基于 PLDA 的多信道多语音说话人确认研究*

许云飞　周若华　颜永红

(中国科学院语言声学与内容理解重点实验室　北京　100190)

摘要: 在 NIST SRE 2012 年评测和实际应用中,可以用说话人的多个语音样本来注册说话人模型,并且这些语音样本取自于各种各样的信道。本文基于 PLDA,尝试了多种打分方法,并提出一种新的得分规整技术,在 NIST SRE 2012 核心测试集上,EER 平均提升 26.0%,MinCost 平均提升 12.4%。

关键词: 说话人识别,PLDA,多语音,得分规整

PLDA for Speaker Verification under Multi – Channel and Multi – Record

XU Yunfei, ZHOU Ruohua, YAN Yonghong

(Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China)

Abstract: In NIST SRE 2012 evaluation and practical applications, multiple recordings, which come from various channel conditions, can be used to train a speaker model. Based on PLDA, this paper will try several score methods and propose one score normalization technique. Equal error rate and minimum cost has been relatively improved 26.0% and 12.4% respectively on NIST SRE 2012 core test corpus.

Keywords: speaker recognition, PLDA, multi – record, score normalization

1　引言

近年来,基于总变化因子(i – vector)[1]技术的说话人识别系统获得显著提升[2-4]。该技术在 GMM 超矢量空间,不区分说话人和信道,而是将两部分当成一个整体建模。通过该技术,帧数不定的语音被映射为固定长度的低维矢量,与此同时尽可能保留了说话人信息[1]。由于 i – vector 是低维的,学者们可用更为复杂的信道补偿技术和打分方法进一步建模。其中,高斯 PLDA[5,11]因其优异的性能而得到学者们的广泛关注。

在 NIST SRE 2012 之前的评测任务中,每个注册说话人只有一句语音样本,PLDA 打分方法可以取得很好的性能[2,5,12,13,14]。在 NIST SRE 2012 评测任务中,每个注册说话人有多句语音样本,并且这些语音样本取自于不同的信道[15]。为了充分利用这些信息,文献[16,17,18]提出了一些有效方法。但是,在 NIST SRE 2012 新任务上,传统的 PLDA 打分方法无法获得较好的性能。因此,本文基于高斯先验 PLDA(以下简称 PLDA),尝试多种打分方法,同时提出一种新的得分规整技术,并分析每种方法对系统性能影响差异的原因。

本文组织如下:第二节简要介绍总变化因子技术;第三节介绍概率线性鉴别分析;第四节介绍复杂信道

本文于 2013 – 10 – 23 收到,2013 – 12 – 16 收到修改稿。

* 基金项目:本论文工作得到国家自然科学基金(批准号:10925419,90920302,61072124,11074275,11161140319,91120001,61271426),中国科学院战略性先导科技专项(面向感知中国的新一代信息技术研究,编号:XDA06030100,XDA06030500),国家 863 计划(资助号:2012AA012503)和中科院重点部署项目(编号:KGZD – EW – 103 – 2)经费资助。

下多语音 PLDA 打分方法和得分规整技术; 第五节为实验及结果; 第六节给出总结。

2 总变化因子

总变化因子技术旨在对说话人和信道同时建模,在 GMM 超向量空间不区分说话人和信道。给定一段语音,与说话人及信道相关的 GMM 超向量 M 由下式表示:

$$M = m + Tw + \varepsilon \tag{1}$$

m 为通用背景模型的超矢量,T 矩阵表示总变化空间,w 矢量是与说话人及信道相关的总变化因子,也就是最终得到的低维矢量 i - vector。ε 表示残差,和 w 均服从高斯分布。

$$w \sim N(0, I) \tag{2}$$

$$\varepsilon \sim N(0, \Sigma) \tag{3}$$

其中,Σ 为对角协方差阵。T 矩阵训练及 w 的计算请参见文献[1,19]。

3 概率线性鉴别分析

忽略 i - vector 的提取机制,PLDA 把它看成由一种生成式模型产生。在 PLDA 框架下,i - vector 的产生过程可以用一个隐藏变量来描述。不同的隐藏变量数目,不同的先验假设构成了不同的 PLDA 模型[2,11,12,20]。假定第 i 个说话人的第 j 个 i - vector 表示为 w_{ij},最常用的 PLDA 模型假设如下:

