

# 采用非监督得分规整和因子分析的说话人确认

郭 武,李轶杰,戴礼荣,王仁华

(中国科技大学电子工程与信息科学系,安徽合肥 230027)

**摘 要:** 在文本无关的说话人确认中,规整算法能够有效地调整测试得分的分布.另外,利用前面已经得到的测试语句的得分来调整规整的参数可以取得更好的效果,这种规整叫做非监督得分规整.在本文中,借用开发集得分来建立说话人和冒认者得分的两个先验高斯分布函数,在实际的测试中,利用最大后验概率准则来对规整的模型参数进行调整.在采用因子分析的情况下,在 NIST 2006 说话人识别测试 1conv4w-1conv4w 数据库上,能够取得等错误率 5.26%.

**关键词:** 说话人确认;联合因子分析;非监督得分规整

**中图分类号:** TN912.34 **文献标识码:** A **文章编号:** 0372-2112(2009)04-0776-04

## Speaker Verification Based on Unsupervised Normalization and Factor Analysis

GUO Wu, LI Yi-jie, DAI Li-rong, WANG Ren-hua

(Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui 230027, China)

**Abstract:** In the text-independent speaker verification, the normalization algorithm can adjust the score distribution. The previous test scores can be used to update the parameters of the normalization, which is defined as unsupervised score normalization in this paper. The scores distributions of the target and impostor in the development corpus are set up as a prior, and the parameters of normalization are updated using the maximum a posteriori (MAP) algorithm in each test process. In the NIST 2006 speaker recognition evaluation (SRE) 1conv4w-1conv4w corpus, the equal error rate (EER) of the system based on the factor analysis and unsupervised score normalization is 5.26%.

**Key words:** speaker verification; joint factor analysis; unsupervised score normalization

### 1 引言

在说话人确认 (speaker verification) 的研究中,利用以前测试语句的信息来动态地调整说话人模型的参数,这叫做非监督的自适应 (unsupervised adaptation),非监督的自适应能够提高说话人识别的性能,对于说话人辨认 (speaker identification) 的实际系统具有更强的意义.从 2006 年开始<sup>[1]</sup>,在 NIST 的说话人识别大赛上,就允许参赛方使用非监督的自适应,对每个模型而言,后面测试语句可以利用前面的测试语句的信息,但是每个模型间不许互相使用,文件的测试列表顺序也不允许改变,如图 1.

非监督自适应最简单的方法是在测试中设置门限,把得分超过门限的测试语句重新加入到模型自适应算法中,重新训练相应的说话人模型<sup>[2,3]</sup>.这种方法有效但是速度非常慢,适合在运算量要求不高的支持向量机<sup>[4]</sup>系统中实现,在混合高斯模型-通用背景模型

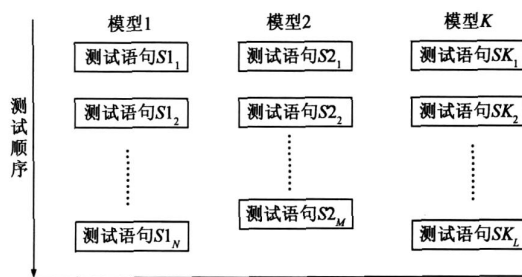


图1 NIST规定的测试顺序

(Gaussian Mixture Models, Universal Back-Ground Model, GMM, UBM) 模型中,会影响计算最大的  $N$  个得分 (TOP-N)<sup>[5]</sup> 快速测试算法,影响系统实时实现.

在本文中,采用另外一种思路来动态利用前面测试语句的信息,也就是仅仅利用测试语句的得分来对规整算法的参数进行调整,而不去更新说话人的模型,把这种算法叫做非监督的得分规整.

