ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.3, March 2007, pp.608−616
DOI: 10.1360/jos180608

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

在线无监督说话人检索中稳健的模型自举算法*

付中华+, 张艳宁

(西北工业大学 计算机学院,陕西 西安 710072)

A Robust Bootstrapping Algorithm of Speaker Models for On-Line Unsupervised Speaker Indexing

FU Zhong-Hua+, ZHANG Yan-Ning

(School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China)

+ Corresponding author: Phn: +86-29-88494848, E-mail: mailfzh@nwpu.edu.cn

Fu ZH, Zhang YN. A robust bootstrapping algorithm of speaker models for on-line unsupervised speaker indexing. *Journal of Software*, 2007,18(3):608−616. http://www.jos.org.cn/1000-9825/18/608.htm

Abstract: A robust bootstrapping framework, which employs Multi-EigenSpace modeling technique based on regression class (RC-MES) to build speaker models with sparse data, and a short-segments clustering to prevent the too short segments from influencing bootstrapping, are proposed in this paper. For a real discussion archived with a total duration of 8 hours, the significant robustness of the proposed method is demonstrated, which not only improves the speaker change detection performance but also outperforms the conventional bootstrapping methods, even if the average bootstrapping segment duration is less than 5 seconds.

Key words: speaker indexing; speaker model; regression class; eigenvoice

摘 要: 基于回归树模型的多特征空间建模方法在回归类内部进行特征音分析,较好地解决了训练数据不足时说话人模型的训练问题,而短语音段聚类策略又进一步避免了过短的语音片断对自举训练的影响.验证实验采用了实际录制的近 8 小时的不同谈话数据.结果显示,即使平均自举片断长度小于 5 秒,新方法依然非常稳健,不仅提高了说话人改变检测的效果,而且优于通常的自举方法.

关键词: 说话人检索;说话人模型;回归类;特征音
中图法分类号: TP391 文献标识码: A

说话人检索技术是根据不同说话人的声音差异对音频数据流进行自动分割、标记和分类的技术,可以用于新闻广播节目的检索[1,2]、音频数据库的自动标记、检索和浏览[3−5]、会议记录的自动生成[6]等,它是多媒体数据库检索和管理的重要组成部分.在许多应用中,音频数据流是连续输入的,没有任何关于说话人人数、声纹特点以及发言起止点的先验知识,只能根据已经得到的数据对数据流进行分割,检测说话人发生改变的时刻,进而对各个发言段进行辨识和分类.这就是在线无监督说话人检索.

在线无监督说话人检索的关键在于如何正确地检测说话人改变的时刻(称为说话人改变检测或说话人分

* Supported by the Science & Technology Research and Development Plan of Shanxi Province of China under Grant No.2005k04-G23 (陕西省科学技术研究发展计划)

Received 2006-07-28; Accepted 2006-11-13

割[3]),以及对发言段的辨识和分类(说话人跟踪或说话人聚类[2]).说话人改变检测方法可以分为基于距离的和基于模型的两类方法.基于距离的方法差别在于选择不同的距离度量,例如,K-L 离散度(Kullback-Leibler divergence)[7]、贝叶斯信息准则[8](Bayesian information criterion,简称 BIC)和广义似然比[9](generalized likelihood ratio,简称 GLR)等.这类方法通过计算两个相邻语音分析段之间的某种距离度量进行假设检验,以确定该相邻语音段是否属于同一个说话人.其优点是无须训练且简单、易行,但是,语音段的滑动步长很难选择,小的步长检测精度高,但计算量会迅速增加.基于模型的方法依赖于事先建立的不同说话人的模型、背景噪声模型甚至是语音和音乐模型,例如从通用背景模型(universal background model,简称 UBM)适应出来的说话人相关混合高斯模型(Gaussian mixture model,简称 GMM)[10],或者抽样说话人模型(sample speaker model,简称 SSM)[11].这类方法采用说话人辨识技术,根据似然比得分对每个语音分析段所属说话人进行标注.其局限性在于需要足够的数据来训练模型,只要模型得到了充分训练,说话人改变检测的精度将大为提高.在实际系统中,基于距离和基于模型的方法常结合使用.通常的做法是[11],先用距离判据对输入音频流进行分割,得到某一说话人(记为 A)的语音数据,然后根据这些数据训练该说话人的模型λ_A;接着对下一个分割得到的语音段,计算其在 λ_A 下的似然得分,并判定该段语音是否属于说话人 A,如果属于则用新数据调整λ_A,反之则建立一个新的说话人模型λ_B;重复上述过程直到语音流结束.这一过程就是说话人模型自举(bootstrapping).