$$w_{ij} = \mu + Vy_i + z_{ij} \tag{4}$$

其中,μ 为所有训练数据的均值,V 矩阵表示说话人空间(本征音矩阵),矢量 y_i 为对应的说话人因子,服从标准高斯分布。z_{ij}表示残差,由一个全角矩阵 D 表示。

$$y_i \sim N(0, I) \tag{5}$$

$$z_{ij} \sim N(0, D) \tag{6}$$

3.1 模型训练

为了应用 PLDA,须在已标注数据集上通过期望最大化方法(EM)估计模型参数 $\lambda = (\mu, V, D)$,初始模型采用随机值。在 E - step,需要估计隐藏变量 y_i 的后验分布。该分布为高斯分布,均值和方差如下:

$$E[y_i] = (J(V^t\Sigma^{-1}V) + I)^{-1}\sum_{j=1}^{J} V^t\Sigma^{-1}(w_{ij} - \mu) \tag{7}$$

$$E[y_i y_i^t] = (J(V^t\Sigma^{-1}V) + I)^{-1} + E[y_i]E[y_i]^t \tag{8}$$

J 对应于第 i 个说话人的训练语音数目。在 M - step,利用 E - step 估计出的所有说话人的均值和方差来更新所有模型参数,如下:

$$\mu = \frac{1}{IJ}\sum_{i,j} w_{ij} \tag{9}$$

$$V = \left(\sum_{i,j}(w_{ij} - \mu)E[y_i]^t\right)\left(\sum_{i,j}E[y_i y_i^t]\right)^{-1} \tag{10}$$

$$D^{-1} = \frac{1}{IJ}\sum_{i,j}(w_{ij} - \mu)(w_{ij} - \mu)^t - VE[y_i](w_{ij} - \mu)^t \tag{11}$$

I 为训练数据里说话人数目总数。

3.2 打分

估计好模型参数后,给定两个 i - vector w_1 和 w_2,其对数似然比由公式计算,其中假设 θ_{tar}表示它们来自同一个说话人,θ_{non}表示它们来自不同的说话人

$$\text{score} = \log\frac{p(w_1, w_2 \mid \theta_{tar})}{p(w_1, w_2 \mid \theta_{non})} = \log N\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma_{tot} & \Sigma_{ac} \\ \Sigma_{ac} & \Sigma_{tot} \end{bmatrix}\right) - \log N\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma_{tot} & 0 \\ 0 & \Sigma_{tot} \end{bmatrix}\right) \tag{12}$$

其中,$\Sigma_{tot} = VV^T + D^{-1}$,$\Sigma_{ac} = VV^T$。因为 μ 代表所有 i - vector 的均值,在测试时,我们可以事先移除,那么令

$\mu=0$,通过 2×2 分块矩阵求逆[21],上式可化简为

$$\text{score} = w_1^T Q w_1 + w_2^T Q w_2 + 2 w_1^T P w_2 \tag{13}$$

公式省略了常数项,P 和 Q 定义如下:

$$Q = \Sigma_{tot}^{-1} - \left(\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac}\right)^{-1} \tag{14}$$

$$P = \Sigma_{tot}^{-1}\Sigma_{ac}\left(\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac}\right)^{-1} \tag{15}$$