另一方面,因子分析<sup>[6,7]</sup>是目前说话人识别中解决

收稿日期:2007-10-15;修回日期:2008-09-25

基金项目:微软基金(No. 07122803);中国科技大学青年教师基金

信道问题的最有效手段之一,可以大幅度地提高说话人识别的性能,获得了广泛地承认,本文也采用了因子分析的手段来提高性能。

## 2 得分规整

在说话人识别系统测试过程中,由于每次测试的得分差异性很大,因此门限的稳定性非常难以确定。得分的差异性来自多方面的。首先,注册的语音由于时间很短,因此语音中的语义信息、信道、环境噪声会不可避免的影响说话人的模型的建立。其次,测试时的数据和训练时数据的不匹配是造成得分分布的差异性的主要原因。综合这两方面的原因,会出现得分分布的以下两种情况的不一致性:

(1) 同一说话人的不一致性(intra speaker variability),同一个人在说话时因为健康、语义、感情、信道、环境、录音的媒质等原因的影响,导致两次说话中出现差异,表现在目标说话人测试时的得分上就会呈现一个概率分布,而不是简单的只有一个固定的数值。

(2) 不同说话人的不一致性(inter speaker variability),实际上判决就是希望把这种不一致性扩大,但是有时这种不一致性在某些人之间却相当小,这也是影响说话人识别门限设置的因素。在与文本无关的说话人识别中这种问题尤为严重,这种不一致性由于不直接影响,因此在算法上很难衡量。

得分规整<sup>[8]</sup>(score normalization)的目的就是希望减小同一说话人的不一致性,扩大不同说话人的不一致性。规整的最基本方法是把测试得分通过式(1)进行规整, $L_s(X)$ 是语句 $X$ 对于说话人 $s$ 的初始得分, $\hat{L}_s(X)$ 是规整后的得分。其中 $\mu_s$ 、 $s$ 是对于说话人 $s$ 的规整参数,这两个参数需要用大量数据进行估计。

$$\hat{L}_s(X) = \frac{L_s(X) - \mu_s}{s} \quad (1)$$

针对具体应用,研究者提出了许多规整算法,典型的有测试规整(test normalization, TNorm),零规整(zero normalization, Znorm),话筒规整(handset normalization, HNorm)<sup>[8]</sup>,在这些规整中,参数都是 $\mu_s$ 、 $s$ ,它们一般采用冒认者或者冒认语音的得分分布来得到,这是由于在实际的说话人识别应用中,用冒认者的模型或者语句很容易获得冒认者得分数据的分布,但是很难获得单个目标说话人的大量得分。

## 3 因子分析

联合因子分析<sup>[6]</sup>是去除信道信息的一种算法,另外,因子分析之后的测试得分采用规整算法能够比基线的 GMM-UBM 系统采用规整得到更大的系统性能提升。在说话人识别实际应用中,一般采用式(2)来估计信道因子空间:

$$m_h(s) = m_{ubm} + y(s) + Ux_h(s) \quad (2)$$

$m_{ubm}$ 为 UBM 模型的均值超矢量<sup>[9]</sup>,  $y(s)$ 认为是特定说话人 $s$ 的与信道无关的均值超矢量相对 $m_{ubm}$ 的偏移, $U$ 代表信道空间, $x_h(s)$ 代表具体某句话中的信道因子。因此式(2)很明确:对每个说话人 $s$ 的第 $h$ 句话,希望尽量把 $Ux_h(s)$ 代表的信道信息去除,只保留与特定说话人 $s$ 相关的信息<sup>[10]</sup>。

## 4 非监督的得分规整

对某个具体的目标说话人的测试中,最初可采用 TNorm 或者 ZNorm 的方法得到冒认者的得分分布,可以得到均值 $\mu$ 和标准差 $\sigma$ 。在每次测试的过程中,可以得到一个新的得分 $s$ ,可以用这个新的得分来更新 $\mu$ 和 $\sigma$ 。但是,一个很重要的问题是需要判断这个新得分 $s$ 到底是不是冒认者的得分,由于没有目标说话人得分的分布,一个最简单的思路是采用固定门限判决的方法,如果得分小于某个门限就认为是冒认者得分,用这个得分来更新分布的 $\mu$ 和 $\sigma$ ,这种方法简单但效果不好。所以关键问题是判决得分 $s$ 到底是不是目标说话人的得分。

我们在第2部分提到,无法得到一个目标说话人的得分分布,但是如果有大量的目标说话人的得分的话,那么用这些得分也可以估计出一个先验分布,用这个分布就可以作为所有目标说话人得分的先验分布 $\pi_{target}$ ;同样,也可以得到所有冒认得分的先验分布 $\pi_{impostor}$ 。下图2绘制的是2005年1conv4w-1conv4w测试中所有冒认者和目标说话人经过 TNorm 之后的得分分布,两个分布都非常接近高斯分布,为简单起见,假设 $\pi_{target}$ 和 $\pi_{impostor}$ 都是一个单高斯的分布,这个假设前提其实也是所有得分规整算法的一个假设。

