doi:10.3969/j.issn.1002-0802.2015.01.020

基于GMM模型的声纹识别模式匹配研究*

于 娴,贺 松,彭亚雄,周 晚
(贵州大学 大数据与信息工程学院,贵州 贵阳 550025)

摘 要:模式匹配是声纹识别的关键问题之一,为了提高识别正确率和识别效率,本文采用GMM模型建模,训练阶段利用EM算法求取参数集,并通过MAP准则实现模式识别。引入LBG算法求取起始参数值,并设计了基于3种方法的联合判决门限决策。实验结果表明GMM模型利用平均值向量和协方差矩阵使它具有更好的模型能力,当高斯混合数为32时识别率达到最高,联合判决门限决策有效降低了误识率与虚警率,并提高了识别效率。

关键词:声纹识别 模式匹配 LBG 高斯混合模型
中图分类号:TP391.4 **文献标志码**:A **文章编号**:1002-0802(2015)01-0097-05

Pattern Matching of Voiceprint Recognition based on GMM

YU Xian, HE Song, PENG Ya-xiong, ZHOU Wan
(College of Big Data & Information Engineering, Guizhou University, Guiyang Guizhou 550025, China)

Abstract: Pattern matching is one of the key problems of voiceprint recognition. In order to improve the accuracy and efficiency of recognition, this paper adopts GMM modeling, applies the EM algorithm to calculate parameter set during the training stage, and achieves pattern recognition via MAP criterion. LBG algorithm is introduced to calculate the initial parameter values, and a combined threshold decision is designed based on 3 methods. Experiment results show that GMM, with mean vector and covariance matrix, enjoys better modeling capability, and reaches the highest recognition rate when the mixed number is 32. The combined threshold decision effectively reduces the false acceptation rate and false alarm rate, and meanwhile, it improves the efficiency of recognition.

Key words: voiceprint recognition; pattern matching; LBG; Gaussian mixture model

0 引 言

随着信息时代的来临,计算机、通信技术等高科技技术在我们的日常生活中随处可见,让我们的生活变得更加便捷与多彩,但随之而来的问题也造成了很多人的困扰。各种卡片必须随身携带,复杂绕口的密码太难记忆,卡片丢失、密码被盗也频繁带来安全隐患和财产损失。而生物识别是生物学和信息学等技术的结合,使得身份鉴定变得更加安全、方便且不需要记忆,帮我们解决了这一难题,它主要是通过运用生理和行为这种与生俱来的特征来实现身份的识别。

声纹识别也属于生物识别,它具有获取方便、使

* **收稿日期**:2014-09-19;**修回日期**:2014-12-20 Received date:2014-09-19;Revised date:2014-12-20
基金项目:用于社区司法矫正的声纹识别系统研究项目(黔科合SY字[2013]3105号);贵州省中药现代化科技产业研究开发专项(黔科合中药字[2013]5066号);贵州省工程技术研究中心建设项目(黔科合G字[2014]4002号)
Foundation Item: The Research Program of Voiceprint Recognition System for Community Judicial Correction (Guizhou Branch of SY[2013]3105); The Special Research and Development of Guizhou TCM Modernization Scientific and Technological Industry (Guizhou Branch of TCM[2013]5066); The Construction Program of Guizhou Engineering Technology Research Center (Guizhou Branch of G[2014]4002)

用简单、识别成本低、可远程操作等优势，被广泛地应用于金融、证券、公安、军队、社保、医疗及其他民用安全认证等领域。当前中国对声纹识别的运用尚处起步阶段，有很广阔的发展前景。声纹识别的主要过程有预处理、特征提取、模式匹配、识别判断，本文主要对声纹识别的模式匹配算法进行研究。

声纹识别模式匹配方法有很多，如动态时间归整(DTW)、人工神经网络(ANN)、隐马尔可夫模型(HMM)、高斯混合模型(GMM)等，由于 DTW 精度难以对正导致识别率低，ANN 训练时间较长，HMM 训练计算量较大，本文选取当前文本无关声纹识别的主流技术高斯混合模型(Gaussian Mixture Model，GMM)作为建模方法。通过 GMM 的离散组合，用均值和协方差矩阵来表示高斯函数，从而得到 GMM[1-2]。由于高斯混合模型 GMM 对语音声学特征分布有较好的拟合特性，基于最大似然决策的 GMM 方法已经成为说话人识别系统的主流方法[3]。它是高斯概率密度函数的延展，因此能够很好地模拟各种形状的密度公布。