化简后,大大提高测试速度。通过对 P 求本征值和对角化,速度还可以进一步提高[13]。

4　复杂信道下多语音 PLDA 打分方法和得分规整技术

NIST SRE 2012 评测任务与较早的评测任务有两点明显区别: ①NIST SRE 2012 评测任务中,注册说话人已知,一句测试语音要么属于且只属于其中一个说话人,要么不属于其中的任何一个说话人; 较早的评测任务中,要求判断一句测试语音是否属于某个说话人 ID,但两个(或多个)说话人 ID 有可能为同一个说话人,我们不知道该信息。②NIST SRE 2012 评测任务中,可以用多句语音样本来训练一个说话人模型,且每个说话人的语音样本数目不等,但同时来源、长度、信道、信号质量等也不同; 较早的评测任务中,每个说话人 ID 对应一句语音样本或几句信道来源相同的语音样本。也就是说,在 NIST SRE 2012 评测任务中,可以利用更多的语音样本来训练说话人模型,但语音样本间的变化也更为复杂; 与此同时,还可以利用说话人唯一这个信息。因此,无论是公式还是文献[11]中的打分方法都无法直接使用。原因在于: 公式只能比较两个 i-vector 之间的相似度,即一个说话人只有一个 i-vector; 每个注册说话人的 i-vector 数目不等,且由于各样本间差异较大,采用文献[11]中的打分方法会造成得分区间不一致,从而带来整体性能下降。

4.1　合并语音

目前,主流的说话人识别系统都是基于统计方法[1,2]。这种统计意义决定着长时语音相对短时语音更加准确地代表说话人信息。这也就是为什么基于 GMM-UBM 的系统,短时语音会严重影响系统性能的主要原因。本质上,i-vector 是由 GMM 统计量而来,语音时长越短,提取出来的 i-vector 越不准确,因而 PLDA 性能越差[22,23]。作为第一次尝试,本文将注册说话人的所有语音样本合并成一个长语音。

4.2　i-vector 求平均

合并语音没有考虑到各注册语音间的多样性(信道,时长,信号质量等)。如果这些语音中存在质量很差的语音,那么会影响说话人模型质量,从而导致该注册说活人的识别率下降; 并且,时长越长,下降越多。为了避免差语音带来整体性能过度下降,可考虑如下 i-vector 平均法。i-vector 为一段语音的 GMM 后验统计量在总变化空间各维度上的映射; 同一个说话人的所有 i-vector 在总变化因子空间上服从高斯分布。这与 PLDA 模型假设一致,因而 PLDA 可以获得较好性能。那么,本文将一个说话人的所有 i-vector 加起来求平均,利用均值 i-vector 代表目标该说话人。即

$$w_i = \frac{1}{J}\sum_{j=1}^{J} w_{ij} \tag{16}$$

4.3　提取说话人因子

i-vector 求平均法只考虑了说话人所有 i-vector 间的平均值,却忽略了表征它们内部变化的方差值。文献[24]提出用 PLDA 提取说话人因子并进行识别的方法,也就是用公式计算测试语音和注册语音的因子,然后计算因子间的余弦距离。在提取说话人因子的时候,同时利用了该说话人所有 i-vector 的均值与方差。原文中,对每一句语音单独提取一个说话人因子,但通过分析发现,该方法还可以将说话人的所有语音放到一起提取一个公共因子。

原文中采用余弦打分计算测试语音和注册语音间的相识度,除了计算量较小这一原因外,作者在本文实验中发现,由于提取后的因子不再与 PLDA 先验假设吻合,后端也就不能使用 PLDA 进行打分。所以本文实验中,采用余弦打分。

4.4 得分平均

虽然提取说话人因子方法利用了说话人所有 i - vector 的均值和方差信息，但由于该因子的后验服从高斯分布，取后验均值的方法忽略了估计该因子时的不确定性。

既然 PLDA 可以有效地比较两个 i - vector 之间的相似度，那么我们可以获取测试语音对说话人的所有注册语音的得分，之后再对这些得分后进行后处理就可以避免以上三种方法中存在的问题。实验表明，得分求平均有一定效果，具体做法是：假定测试语音对第 i 个说话人的所有语音得分为 $s_{i1}, s_{i2}, \cdots, s_{in_i}$，那么最终得分为

$$s_i = \frac{1}{n_i} \sum_{j=1}^{n_i} s_{ij} \tag{17}$$

4.5 指数得分平均

注意到 PLDA 给出的分数是以 e 为底的对数似然比，直接把分数求平均有一个风险：只要注册说话人的样本语音中存在一句语音引起漏检，并且该分数是一个较大的负数，就会引起整体得分偏低，最终导致所有目标测试语音对该说话人的漏检。通过指数求平均法可以有效抑制该问题，具体做法如下