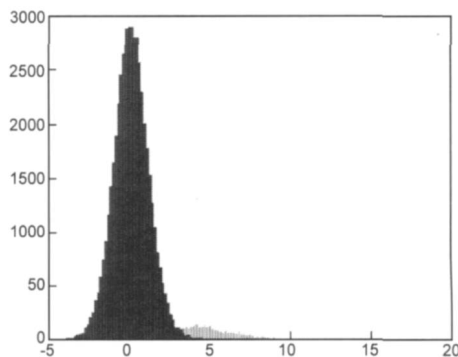


图2 NIST 2005年1conv4w-1conv4w所有得分分布

在测试的过程中,利用 $\pi_{target}$ 和 $\pi_{impostor}$ 可算出得分 $s$ 属于冒认者还是目标说话人的概率,如式(3):

$$P_{target} = \frac{p(s|\pi_{target})}{p(s|\pi_{target}) + p(s|\pi_{impostor})} \quad (3)$$

$$P_{impostor} = 1 - P_{target}$$

式(3)中有两个假设,首先认为单个目标说话人的得分

分布与所有说话人的得分分布一致,由于所有的得分是首先经过了 TNorm 规整的,按照理论上说,冒认者得分的分布是一个标准正态分布,而目标说话人的得分分布是无法获得的,因此这个假设可以认为近似成立.第二个假设  $p_{\text{target}}$  和  $p_{\text{impostor}}$  两个分布是等概率分布的,图 2 中可以看出 NIST 测试中这  $p_{\text{impostor}}$  分布的概率要远远大于  $p_{\text{target}}$  分布,但是实际中影响性能的是  $S$  处在两个先验的得分分布重叠区域,而在这种情况下对先验概率进行加权意义就不大了,因此式(3)采用的是等概率分布.采用式(4)对规整的参数  $\mu$  和  $\sigma$  进行调整,具体调整的算法采用 MAP,在 MAP 过程中,因为只有一个点,调整  $\sigma$  是不合适的,只调整均值参数  $\mu$ .

$$\mu^{\text{new}} = \frac{P_{\text{impostor}}}{P_{\text{impostor}} + 1} S + \frac{1}{P_{\text{impostor}} + 1} \mu^{\text{old}} \quad (4)$$

参数是 MAP 的相关因子,在实际中取 12 ~ 100 都能够取得效果提高,在我们的实验中取 80 能够取得最好的实验结果.

在得到了更新的参数之后,采用最简单的 ZNorm 的规整算法对每次测试的得分进行规整,对下一次测试得分  $S$  采用新参数  $\mu^{\text{new}}$  进行规整,如式(5):

$$\hat{S} = \frac{S - \mu^{\text{new}}}{\sigma} \quad (5)$$

## 5 实验配置和结果

### 5.1 数据库及评测方法

本文采用 NIST 2006 年说话人识别比赛的 1conv4w-1conv4w 作为实验,总计有 810 个目标说话人,有 3700 多句话作为测试语句,去除无效的测试外,总计有 51448 次测试,其中目标说话人的测试有 3612 次,冒认者测试有 47836 次.测试语音和训练语音为 5 分钟左右的对话中的语音抽出的一个声道的声音,实际中有语音部分大约有 2.5 分钟左右.

采用 NIST 2004 数据库作为因子分析中信道估计的开发集,该数据库中总计有 310 个说话人,选择训练和测试集中有三种以上不同信道信息的说话人的语音作为信道子空间估计的训练数据,经过筛选后,选择 223 人,2320 段语音,平均每人大约有 10 段说话.我们就用这些语音来训练信道矩阵  $U$ .

采用 NIST 2005 1conv4w-1conv4w 数据库中的数据作为得分规整的开发集,另外采用 NIST 2004 1conv4w 训练集作为 TNorm 的数据库.

