基于概率修正 PLDA 的说话人识别技术*

李琳[1]，万丽虹[1]，洪青阳[1]*，张君[1]，李明[2]

1. 厦门大学信息科学与技术学院，厦门 361005；
2. 中山大学卡内基梅隆大学联合工程学院，广州 510006

文 摘：为减弱注册语音与测试语音时长不一致对说话人识别性能的负面影响，提出一个概率修正 PLDA 建模方法。根据语音时长自适应改变传统 PLDA 模型中 i-vector 的概率分布函数，提高 PLDA 对每个说话人每段语音的时长表征能力，以增强说话人类别的区分度。为验证基于概率修正 PLDA 模型的有效性，进行了 NIST SRE10 core-core 测试集在三种不同时长的评测实验，以及 NIST 2014 i-vector machine learning challenge 两个测试任务。结果表明，相较于传统的 PLDA 训练模型，通过语音时长的约束提高了说话人识别性能。

关键词：高斯 PLDA；i-vector；语音时长；概率修正；说话人识别

中图分类号：TN912.34

传统说话人识别技术从语音样本提取特征参数，并利用说话人特征的差异性建立分类模型，如高斯混合模型（Gaussian Mixture Model，GMM）[1]，以区分目标说话人和冒充说话人。然而，说话人特征差异性的表征能力受到说话人情绪、背景噪声、语音时长、采集设备等因素的制约，直接影响了现有说话人识别技术的识别效果。

在实际应用中，较频繁出现参考语料与测试语料的录制信道不同和时长不一致的情况。采用 Eigenvoice, Eigenchannel, Joint Factor Analysis 等说话人-信道联合模型[2-3]对 GMM 均值超向量进行信道无关的说话人因子分析，一定程度上削弱了信道差异对说话人识别性能的影响。基于 i-vector 的说话人识别系统[4]使用有害因子投影（nuisance attribute projection,, NAP）、线性区分性分析（linear discriminant analysis, LDA）、类内协方差归一化（within-class covariance normalization, WCCN）,或概率线性区分性分析（probabilistic Linear discriminant analysis ,PLDA)[5-6]等区分技术更好地解决了信道不匹配问题。由于时长信息和信道信息、音素信息一样，是随着语音段的录制而存在着，但是传统 GMM 建模方法一定程度上模糊了每个语音样本的时长信息。虽然完全变化因子 i-vector 的提取过程考虑了时长的影响，采用了与语音样本帧数的倒数相关的概率分布函数，但单纯使用 i-vector 作为新型声学特征和 PLDA 作为区分模型的说话人识别系统在时长不一致及短语音情况下仍会出现明显的性能下降[7]。近年来，学者们开始针对时长不一致问题展开一系列的研究。Kenny 等[8]人将时长信息作为信道信息的附加补偿，在说话人-信道空间建模时多设置了一组表征时长的信道偏移参量，在 NIST SRE10 core-core 测试中，将 EER 由 6.8%降为 5.9%，以增加 PLDA 训练过程中似然函数的计算复杂度为代价，换取识别性能对样本时长的鲁棒性。Hasan 等[9]假设样本时长为 i-vector 变量空间中的加性噪声，提出 3 种优化方法：①采用同一语料多种时长样本进行 PLDA 建模；②在分数域构建 QMF 函数，加入时长信息的调节作用；③使用时长方差规整得到新的 i-vector 变量。经过 NIST SRE12 的评测结果分析得到，第 2 种方法对短语音的识别效果最显著。Kanagasundaram 等[10]提出时长方差规整算法（SUVN），在 i-vector 特征域中，结合 SUVN，LDA 以及 PLDA 等补偿信道差异性和时长变化性。

本文首先将 i-vector 向量进行白化和归一化处理[11]，建立 i-vector 的标准高斯分布。然后，引入语音样本的时长信息，将其作为每个说话人每个 i-vector 在 PLDA 模型中的方差调节因子，描述每个 i-vector 向量由时长不

