ISSN1000-9825 Journal of Software 软件学报 2000,11(6):798~803

与文本无关的说话人自适应确认方法*

张怡颖 朱小燕 张钹

(清华大学计算机科学与技术系 北京 100084)

(清华大学智能技术与系统国家重点实验室 北京 100084)

E-mail: zxy-dcs@mail.tsinghua.edu.cn

摘要 该文提出一种与文本无关的自适应说话人确认方法.此自适应方法基于作者所提出的用全局说话人模型标准化似然得分值进行说话人确认的方法,以解决此方法应用于实际系统时存在的训练时间较长的问题,从而缩短新用户注册系统的等待时间,使新用户能够在较短的时间内开始系统的使用.实验结果充分说明了此方法的有效性;当系统有30个用户时,新用户的注册速度加快了12倍.

关键词 说话人确认,似然得分标准化,自适应方法,最大化似然概率,高斯混合模型.

中图法分类号 TP391

说话人确认是根据说话人的语句来确定是否与所声称的参考说话人相符,这种确认只有接受和拒绝两种可能[1,2].GSMSV(speaker verification method based on the global speaker model)方法通过建立全局说话人模型对传统的似然得分值进行标准化[3].该方法具有如下特点:(1) 突出说话人之间的差距;(2) 提高系统区分参考说话人的能力;(3) 能够较容易地检测外来伪装者;(4) 确认速度快;(5) 系统能够适应于不同长度的语音输入.

GSMSV方法在系统识别率和确认速度等方面均显示出极大的优势.然而,随着系统参考说话人数目的增加,新用户的注册等待时间也成比例地增加.当系统的合法使用者较多时,此等待时间将可能无法忍受,因此影响了GSMSV方法的实时应用.本文的自适应GSMSV方法(adaptive GSMSV,简称AGSMSV)即是针对这一问题而提出来的.由于注册等待时间主要消耗在重新训练全局说话人模型上,该方法从这里着手,通过近似估计和在重估公式中引入权重系数快速更新全局说话人模型,从而使新用户能够在尽可能短的时间内使用系统.

1 利用全局说话人模型的确认方法

假设系统共有N个参考说话人,他们的模型分别为$\lambda_1,\ldots,\lambda_i,\ldots,\lambda_N$,其中$\lambda_i$由最大化似然概率$P(Y_i|\lambda)$获得,$Y_i$是第$i$个参考说话人的训练数据.GSMSV方法在这$N$个说话人模型之外,附加一个全局说话人模型$\lambda_{GSM}$,它由最大化似然概率$\prod_{i=1}^{N}P(Y_i|\lambda_{GSM})$获得,即$\lambda_{GSM}$的训练数据包括所有参考说话人的数据.于是在GSMSV方法中共有$N+1$个说话人模型,其中λ_{GSM}代表了多个说话人的共同统计特性.

令$S_i(X)$为对所声称的参考说话人i的输入语音X进行标准化的似然得分,其计算方法如下:

$$S_i(X)=P(X|\lambda_i)-P(X|\lambda_{GSM}). \tag{1}$$

由于λ_{GSM}从所有参考说话人的训练数据得到,它包含了说话人之间在发音、背景环境以及文本内容等方面的共同特性,因而删除这些信息将会避免次要因素的干扰,强调说话人之间的差异.于是,GSMSV方法的决策规则是

* 本文研究得到国家自然科学基金(No.69823001)资助.作者张怡颖,女,1971年生,博士生,主要研究领域为说话人识别,语音识别.朱小燕,女,1957年生,博士,副教授,主要研究领域为人工神经网络,模式识别,语音处理.张钹,1935年生,教授,博士生导师,中国科学院院士,主要研究领域为人工智能,计算机应用.

本文通讯联系人:朱小燕,北京 100084,清华大学计算机科学与技术系

本文1999-02-04收到原稿,1999-06-09收到修改稿

$$S_i(X)\begin{cases}>\eta, & \text{接受对参考说话人 } i \text{ 的身份确认}\\ \leqslant\eta, & \text{拒绝对参考说话人 } i \text{ 的身份确认}\end{cases}, \tag{2}$$

其中 η 是阈值.为了避免计算中的溢出,对似然得分采用对数运算,于是决策规则变为

$$\log P(X|\lambda_i)-\log P(X|\lambda_{GSM})\begin{cases}>\eta', & \text{接受对参考说话人 } i \text{ 的身份确认}\\ \leqslant\eta', & \text{拒绝对参考说话人 } i \text{ 的身份确认}\end{cases}, \tag{3}$$

其中 η' 是阈值.

