文章编号:1005-1228(2006)02-0048-04

基于 GMM 的说话人辨认系统及其改进

谢 霞,李 宏,郑 俊
(中南大学信息科学与工程学院,湖南 长沙 410083)

摘 要:建立声学模型是说话人识别技术的重要环节。文章介绍了一种改进的 GMM 算法,将基于样本和核的相似性度量的动态聚类算法与传统高斯混合模型结合起来进行建模,识别辨认时,对语音帧得分进行加权处理。实验表明:改进后的与文本无关的说话人辨认系统无论是在建模时间还是识别效率上都要高于传统的基于 GMM 的说话人辨认系统。

关键词:混合高斯模型;帧似然概率;聚类算法
中图分类号:TN912.34 文献标识码:A

GMM-based speak identification system and its improvement

XIE Xia, LI Hong, ZHENG Jun
(School of Information Science and Engineering,
Central South University, Changsha, Hunan 410083, China)

Abstract: Characteristic modeling is an important link in technology of speaker identification. This article introduces an improved algorithm of GMM which combines classical GMM with clustering algorithm based on similarity measure between stylebook and nuclear for modeling and weighting speech frames' score while recognizing. From the results, we can conclude that the improved uncertain text-speaker-recognition system has higher performance than classic system both on modeling velocity and recognition rate.

Key words: Gaussian Mixture Model; frame likelihood; clustering algorithm

0 引言

说话人识别是语音识别的一种,它是根据人发出的声音,利用计算机来辨认说话者,或对人进行身份验证。该项技术可用于刑侦破案、机要保密、语音加密口令、指挥系统、电子语音锁、玩具和家用电器等。说话人识别包括系统训练和系统测试两个阶段,当训练与测试采用相同语句时,称为与文本有关的说话人识别,反之称为与文本无关的说话人识别。与文本无关的说话人识别已成为当前说话人识别研究的重点。目前,在文本无关说话人识别中常用的说话人识别方法有矢量量化法(Vector quantization:VQ)和高斯混合模型法(Gaussian Mixture Model:GMM)。矢量量化法为每个说话人建立语音码书,并运用矢量距离累加计算进行匹配,由于没有充分利用语音信号的统计特征,因此识别性能一般,鲁棒性不够好。GMM 模型以多个正态分布逼近语音信号的实际统计分布,并运用 Bayes 分类器进行识别,这一模型不仅通过 EM 算法较好地解决了语音信号统计分布特

收稿日期:2006-01-07
作者简介:李 宏(967-),男,湖南长沙人,副教授,主要从事模式识别,数字信号处理的研究;谢 霞(980-),女,江西人,研究生,主要从事语音信号处理方面的研究。

征的估计问题,而且分类器简单可行,因此,近 10 年来,基于高斯混合模型的方法受到了研究者的普遍重视。本文的主要研究工作是对传统的 GMM 说话人识别系统进行改进。传统的 GMM 算法的初始化是基于 K 均值的,识别时通过计算出各模型的帧似然概率进行判别。本文利用基于样本和核的相似性度量的动态聚类算法对初始化进行了改进[1],用它所得到的初始化数据比较接近实际模型的均值,减少了 GMM 模型初始化的盲目性,提高了 GMM 算法的收敛速度。识别时,针对个别破坏帧的影响,对帧似然概率进行加权变换,以提高识别效率[2]。

1　传统基于 GMM 算法的说话人辨认系统

GMM 可以表示为若干个高斯概率密度的线性组合。M 阶 GMM 的概率密度函数如下:

$$P(o|i,\lambda)=\sum_{i=1}^{M}P(o|i,\lambda)=\sum_{i=1}^{M}c_iP(o|i,\lambda),\quad \sum_{i=1}^{M}c_i=1 \tag{1}$$

