基于说话人分类技术的分级说话人识别研究

刘文举[1], 孙 兵[1,2], 钟秋海[2]
(1. 中科院自动化所模式识别国家重点实验室, 北京 100080; 2 北京理工大学自动控制系, 北京 100081)

摘 要: 识别正确率和抗噪性能固然是说话人识别的研究重点, 但识别响应速度也是决定系统实用化的关键所在. 本文成功地提出了基于说话人分类技术的分级说话人辨识方法, 极大地提高了系统运行速度, 随着注册说话人数的增多, 较之传统的说话人辨识方法, 其优势更加明显. 同时在说话人确认中, 该方法的使用, 进一步提高了确认的正确率, 有效地降低了错误接受和错误拒绝率. 本文提出的可信度打分方法, 也一定程度上改进了系统的性能. 实验表明: 基于说话人分类技术的说话人辨识方法使系统的运行速度平均提高了 3.5 倍, 对说话人确认等误识率和最小误识率平均下降了 53.75%.

关键词: 说话人辨识; 说话人确认; 说话人分类; Cohort 集; 可信度打分
中图分类号: TN912.34 文献标识码: A 文章编号: 0372-2112 (2005) 05-1230-04

Research on Hierarchical Speaker Recognition Based on Speaker Clustering Technology

LIU Wen-ju[1], SUN Bing[1,2], ZHONG Qiu-hai[2]
(1. *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080, *China*;
2. *Department of Automatic Control, Beijing Institute of Technology, Beijing* 100081, *China*)

Abstract: Recognition correct rate and noise robust property are indeed important for speaker recognition research, but the response rate of recognition is also a key factor for a speaker recognition system when applied in the real world. Owing to this, we propose a novel speaker identification approach based on speaker clustering, namely Hierarchical Speaker Identification (HSI). It can increase the running speed greatly for speaker identification systems, and the more the number of registered speakers is, the faster the HSI system runs than the Conventional Speaker Identification (CSI) system. Simultaneously, its counterpart for speaker verification based on speaker clustering, can reduce the rates of false rejection and false acceptance efficiently to improve the capability of verification. A new method is also presented here called reliability scoring. The experiments show that speaker clustering based algorithms can run faster 3.5 times than original approach for the speaker identification and is 53.75% deduction of equal or minimal error rates for the speaker verification on average.

Key words: speaker identification; speaker verification; speaker clustering; cohort set; reliability scoring

1 引言

语音信号是人类日常交流中最重要的一种媒体, 其中包含有异常丰富的信息, 如: 文本内容或语义, 声纹特征, 所持的语种或方言, 情绪, 健康状况等. 语音识别的目的是识别出一段语音中的文本内容, 而说话人识别则是识别出一段语音讲话者的个人身份[1].

说话人识别任务有多种类型, 一般说来, 可分为说话人辨识(Speaker Identification), 说话人确认(Speaker Verification)和说话人探测/跟踪(Speaker Detection/Tracking). 基于声纹的说话人辨识其实是一个传统的模式识别问题, 就是用待测者的声纹去匹配已经训练好的声纹模板, 然后判断使用者是哪一个合法用户. 而说话人确认则是判断该使用者是否为合法用户. 说话人探测是指对一段包含多个说话人的语音, 要正确标注在这段语音中说话人切换的时刻.

在说话人辨识实验中, 我们发现随着注册说话人数的增加, 一次辨识所花费的时间随之直线上升, 很明显, 每一次辨识需要将测试语音去匹配所有说话人的声纹模型, 然后找出最相近的模型对应的说话人作为辨识结果, 这样必然导致模型数(注册人数)越多, 所花时间越长, 注册人数达到一定数量后, 系统就很难进行实时响应, 这样的系统辨识率再高, 也不能满足实用要求. 因此, 本文基于说话人分类技术的分级说话

收稿日期: 2004-09-17; 修回日期: 2004-11-29
基金项目: 国家自然科学基金(No. 60172055, No. 60121302); 北京市自然科学基金(No. 4042025)

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

人辨识方法正是以这些考虑为出发点提出的, 这种方法在基本不降低或略有降低辨识率的同时, 可以大大降低辨识系统时间复杂度, 提高识别响应速度, 在此基础上, 我们还成功地将该方法用于说话人确认中, 在未增加一次确认的平均时间长度的同时, 明显提高了正确识别率.

2　说话人分类算法

说话人分类是语音识别中自适应方法的一种[2]. 这里有两个关键问题: 第一个问题是如何将参与训练的所有说话人分成数量不大的若干类别, 这被称作为"说话人聚类". 类别的数量过大将使参数的存储量过大, 而类别数太小又将起不到分类的作用. 第二个问题就是如何根据未知说话人的一段语音来判断他(或她) 所属的类别.