* **基金项目：**国家自然基金(No.61105026).
作者简介：李琳（1982— ），女，博士（博士），副教授，lilin@xmu.edu.cn.
通讯作者：洪青阳，qyhong@xmu.edu.cn

同而产生的信息熵：样本时长越短，携带的说话人信息越少，偏离高斯分布均值的程度越大。最后，采用最大期望（expectation maximization, EM）算法实现对开发集 i-vector 向量分布概率函数的最大似然估计，建立起一个受语音时长约束的概率修正 PLDA 模型（modified-prior PLDA）。本文分别在 NIST SRE10 core-core 测试集（女性部分）和 NIST 2014 i-vector machine learning challenge 的评测任务中验证了概率修正 PLDA 模型的有效性。

1 基线系统

将联合因子分析（JFA）算法中说话人因子分量和信道因子分量同时映射到一个低维空间，使用基于 Baum-Welch 统计量对 GMM 均值超向量进行降维处理得到一固定长度的完全因子向量 i-vector，即每一段语音样本均可表示为一个 i-vector。

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

式中：$\boldsymbol{M}$ 为 GMM 均值超向量；$\boldsymbol{m}$ 为一个与说话人和信道无关的均值超矢量，$\boldsymbol{T}$ 为低秩的全局差异空间矩阵，$\mathbf{x}$ 则表示一个满足标准正态分布 N(0,$\boldsymbol{I}$)的随机向量，即 i-vector。

假设 x_{tar} 和 x_{tst} 分别代表目标说话人和测试语音所对应的 i-vecor。本文的基线系统将采取余弦距离值(CDS)作为基线系统的决策分数

$$s_{baseline} = \frac{< \boldsymbol{x}_{tst}, \boldsymbol{x}_{tar} >}{\|\boldsymbol{x}_{tst}\| \|\boldsymbol{x}_{tar}\|} \tag{2}$$

2 标准高斯 PLDA 系统

给定一组来自 N 个说话人的 i-vector 向量 $\{x_{ij}, i = 1,...,N,\ j = 1,2,...,M_i\}$（其中，每个说话人有 M_i 条语音样本），每个 i-vector 经过白化和归一化处理，满足标准高斯分布。进一步，将 i-vector 分解为确定信号和随机噪声，则得到其 PLDA 模型：

$$\boldsymbol{x}_{ij} = \boldsymbol{\mu} + \boldsymbol{\varphi}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_{ij} \tag{3}$$

式中：$\boldsymbol{\mu}$ 代表来自开发集所有 i-vector 向量的均值；$\boldsymbol{\beta}_i$ 是第 i 个说话人的说话人因子，满足标准正态分布 N(0,I)；矩阵 $\boldsymbol{\varphi}$ 是固定维度的说话人子空间；残差 $\boldsymbol{\varepsilon}_{ij}$ 包含信道因子，服从均值为 0，协方差矩阵为 $\boldsymbol{\Sigma}$ 的正态分布。

利用一定规模的语音样本开发集，使用 EM 算法估计出 PLDA 参数集{ $\boldsymbol{\mu}$，$\boldsymbol{\varphi}$，$\boldsymbol{\Sigma}$ }。一般采用对数似然比作为标准高斯 PLDA 的决策分数：

$$s_{G\text{-}PLDA} = \log \frac{p(\boldsymbol{x}_{tar}, \boldsymbol{x}_{tst} \mid H_S)}{p(\boldsymbol{x}_{tar}, \boldsymbol{x}_{tst} \mid H_d)} \tag{4}$$