要正确地标注音频数据流,说话人模型必须用足够的数据进行充分训练.然而在在线无导师的情况下,这一条件几乎无法满足,因为只有实时输入过程中已经得到的数据才能够用于模型训练.例如在电话交谈或会议中,不同说话人的发音片断长度差异极大,而且大多数片断的持续时间很短.在本文实验所采用的实测数据中,发音片断短的只有 1 秒,长的可达 2 分钟,持续时间小于 10 秒的发音片断占 75%.在这种条件下,训练数据往往不足以充分训练说话人模型,即使勉强训练得到了一个粗略的模型,也会因为无监督训练的不确定性对后续语音段的标定产生误判.为此,文献[11]指出,每个说话人持续发言 1 分钟左右时,所产生的语音数据才能够很好地训练或适应该说话人的模型,进而有效地分辨不同说话人的声音.而文献[2]则丢掉了小于 3 秒的发音段,并采用了一种增量式的 GMM 训练方法以得到有效的说话人模型.为了解决发音段长短差异影响说话人模型的训练,文献[12]应用了基于 BIC 的短段聚类方法,较好地解决了离线(off-line)无监督情况下的说话人模型训练的问题.本文的研究即着眼于在线无监督情况下稳健、有效的说话人模型自举方法.

如何在有限的训练数据下估计说话人模型参数是说话人识别的一个难点,目前的解决方法有特征音方法(eigenvoice)法和极大似然线性回归方法(maximum likelihood linear regression,简称 MLLR).在前期的研究中,我们提出了一种较好的基于回归类的多特征空间建模方法(multi-eigenspace modeling based on regression class,简称 RC-MES)[13],该方法能够把语音信号中包含的说话人差异和音素差异分离,提高了训练数据有限时模型的分辨能力.结合这种方法,本文提出了一种在线无监督说话人检索中稳健的说话人模型自举方法,此外,为了避免过短的语音段对自举过程的影响,新方法还采用了基于 GLR 的短段聚类策略.

本文第 1 节给出 RC-MES 建模方法以及参数估计公式.第 2 节提出稳健的说话人模型自举方法框架,其中,第 2.1 节介绍改进的 GLR 说话人改变检测方法,第 2.2 节给出模型自举和语音短段聚类方法.第 3 节是实验过程和实验结果讨论.第 4 节总结全文.

1 基于 RC-MES 的说话人模型适应

在许多应用中,用于模型训练的数据往往非常有限,因此,利用有限的数据尽可能稳健地训练说话人模型已经是说话人识别的一个难题.为了减少需要估计的模型参数数量,Kuhn 等人提出了特征音方法[14].该方法认为,每个说话人的模型都可以由其在特征空间(eigenspace)中的投影来表示,该特征空间是由事先训练的特征音张成的.因此,说话人模型的训练就变成了投影参数的估计,这在很大程度上减少了待估计的参数个数.

特征空间的目的是表现不同说话人之间的差异,即假定沿特征音的方向,不同说话人的差异最大.但事实上,不同说话人语音的声学差异不仅仅是由说话人自身原因造成的,不同音素的声学差异也是另一个重要的因素.因此,只有当不同的说话人发同一个音时,语音中的声学差异才完全体现了说话人自身的发音差异,而这一

差异正是说话人辨识的依据.基于回归类的多特征空间建模方法(RC-MES)采用了 MLLR 中回归类的思想,将上述特征音方法运用到每个回归类当中.

MLLR 是说话人适应中的关键技术,它解决了说话人适应过程中某些没有对应适应数据的模型分量的适应问题.这种方法通过共享变换矩阵的方式将若干模型分量进行捆绑,组成回归类.变换矩阵由回归类内有对应适应数据的模型分量估计得出,并由整个回归类内所有的分量共享.当适应数据充足时,每一个回归类仅对应一个音素,如果适应数据不足,则根据适应数据的数量将音素集按照某种距离准则分割成若干回归类,并保证每个回归类都有足够的适应数据.所有可能的音素集划分即组成了回归类树(regression class tree).回归类本身实际上意味着在回归类内部的各分量之间的差异较小,而不同回归类之间的差异较大.因此,在回归类内部进行特征音分析将会克服通常单一的特征音法混淆音素差异和说话人差异的不足.

基于 RC-MES 的适应方法包括两个阶段:离线(offline)过程和在线(online)过程.

Offline 过程:

(1) 为每个音素建立一个说话人无关(speaker independent,简称 SI)模型(由 GMM 表示),并将全部高斯分量按照分量间离散度[7]进行聚类以组织成回归类树;

(2) 针对所有的回归类划分,不失一般性,假设划分 A 把所有高斯分量分成了 S 个回归类,将每个回归类内的高斯分量组合成 GMM,变成 S 个关于回归类的 SI 模型;

(3) 为 R 个参考说话人各建立 S 个回归类级的说话人相关(speaker dependent,简称 SD)模型,即按照似然得分将各人特征分配到各回归类内,再用这些适应数据将回归类的 SI 模型适应成与各个参考人对应的回归类级的 SD 模型.对于划分 A 共有 $R*S$ 个回归类级 SD 模型;

(4) 对划分 A 中每个回归类进行特征音分析,得到 S 组基$[e_i(0),e_i(1),\ldots,e_i(k)],i=1,\ldots,S$.