### 5.2 特征参数提取

本文采用的是 39 维的 MFCC 参数,对于 MFCC 参数提取,语音信号先去直流,预加重(因子为 0.97),经过帧宽 25ms,帧移是 10ms 的汉明窗.在抽取 MFCC 特征参数的同时,采用基于能量的寂静帧检测算法去除静音

帧.抽取 0 - 12 维 MFCCs,总计为 13 维,特征参数通过倒谱减均值(cepstrum mean subtraction, CMS)和倒谱低通滤波(RelAtive SpecTrA, RASTA)<sup>[5]</sup>去除信道卷积噪声,通过一阶差分、二阶差分总计构成 39 维,最后特征通过短时高斯化以提高识别率.

### 5.3 系统描述

首先实现了一个基线系统,这个系统是表 1 中的系统 1.该系统采用标准的 2048 高斯的 GMM-UBM<sup>[5]</sup>算法,得分采用 Tnorm 规整.

另外,构建了系统 2 到系统 6,这些系统都是采用 512 高斯的 GMM 系统,采用因子分析的算法去除信道信息,具体实现如下:

首先采用 NIST 2004 年训练集的数据用期望最大化(expectation Maximization, EM)算法训练一个 UBM 模型,在本文中 GMM 采用 512 个高斯能够取得最好的识别性能,分别为男声、女声单独训练一个 256 个高斯拼接起来形成一个 512 的 UBM 系统.利用因子分析估计出信道空间,在实验中信道因子数  $R_c$  取 30,在模型训练和测试中去除信道影响.

系统 2 为不做得分规整的因子分析系统,系统 3 是在系统 2 基础上采用 2005 年 1conv4w 训练集中所有的语句分性别做 Znorm 规整的系统.

系统 4、5、6 都需要先对所有的得分做 Tnorm, Tnorm 的数据为 NIST 2004 年 1conv4w 训练集中所有的 246 个男声语句和 370 个女声语句.系统 6 中的所有目标说话人得分的先验分布  $p_{\text{target}}$  和冒认者得分先验分布  $p_{\text{impostor}}$  采用 NIST 2005 年 1conv4w-1conv4w 测试中的所有测试得分来估计.

### 5.4 结果

本文采用等错误率(Equal Error Rate, EER)、最小检测代价函数(Minimum Detection Cost Function, MinDCF)和 DET(Detection Error Tradeoff)曲线<sup>[1]</sup>三个参数来衡量系统性能,各系统的 EER 和 minDCF 如下表 1 所示.

表 1 2006 年 1conv4w-1conv4w 不同系统实验结果

| 序号 | 系统描述                          | EER    | MinDCF |
|----|-------------------------------|--------|--------|
| 1  | GMM-UBM 基线系统,高斯数为 2048, TNorm | 9.0 %  | 0.042  |
| 2  | 基于因子分析的 GMM-UBM               | 7.07 % | 0.032  |
| 3  | 基于因子分析的 GMM-UBM, ZNorm        | 6.60 % | 0.031  |
| 4  | 基于因子分析的 GMM-UBM, TNorm        | 5.89 % | 0.028  |
| 5  | 基于因子分析的 GMM-UBM, TZNorm       | 5.81 % | 0.028  |
| 6  | 基于因子分析的 GMM-UBM, 非监督的得分规整     | 5.26 % | 0.026  |

另外,把系统 1、2、5、6 的 DET 曲线绘制出来,如图 3 所示.

从 EER、minDCF 和 DET 的结果来看,系统 6 相对于基线系统在 EER 上有 41 % 的降低,相对于单独做

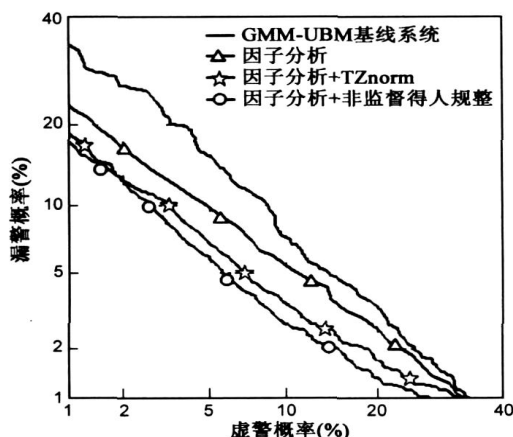


图3 系统1、2、5、6的DET曲线

TNorm的系统4等错误率由5.89%降低到5.26%,考虑到5.89%已经是一个识别率很高的系统,提高也是很可观的,优于NIST 2006年评测第一名STBU<sup>[4]</sup>的5.4%和LPT<sup>[2]</sup>的5.9%的结果,而且相对STBU系统而言,在前端特征参数上没有进行任何复杂的处理技术。