$$s_i = \log\left(\frac{1}{n_i} \sum_{j=1}^{n_i} e^{s_{ij}}\right) \tag{18}$$

从公式(17)可以看出，如果说话人的语音样本中存在个别质量较差的，会导致目标测试语音对它们打分很低，但由于公式(18)中在求和时，以 e 为底的指数对负数不敏感，也就是质量较差的语音样本在整体得分中所占的比重大大减小，以至于可以忽略，从而避免了目标测试语音对该说话人漏检。

4.6 得分取中值

因为指数求和时，只要存在一个正分，那么整体得分就为正。也就是指数得分平均法可以避免因个别语音引起的漏检，但无法避免个别语音引起的虚警。在 NIST SRE 2012 评测所指定的注册说话人中，某些说话人只给了 interview 类型下的语音样本，虽然数量众多，但是有很多信号质量不好。此时，取中值在某种程度上可以减少虚警。具体做法是，将测试语音对注册说话人的所有语音得分进行排序，取排在中间位置的那个分数代表整体得分。

4.7 指数得分平均 + 去最大最小

受得分取中值思想启发，在注册说话有很多语音样本时，不必使用所有得分。在指数得分平均前，先对所有得分进行排序，去掉最大分和最小分。这么做，在一定程度上能同时减少虚警和漏检。

4.8 对数似然比得分规整

恰当的得分规整技术可以有效提升系统性能。然而，无论是传统的 Znorm Tnorm [25] 还是最近提出的 Snorm [26]，对 PLDA 系统都没有帮助 [2,13]。

这里提出一种得分规整方法，旨在：①测试语音属于所有注册说话人中的某一个时，拉开最高得分和次高分之间的距离；②测试语音不属于任何一个注册说话人时，将所有得分拉低。

假设总共有 N 个说话人，测试语音对第 i 个说话人的得分为 s_i，那么规整后的分数为

$$s_i := \frac{L^{s_i}}{\sum_{j=1}^{N} L^{s_j}} \tag{19}$$

L 为一大于 1 的数，可根据开发集调整得到，一般来说不应过大。L 的大小可以控制最高分和次高分之间的距离。考虑到 PLDA 得分为对数似然比，本文如下使用得分规整公式

$$s_i := \log \frac{L^{s_i}}{\sum_{j=1}^{N} L^{s_j}} \tag{20}$$

5 实验及结果

本文所有的实验都是在 NIST SRE 2012 核心测试集上进行的。

5.1　前端处理

实验中所使用的特征为 60 维 MFCC 特征,其基本特征由 19 维的基本倒谱系数和一维能量构成; 然后对基本特征做一阶差分和二阶差分。在提取特征前,采用 BUT Hungarian phoneme recognizer [27] 对数据进行语音/静音切割; 对于麦克风数据,由于存在串扰,需要采用 B 信道的信息对 A 信道进行处理 [28] 。然后按窗长 25ms、窗移 10ms 提取 60 维的 MFCC 特征,最后使用倒谱均值减(CMN) 及倒谱方差规整技术(CVN) 对特征进行规整。

5.2　模型训练及数据选取

本文所有模型都是性别相关的。使用 NIST SRE 2004、2005、2006 电话语音训练 GMM - UBM,高斯数为 2048。使用 NIST SRE 2004、2005、2006、2008、2010、Switchboard、Fisher 所有语音来训练总变化矩阵; 最终得到 600 维的 i - vector。PLDA 训练数据数据来自 NIST SRE 2006、2008、2010 其中 male 8131 句话,female 12452 句话。

5.3　实验结果

本文采用等错率(EER) 和 NIST SRE 2012 规整后的最小检测错误代价(MinCost) [15] 两种指标对说话人识别系统进行评价。把 i - vector 送入 PLDA 系统前,先对其做长度规整 [11] 预处理。