在语音识别中, 有多种方法用于说话人分类[3], 性别相关分类方法是使用最多的说话人聚类方法, 但这些方法并不能很好地适用于我们所面临的问题中, 对此在 ISODATA 算法[9] 的基础上, 我们进行了有效的改进, 使其可以很好地解决说话人分类中的两类问题, 改进算法的详细过程和步骤可参考我们已发表的文章[4].

2.1　初始分类

在运行分类算法之前, 需先进行初始分类, 设其分为 K 类, 本文采用的方法如下:

(1) 从 N 个注册说话人的模型中, 任选一个模型 λ, 并令类模型 $\lambda_1^M = \lambda$;

(2) 分别计算其余 $N-1$ 个说话人模型到类模型 λ_1^M 的距离, 并按升序的排列方式, 对这些距离值排序;

(3) 取排在第 $[\frac{N}{K}k]$ 位处对应说话人的模型 λ', 并令 $\lambda_k^M = \lambda'$, ($k = 2, 3, \cdots, K$), 其中 $[\cdot]$ 表示取整函数.

2.2　模型间的距离

在进行说话人分类过程中, 我们需要比较说话人之间的相似性, 所以需要定义一种说话人模型间的距离. 也就是说如果两者语音特征越接近, 那么他们模型间所定义的距离就越小. 一般采用文章[5]中的方法, 对两个模型及其相应数据进行轮换计算, 但这一方法缺点在于计算资源的消耗. 方法[6] 中提出的利用模型本身的参数直接计算其距离而不用再导入语音数据进行计算, 在快速的同时可以很好地满足说话人分类的要求.

$$\varepsilon = (\mu_1^i - \mu_2^j)^2 \tag{1}$$

$$d_{ij} = \frac{\sigma_1^i}{\sigma_2^j} + \frac{\sigma_2^j}{\sigma_1^i} + \frac{\varepsilon}{\sigma_2^j} + \frac{\varepsilon}{\sigma_1^i} \tag{2}$$

d_{ij} 表示 Gaussian mixture models(GMM)[7] 中模型 1 中分量 i 与模型 2 中分量 j 的距离. μ_1^i, μ_2^j 和 σ_1^i, σ_2^j 分别表示模型 1 和模型 2 第 i 和 j GMM 混合分量的均值和方差.

$$d(\lambda_1, \lambda_2) = \sum_{i=1}^{H} w_i^1 \min_j d_{ij} + \sum_{j=1}^{L} w_j^2 \min_i d_{ij} \tag{3}$$

这里 $d(\lambda_1, \lambda_2)$ 就表示混合数为 H 的 GMM 模型 λ_1, 与混合数为 L 的 GMM 模型 λ_2 之间的距离, w_i^1, w_j^2 分别表示对 GMM 模型 λ_1 和 λ_2 之间第 i 和 j 个距离分量赋予的权值.

有两点需要说明:

(1) 分类数 K 值选取. 根据本文说话人实验表明, 类中说话人的数目在 5~10 之间可以取得最佳的辨识和确认效果, 而 K 值由判断类中的说话人数目在最佳范围来自适应获得;

(2) 模型间距离 d_{ij} 的应用. 上述定义的模型间距离是说话人分类算法中的主要计算依据, 目标是针对训练好的说话人或说话人类 GMM 模型.

3　基于说话人分类技术的分级说话人辨识

对传统说话人辨识而言, 设有 N 个说话人, 对应的 GMM 模型分别为 $\lambda_1, \lambda_2, \cdots, \lambda_N$. 辨识的最终目的是对一个观测语音序列 X, 找到使之有最大后验概率的模型所对应的说话人 λ_S, 即:

$$S = \arg\max_{1 \leqslant k \leqslant N} P(\lambda_k | X) = \arg\max_{1 \leqslant k \leqslant N} \frac{P(X|\lambda_k)P(\lambda_k)}{P(X)} \tag{4}$$

假定每个说话人出现的概率为等概率, 且因 $P(X)$ 对每一个说话人都是一样的, 上式可简化为:

$$S = \arg\max_{1 \leqslant k \leqslant N} P(X|\lambda_k) \tag{5}$$

对于分级说话人而言, 首先在训练过程中, 我们对注册说话人进行分类, 其中每一类由语音特征比较接近的说话人组成, 再利用这一类说话人的语音数据训练得到同一类模型, 假定最终共得到 K 个类模型.

