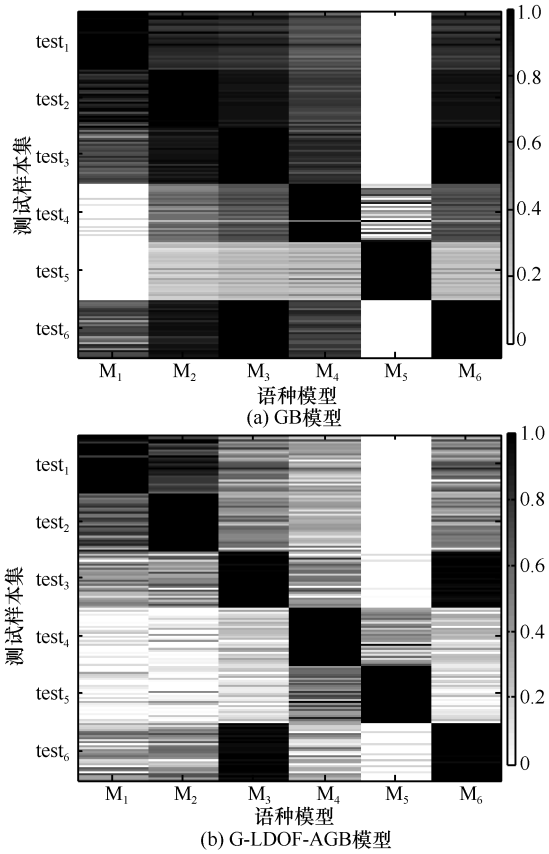


图 6　测试样本集在 6 种易混淆语种上的得分

6　结束语

本文针对语种识别任务中，语种数据类内存在的复杂多样性，提出一种基于 LDOF 准则的自适应高斯后端语种识别方法。与已有的将训练样本划分的识别方法不同，该方法动态地根据不同的测试样本选择与之特性相近的训练样本，进而自适应地调整训练好的高斯模型。定义 LDOF 准则，保证了选出的训练样本子集和测试样本具有类似属性，从而在每一次测试中，训练模型能更好地匹配于测试样本。该方法具有顽健性强、复杂度低的优点。通过在 NIST LRE 2009 的 6 个易混淆语种任务集的识别结果充分说明了该方法的有效性。

参考文献：

[1]　蒋兵. 语种识别深度学习方法研究[D]. 合肥: 中国科学技术大学, 2015.
JIAN B. Deep learning based spoken language identification[D]. Hefei: University of Science and Technology of China, 2015.

[2] DEHAK N, KENNY P, DEHAK R, et al. Front-end factor analysis for speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(4): 788-798.

[3] DEHAK N, TORRES-CARRASQUILLO P A, REYNOLDS D A, et al. Language recognition via i-vectors and dimensionality reduction[C]//The 12th Annual Conference of the International Speech Communication Association (Interspeech). 2011: 857-860.

[4] MARTINEZ D, PLCHOT O, BURGET L, et al. Language recognition in iVectors space[C]//The Interspeech 2011, Conference of the International Speech Communication Association. 2011: 861-864.

[5] PENAGARIKANO M, VARONA A, DIEZ M, et al. Study of different backends in a state-of-the-art language recognition system[C]//Interspeech. 2012: 2049-2052.

[6] 杨绪魁, 屈丹, 张文林. 正交拉普拉斯语种识别方法[J]. 自动化学报, 2014, 40(8): 1812-1818.
YANG X K, QU D, ZHANG W L. An orthogonal laplacian language recognition approach[J]. Acta Automatica Sinica, 2014, 40(8): 1812-1818.

[7] LIU G, HASAN T, BORIL H, et al. An investigation on back-end for speaker recognition in multi-session enrollment[C]// 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013: 7755-7759.

[8] VAN L D A, BRUMMER N. Channel-dependent GMM and multi-class logistic regression models for language recognition[C]// 2006 IEEE Odyssey-The Speaker and Language Recognition Workshop. IEEE. 2006: 1-8.

[9] BENZ M F, GAUVAIN J L, LAMEL L. Language score calibration using adapted Gaussian back-end[C]//Interspeech 2009. 2009: 2191-2194.

[10] SENOUSSAOUI M, KENNY P, BRÜMMER N, et al. Mixture of PLDA models in i-vector space for gender-independent speaker recognition[C]//Interspeech. 2011: 25-28.

[11] KANAGASUNDARAM A, VOGT R J, DEAN D B, et al. PLDA based speaker recognition on short utterances[C]//The Speaker and Language Recognition Workshop (Odyssey 2012). ISCA, 2012.

[12] SARKAR A K, MATROUF D, BOUSQUET P M, et al. Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification[C]//Interspeech. 2012: 2662-2665.

[13] WANG M G, SONG Y, JIANG B, et al. Exemplar based language recognition method for short-duration speech segments[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 7354-7358.

[14] SONG Y, HONG X, JIANG B, et al. Deep bottleneck network based i-vector representation for language identification[C]. Interspeech 2015. 2015:398-402.

[15] 洪新海, 宋彦, 蒋兵, 等. 采用 DBN 的 TV 改进方法在语种识别中的应用[J]. 信号处理, 2015, 31(9): 1152-1158.
HONG X H, SONG Y, JIANG B, et al. Improved total variability modeling method using deep bottleneck network for language identification [J]. Journal of Signal Processing, 2015, 31(9): 1152-1158.

[16] 王梦鸽. 短时语种识别若干问题研究[D]. 合肥: 中国科学技术大学, 2014.
WANG M G. Research on problems in spoken language identification with short-duration segments [D]. Hefei: University of Science and Technology of China, 2014.

[17] ZHANG K, HUTTER M, JIN H. A new local distance-based outlier detection approach for scattered real-world data[M]//Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2009: 813-822.

[18] BISWAS S, ROHDIN J, SHINODA K. I-vector selection for effective PLDA modeling in speaker recognition[C]//Proceedings Odyssey 2014-The Speaker and Language Recognition Workshop. 2014: 100-105.

[19] VAN DER M L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(2605): 2579-2605.

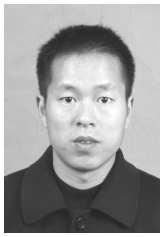
[20] MARTIN A F, PRZYBOCKI M A. NIST 2003 language recognition evaluation[C]//Interspeech. 2003.

[21] MARTIN A F, GREENBERG C S. The 2009 NIST language recognition evaluation[C]//Odyssey. 2010: 30.

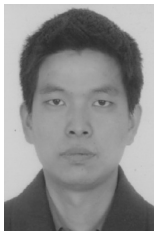
作者简介：

叶中付（1959-），男，安徽桐城人，博士，中国科学技术大学教授、博士生导师，主要研究方向为语音信号处理、阵列信号处理、雷达信号处理和图像分析与处理。

戚婷（1993-），女，安徽淮南人，中国科学技术大学硕士生，主要研究方向为语种识别。



李赛峰（1980-），男，江西萍乡人，中国科学技术大学博士生，主要研究方向为通信信号处理和语音信号处理。



宋彦（1972-），男，安徽合肥人，博士，中国科学技术大学副教授，主要研究方向为语种识别和基于内容的音/视频分析与检索。