1 GMM 模型算法

1.1 参数训练

GMM 中的参数是利用训练样本 $\{x_1, x_2, \cdots, x_m\}$ 通过计算 $p(x,z)$ 的最大似然估计的方法得到，m 为高斯混合密度的混合数，z 为隐含随机变量。这种最大似然估计可以利用期望值最大化算法(Expectation MaXimization Algorithm，EM)，通过迭代得到[4]。其具体步骤如下：

$p(x,z)$ 的最大似然估计

$$l(\mu,\Sigma)=\sum_{i=1}^{m}\log p(x;\mu,\Sigma)=\sum_{i=1}^{m}\log\sum_{z}p(x,z;\mu,\Sigma) \tag{1}$$

根据 EM 公式，上式等于

$$\sum_{i=1}^{m}\log\sum_{z_i}Q_i(z_i)\log\frac{p(x_i,z_i;\mu,\Sigma)}{Q_i(z_i)}=\sum_{i=1}^{m}\sum_{j=1}^{k}w_i^{(j)}\log\cdot\frac{\frac{1}{(2\pi)^{n/2}|\Sigma^{(j)}|^{1/2}}\exp\left(-\frac{1}{2}(x_i-\mu^{(j)})^{\mathrm{T}}\sum{}_{-1}^{j}(x_i-\mu^{(j)})\right)}{w_i^{(j)}} \tag{2}$$

式中，$i=1,2,\cdots,m$，$Q_i(z_i)$ 为混合权值，且 $\sum_z Q_i(z)=1$，$Q_i(z)\geqslant 0$，$\mu^{(j)}$ 为均值矢量，$\Sigma^{(j)}$ 为协方差矩阵，这里取为对角矩阵，$w^{(j)}$ 为混合权重，且

$$w_i^{(j)}=Q_i(z_i=j)=p(z_i=j\mid x_i;\mu,\Sigma) \tag{3}$$

宽窄、走向和函数形状的中心等这些密度函数的特性都由这些参数确定。这里用 λ 来表示它们的集合，$\lambda=\{w^{(i)},\mu^{(i)},\Sigma^{(i)}\}$，它可以用来表示一个完整的高斯混合密度函数。每一个训练模型都用一个唯一的 λ 来表示。

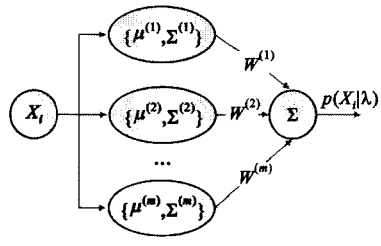

图 1 GMM 模型

Fig. 1 GMMdiagram

式(2)固定 $\Sigma^{(j)}$，对 $\mu^{(j)}$ 求导后等于 0，可得

$$\mu^{(j)}=\frac{\sum_{i=1}^{m}w_i^{(j)}x_i}{\sum_{i=1}^{m}w_i^{(j)}} \tag{4}$$

同理可得

$$\Sigma^{(j)}=\frac{\sum_{i=1}^{m}w_i^{(j)}(x_i-\mu^{(j)})(x_i-\mu^{(j)})^{T}}{\sum_{i=1}^{m}w_i^{(j)}} \tag{5}$$

EM 算法的基本思想是把初始模型与新模型 λ^*，按照式(6)重复迭代，

$$p(x\mid\lambda^*)\geqslant p(x\mid\lambda) \tag{6}$$

直到它们满足

$$\delta=\{p(x\mid\lambda^*)-p(x\mid\lambda)\}\leqslant\eta \tag{7}$$

时为止，这时的 λ^* 为最优值。

这就是 EM 迭代算法估计 GMM 参数的过程，通常情况下要得到一个稳定的 GMM 需要经过五到十次的迭代。

1.2 GMM 模型识别

GMM 的具体步骤是：在训练阶段，根据最大似然估计准则从语音特征矢量中找出一个使得 m 个 x_i 的平均概率最大的参数集 λ；而在识别阶段，则是根据最大后验概率准则(Maximum A Posterior，MAP)[5]，找出使识别语音概率最大的 λ_i 作为识别结果，则由贝叶斯理论，最大后验概率可以写成

$$\arg\max_{1\leqslant i\leqslant m}p(\lambda_i\mid x)=\arg\max_{1\leqslant i\leqslant m}\frac{p(x\mid\lambda_i)p(\lambda_i)}{p(x)} \tag{8}$$