在辨识过程中, 首先利用测试语音对训练中生成的类模型进行辨识, 也就是说在 K 个类模型中找到与测试语音最接近的类模型, 这个过程我们称之为类辨识.

$$k = \arg\max_{1 \leqslant k \leqslant K} P(X|\lambda_k^M) \tag{6}$$

假定判定语音属于类别 k, 在所属该类别的说话人模型中 $\{\lambda_{k1}, \lambda_{k2}, \cdots, \lambda_{kn_k}\}$, 找到与测试语音最接近的说话人所对应的模型, 这个过程我们称之为类内说话人辨识, 该辨识结果我们认定为系统的辨识结果.

$$S = \arg\max_{1 \leqslant i \leqslant n_k} P(X|\lambda_{ki}) \tag{7}$$

可见, 该辨识过程分为类辨识和类内说话人辨识两个阶段, 因而我们称其为基于说话人分类技术的分级说话人辨识.

与传统说话人辨识相比, 基于说话人分类技术的分级说话人辨识不需要计算测试语音对所有注册说话人模型求后验概率, 即前者需计算 N 个后验概率, 而后者只须计算 $(K + n_k)$ 个后验概率(设注册说话人数为 N, 分为 K 类, 其中第 k 类包含 n_k 个说话人). 因此, 该方法的使用, 大大降低了识别过程中的计算量, 在未明显降低辨识率的同时, 使辨识系统的运行速度更快.

4　基于说话人分类技术的分级说话人确认

说话人确认相当于两类的说话人辨识问题, 正模型(即目标说话者模型) 表征的是目标说话人的声学特征, 而反模型(即冒认者模型) 理论上就是表征目标说话人在整个声学特征空间中的补集. 反模型分为两大类: 一类是背景模型(Background Model), 又称 Cohort 反模型; 另一类是全局背景模型(Universal Background Model 或 World Background Model) 反模型[8].

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

对于说话人确认技术,如何为一个说话人训练其反模型,是解决问题的关键所在.全局背景模型由于其涵盖太广,实际算法中是无法实现的,因此在实验中,我们采用 Cohort 反模型作为相应的 Baseline.

4.1 Cohort 反模型

Cohort 反模型基本思想是,针对目标说话人 C 从已知注册用户集中按某种规律挑选出有限个说话人来组成 Cohort 集 B,用 B 中的用户模型来近似理想中的反模型 $\lambda_{\bar{C}}$,定义测试语音在反模型中的似然打分为:

$$\log P(X|\lambda_{\bar{C}}) = \log\left\{\frac{1}{|B|}\sum_{b=1}^{|B|}P(X|\lambda_b)\right\} \tag{8}$$

判决打分 $$\Delta(X) = \log P(X|\lambda_C) - \log P(X|\lambda_{\bar{C}}) \tag{9}$$

显然,对于系统预先设置好的阈值,打分高于阈值,系统则接受;打分低于阈值,则拒绝.

4.2 预筛选

在 Cohort 集的选取上,我们选用与目标说话人比较相似的用户组成 Cohort 集.选用这样的 Cohort 集的前提是:冒充者力图模仿目标说话人的发音特征,即为合作型的冒充者,这时采用和说话人最相似的用户来组成 Cohort 集[5].对于那些与目标说话人差别很大,甚至性别也不相符的冒认者,如果是仅仅采用这样的 Cohort 集,会导致认证系统对于此类冒充者的确认极为脆弱.

为了有效解决这样的问题,我们在确认过程之前,预先进行类辨识,也就是说首先找到与当前说话人最相近的类,如果当前说话人与目标说话人不属于同一类,那么认定该说话人为冒充者,如果属于同一类,再利用 Cohort 集对相似人进行确认.这样可以很好地避免因为冒充者的语音,与目标说话人语音特征以及该目标说话人 Cohort 集中的说话人的语音特征相差甚远而进行的冗余计算,也可以避免了当冒充者的语音特征与目标说话人以及其背景模型的特征差异较小时,而使得系统确认异常脆弱的问题.

4.3 可信度打分

这里我们提出可信度打分的概念,表示如下:

$$\Lambda(X) = \frac{\log P(X|\lambda_C) - \log P(X|\lambda_{\bar{C}})}{-\log P(X|\lambda_C)} \times 100\% \tag{10}$$

易知:$P(X|\lambda_C)$ 取 log 后,为负值,$P(X|\lambda_C)$ 越小,$-\log P(X|\lambda_C)$ 越大,$\Delta(X)$ 越小.因此,对上述公式度量含义可以得到清晰的解释:如果对于同一个目标说话人,即使测试的两段语音的判决打分值 $\Delta(X)$ 相等,根据可信度度量的判决打分 $\Lambda(X)$,其中对目标说话人模型概率得分值低的测试语音,其可信度度量的判决打分 $\Lambda(X)$ 就低,这样就认定它属于目标说话人的可信度低于概率得分值高的语音.由此可见:利用可信度度量打分的方法在原理上一定会使系统性能进一步得到改善.