Online 过程:

(1) 获得新的说话人数据,根据适应数据的数量确定一个回归类的划分 A',对应 S'个回归类,划分原则是保证每个类都有足够的适应数据;

(2) 对于 S'个回归类的 SI 模型,按照极大似然原则将新数据划分到各个回归类当中;

(3) 在每个回归类内部,根据所分到的适应数据估计 S'组权重$[w_i(1),w_i(2),\ldots,e_i(k)],i=1,\ldots,S'$;

(4) 在每个回归类内部进行迭代至各类的似然均达到最大,然后合成新的说话人的 S'个回归类级 GMM;

(5) 将所有的 S'个回归类级 GMM 混合,构成新说话人的最终 GMM.

上述说话人的建模过程实际上是各个特征音对应权重,GMM 中高斯分量权重以及 GMM 协方差矩阵的估计问题.通常采用极大似然特征分解方法(maximum likelihood eigen-decomposition,简称 MLED)来估计特征音权重,GMM 的高斯分量权重和协方差矩阵则是直接取自 SI 模型,当特征音权重估计完成后,再根据新数据采用 EM 算法迭代一次以校正高斯分量权重和协方差矩阵.

第 S 个回归类的特征音矢量集合可以写成

$$e^s(j)=\left[e_0^s(j)^T,e_1^s(j)^T,...,e_m^s(j)^T,...,e_M^s(j)^T\right],j=0,1,...,K \tag{1}$$

其中,$e_m^s(j)$ 是该回归类内第 m 个高斯分量的均值矢量,K 是特征维数.特征音矢量的权重可以用下式迭代估计(具体推导过程参见文献[15])

$$\sum_{t=1}^{T}\sum_{m=1}^{M}r^{(m)}(t)\cdot\left[e_m^s(i)\right]'\cdot C_m^{-1}\cdot o_t^{(s)}=\sum_{t=1}^{T}\sum_{m=1}^{M}r^{(m)}(t)\cdot\left\{\left[\sum_{k=0}^{K}w_s(k)\cdot e_m^s(k)\right]'\cdot C_m^{-1}\cdot e_m^s(i)\right\} \tag{2}$$

其中,$o_t^{(s)}$ 是 t 时刻属于回归类 S 的观测矢量,C_m^{-1} 是回归类 S 中第 m 个高斯分量的逆协方差矩阵,$r^{(m)}(t)$是相应的状态停留概率,即

$$r^{(m)}(t)=P(i_t=m\mid o_t^{(s)},\lambda_s)=p_m^s\cdot b_m^s(o_t^s)\Big/\sum_{k=0}^{M}p_k^s\cdot b_k^s(o_t^s) \tag{3}$$

其中,p_m^s 和 $b_m^s()$ 是回归类 S 中第 m 个高斯分量的权重和观测概率,λ_s 是回归类 S 的模型参数.

在式(2)中包含有 k+1 个未知数(回归类 S 中特征音的权重)的 k+1 个方程,可以迭代求解.根据估计出的特征音权重和 Offline 中计算的特征音可以得到新说话人的超均值矢量,由此可以得到各个回归类内各高斯分量的均值矢量.将这些回归类捆绑到一起可以构成新说话人 GMM 模型,最后再根据新数据用 EM 算法对新模型进行一次迭代,以校正高斯分量权重和协方差矩阵.

2 说话人改变检测和模型自举

本文提出的在线无监督说话人检索方法的整体框架如图 1 所示.整个框架的核心部分包括:稳健的说话人改变检测,每个说话人初始模型的自举,根据新的语音段适应相应说话人模型以及语音段的聚类.在实际系统设计中,诸如特征提取以及噪声削减等前端信号处理模块也是影响系统性能的重要组成部分,但是在本文中并不涉及这部分内容.