为了进一步提高与文本无关的说话人确认系统的适应性,减轻受不同输入语音持续时间的影响,我们将似然得分值用持续时间再次标准化,于是决策规则(3)进一步完善为

$$\frac{\log P(X|\lambda_i)-\log P(X|\lambda_{GSM})}{T_X}\begin{cases}>\eta'', & \text{接受对参考说话人 } i \text{ 的身份确认}\\ \leqslant\eta'', & \text{拒绝对参考说话人 } i \text{ 的身份确认}\end{cases}, \tag{4}$$

其中 T_X 为持续时间,η''为阈值.

2 自适应说话人确认方法

2.1 问题的提出

在实时说话人确认系统中,存在对如下功能的广泛要求:当新用户加入系统时,用户常常希望能在尽可能短的时间内使用系统.GSMSV 方法在计算新用户的说话人模型的同时,更新全局说话人模型 λ_{GSM},使其适用于新用户.由于更新 λ_{GSM}需要所有参考说话人的语音数据参加训练,因此新用户的注册等待时间较长,而且随着系统用户数目的增加,等待时间成正比增长;当系统用户数量庞大时,此等待时间将可能变得无法忍受.

我们利用 GSMSV 方法实现了一个说话人确认系统,并依次加入用户,每个用户的注册语音长度为 10s.我们将新用户的注册等待时间随系统现有用户数目的变化记载下来,列于表 1 中.从表 1 我们看到,当系统用户较少时,新用户的等待时间是可以接受的.例如,当系统包含 10 个用户时,用户等待时间约为 4min.若系统具有 100 个用户,则新用户的等待时间大约是 40min.对于实时应用来说,这样很容易使用户不耐烦,影响用户接受系统的感觉.当系统用户进一步增多或者注册语音时间加长时,注册时间将更令人难以接受.本文的自适应 GSMSV 方法即是针对这一问题提出的.由于 GSMSV 对新用户的训练时间主要消耗于计算全局说话人模型 λ_{GSM}各参数,因此,自适应 GSMSV 方法从更新 λ_{GSM}着手以缩短训练时间.

Table 1 The registration time for a new user

表 1 新用户的注册等待时间

Number of current users①	1	2	5	10	20	30	50	100
Waiting time (s)②	46	72	143	263	500	730	1 211	2 350

①系统中现有用户个数,②等待时间(秒)

2.2 自适应 GSMSV 方法

说话人模型采用高斯混合模型 GMM(Gaussian mixture model).GSMSV 方法对于 λ_{GSM}参数的估计类似于采用最大似然概率准则[4]的半连续隐马尔可夫模型的参数估计.假设当前有 N 个用户,他们的训练数据在转换成特征矢量序列后表示为 $\{y_1^{(i)},y_2^{(i)},\ldots,y_k^{(i)},\ldots,y_{T(i)}^{(i)}\}$,$i=1,2,\ldots,N$,其中 i 代表第 i 个用户,$T(i)$为其训练数据的特征矢量的个数.第$(N+1)$个用户的训练数据表示为 $\{y_1^{(N+1)},y_2^{(N+1)},\ldots,y_k^{(N+1)},\ldots,y_{T(N+1)}^{(N+1)}\}$.令 λ_{GSM}的参数为

$$\lambda_{GSM}=((c_1^{GSM},\mu_1^{GSM},\Sigma_1^{GSM}),\ldots,(c_k^{GSM},\mu_k^{GSM},\Sigma_k^{GSM}),\ldots,(c_M^{GSM},\mu_M^{GSM},\Sigma_M^{GSM})).$$