表 1 和表 2 中各系统性能第一个为 EER(%) ,第二个为 MinCost,粗体表示性能最好。为了对比,本文从注册说话人中随机选取一句语音样本代表该说话人作为基线结果。比较表一、二中第 1 行与第 2 行发现,“合并语音”后,系统性能反而有所下降。原因在于: 合并较差的语音,会影响整个语音的质量; 因为 i - vector 是基于 GMM 后验统计量,故估计出来的 i - vector 不准确,从而导致 PLDA 无法对说话人准确建模。虽然“i - vector 求平均”与“合并语音”两个方法性能相当,但它给我们提供了另外一个思路: 既然每个 i - vector 都估计准确,那么通过寻找一个更加准确的映射函数,可以将多个 i - vector 合并为以一个,既避免了“合并语音”中存在的问题,还可以将模型简单化。‘提取说话人因子”方法考虑了各语音间的多样性,在得分规整前,仅仅利用余弦打分,整体性就可达到最好。‘得分平均’、‘指数得分平均’、‘得分取中值”三种方法各有长短,性能相当。“去最大最小分”具有同时规避虚警和漏检的作用,结合“指数得分平均”后,系统性能性能得到一定提升。

表 1　各系统的性能

打分方法	condition 1	condition 2	condition 3	condition 4	condition 5
随机选取	7.456/0.472	4.327/0.402	8.128/0.518	7.121/0.528	5.301/0.408
合并语音	7.629/0.500	7.081/0.442	8.855/0.507	8.908/0.551	8.215/0.459
i - vector 求平均	7.559/0.489	6.935/0.446	8.673/0.493	9.211/0.566	7.983/0.464
提取说话人因子	**5.765/0.377**	**2.677/0.300**	**7.172/0.422**	**4.221**/0.423	**3.219**/0.300
得分平均	6.558/0.392	2.907/0.288	7.400/0.456	5.407/0.419	3.580/0.292
指数得分平均	6.593/0.466	2.747/0.300	7.765/0.469	5.156/0.439	3.403/0.303
得分取中值	6.593/0.403	3.060/0.306	7.400/0.464	5.546/0.438	3.787/0.306
指数得分平均 + 去最大最小	6.489/0.451	2.692/**0.271**	7.426/0.463	4.891/**0.412**	3.296/**0.286**

表 2　对数似然比得分规整后各系统的性能

打分方法	condition 1	condition 2	condition 3	condition 4	condition 5
随机选取	7.000/0.465	2.788/0.362	7.400/0.516	5.700/0.460	3.348/0.354
合并语音	7.174/0.462	5.505/0.427	7.631/0.456	7.375/0.508	5.949/0.426
i - vector 求平均	6.904/0.444	5.386/0.437	7.553/0.441	7.581/0.530	5.820/0.430
提取说话人因子	21.023/0.838	29.388/0.719	24.096/0.854	29.434/0.753	28.925/0.699
得分平均	5.420/0.331	1.686/0.265	6.206/0.417	4.111/0.377	1.932/0.231
指数得分平均	5.626/0.397	1.402/0.242	6.882/0.441	4.111/0.343	1.621/0.227
得分取中值	5.358/**0.328**	1.770/0.280	6.258/**0.408**	4.293/0.422	2.112/0.232
指数得分平均 + 去最大最小	**5.178**/0.353	**1.333/0.223**	**6.077**/0.435	**3.721/0.319**	**1.541/0.178**

表 2 给出了各系统得分规整后的性能。其中，“提取说话人因子”方法中，由于其得分不是对数似然比，且分数值都较小（ <1，余弦打分决定的），采用本文提出的得分规整技术后性能下降很多。得分规整后，“得分取中值”方法在条件一、三上的 MinCost 最好，在其他情况下，都是“指数得分平均 + 去最大最小分”最好。在某种意义上，MinCost 可以看成带权重的 EER，MinCost 对于虚警的惩罚高于漏检（ NIST SRE 2008，1:0.01；NIST SRE 2010，1:0.001； NIST SRE 2012，1:0.01 ~0.001）。“指数得分平均”虽然可以抑制漏检，但无法避免虚警；“去最大最小分”可以弥补虚警问题，但不彻底。因为 interview 的语音质量不如 telephone 的语音质量，因而在一、三两个 interview 测试条件上，即便“指数得分平均 + 去最大最小分”的 EER 比“得分取中值”好，但后者的 MinCost 却更好。