(1)式中,λ 为 GMM 模型的参数集;O 为 K 维的声学特征矢量。i 为隐状态号,也就是高斯分量的序号。M 阶 GMM 就有 M 个隐状态;c_i 为第 i 个分量的混合权值,其值对应为隐状态 i 的先验概率。(1)式中 $P(o|i,\lambda)$ 为高斯混合分量,对应隐状态 i 的观察概率密度函数,一般采用 K 维单高斯分布函数,如(2)式所示:

$$P(o|i,\lambda)=N(o,\mu_i,\sum_i)=\frac{1}{(2\pi)^{k/2}|\sum_i|^{1/2}}\times\exp\left[-\frac{(o-\mu_i)^T\sum_i^{-1}(o-\mu_i)}{2}\right] \tag{2}$$

(2)式中,μ_i 为均值矢量;$\sum_i$ 为协方差矩阵,i=1,2,…,M。因此(1)式可以理解为,M 阶 GMM 是用 M 个单高斯分布的线性组合来描述,即 GMM 参数集 λ 可由各均值矢量、协方差矩阵及混合分量的权值组成,表示成如下三元组的形式:

$$\lambda=\{c_i,\mu_i,\sum_i;i=1,\cdots,M)\} \tag{3}$$

(3)式中,协方差矩阵 $\sum_i$ 可以取普通矩阵,也可以取对角阵。

为说话人建立 GMM 模型,实际上就是通过训练估计 GMM 模型的参数,常用的方法是最大似然估计方法。最大似然估计的目的是在给定训练矢量集的情况下,寻找合适的模型参数,使 GMM 模型的似然函数值最大。假设可用的训练矢量集为 $O=\{o_1,o_2,\cdots,o_3\}$,则高斯混合模型的似然函数可表示为

$$P(O|\lambda)=\prod_{t=1}^{T}P(o_t|,\lambda) \tag{4}$$

由于似然函数 $P(O|\lambda)$ 和参数集 λ 是很复杂的非线性函数关系,不易用通常办法找到其极大值点,可以采用 EM 算法来估计高斯混合模型的参数 λ[3]。给定一个说话人的训练语音,EM 算法的计算过程是从一个初始模型开始,每次迭代地估计出一个新的模型参数 $\bar{\lambda}$,使 $P(O|\lambda)\le P(O|\bar{\lambda})$,然后再以 $\bar{\lambda}$ 作为模型的参数开始下一次的迭代,这样反复迭代,直到满足收敛条件。训练完 GMM 模型参数后,就能够进行说话人识别了。

对于 N 个人的说话人识别系统,其中每一个说话人用一个 GMM 模型来代表,记为 $\lambda_1,\lambda_2,\cdots\lambda_N$。在识别阶段,假设测试语音的特征矢量序列为 $O=\{o_1,o_2,\cdots,o_T\}$,则该人为第 n 个人的后验概率为:

$$P(\lambda_n|O)=\frac{P(O|\lambda_n)P(\lambda_n)}{P(O)}=\frac{P(O|\lambda_n)P(\lambda_n)}{\sum_{m=1}^{N}P(O|\lambda_m)P(\lambda_m)} \tag{5}$$

(5)式中,$P(\lambda_n)$ 为第 n 个人说话的先验概率;$P(O)$ 为所有说话人条件下特征矢量集 O 的概率;$P(O|\lambda_n)$ 为第 n 个人产生特征矢量集 O 的条件概率。识别结果由最大后验概率准则给出[4],即:

$$n^*=\arg\max_{1\le n\le N}P(\lambda_n|O) \tag{6}$$

式中,n^* 表示识别判决结果。一般情况下,每个人说话的先验概率设为相等,即 $P(\lambda_n)=\frac{1}{N}$,n=1,2,…,N。此外,对于每个说话人(5)式中的 $P(O)$ 都相等。这样,识别结果也可以写成:

$$n^*=\arg\max_{1\le n\le N}P(O|\lambda_n) \tag{7}$$

这时,最大后验概率准则就转化成了最大似然准则。假设各帧特征独立时有:

$$n^*=\arg\max_{1\le n\le N}P(O|\lambda_n)=\arg\max_{1\le n\le N}\prod_{t=1}^{T}P(o_t|\lambda_n) \tag{8}$$