其中,μ_k^{GSM},Σ_k^{GSM} 分别为第 k 个高斯密度函数的均值矢量和协方差矩阵;c_k^{GSM} 是相应的权重;M 是混合分量的个数.采用迭代算法的全局说话人模型参数重估公式如下:

$$\hat{c}_j^{GSM}=\frac{\sum_{n=1}^{N+1}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)}{\sum_{n=1}^{N+1}\sum_{t=1}^{T(n)}\alpha_t^{(n)}\cdot\beta_t^{(n)}},\quad j=1,2,\ldots,M, \tag{5}$$

$$\hat{\mu}_j^{GSM}=\frac{\sum_{n=1}^{N+1}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)y_t^{(n)}}{\sum_{n=1}^{N+1}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)},\quad j=1,2,\ldots,M,\tag{6}$$

$$\hat{\Sigma}_j^{GSM}=\frac{\sum_{n=1}^{N+1}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)\cdot(y_t^{(n)}-\hat{\mu}_j^{GSM})(y_t^{(n)}-\hat{\mu}_j^{GSM})^T}{\sum_{n=1}^{N+1}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)},\quad j=1,2,\ldots,M,\tag{7}$$

$$\theta_j^{(n)}(t)=\begin{cases}c_j^{GSM}p_j[y_1^{(n)}]\beta_1^{(n)} & t=1\\ c_j^{GSM}p_j[y_t^{(n)}]\alpha_{t-1}^{(n)}\beta_t^{(n)} & t=2\sim T(n)\end{cases},\tag{8}$$

$$\alpha_t^{(n)}=\begin{cases}p[y_t^{(n)}]\alpha_{t-1}^{(n)} & t=2\sim T(n)\\ p[y_1^{(n)}] & t=1\end{cases},\tag{9}$$

$$\beta_t^{(n)}=\begin{cases}p[y_{t+1}^{(n)}]\beta_{(t+1)}^{(n)} & t=1\sim(T(n)-1)\\ 1 & t=T(n)\end{cases}.\tag{10}$$

在式(5)～(7)中,c_j^{GSM},$\hat{\mu}_j^{GSM}$和$\hat{\Sigma}_j^{GSM}$是当前迭代的最新值,c_j^{GSM},μ_j^{GSM}和Σ_j^{GSM}是上一次迭代的相应值,此迭代过程的初值设置由分段K平均算法得到[5]. $p_j[y_t^{(n)}]$和$p[y_t^{(n)}]$的计算如下

$$p_j[y_t^{(n)}]=\frac{1}{(\sqrt{2\pi})^{D/2}\cdot(|\Sigma_j^{GSM}|)^{-1/2}}\cdot\exp\left(-\frac{1}{2}(y_t^{(n)}-\mu_j^{GSM})^T\Sigma_j^{GSM-1}(y_t^{(n)}-\mu_j^{GSM})\right),\tag{11}$$

$$p[y_t^{(n)}]=\sum_{j=1}^{M}c_j^{GSM}\cdot p_j[y_t^{(n)}].\tag{12}$$

自适应GSMSV方法在新用户到来时对于λ_{GSM}参数的更新一步完成,无需进行迭代,更新的初值设置为前一个用户对λ_{GSM}的更新值,其对λ_{GSM}的各参数估计如下:

$$\hat{c}_j^{GSM}=\frac{(1-\rho)\cdot\sum_{n=1}^{N}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)+\rho\cdot\sum_{t=1}^{T(N+1)}\theta_j^{(N+1)}(t)}{(1-\rho)\cdot\sum_{n=1}^{N}\sum_{t=1}^{T(n)}\alpha_t^{(n)}\cdot\beta_t^{(n)}+\rho\cdot\sum_{t=1}^{T(N+1)}\alpha_t^{(N+1)}\cdot\beta_t^{(N+1)}},\quad j=1,2,\ldots,M,\tag{13}$$

$$\hat{\mu}_j^{GSM}=\frac{(1-\rho)\cdot\sum_{n=1}^{N}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)y_t^{(n)}+\rho\cdot\sum_{t=1}^{T(N+1)}\theta_j^{(N+1)}(t)y_t^{(N+1)}}{(1-\rho)\cdot\sum_{n=1}^{N}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)+\rho\cdot\sum_{t=1}^{T(N+1)}\theta_j^{(N+1)}(t)},\quad j=1,2,\ldots,M,\tag{14}$$

$$\hat{\Sigma}_j^{GSM}=\frac{(1-\rho)\cdot A+\rho\cdot B}{(1-\rho)\cdot\sum_{n=1}^{N}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)+\rho\cdot\sum_{t=1}^{T(N+1)}\theta_j^{(N+1)}(t)},\quad j=1,2,\ldots,M,\tag{15}$$

$$A=\sum_{n=1}^{N}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)\cdot(y_t^{(n)}-\hat{\mu}_j^{GSM})(y_t^{(n)}-\hat{\mu}_j^{GSM})^T,\tag{16}$$

$$B=\sum_{t=1}^{T(N+1)}\theta_j^{(N+1)}(t)\cdot(y_t^{(N+1)}-\hat{\mu}_j^{GSM})(y_t^{(N+1)}-\hat{\mu}_j^{GSM})^T.\tag{17}$$