通过比较表 1 和表 2 对应行发现，使用本文所提得分规整技术显著提升了识别性能。平均提升是指，将各系统的性能提升求平均。在 EER 上，5 个测试条件平均提升 13.67%、38.49%、14.50%、21.53%、41.59%；MinCost 上看，5 个测试条件平均提升 13.00%、9.51%、7.42%、11.39%、20.65%。

表 3 训练数据对 PLDA 系统的性能影响

训练数据	condition 1	condition 2	condition 3	condition 4	condition 5
注册说话人语音	8.630/0.519	5.740/0.416	9.190/0.535	6.940/0.454	6.700/0.415
大量语音	6.489/0.451	2.692/0.271	7.426/0.463	4.891/0.412	3.296/0.286

从表 3 中发现，训练数据并不是越多越好。“大量语音”包括“注册说话人语音”和 NIST SRE 2004 及 switchboard 2p2 语音。“注册说话人语音”为 NIST SRE 2012 评测中所指定的目标说话人的训练语音。由于信道变化差异较大，训练语音和测试语音越匹配，性能越好。

6 结束语

本文介绍了总变化因子技术以及 PLDA 系统。提出了七种 PLDA 打分方法和一种得分规整技术。将注册说话人的所有语音拼接，混入质量较差的语音会带来性能下降。“虽然 i - vector 求平均”性能与合并语音”性能相当，但提供了简化模型的思路。“提取说话人因子”有较好的性能，但没有考虑因子的不确定性，并且无法使用本文提出的得分规整技术。“得分平均”无法避免漏检问题。“指数得分平均”可以避免漏检但无法避免虚警。“得分取中值”具有抑制虚警作用。“去最大最小分”可同时避免虚警和漏检，结合“指数得分平均”有一定容错能力，相对更为鲁棒。

对数似然比得分规整显著提升了说话人识别性能。在信道分布较为复杂时，用匹配的数据训练 PLDA 可以较大提升系统性能。

参 考 文 献

[1] N Dehak，P Kenny，R Dehak，et al. Front - End Factor Analysis For Speaker Verification [J]. IEEE Transactions on Audio，Speech and Language Processing，2011，19(4)：788 - 798

[2] P Kenny. Bayesian speaker verification with heavy tailed Priors. Brno，Czech Rebublic：Proceedings of Odyssey Speaker and Language Recognition Workshop，2010.

[3] N Brummer. EM for Probabilistic LDA. https：//sites. google. com/site/nikobrummer，Feb. 2010.

[4] M Senoussaoui，P Kenny，N Brummer，et al. Mixture of PLDA models in i - vector space for gender independent speaker recognition. Florence，Italy：Proceedings of International Conference on Speech Communication and Technology，Aug. 2011.

[5] P Matejka，O Glembek，F Castaldo，et al. Full - covariance UBM and heavy - tailed PLDA in i - vector speaker verification. Prague，Czech Republic：Proceedings of IEEE International Conference on Acoustics，Speech，and Signal Processing，2011：4536 - 4539

[6] L Burget，O Plchot，S Cumani，et al，Discriminatively trained probabilistic linear discriminant analysis for speaker verification. Prague，Czech Republic：Proceedings of IEEE International Conference on Acoustics，Speech，and Signal Processing，2011：4832

－4835

[7] N Dehak,R Dehak,J Glass,et al. Cosine similarity scoring without score normalization techniques. Brno,Czech Rebublic: Proceedings of Odyssey Speaker and Language Recognition Workshop,2010.