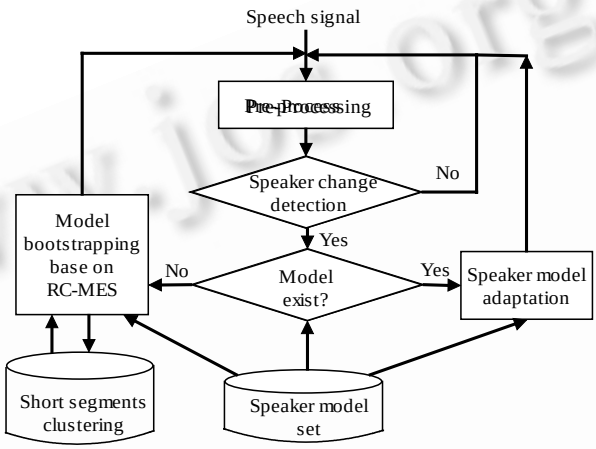

Fig.1 Flow diagram of the proposed unsupervised speaker indexing

图 1 在线无监督说话人检索流程图

为了保障说话人改变检测的精度,同时有效地控制计算量复杂度,本文在改进 GLR 判据的基础上,采用了固定分析段长和可变滑动步长相结合的分段方法.在说话人模型自举时,采用了短语音段聚类策略,以及在此基础上的 RC-MES 的说话人建模方法.

2.1 说话人改变检测

稳健的说话人改变检测是说话人建模和聚类的首要前提.由于在线无监督情况下没有有关说话人的任何先验知识,因此,语音流中说话人改变的时间点检测只能采用基于距离的方法,例如 K-L 离散度、BIC 或 GLR 等.但是随着声学条件和环境噪声的改变,用于假设检验的参数以及阈值等都需要根据经验进行调整,而且为了提高说话人改变点的检测精度,需要很小的分析窗滑动长度,其后果是计算复杂度迅速增高.因此,本文采用了稳健的基于 GLR 的改变点检测方法[16,17]与局部搜索算法[11]相结合的方案.

假设有两个特征矢量集合 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$,分别从当前分析窗内两个相邻语音段经特征提取得到,记为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}$,$\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_N\}$,其中,$N$ 是语音段内的帧数.分别对 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 集合中的数据运用极大似然(maximum likelihood,简称 ML)建立一个高斯模型,记为 λ_X 和 λ_Y.令 $\boldsymbol{Z}$ 为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的并,对 $\boldsymbol{Z}$ 中的数据运用 EM 算法建立一个包含两个分量的 GMM 模型,记为 λ_Z.于是,两个相邻语音段之间的 GLR 距离 d_{GLR} 可以表示为

$$d_{GLR}=(L(\boldsymbol{X}|\lambda_X)+L(\boldsymbol{Y}|\lambda_Y))-L(\boldsymbol{Z}|\lambda_Z) \tag{4}$$

其中,$L()$为对数似然得分函数,即

$$L(\boldsymbol{Z}\mid\lambda_Z)=\sum_{i=1}^{N}\log p(\boldsymbol{x}_i\mid\lambda_Z)+\sum_{i=1}^{N}\log p(\boldsymbol{y}_i\mid\lambda_Z) \tag{5}$$

$$L(\boldsymbol{X}\mid\lambda_{\boldsymbol{X}})=\sum_{i=1}^{N}\log p(\boldsymbol{x}_i\mid\lambda_{\boldsymbol{X}}),L(\boldsymbol{Y}\mid\lambda_{\boldsymbol{Y}})=\sum_{i=1}^{N}\log p(\boldsymbol{y}_i\mid\lambda_{\boldsymbol{Y}}) \tag{6}$$

其中,$p(\boldsymbol{x}_i|\lambda_{\boldsymbol{X}})$表示为特征矢量 $\boldsymbol{x}_i$ 在模型 $\lambda_{\boldsymbol{X}}$ 下的似然得分,$p(\boldsymbol{y}_i|\lambda_{\boldsymbol{Y}})$等类同.上述方法通过为集合 $\boldsymbol{Z}$ 构造含有两个分量的 GMM 模型,使得两个语音段模型 $\lambda_{\boldsymbol{X}}$ 和 $\lambda_{\boldsymbol{Y}}$ 的参数个数之和与整个分析窗模型 $\lambda_{\boldsymbol{Z}}$ 的参数个数相同,从而避免了式(4)中需要额外加入与参数个数差异有关的惩罚因子.可以证明[16],如果 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分别属于两个不同的分布,则有

$$0.0<d_{GLR}\leq N\log 2.0 \tag{7}$$

如果 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 具有相似的概率密度函数,则有

$$d_{GLR}\leq 0 \tag{8}$$

因此,当相邻语音段间的 d_{GLR} 大于 0 时,则认为说话人发生了改变.

事实上,说话人自身的变化以及单个高斯模型的过于简化,都会使基于距离的方法出现漏检和误检.漏检是指错过正确的改变点,即把不同人的语音段当成一个人的语音段,用这些数据去适应某个人的语音模型将会导致模型失效.误检是在没有改变点的地方检测出改变点,即把一个人的语音段当成不同人的语音段.这类错误会产生许多短语音段,但可以通过后继的语音段聚类进行一定程度的校正.为了提高系统的性能,我们选择了比理论值要小的阈值,记为 θ_{GLR}.这样做可以保持较小的漏检率,避免把不同人的语音混在一起训练模型.尽管会增加误检率,但是说话人模型不会受到较大的影响,只是训练的数据有所减少,而且增加的误检率还有可能通过后继的语音段聚类进行修正.