2 改进的初始化方法

传统 K 均值算法的一个缺点是只采用均值作为一个类的代表。这只有当类的自然分布为球状或接近于球状时，即每类中各分量的方差接近于相等时，才可能有较好的效果。而对于语音这样的数据，各分量方差不等而是呈椭圆状的正态分布，K 均值算法作为初始化算法效果不是很好。为了解决这个问题，在初始化过程中，采用基于样本和核的相似性度量的动态聚类算法。定义一个核 $k_j(x,v_j)$ 来表示一个类 c_j，其中 v_j 是定义 k_j 的一个参数集。核 k_j 可以是一个函数，一个点集或其它适当的分类模型。规定一个样本 y 与核 k_j 之间某种度量 $\triangle(x,k_j)$ 来表示样本 y 与类 c_j 之间的相似程度。本系统以样本的统计估计值为参数的正态函数作为核函数[4]，即：

$$k_j(x,v_j)=\frac{1}{(2\pi)^{k/2}|\sum_j|^{1/2}}\times \exp\left[-\frac{(x-m_i)^T\sum_j^{-1}(x-m_i)}{2}\right] \quad (9)$$

这里的参数集 v_j 为：

$$v_j=\{m_j,\sum_j\} \quad (10)$$

其中 m_j 为样本均值，$\sum_j$ 为样本协方差矩阵。相似性度量为：

$$\triangle(x,k_j)=\frac{1}{2}(x-m_j)^T\sum_j^{-1}(x-m_j)+\frac{1}{2}\log|\sum_j| \quad (11)$$

采用正态形式的核函数有利于拟合混合密度中的各高斯分布。类似于 C 均值算法，定义准则函数为：

$$J_k=\sum_{j=1}^{C}\sum_{y\in c_j}\triangle(x,k_j) \quad (12)$$

算法应使 J_k 最小。

初始化步骤如下：

(1)选择初始划分，将样本集任意分成 m 类，并确定每类的初始核 k_j，j=1,2,…,m。

(2)按照下列规则

$$\triangle(x,k_j)=\min_k\triangle(x,k_j),k=1,2,\cdots,m \quad (13)$$

则 $y\in c_j$

(3)重新修正核 k_j，若核 k_j 保持不变，则算法终止；否则转步骤(2)。

初始化后将 m_j 作为 GMM 模型中第 j 个高斯分量的初始均值，$\sum_j$ 为第 j 个高斯分量的初始协方差矩阵。各类中语音帧数目与总的语音帧数目的比值作为各类的初始概率。

3 帧似然变换加权

在基于 GMM 的说话人识别中，一般说来与测试语音同类的目标模型得分高的帧要多于其它非目标模型[4]。但通过观察发现由于说话人的各项特征长时间变动或者受噪声等干扰的影响，某些测试帧对于非目标模型的得分反而大于目标模型的得分，称为破坏帧，且个别坏帧对于非目标模型的得分还可能极高，而对于目标模型的得分极低。无形之中，非目标模型的得分被拉近甚至可能超过目标模型，引起误判，从而设想将每帧对各模型的得分进行加权。

具体加权算法如下：假设测试语音 $O=\{o_1,o_2,\cdots,o_T\}$ 各帧对某个语音模型 λ 的得分为：

$$P_t=(o_t|\lambda_n),t=1,2,\cdots,T \quad (14)$$

将各帧的得分按照由大到小的顺序排列为 $p'_1,p'_2,\cdots,p'_T$。排序后再进行升半正弦加权，则该测试语音对模型 λ_n 的总分为：

$$n^*=\prod_{t=1}^{T}\sin(\pi^*\frac{t}{T})p'_t \quad (15)$$

这样处理后对打分极低和极高帧的影响起了一定的抑制作用，从而提高了识别效率。

4 实验结果及分析

实验语音数据取自普通实验室环境，录取 30 个人的语音数据，其中男性、女性各 15 人，每人任选文本朗读两段两分钟的文字，一段作训练用，一段作测试用。录音采样频率为 8KHz，量化精度为 16bit。语音由话筒接收，经声霸卡转为数字化信号，再对其作 $H(z)=1-0.96z^{-1}$ 高频提升。语音经过预处理之后，取帧宽为 256 点(29ms)，帧移为 128 点(9.5ms)，提取 12 阶的 LPCC 作为特征参数矢量。GMM 模型的混合数取 16。模板训练时，当两次迭代的值之差小于 5×10^{-5} 时，则认为满足迭代条件；当迭代 40 次仍不满足迭代条件时，则退出迭代。