[8] S Cumani,N Brummer,L Burget,et al. Fast discriminative speaker verification in the i－vector space. Prague,Czech Republic: Proceedings of IEEE International Conference on Acoustics,Speech,and Signal Processing,2011: 4852－4855

[9] J Villalba,N Brummer. Towards fully Bayesian speaker recognition: Integrating out the between speaker covariance. Florence,Italy: Proceedings of International Conference on Speech Communication and Technology,Aug. 2011.

[10] T Stafylakis,P Kenny,M M Senoussaoui,et al. Preliminary investigation of Boltzmann machine classifiers for speaker recognition. Biopolis,Singapore: Proceedings of Odyssey Speaker and Language Recognition Workshop,2012.

[11] Simon J D Prince,James H Elder. Probabilistic linear discriminant analysis for inferences about identity. Rio de Janeiro,Brazil: Proceedings of International Conference on Computer Vision,2007: 1－8

[12] Yang Hai,Liang Chun－Yan,Xu Yun－Fei,et al. Sparse probabilistic linear discriminant analysis for speaker verification. Portland,Oregon: Proceedings of International Conference on Speech Communication and Technology,2012.

[13] D Garcia Romero,C Espy Wilson. Analysis of i－vector length normalization in speaker recognition systems. Florence,Italy: Proceedings of International Conference on Speech Communication and Technology,2011.

[14] P M Bousquet,A Larcher,D Matrouf,et al. Variance－spectra based normalization for i－vector standard and probabilistic linear discriminant analysis. Biopolis,Singapore: Proceedings of Odyssey Speaker and Language Recognition Workshop,2012.

[15] NIST 2012 Speaker Recognition Evaluation Plan. http://nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan－v17－r1.pdf. 2012.

[16] Hanwu Sun,Kong Aik Lee,Bin Ma. Anti－model kl－svm－nap system for NIST SRE 2012 evaluation. Lyon,France: Proceedings of International Conference on Speech Communication and Technology,2013.

[17] Xiao Fang,Najim Dehak,James Glass. Bayesian distance metric learning on i－vector for speaker verification. Lyon,France: Proceedings of International Conference on Speech Communication and Technology,2013.

[18] Bengt J Borgstrom,Alan McCree. Discriminant trained Bayesian speaker comparison of i－vectors. Lyon,France: Proceedings of International Conference on Speech Communication and Technology,2013.

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

[20] Niko Brummer,Edward de Villiers. The speaker partitioning problem. Brno,Czech Republic: Proceedings of Odyssey Speaker and Language Recognition Workshop,2010.

[21] Christopher M Bishop. Pattern recognition and machine learning [M]. Singapore: Springer,2006.

[22] Sandro Cumani,Oldrich Plchot,Pietro Laface. Probabilistic linear discriminant analysis of i－vector posterior. Lyon,France: Proceedings of International Conference on Speech Communication and Technology,2013.

[23] Patrick Kenny,Themos Stafylakis,Pierre Ouellet,et al. PLDA for speaker verification with utterances of arbitrary duration. Lyon,France: Proceedings of International Conference on Speech Communication and Technology,2013.

[24] N Dehak,Z Karam,D Reynolds,et al. A channel－blind system for speaker verification. Prague,Czech Republic: Proceedings of IEEE International Conference on Acoustics,Speech,and Signal Processing,2011.

[25] R Auckenthaler,M Carey,H Lloyd－Thomas. Score normalization for text－independent speaker verification systems [J]. Digital Signal Processing,2000,10(1－3): 42－54

[26] N Brummer,A Strasheim. AGNITIO's Speaker Recognition System for EVALITA 2009. Italy: The 11th Conference of the Italian Association for Arti? cial Intelligence,2009.

[27] P Schwarz,M Pavel,J Cernocky. Hierarchical structures of neural networks for phoneme recognition. Toulouse,France: IEEE International Conference on Acoustics,Speech,and Signal Processing,2006.

[28] Luciana Ferrer,Yun Lei,Mitchell McLaren,et al. SRI NIST 2012 SRE System Description. 2012.

作者简介

许云飞,1988 年 1 月,男,博士在读,说话人识别。

周若华,1972 年 1 月,男,研究员,音乐信号处理。

颜永红,1967 年 3 月,男,研究员,语音识别。