式中：H_s 表示测试语音来自同一说话人的假设条件，H_d 表示测试语音来自冒充者的假设条件。

3 时长约束的概率修正 PLDA 模型

文献[9]中通过分析语音样本所包含的音素（phonemes）统计量与语音时长（5s，10s，20s，40s 和全时长）的关系，发现，音素的数量随着语音时长的减小而呈指数递减，而当语音时长增加到一定长度时，如时长在 1min 以上，音素的统计量将保持不变。由此可见，语音的时长对说话人识别性能具有不容忽视的影响。对于同一说话人，语音时长越短，对应 i-vector 的 PLDA 模型将趋向于产生越大的协方差。

3.1 高斯分布函数的修正

考虑语音样本时长的影响力，本文假定公式（3）中的 $\boldsymbol{\varepsilon}_{ij}$ 将服从一个新的正态分布：

$$\boldsymbol{\varepsilon}_{ij} \sim N(\boldsymbol{\varepsilon}_{ij} \mid 0, \boldsymbol{\Sigma}) \Rightarrow N(\boldsymbol{\varepsilon}_{ij} \mid 0, \left(\frac{L_{ij}}{\alpha}\right)^{-\lambda} \boldsymbol{\Sigma}) \tag{5}$$

式中：L_{ij} 代表第 i 个说话人第 j 段语音样本的时长，可用帧数表示，α 和 λ 为调节参数，刻画语音时长对分布函数的影响程度。

已知开发集中有 N 个说话人，每个说话人有 M_i 个语音样本，即 $i = 1,...,N,\ j = 1,2,...,M_i$，设定 $\boldsymbol{\eta}_{ij}$ 代表 i-vector 向量的一阶统计量 x_{ij} - μ，则后验概率 $P(\boldsymbol{\eta}_{ij} \mid \boldsymbol{\beta}_i)$ 为

$$P(\boldsymbol{\eta}_{ij} \mid \boldsymbol{\beta}_i) = N(\boldsymbol{\eta}_{ij} \mid \varphi\boldsymbol{\beta}_i, \left(\frac{L_{ij}}{\alpha}\right)^{-\lambda} \boldsymbol{\Sigma}) \tag{6}$$

设定 $\boldsymbol{F}_i$ 是第 i 个说话人一阶统计量的均值，如下所示：

$$\boldsymbol{F}_i = \frac{\sum_{j=1}^{M_i} \boldsymbol{\eta}_{ij}}{M_i} = \boldsymbol{\varphi}\boldsymbol{\beta}_i + \frac{\sum_{j=1}^{M_i} \boldsymbol{\varepsilon}_{ij}}{M_i} \tag{7}$$

且服从正态分布 $\mathrm{N}(\boldsymbol{F}_i \mid \boldsymbol{\varphi}\boldsymbol{\beta}_i, \frac{\boldsymbol{\Sigma}}{M_i'})$，其中

$$M_i' = \left(\frac{\sum_{j=1}^{M_i} (\frac{L_{ij}}{\alpha})^{-\lambda}}{M_i^2} \right)^{-1} \tag{8}$$

引入中间变量 $\boldsymbol{K}$,

$$\boldsymbol{K} = \boldsymbol{I} + M_i' \boldsymbol{\varphi}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\varphi} \tag{9}$$

根据贝叶斯法则，可计算得到后验概率

$$P(\boldsymbol{\beta}_i \mid \boldsymbol{F}_i) = N(\boldsymbol{\beta}_i \mid \boldsymbol{K}^{-1}\boldsymbol{\varphi}^T\boldsymbol{\Sigma}^{-1}M_i'\boldsymbol{F}_i, \boldsymbol{K}^{-1}) \tag{10}$$

3.2 EM 迭代

采用 EM 算法以估计得到 PLDA 模型参数，本质上是进行极大似然估计求解含有隐变量的概率模型参数。在每一次迭代中，在 E-step 先求出给定训练数据下隐变量的期望，然后在 M-step 将这个期望最大化。通过迭代逐渐收敛，达到局部最优值。