又因为 $p(x)$ 没有先验知识且为无条件概率，上式可化简为求模型对数据的先验概率，即

$$S=\arg\max_{\lambda}(p(x\mid\lambda)) \tag{9}$$

GMM之所以在声纹识别中运用普遍是因为它是m个高斯函数的加权平均,能够用一定量的高斯函数拟合任意语音的特征分布。

2　求取初始点

通过验证,选取不同的起始参数,会大大影响EM算法的识别率和迭代速度,因此,为了提高识别率,选取一个好的初始点是必不可少的。常用的求取初始点的算法有LBG算法、K-均值算法等,由于LBG算法压缩比大且失真较小,而K-均值算法对数据集中的孤立点较为敏感,少量的孤立点数据就会严重影响到聚类结果,因此本文选取目前码本训练性能比较好的LBG(LINE-BUZO-GRAY)算法[6]做为GMM训练中寻求初始点的方法。LBG算法步骤:

1)用训练向量的均值做为向量集的质心,并将向量集按照式(10)所示的方法分裂成双倍的数量。

$$\begin{aligned} o_K^+ &= o_K(1+\varepsilon) \\ o_K^- &= o_K(1-\varepsilon) \end{aligned} \tag{10}$$

其中,o为向量集的质心,$\varepsilon=0.05$,表示分裂参数。

2)测量每个训练向量的欧氏距离,找出与其距离最短的质心o_l,将向量分别与它们的o_l分配到一个集合中去。

3)再用每个新集合的均值作为其新的质心。

4)不断执行第2、3步直到前后两次的训练向量与其o_l的距离和的总体之差$\| J_n - J_{n-1} \|$小于临界值$\Omega=0.01$为止。

5)不断执行步骤1直到向量集达到我们所需的数量,它们的均值就是我们所需要的量化结果。

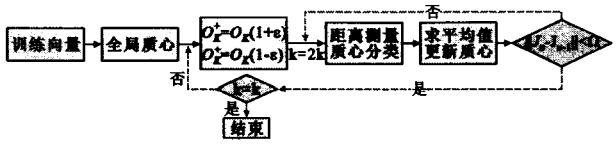

图2　LBG算法流程

Fig. 2 LBG algorithm flow chart

3　判决门限

目前大部分声纹识别研究都局限在对某方法的有效性及对算法的局部改进上,这些方法都因为只侧重于对某方面的研究而产生了片面性。实际上,语音具有各种各样的特征,为了提高语音的识别率,本文使用联合判决门限对语音做出识别判断。既是先利用短时平均能量、短时平均过零率作为初步检测,再用GMM作精确检测的序贯识别[7]。声纹识别可用两个重要的参量来表示其识别性能,误识率和虚警率。误识率是指把待测语音中的伪冒者错误判定为与参考模板中某样本相匹配的情况所占的比例;虚警率是指拒识待测语音中正确语音段的情况所占的比例[8]。而不同的判决门限可以调节误识率和虚警率以达到相应的识别要求。

1)短时平均能量定义为

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)\, w(n-m)]^2 \tag{11}$$

其中,$\{x(m)\}$为输入信号序列,$w(n-m)$为汉明窗

$$W(n,\alpha) = (1-\alpha) - \alpha\cos\frac{2\pi n}{N-1} \quad 0 \leqslant n \leqslant N-1 \tag{12}$$

不同的α值会产生不同的汉明窗,本文取$\alpha=0.46$。

计所有样本的短时平均能量的均值为e_μ,最大和最小极值分别计为$e_{\max}$、$e_{\min}$,则$|e_{\max}-e_\mu|$和$|e_{\min}-e_\mu|$这两个距离中较大的一个为低门限值e_l,较小的一个为高门限值e_h。