## 6 结论和分析

本文相对于传统的说话人得分规整算法而言,在测试得分中,采用MAP的方法动态地更新规整参数,实际上可以认为是一种动态的ZNorm;与因子分析相结合,取得了非常好的识别性能提高。

可以看出,在TNorm之后作ZNorm,也就是实验中的系统5中的TZNorm相对仅做TNorm的系统,性能提高很小;而非监督的得分规整能够提高很大,这主要还是由于ZNorm中选择的数据和实际测试中的数据分布不一样,而采用动态的参数更新之后,使规整的参数与测试实际环境下的冒认者分布趋于一致,因此规整后能够取得明显的性能提高。

本文推荐的算法与对说话人模型进行非监督自适应方法比较,好处在于基本上不会额外再耗费计算量,进一步的工作可以将模型非监督的自适应和得分的非监督的规整都结合起来,相信能够进一步提高性能。

## 参考文献:

- [1] NIST. The NIST Year 2006 Speaker Recognition Evaluation Plan [OL]. [http://www.nist.gov/speech/tests/spk/2006/sre06\\_evalplan\\_v9.pdf](http://www.nist.gov/speech/tests/spk/2006/sre06_evalplan_v9.pdf).
- [2] C Vair, D Colibro, F Castaldo, et al. Loquendo-politecnico di Torino's 2006 NIST speaker recognition evaluation system [A]. Proceedings of INTERSPEECH [C]. Antwerp, Belgium, 2007. 1238 - 1241.
- [3] A Preti, J F Bonastre, D Matrouf. Confidence measure based unsupervised target model adaptation for speaker verification [A]. Proceedings of INTERSPEECH [C]. Antwerp, Belgium, 2007. 754 - 757.

- [4] P Matějka, L Burget, P Schwarz, O Glembek, et al. STBU system for the NIST 2006 speaker recognition evaluation [A]. Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) [C]. Honolulu, Hawaii, USA, 2007. 4:221 - 224.
- [5] D A Reynolds, T F Quatieri, R B Dunn. Speaker verification using adapted Gaussian mixture models [J]. Digital Signal Processing, 2000, 10(1 - 3): 19 - 41.
- [6] P Kenny, G Boulianne, P Ouellet, P Dumouchel. Speaker and session variability in GMM-based speaker verification [J]. IEEE Transactions on Audio, Speech and Language Processing, 2007, 15(4): 1448 - 1460.
- [7] R Vogt, B Baker, S Sridharan. Modeling session variability in text-independent speaker verification [A]. Proceedings of INTERSPEECH [C]. Lisbon, Portugal, 2005. 3117 - 3120.
- [8] Frédéric Bimbot, Jean-François Bonastre. A tutorial on text-independent speaker verification [J]. EURASIP Journal on Applied Signal Processing, 2004, (4): 430 - 451.
- [9] W M Campbell, D E Sturim, D A Reynolds, et al. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation [A]. Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) [C]. Toulouse, France: IEEE Press, 2006. 1: 97 - 100.
- [10] Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso. Compensation of nuisance factors for speaker and language recognition [J]. IEEE Transactions on Audio, Speech and Language Processing, 2007, 15(7): 1969 - 1978.

## 作者简介:



郭 武 男, 1973 年 12 月出生于湖南桃江。1995 年、1999 年和 2007 年在中国科技大学分别获得工学学士、硕士和博士学位。现任中国科技大学讲师, 主要研究方向为说话人识别。  
E-mail: guowu@ustc.edu.cn



李轶杰 男, 1985 年 3 月出生于湖南岳阳。2006 年在中国科技大学获得工学学士学位, 现为硕士研究生, 从事因子分析的研究。

戴礼荣 男, 1962 年 7 月出生于安徽。1997 年在中国科技大学获得博士学位。现任中国科技大学教授、博士生导师, 主要研究方向: 语音信号处理、说话人与语种识别、多媒体通信。

王仁华 男, 1943 年 8 月出生。教授、博士生导师, 主要研究方向: 语音合成、语音识别。