（1）E-step：在给定观测数据和当前参数下对未观测数据 $\boldsymbol{\beta}_i$ 的条件概率分布 $P(\boldsymbol{\beta}_i \mid \boldsymbol{F}_i)$ 的期望值进行估算，即

$$E(\boldsymbol{\beta}_i) = \boldsymbol{K}^{-1}\boldsymbol{\varphi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\varphi} \tag{11}$$

又由期望相关公式可以得到：

$$E(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^T) = E(\boldsymbol{\beta}_i)E(\boldsymbol{\beta}_i)^T + \boldsymbol{K}^{-1} \tag{12}$$

（2）M-step：根据最大似然估计原理，对 $\prod_{i,j} P(\boldsymbol{x}_{ij}, \boldsymbol{\beta}_i)$ 求最大值，采用对数化简

$$\begin{aligned} &\max[\log \prod_{i,j} P(x_{ij}, \beta_i)] \\ &= \max \sum_{i,j} [\log P(x_{ij} \mid \beta_i) + \log P(\beta_i)] \end{aligned} \tag{13}$$

将 $P(\boldsymbol{x}_{ij} \mid \boldsymbol{\beta}_i)$ 和 $P(\beta_i)$ 的高斯分布概率密度函数代入公式（13），再分别对 $\boldsymbol{\Phi}$ 和 $\boldsymbol{\Sigma}$ 求导，整理得到

$$\boldsymbol{\varphi} = \frac{\sum_{i=1}^{N} (\sum_{j=1}^{M_i} (\frac{L_{ij}}{\alpha})^{\lambda} \boldsymbol{\eta}_{ij}) E(\boldsymbol{\beta}_i)^T}{\sum_{i=1}^{N} (\sum_{j=1}^{M_i} (\frac{L_{ij}}{\alpha})^{\lambda}) E(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^T)} \tag{14}$$

$$\boldsymbol{\Sigma} = \frac{\sum_{i,j} [(\frac{L_{ij}}{\alpha})^{\lambda} \boldsymbol{\eta}_{ij} (\boldsymbol{\eta}_{ij}^T - E(\boldsymbol{\beta}_i)^T \boldsymbol{\varphi}^T)]}{\sum_{i=1}^{N} M_i} \tag{15}$$

为得到对 $\boldsymbol{\Phi}$ 和 $\boldsymbol{\Sigma}$ 的最佳估算，需要经过 E-step 和 M-step 的不断迭代，当公式（13）计算得到的数值增长速度小于 1×10^{-3}，则停止迭代。

4 实验数据分析

4.1 评测数据

本文分别参考 ALIZE 开发包[12]和文献[11]提供 PLDA 开源代码，实现了 3 个说话人识别系统：基于 i-vector+CDS 的基线系统（简称“基线系统”），i-vector+PLDA 识别系统（简称“PLDA 系统”）和 i-vector+概率修正 PLDA 识别系统（简称“概率修正系统”）。采用 32 维 MFCC，训练 1024 阶的 UBM-GMM，i-vector 维数为 400，PLDA 说话人因子维数为 120。

为验证本文提出的概率修正 PLDA 模型的有效性，我们采用 NIST SRE10 core-core 测试集（女性）和 NIST 2014 i-vector machine learning challenge 测试集进行识别性能评估。

1） NIST SRE10 core-core 测试数据准备

UBM 训练数据：NIST2004、2005 年女性数据共 11370 条语音；

T 矩阵训练数据：NIST2004、2005、2006、2008 年女性数据共 20348 条语音；

PLDA 训练数据：与 T 矩阵训练数据相同语音提取的 i-vector；

core-core 测试条件下：

（1）模型—NIST SRE10core 女性数据，共训练模型 290 个；

（2）测试—NIST SRE10core 女性数据，共提供测试样本 357 个；

进行确认测试 355 次，冒认测试 15958 次。

2） 时长不匹配评测实验数据准备

将 NIST SRE10 core-core 测试集中的测试语音分别随机截短至 20s 和 10s，对应的 UBM、T、PLDA 模型、训练模型和测试次数不变。

3） NIST 2014 i-vector machine learning challenge 测试数据准备

NIST 2014 i-vector challenge 组委会从历年的 NIST SRE 数据库中提取 600 维的 i-vector 数据，分别组成开发集、模型集和测试集。开发集包含 4,959 个说话人共 36,573 个 i-vector，可用于 PLDA 模型训练；模型集包含 1,306 个说话人，每个说话人有 5 条 i-vector；测试集则有 9,634 个 i-vector。测试任务分成两个部分：progress 测试和 evaluation 测试。