实验一：建模时长不同，在满足迭代终止条件时，比较采用 K 均值初始化和采用新的初始化方

法的迭代次数和模型训练时间见表 1。

从表 1 中可以看出，当建模训练数据小于 5s,两种方法的迭代次数都超过了 40 次;当训练语音数据增多时,虽然总的训练时间增多,但是迭代次数却减少了。特别是采用新的初始化方法后，迭代次数会有大幅度减少，训练时间也会比改进前少。

表 1　迭你次数和建模时长比较

建模时长(s)	采用均值初始化		采用新的初始化方法	
	迭代次数	建模时间(s)	迭代次数	建模时间(s)
5	40	1.5	40	1.5
15	30	4	25	3
20	28	7	18	5
30	25	8	15	6

实验二:当采用不同的时长建模时,采用新的初始化方法以及对帧似然概率加权后与传统 GMM 方法在误识率上的比较见表 2。测试时,从测试语音段中为每个人选取 10 段 8s 语音，每段语音内容不同,30 个人共有 300 段测试语音,假设有 M 段语音没有判断为对应的说话人,则误识率为 M/300。

表 2　建模时长与误识率的比较

建模时长(s)	误识率		
	采用 K 均值初始化(%)	采用新的初始化方法(%)	对帧似然概率加权(%)
5	11.0	10.6	9.6
15	9.3	9.0	6.6
25	7.6	6.3	4.6
30	5.0	4.6	2.3

从表 2 中可以看出，当训练模板采用的语音较少时(仅 5s),识别系统的误识率较高。采用 K 均值初始化的方法误识率最高，采用新的初始化方法后的误识率减小,但不大;采用对帧似然概率加权的方法误识率最低，尤其在建模语音数据大于 25s 时。

在实验中发现:GMM 模型混合数对识别的影响不大，当混合数取 8 时已经能达到较好的识别效果。取 16 时效果最好,混合数取得过大对识别率提高不大,却使计算量增大很多。另高斯混合模型中当协方差矩阵 $\sum_i$ 取对角阵时也可以使计算量减少,提高建模速度也不会使识别率降低。但要注意说话人的语音一定要自然，不可故意非正常的发出,这样会使识别率大大降低。此外说话人的健康状态也影响识别效果，当说话人感冒或是喉炎时,声音发生变化,因此会使识别系统的性能下降。

5　小结

本文在传统的基于 GMM 的与文本无关的说话人识别系统上加以改进，将基于样本和核函数的动态聚类方法用于初始聚类，识别时对帧似然概率进行加权,实验表明:本说话人识别系统在减少建模时间和提高识别效率上都取得了一定的效果。

参考文献：

[1] Zhang Lei, Han Jiqing, Wang Chengfa. A novel weighted likelihood measure for speech recognition under G-Force [A].7th joint conference on information science. USA:North Carolina,2003,692-696.

[2] T.F.Quatieri, D.A. Reynolds,G.C. O'Leary. Handset Nonlinearity Estimation with Application to Speaker Recognition [J]. IEEE Trans. Speech and Audio Processing, 2000, 8(5):567-584.

[3] 张磊,韩纪庆,郑铁然.语音信号处理[M].北京:清华大学出版社,2004.

[4] K. Markov. S. Nakgawa. Text-independent Speaker Recognition System Using Frame Level Likelihood Processing [R]. Technical Report of IEICE, 1996, SP96-17. 37-44.

[5] 边祺,张学工.模式识别[M].北京:清华大学出版社,2000.

[6] 戴红霞,赵力.采用帧概率变换的与文本无关说话人识别系统的实现[J].电声技术,2004(9):40-41.