在说话人改变点检测中,分析窗的长度和窗滑动步长是影响检测精度的重要因素,步长小则检测精度高,但是计算量会迅速增大.为了简化计算并保持一定的精度,我们采用了固定分析窗长度和可变移动步长的策略,即局部搜索算法(localized search algorithm,简称 LSA)[11].分析窗首先按照大步长滑动,直到分析窗内两个相邻语音段之间的距离 d_{GLR} 小于阈值,这说明当前分析窗内发生了说话人改变.接着,分析窗再按照小步长滑动,以便较好地分辨说话人改变的时间点.检测出改变点之后,分析窗又恢复为大步长.在我们的实验中,大步长取 2 秒,小步长取 0.2 秒.

2.2 说话人模型自举与短语音段聚类

基于距离的说话人改变点检测把语音数据流分成若干语音段,而且按照距离准则,假定每个语音段是来自不同的说话人.这样,就可以依据这些语音段训练对应说话人的模型,或者对已有模型进行适应.说话人模型自举过程实际上表明,每一个新说话人模型的建立,都完全依赖于该说话人的第 1 个语音片断.然而在现实的应用中,仅仅用这些数据来训练一个说话人模型往往是不够的.例如,在我们的实验数据中,每个发言者的语音片断长度相差非常大,短的仅有 1 秒,长的可达 120 秒,而且交谈的开始和结束往往大都是一些很短的语音片断.在这样的条件下,即使按照模型训练步骤建立了说话人模型,这种粗略的模型也会给后续的语音段划分造成恶劣影响.

在说话人辨识技术中,说话人模型常用 GMM 描述.通常要训练一个较好的 GMM 模型需要 80 秒~100 秒语音数据(与模型参数数量有关)[18].在说话人确认中,常用的说话人模型训练方法通常都采用了说话人适应技术,即根据各个说话人的语音数据从一个基准模型(如 UBM[10]或 SSM[11])训练适应来的.在在线无监督说话人检索中,实现没有先验的数据来训练模型,只有数据流分析中当前已获得数据能够用于训练说话人模型,而这些数据是非常有限的.为此,我们应用了 RC-MES 方法来解决有限数据条件下说话人模型的训练问题.另外,为了保证模型训练的高可靠性,避免过短的语音段用于训练,我们还采用了短语音短聚类的方法,即利用说话人改变检测过程对同属某一说话人的短语音段进行聚类,直到该类数据数量足够进行 RC-MES 训练为止.

设集合 $S=\{s_1,s_2,\ldots,s_K\}$ 对应 K 个说话人的语音数据类,类 i 和类 j 之间的 GLR 距离记为 $d_{GLR}(s_i,s_j)$,满足 $d_{GLR}(s_i,s_j)>\theta_{GLR}$,且每个类内的数据长度均小于模型自举长度,记为 $\theta_{RC\text{-}MES}$.设新检测出的语音段为 s_{new},计算 s_{new} 与集合 S 内每一个语音类之间的 GLR 距离 $d_{GLR}(s_{new},s_i)(1\leq i\leq K)$,如果存在某个类 $j(1\leq j\leq K)$,使得 $d_{GLR}(s_{new},s_j)<\theta_{GLR}$,则将 s_{new} 归并到数据类 j 内,如果不存在,则将 s_{new} 作为集合 S 内的一个新元素.经过这一聚类过程之后,检查集

合 S 内发生变化的数据类的数据长度,如果长度超过 $\theta_{RC\text{-}MES}$,则从集合 S 内取出该元素进行 RC-MES 训练,以建立相应说话人的模型.

说话人模型自举和短语音段聚类算法可以描述如下:

(1) 使用 GLR 距离找到一个新的语音段 s_{new}.

(2) 计算 s_{new} 在已有的说话人模型下的似然得分.

(3) 如果最高得分(假设相对于说话人模型 λ_i)高于辨识阈值 θ_{id},则依据 s_{new} 数据用最大后验方法适应模型 λ_i,转(1);反之,按照上面的短语音聚类方法将 s_{new} 归到集合 S 内.

(4) 检查集合 S 内各个语音类的数据长度.

(5) 如果某一元素设为 s_j,数据长度超过模型自举长度 $\theta_{RC\text{-}MES}$,则依据 s_j 中的数据采用 RC-MES 训练一个新的说话人模型,同时将 s_j 从集合 S 中删除.

(6) 转(1).