4.2 调节参数 α 和 λ 的选择

公式（5）定义了时长约束下的说话人因子分布概率函数，可见，调整 α 和 λ 的取值，将改变说话人因子的概率分布。

为简化计算复杂度，在本文实验中，α 取开发集所有 i-vector 的时长均值。确定 α 后，再微调 λ 的取值，观察系统识别性能，发现：当 λ 取值在 0.4-0.8 之间，说话人识别系统将获得最显著的识别效果。

4.3 性能对比

为验证概率修正 PLDA 模型对时长变化的鲁棒性，本文将 NIST SRE10 core-core 测试集的测试数据进行截短至 20s 和 10s，分别进行不同时长的评测任务。本文采用等错率（equal error rate ，EER）和最小决策代价函数（minimum decision cost function，minDCF）作为说话人识别系统的评测准则，并对 minDCF 进行 norm 规整得到 Cnorm[13]。

表 1 NIST SRE10 core-core 评测 EER 值（不同时长）

时长/s	基线系统	PLDA 系统	概率修正系统
全时长	7.61%	3.66%	3.38%
20	12.39%	6.47%	6.21%
10	17.62%	9.80%	9.29%

表 2 NIST SRE10 core-core 评测 Cnorm 值（不同时长）

时长/s	基线系统	PLDA 系统	概率修正系统
全时长	0.2824	0.1901	0.1898
20	0.5034	0.3195	0.2883
10	0.6054	0.4145	0.4366

表 1 和表 2 分别列出了不同时长情况下，基线系统、PLDA 系统和概率修正系统这三个识别系统在 NIST SRE10 core-core 测试集（女性）上的评测结果。可看到，随着测试语音时长变短后，3 种系统的识别性能都有一定幅度的下降，其中，基线系统的识别性能下降最严重，而概率修正系统则表现地相对鲁棒。在同一时长情况下，概率修正系统取得更低的 EER 值，大部分情况下可以获得更小的 minDCF 值。只有在时长为 10s 的评测任务中，概率修正系统的 minDCF 值略高于 PLDA 系统，出现了类似于文献[8]的实验情况，值得进一步研究探讨。

NIST 2014 i-vector machine learning challenge 提供的每个 i-vector 都包含原始语音的段长信息，有利于应用概率修正 PLDA 系统验证性能。采用 EER 和 minDCF[14]作为说话人识别系统的评测准则。

表 3 NIST 2014 i-vector challenge 评测 EER 值

评测集	基线系统	PLDA 系统	概率修正系统
Progress	5.16%	3.27%	3.15%
Evaluation	4.49%	3.14%	3.12%

表 4 NIST S2014 i-vector challenge 评测 minDCF 值

时长情况	基线系统	PLDA 系统	概率修正系统
Progress	0.3859	0.3189	0.3081
Evaluation	0.3782	0.3076	0.2966

由表 3 和表 4 观察，发现，在 progress 测试任务中，概率修正 PLDA 系统的 EER 减少了 3.8%，minDCF 获得 3.5%的改进。在 evaluation 测试任务中，概率修正 PLDA 系统性能同样取得一定程度的改进。

5 结 语

鉴于传统 PLDA 模型缺乏对时长信息的利用，本文提出一种新的 PLDA 模型，在标准高斯 PLDA 建模过程中，利用时长信息控制说话人因子的概率分布参数，从而加强说话人识别系统对时长因素影响的鲁棒性。

参 考 文 献

[1] D Reynolds, TQuatieri, and R Dunn. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Process., 2000, 10(1-3): 19-41.