在式(13)～(15)中,$\theta_j^{(n)}(t)$,$\alpha_t^{(n)}$和$\beta_t^{(n)}$的计算分别与式(8)～(10)相同. ρ是一个权重系数,用来衡量新用户的注册语音对更新λ_{GSM}参数所起的作用. ρ值越大,新训练数据的贡献越重要. 由于自适应GSMSV对λ_{GSM}的更新从前一个用户的更新值开始,ρ的选择至关重要. 如果ρ值过小,当用户较多时,新数据所起的作用将被原有语音数据淹没,λ_{GSM}参数没有明显变化,那么系统对新用户的识别率将比较低. 如果ρ值过大,则系统将会降低对于老用户的识别率.

与原始GSMSV方法相比,自适应GSMSV方法具有如下特点:(1) 自适应GSMSV方法对λ_{GSM}参数的更新是一个一步完成的过程,即采用式(13)～(17)对于λ_{GSM}参数进行一次更新,而原始GSMSV方法的更新是一个迭代过程;(2) 自适应GSMSV方法对λ_{GSM}参数的更新是在前一个用户值的基础之上进行的,而原始GSMSV方法的迭代过程由十分耗时的K平均算法获得的初值开始. 自适应GSMSV方法的这两个特点使得其注册过程

计算开销大幅度降低,从而能够满足实时要求.

3 实验和讨论

3.1 数据库和实验设置

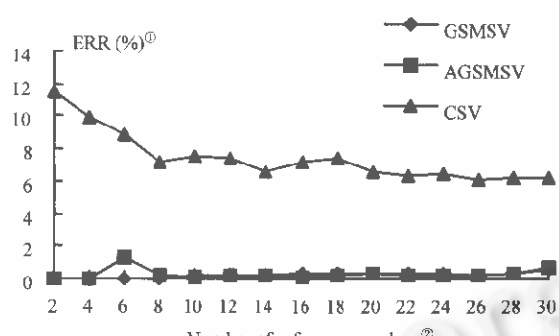
下面的实验所用到的数据来自国家863委员会提供的普通话语音数据库863Bag.其中50个人(25男,25女),每个人50句语音被本文采用.每个人的语音内容各不相同,每句话的持续时间从1.2s～5.6s不等.

选择30个人(15男,15女)作为参考说话人,其他20人作为外来的伪装者.训练数据为每个人15句话,每次测试一句话,每人提供35个测试句子.对参考说话人语音数据的测试构成闭合集测试,某参考说话人语音对其他参考说话人构成伪装者的语音;对外来伪装者语音数据的测试构成开放集测试.每个参考说话人的训练语音的平均持续时间为60s,每个测试语音的平均持续时间为3.5s.

所有语音以16KHz,8位采样.特征使用16维倒谱+16维差分倒谱+差分能量,帧长16ms,帧移8ms.说话人模型中高斯密度函数的个数为64,采用各参考说话人的平均等差错率作为衡量不同方法性能的标准.

3.2 测试结果

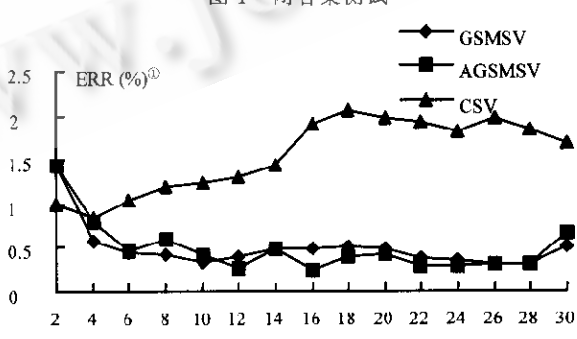
我们用不同数目的参考说话人来测试自适应GSMSV方法的有效性.实验起始于两个参考说话人(1男,1女),每次测试加入两个参考说话人(1男,1女),对λ_{GSM}参数更新两次,即分别用新女声数据和新男声数据更新,然后用这两次更新后的λ_{GSM}参数进行确认测试.GSMSV方法、自适应GSMSV方法和利用传统似然得分的说话人确认方法(speaker verification method with the conventional likelihood score,简称CSV)的测试结果显示于图1和图2之中.