其中,初始化时没有一个说话人模型.因此,似然得分为 0 且初始集合 S 为空.

3 实验结果与讨论

本文说话人检索实验中所用到的语音数据均来自不同时间采录的实验室环境下的会议、研讨会、讨论等,总共包含 8 个小时的有效语音数据.各个场景中说话人人数为 2 人~5 人.单一说话人发言长度为 1 秒~120 秒,发言长度少于 10 秒的语音段占总数的 75%.所有的音频采样格式均为 16kHz 采样,16 比特量化,单声道,采样数据首先经过权重为 0.97 的预加重处理.语音端点检测采用自适应短时能量方法.语音特征矢量共 26 维,包括 12 维 Mel 倒谱系数(Mel-frequency cepstrum coefficients,简称 MFCC)、能量以及它们的差分特征 ΔMFCC 特征.语音短时分析采用 30 毫秒的海明窗,窗移为 10 毫秒.

RC-MES 训练的离线操作包括为每个音素建立一个 6 分量的 GMM,使用离散度准则建立回归类树,以及在每个回归类内部构建特征音集合.为了保障特征音训练的独立性以及参考说话人的广泛代表性,我们从 TIMIT 数据库[19]中随机抽取了 100 个说话人的语音数据(TIMIT 标准数据库为英文语料库,说话人检索和辨识主要是分辨不同人的声音,与具体的语言无关),这些数据按照极大似然原则被分配到各个回归类中,并在回归类内部训练 RC-MES 需要的说话人无关模型和特征音模型.

实验共分两部分,首先是说话人改变点检测实验,用来选择合适的 GLR 距离阈值 θ_{GLR},该阈值会影响说话人改变点检测的误检率和漏检率.作为比较,我们还用有监督说话人检索对说话人改变点的检测结果进行校正.这是因为 GLR 检测可能把本来没有发生改变的地方错当成改变点(即误检率),也可能错过真正的改变点(即漏检率).前者会把本来属于同一说话人的语音段分成几个较短的语音段.而随后进行的说话人检索有聚类步骤,能把这些语音段进行合并,从而改善误检率.这里有监督的含义是事先已知说话人人数,手工切分有声/无声段,同时保障各说话人模型经过充分训练.这样做的目的为了让实验集中反映距离阈值 θ_{GLR} 与两种错误率的关系.该阈值的理论值是 0,但是为了避免真改变点被漏掉,我们试图选择一个稍小一点的阈值.因为误检率可以通过后续的说话人检索过程进行一定程度的校正.我们分别用 PRC(precision)和 RCL(recall)来反映两种错误率,PRC 越大则误检率越小,RCL 越大则漏检率越小,分别定义如下:

$$PRC = \frac{\text{检测出的正确改变点数量}}{\text{检测出的全部改变点数量}} \tag{9}$$

$$RCL = \frac{\text{检测出的正确改变点数量}}{\text{全部的正确改变点数量}} \tag{10}$$

为了比较综合效果,我们采用了 F 测度[16]:

$$F = \frac{2.0 * PRC * RCL}{PRC + RCL} \tag{11}$$

说话人改变检测的结果如图 2 所示.不难看出,在同一阈值下,经过说话人检索,RCL 基本保持不变,而 PRC 有一定程度的提高.这说明说话人检索过程不会改变漏检率,但是能够降低误检率.另外,当说话人改变点检测

阈值θ_{GLR}略为降低时,*RCL* 有所增高,而 *PRC* 则有一定程度的下降,但是经过说话人检索,*PRC* 有了一定程度的恢复.例如,当θ_{GLR}从 0 降到−0.1 时,*RCL* 从 0.69 增加到 0.78,而 *PRC* 则从 0.67 降低到 0.55,但是经过说话人检索中的聚类操作,*PRC* 值又上升至 0.74.综合来看,经过检索之后,θ_{GLR}=0 时的 *F* 测度为 0.74,而θ_{GLR}=−0.1 时该平均值为 0.76,前者略低于后者,因此,在后面的说话人检索实验中,我们选择了θ_{GLR}=−0.1.