5 实验结果比较和分析

5.1 基本实验系统介绍

在实验中,我们采用 GMM 模型作为说话人模型,模型混合数为 64,数据来源于 863 语音库,每一个说话人的模型是由 15 个(每个约 4 秒长度)文件训练得到.语音特征参数为:去除空白语音段,使用 12 阶 MFCC,做能量归一化,不做倒谱均值归一化,采用静态特征和一二阶差分共 39 维.实验平台为 PC 机,CPU 为 Intel 公司的 Pentium4,主频 2G,内存 256M.

5.2 说话人辨识

我们取 863 库中的 40 个女性作为测试对象,每位女性的 18 个文件作为测试语音,其中 10 个用于实验Ⅰ,另外 8 个用于实验Ⅱ,每两个文件(约 8 秒长度)用于一次辨识.我们用 CSI(Conventional Speaker Identification)和 HSI(Hierarchical Speaker Identification)分别表示传统方法说话人辨识方法和本文提出的分级说话人辨识方法;实验中引入的 CSIT、CSIR 和 HSIT、HSIR 分别表示上述试验条件下传统方法说话人辨识所用时间、正确辨识率和分级说话人辨识所用时间、正确辨识率.首先来看一下本文提出的 HSI 分级说话人辨识方法节省辨识计算时间的有效性,见表 1.

表 1 一次辨识所花费时间

一次辨识耗时实验	注册人数	CSIT	HSIT
实验 1(k= 6)	40	4.515(s)	1.367(s)
实验 2(k= 9)	82	8.523(s)	1.912(s)
实验 3(k= 16)	160	16.232(s)	2.835(s)

由表 1 可知,随着注册人数的增加,CSI 一次辨识所花时间大幅增长,而 HSI 的增长却相对较少.当注册人数为 40 时,HSI 花费时间占 CSI 的 30.3%,而当人数增加到 160 时,所费时间只占到 CSI 的 17.5%,总体来说,在我们的实验中,基于说话人分类技术说话人辨识方法比传统方法的运行速度提高了 3.5 倍,大大改善了传统说话人辨识系统的性能.

表 2 中 HSIR(1)表示最可能的 1 个类作为候选类,HSIR(2)表示采用两个最为可能的类作为候选类.

表 2 辨识结果

辨识试验(k 取值)	CSIR	HSIR(1)	HSIR(2)
实验Ⅰ(6)	98.75%	98.75%	98.75%
实验Ⅱ(6)	99%	98.5%	99%

如表 2 所示,之所以 HSI(1)的正确辨识率略低于 CSI,通过分析我们发现存在测试语音原本属于 A 类,但是系统却错误认定为其属于 B 类,也就是说类辨识会存在一定的误差.如果在类辨识过程中我们将两个最可能类作为候选类,这两个类中的所有说话人都作为下一步辨识的候选说话人,我们发现此时 HSI(2)的识别率与传统说话人辨识相同.尽管 HSI(2)运行速度要略微慢于 HSI(1),但它仍明显快于传统的说话人辨识方法.

5.3 说话人确认

我们取 863 库中的 82 名女性作为测试对象,每名女性都选择 25 个句子进行自身登录,另外 3 个句子去尝试登录其余 81 个用户,另外采用 Intel 语音数据库中持北京口音的另外 30 名女性的每人 5 个测试句子.因此,测试共进行了 25* 82+ 3 * 81* 82+ 30* 5* 82= 34276 次,其中自身登录了 25* 82= 2050 次,自身登录与冒认登录比例约为 1:15.7.

在表 3 中,Cohort 1 这一栏表示由五个与目标说话人最相似和五个最不相似的说话人组成 Cohort 集的确认结果,Cohort

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 3　不同条件和方法下的确认结果

	Cohort I	Cohort II	HSV	Cohort I + 可信度	Cohort II + 可信度	HSV + 可信度
等错误率	0.3384%	0.5310%	0.2042%	0.3267%	0.5164%	0.1867%
最小错误率	0.5718%	0.8928%	0.3442%	0.5426%	0.8286%	0.3238%

II 这一栏表示由十个与目标说话人最相似的说话人组成的 Cohort 集的确认结果, HSV 表示本文提出的分级说话人确认(Hierarchical Speaker Verification), 其余表示确认过程中使用了可信度打分, 它们对应栏下面的值表示使用该方法的确认结果. 由于固定的比例关系找到等错误率是比较困难的, 表中列出的等错误率是在两类错误(错误拒绝率和错误接受率) 充分接近基础上的算术平均值. 实验中说话人分类的数目为 9, 由表 3 可见: 本文提出的基于说话人分类的说话人确认 HSV 方法比传统方法等误识率和最小误识率平均下降了 53%、. 而 HSV 结合可信度方法后比传统方法等误识率和最小误识率平均分别下降了 56% 和 53%.