①等差错率,②参考说话人数目.

Fig. 1 Closed set test

图1 闭合集测试



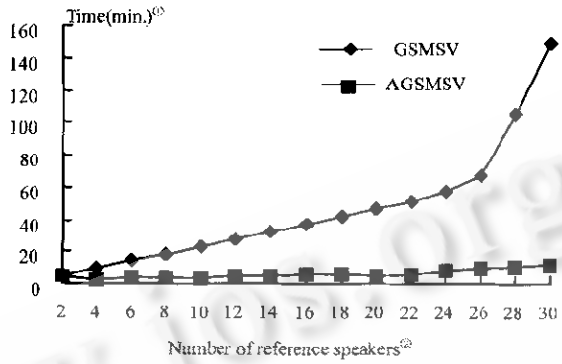
①等差错率,②参考说话人数目.

Fig. 2 Open set test

图2 开放集测试

我们看到,对于闭合集和开放集测试,GSMSV 方法和自适应 GSMSV 方法的等差错率都远远低于传统方法的相应值.自适应 GSMSV 方法的等差错率与 GSMSV 方法相近,只是偶尔略高于 GSMSV 方法.

当新用户两两(1 男,1 女)加入系统时,训练时间随参考说话人数目的变化显示于图 3 中.当系统包含 30 个参考说话人时,GSMSV 的等待时间是自适应 GSMSV 方法的 12 倍.



①时间(分),②参考说话人数目.

Fig. 2　Change of training time with the number of users

图 3　训练时间随人数变化

以上实验充分说明了自适应 GSMSV 方法的有效性和实用性.与原始 GSMSV 方法相比,自适应 GSMSV 方法大大降低了新用户的注册等待时间,但是并没有影响系统的等差错率,甚至在某些情况下,其等差错率低于原始 GSMSV 方法.

4　结　论

本文提出了一种自适应的 GSMSV 方法,以解决原始 GSMSV 方法在实时应用中所面临的新用户注册时间过长的问题.这种自适应方法对于全局说话人参数的调整不再是一个迭代过程,而是通过权重系数的引入成为一步更新过程.此权重系数能有效地体现新用户数据对全局说话人模型的更新所起的作用,因此能充分调整系统,使其在不降低对原用户的性能的前提下提高对新用户的适应性.此外,由于初值的选择方法不同,自适应 GSMSV 方法进一步节省了训练开销.以上技术的引入,为 GSMSV 方法应用于实时系统创造了有利条件.

参考文献

1　Soong A E, Rosenberg L R, Rabiner *et al*. A vector quantization approach to speaker recognition. AT&T Technology Journal, 1987,66(2):14～26

2　Tishby N. On the application of mixture AR hidden Markov models to text independent speaker recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 1991,39(3):563～570

3　Zhang Yi-ying, Zhu Xiao-yan, Zhang Bo. A new speaker verification method with global speaker model and likelihood score normalization. Journal of Computer Science and Technology, 2000,15(2):184～193

4　Liporace L A. Maximum likelihood estimation for multivariate observations of Markov sources. IEEE Transactions on Information Theory, 1982,IT-28(5):729～734

5　Rabiner L R, Juang S E, Sondhi M M. Recognition of isolated digits using hidden Markov models with continuous mixture densities. AT&T Technology Journal, 1986,64(6):1211～1222

An Adaptive Method for Text-Independent Speaker Verification

ZHANG Yi-ying ZHU Xiao-yan ZHANG Bo

(*Department of Computer Science and Technology Tsinghua University Beijing 100084*)
(*State Key Laboratory of Intelligent Technology and Systems Tsinghua University Beijing 100084*)

Abstract In this paper, a novel adaptive text-independent speaker verification method is proposed. This adaptive method is based on the previous work, which uses global speaker model to normalize the likelihood score, and solves one problem of the previous method, i.e., the training time is to long. As a consequence, the waiting time for a new registration is shortened so that a new user can use the system in a short period. The experimental results fully demonstrate the effectiveness of this novel method. When the system has 30 users, the registration time for a new user is accelerated 12 times.

Key words Speaker verification, likelihood score normalization, adaptive method, maximial likelihood probability, Gaussian mixture model.