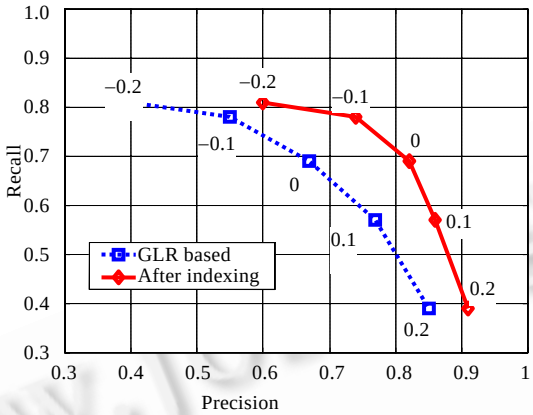

Fig.2 Recall-Precision tradeoff of speaker change detection with different θ_{GLR}

图 2 采用不同θ_{GLR}时说话人改变检测的 Recall-Precision 对照图

第 2 部分实验旨在比较文中所提出的方法与传统说话人模型自举方法的效果.我们比较的传统方法有基于 UBM 的说话人检索方法和一般的特征空间方法.基于 UBM 的说话人检索方法常被用作对比的基本方案[4,5,11],这种方法源自说话人模型训练,在说话人识别中有非常成功的应用[10].特征空间方法则是目前最稳健的说话人适应技术[7,14],能够较好地解决训练时间不足条件下的说话人识别和检索问题[12].在 UBM 方案中,UBM 包含了 512 个经过完全训练的高斯分量,每一个说话人模型由 16 个高斯分量的 GMM 构成,说话人模型都是用 MAP 方法从 UBM 中适应而成.在特征空间方案中,我们运用了 40 个全局特征音矢量.在基于 RC-MES 的方法中,这些特征音是分散在各个回归类内部的,说话人模型自举长度阈值$\theta_{RC\text{-}MES}$为 10 秒.

我们比较了几种方法在不同平均自举语音段长度下的说话人检索精确度,因为不同平均自举语音段长度对说话人模型的训练有直接的影响.为了得到不同自举语音段长度的数据,我们事先对语音流数据进行了标定,经过删减拼接,使得每个说话人的第 1 段发言长度(即平均自举语音段长度)达到要求的时间.实验结果如图 3 所示,其中,横轴表示平均自举语音段的长度,纵轴表示说话人检索的精确度,参见式(11).从图中可以看出,随着平均自举语音段长度的减少,几种方法的说话人检索精确度都有不同程度的下降.其中,UBM 方案下降得最快,从 30 秒~5 秒精确度下降了 1 倍多,说明这种方法对训练数据的数量要求很苛刻,不适合应用于说话人检索应用.特征空间方法明显好于 UBM 方案,但本文方法表现得最为稳健,尤其是当平均自举语音段长度达到 30 秒时,基于 RC-MES 的新方法效果依然好于其他两者.原因一方面在于短语音段聚类能够较有效地控制自举长度,避免过短的语音片段直接用于训练;另一方面,采用特征音来训练说话人模型需要的训练样本少得多,而在回归类内部进行特征音分析能够更好地区分不同说话人的声音.这说明基于 RC-MES 的适应技术能够更好地适应自举语音段长度的变化,即使在自举长度较短时(<5 秒),也依然能够保持较好的检索性能.

4 结 论

说话人检索技术是对语音数据流进行自动分割标定的技术.由于许多应用都无法事先得到任何关于说话人人数、声纹特点以及发言起止点的先验知识,因此只能进行在线无监督说话人检索.在这种检索技术中,如何正确地检测说话人改变时刻,以及根据有限的数据训练说话人模型都是关键的技术难点.通常的说话人检索技术都需要有足够的自举语音数据,或者说对自举语音段的长度非常敏感.然而在实际应用中,每个说话人的发言

长度差异非常大,而且大都持续时间很短,从而导致检索性能不够稳健.本文提出了一种稳健的说话人模型自举算法,该算法采用 RC-MES 技术较好地解决了训练数据不足时说话人模型的训练问题.同时,为了避免过短的语音段影响说话人模型自举过程,我们还采用了短语音段聚类方法.这种方法能够根据 GLR 距离将短的语音段合并,直到其长度足以进行 RC_MES 训练为止.

为了验证新方法的稳健性和有效性,我们进行了说话人改变检测实验和说话人检索实验.实验数据包括大约 8 小时的各种会谈讨论语音数据,其中,个人发言长度小于 10s 的占 75%.在说话人改变检测实验中,我们采用了比理论值略小的距离门限.结果显示,经过说话人检索的校正后,检测效果要好于理论门限.另外,说话人检索实验的结果表明,本文的方法不仅好于其他常用的自举方法,而且在不同的平均自举片断长度下均保持了很好的稳健性.

为了更进一步地验证新方法的效果,我们将在 NIST 数据库基础上进行进一步的实验和评测.