[2] P Kenny, G Boulianne, and P Dumouchel. Eigenvoice modeling with sparse training data[J]. IEEE Trans. Speech and Audio Process., 2005,13(3):345-354.

[3] P Kenny, G Boulianne, P Ouellet, et al. Joint factor analysis versus Eigenchannels in speaker recognition[J]. IEEE Trans. Audio Speech Lang. Process., 2007, 15(4):1435-1447.

[4] N Dehak, P Kenny, R Dehak, et al. Front-end factor analysis for speaker verification[J]. IEEE Trans. on Audio Speech Lang.Process., 2011, 19(4): 788- 798.

[5] S Prince and J Elder. Probabilistic linear discriminant analysis for inferences about identity[C].in Proc.Computer Vision, 2007, pp:1-8, Rio de Janeiro, Brazil.

[6] S Cumani, O Plchot and P Laface. On the use of i-vector posterior distributions in probabilistic linear discriminant analysis[J]. IEEE Trans. on Audio Speech Lang.Process., 2014, 22(4): 846- 857.

[7] A Sarkar, D Matrouf, P Bousquet, et al. Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification[C]. in Proc. InterSpeech, 2012, pp: 2661-2664, Portland, Oregon, USA.

[8] P Kenny, T Stafylakis, P Quellet, et al. PLDA for speaker verification with utterances of arbitrary duration[C]. in Proc. Acoustics, Speech and Signal Processing, 2013(ICASSP), pp: 7649 – 7653, Vancouver, Canada.

[9] T Hasan, R Saeidi, J Hansen, et al.Duration mismatch compensation for i-vector based speaker recognition systems[C].in Proc. Acoustics, Speech and Signal Processing, 2013(ICASSP), pp:7663 – 7667, Vancouver, Canada.

[10] A Kanagasundaram, D Dean, S Sridharan, et al. Improving short utterance i-vector speaker verification using utterance variance modeling and compensation techniques[J]. IEEE Trans. Speech Communication, 2014, 59: 69-82.

[11] D Garcia-Romero and C Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems[C].in Proceedings of Interspeech, 2011,pp:249-252, Florence, Italy.

[12] ALIZE.http://www.signalprocessingsociety.org/technicalcommittees/list/sl-tc/spl-nl/2013-05/ALIZE/

[13] NIST. The NIST 2010 Speaker Recognition Evaluation Plan. http://www.itl.nist.gov/iad/mig/tests/spk/2010/index.html.

[14] NIST. The 2013-2014 speaker recognition ivector machine learning challenge. https://ivectorchallenge.nist.gov, 2014.

Modified-prior PLDA Based Speaker Recognition

Li Lin[1], Wan Lihong[1], Hong Qingyang[1], Zhang Jun[1], Li Ming[2]

1. School of Information Science and Technology，Xiamen University，Xiamen 361005，China;
2. SYSU-CMU Joint Institute of Engineering, Sun Yat-Sen University, China

Abstract：To release the negative impact on the performance of speaker recognition systems due to the duration mismatch between enrollment utterance and test utterance, a modified-prior PLDA method is proposed. The probability distribution function of i-vector was modified by incorporating the covariance matrix with duration of each utterance of each speaker during the PLDA training, which further improved the discriminant capability of speaker classification. To evaluate the robustness of the proposed modified-prior PLDA method, extensive experiments were performed on NIST SRE10 core-core task (female part) in duration mismatch conditions and NIST 2014 i-vector machine learning challenge. Experimental results demonstrated that the duration-based modified-prior PLDA method achieved better compared with the traditional PLDA.

Keywords：Gaussian PLDA; i-vector; duration; modified-prior; speaker recognition