为了尽量避免在类辨识时, 出现错误拒绝自身登录说话人所在的类, 我们采用两个可能性最大的类作为下一步确认的候选类. 结果表明, 基于说话人分类技术的分级说话人确认的正确率明显优于传统的说话人确认的情况. 尽管我们在确认之前增加了类辨识, 对冒认者进行预筛选这一过程, 少部分会增加一次确认所需时间, 但平均而言, 分级说话人确认所花费时间都少于、最多等于传统说话人确认的耗时, 因为大量冒认者在类辨识时, 就被系统拒绝了, 而无需下一步再进行确认. 最后, 本文提出的可信度打分方法的应用, 在具体的实验中是采用浮动阈值的方法, 尽管它可以降低冒认者的打分, 但对系统性能只略有改善. 如果采用固定阈值, 效果将可能更明显.

6　总结与展望

在本文中, 我们探讨了基于说话人分类技术的分级说话人识别方法. 在说话人辨识中, 分类技术的使用, 在没有明显降低辨识率的情况下, 极大地加快了系统辨识的速度, 随着注册人数的增加, 优势则更加明显. 在类模型的建立上, 寻找更为合适的聚类方法以及运用说话人自适应技术不仅能加快模型的训练速度, 而且能够使类模型更为鲁棒地表征该类说话人语音特征, 这将是我们进一步的工作.

在说话人确认方面, 由与目标说话人特征相似的说话人组成 Cohort 集, 可以很好地区分与目标说话人特征相近的冒认者; 对于与目标说话人特征差距较大的冒认者确认时, 这里通过运用类辨识的方式, 预先拒绝这类冒认者; 同时, 通过采用两个可能性最大的类作为辨识结果, 可以有效地减少错误拒绝率. 实验证明: 基于说话人分类技术以及可信度打分方法的说话人确认方法, 能够有效地提高说话人确认的正确率.

参考文献:

[1] Douglas A Reynolds. An overview of automatic speaker recognition technology[A]. Proc ICASSP[C]. Orlando, Florida, USA: IEEE, 2002. 4072- 4075.

[2] Yuqing Gao, et al. Speaker adaptation based on pre-clustering training speakers[A]. Proc Eurospeach[C]. Rhodes, Greece: ESCA, 1997. 2095 - 2098.

[3] Ernest J Pusateri. Rapid speaker adaptation using speaker clustering [A]. Proc ICSLP' 2002[C]. Denver, Colorado, USA: ISCA, Sept. 2002. 61- 64.

[4] Bing Sun, et al. Hierarchical speaker identification using speaker clustering[A]. Proc NLP-KE' 2003[C]. Beijing, China: IEEE, 2003. 299 - 304.

[5] Douglas A Reynolds. Comparison of background normalization methods for text-independent speaker verification[A]. Proc Eurospeach[C]. Rhodes, Greece: ESCA, 1997. 963- 966.

[6] Homayoon S M Beigi, et al. A distance measure between collections of distributions and its application to speaker recognition[A]. Proc ICASSP[C]. Seattle, Washington, USA: 1998. 753- 756.

[7] Douglas A Reynolds, et al. Robust text-independent speaker identification using Gassian mixture speaker models[J]. IEEE Trans on Speech and Audio Processing, 1995, 3(1): 72- 83.

[8] Douglas A Reynolds, et al. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, 10(1): 19- 41.

[9] 边肇祺, 张学工. 模式识别[M], 北京: 清华大学出版社, 2000. 235- 239. Bian Zhaoqi, et al. Pattern Recognition[M], Beijing, China: Tsinghua University Press, 2000. 235- 239.

作者简介:

刘文举　男, 1960 年 4 月生于北京, 获人工智能与智能控制方向博士学位, 现在中科院自动化所模式识别国家重点实验室工作, 主要研究领域为汉语连续语音识别、语音合成、声音转换算法、神经元网络算法、马尔柯夫类模型实用快速算法及人工智能方法进行规划、决策等. E-mail: lwj@nlpr.ia.ac.cn.

孙　兵　男, 1976 年 10 月生于江苏, 模式识别与智能系统专业硕士研究生, 研究方向为说话人识别.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net