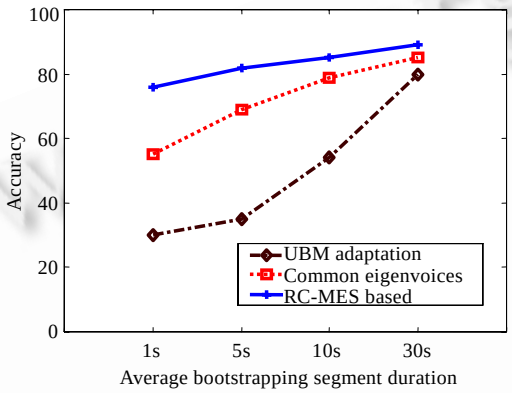

Fig.3 Speaker indexing accuracy for various bootstrapping approaches with different bootstrapping segment duration

图 3 不同自举方法在自举语音段长度不同时的说话人检索精确度

References:

[1] Delacourt P, Wellekens CJ. DISTBIC: A speaker-based segmentation for audio data indexing. Speech Communication, 2000, 32(1-2):111−126.

[2] Lu L, Zhang HJ. Unsupervised speaker segmentation and tracking in real-time audio content analysis. Multimedia Systems, 2005, 10(4):332−343.

[3] Sancho SS, Ascensión GA, José MLM, Carlos BC. Offline speaker segmentation using genetic algorithms and mutual information. IEEE Trans. on Evolutionary Computation, 2006,10(2):175−186.

[4] Meignier S, Moraru D, Fredouille C, Bonastre JF, Besacier L. Step-By-Step and integrated approaches in broadcast news speaker diarization. Computer Speech and Language, 2006,20(1-2):303−330.

[5] Aronowitz H, Burshtein D, Amir A. Speaker indexing in audio archives using Gaussian mixture scoring simulation. In: Bengio S, Bourlard H, eds. Proc. of the 1st Int'l Workshop on Machine Learning for Multimodal Interaction. LNCS 3361, Heidelberg: Springer-Verlag, 2005. 243−252.

[6] Anguera X, Wooters C, Peskin B, Aguilo M. Robust speaker segmentation for meetings: The ICSI-SRI spring diarization system. In: Renals S, Bengio S, eds. Proc. of the 2nd Int'l Workshop on Machine Learning for Multimodal Interation. LNCS 3869, Heidelberg: Springer-Verlag, 2005. 402−414.

[7] Campbell JP. Speaker recognition: A tutorial. Proc. of the IEEE, 1997,85(9):1437−1462.

[8] Chen SS, Gopalakrishnan PS. Clustering via the Bayesian information criterion with applications in speech recognition. In: Acero A, Hon HW, eds. Proc. of the 1998 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing, vol.2. Seattle, Washington: IEEE, 1998. 645−648.

[9] Gish H, Schmidt N. Text-Independent speaker identification. IEEE Signal Processing Magazine, 1994,11(4):18−32.

[10] Reynolds DA, Quatieri TF, Dunn RB. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 2000, 10:19−41.

[11] Kwon S, Narayanan S. Unsupervised speaker indexing using generic models. IEEE Trans. on Speech and Audio Processing, 2005, 13(5):1004−1013.

[12] Nishida M, Kawahara T. Speaker model selection based on the Bayesian information criterion applied to unsupervised speaker indexing. IEEE Trans. on Speech and Audio Processing, 2005,13(4):583−592.

[13] Fu ZH, Zhao RC. Speaker modeling technique based on regression class for speaker identification with sparse training. In: Li SZ, *et al*. eds. Proc. of the Sinobiometrics 2004. LNCS 3338, Heidelberg: Springer-Verlag, 2004. 610−616.

[14] Kuhn R, Junqua JC, Niedzielski NP. Rapid speaker adaptation in eigenvoice space. IEEE Trans. on Speech and Audio Processing, 2000,8(6):695−706.

[15] Fu ZH. Research on robustness of speaker recognition system [Ph.D. Thesis]. Xi'an: Northwestern Polytechnique University, 2004 (in Chinese with English abstract).

[16] Ajmera J, McCowan I, Bourland H. Robust speaker change detection. IEEE Signal Processing Letters, 2004,11(8):649−651.

[17] Lu J, Mao B, Sun ZX, Zhang FY. An improved speaker based speech segmentation algorithm. Journal of Software, 2002,13(2): 274−279 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/13/274.pdf

[18] Reynolds DA, Rose RC. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. on Speech and Audio Processing, 1995,3(1):72−83.

[19] Garofolo J, *et al*. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. National Institute of Standards and Technology, 1993.

附中文参考文献:

[15] 付中华,赵荣椿.说话人识别系统鲁棒性研究[博士学位论文].西安:西北工业大学,2004.

[17] 卢坚,毛兵,孙正兴,张福炎.一种改进的基于说话人的语音分割算法.软件学报,2002,13(2):274−279. http://www.jos.org.cn/1000-9825/13/274.pdf

付中华(1977－),男,湖北十堰人,博士,CCF高级会员,主要研究领域为说话人辨识/确认,语音信号处理,说话人定位及跟踪.

张艳宁(1968－),女,教授,博士生导师,主要研究领域为智能信息处理,数据挖掘,模式识别,计算机视觉.