2)短时平均过零率为

$$Z_n = \sum_{m=-\infty} |\mathrm{sgn}[x(n)] - \mathrm{sgn}[x(n-1)]|\, w(n-m) \tag{13}$$

取

$$w(n) = \begin{cases} \dfrac{1}{2N}, & 0 \leqslant n \leqslant N-1 \\ 0, & \text{其他} \end{cases} \tag{14}$$

用与1)相同的方法得到均值z_μ,及高、低门限值z_h、z_l。

3)计算训练样本与得到的GMM模型的相似程度,分别取最小和最大的作为高、低门限值s_h、s_l。

把上面3种方法分为两级,第1级由短时能量与过零率共用判决,第2级由GMM模型最大后验概率来完成,判决方法如图3所示。

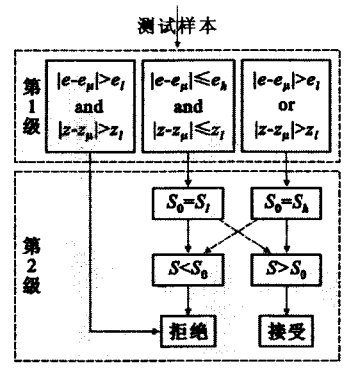

图3　联合判决门限决策

Fig. 3 Combined threshold decision

当$|e-e_\mu|\leqslant e_h$，且$|z-z_\mu|\leqslant z_h$时，说明该测试样本类似于训练模板，则把第 2 级高斯混合模型的门限值设为低门限阈值($\eta_0=\eta_l$)，以此来降低虚警率；当$|e-e_\mu|\geqslant e_l$，且$|z-z_\mu|\geqslant z_l$时，说明该测试样本与训练模板有很大差距，因此可以直接省掉高斯混合模型识别，设定其判决结果为拒绝；其他情况则认为仅通过第 1 级识别无法做出明确判断，所以将第 2 级高斯混合模型识别的门限值设为高门限阈值($\eta_0=\eta_h$)，以此来降低误识率。

联合判决门限不仅可以让三种方法互补，且在第一部就可排除距离模板最偏远的测试样本，缩小了需 GMM 检测的样本范围。且由于短时平均能量和过零率都已在预处理阶段得到，无需在识别阶段再重复计算，有效地降低了计算量。

4 实验结果及分析

实验选用的是 GMM 的说话人识别系统，语音采样率为8 000 Hz，帧长为20 ms，语音参数为 16 维 Mel 频率倒谱系数。实验语料来自 50 人（男女各 25 人），每人 30 条语音样本，每个样本时间为2 s。

4.1 GMM 模型识别结果

下图为分别运用 LBG 算法及 K-均值算法求取初始点的情况下，不同高斯混合数时 GMM 模型的识别结果对比。

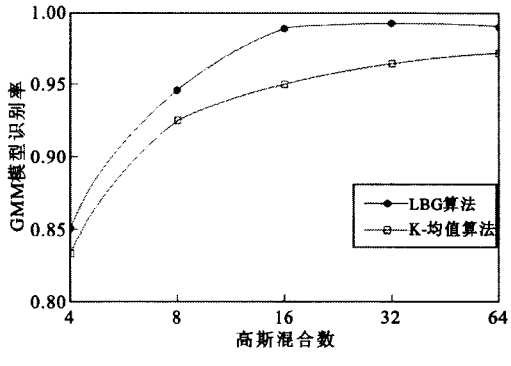

图 4 GMM 实验结果

Fig. 4 Experiment results of GMM

从图 4 可以看出，在其他条件相同的情况下，运用 LBG 算法求取初始点比运用 K-均值算法具有更高的识别率。GMM 模型的识别率会随着高斯混合数的增大，先升高再降低，当高斯混合数为 32 时识别率达到最高。

4.1 选择不同门限对识别率的影响

下图为分别以 GMM 模型低门限值(s_l)、高门限值(s_h)及联合判决决策作为判决门限时的误识率和虚警率柱状图。

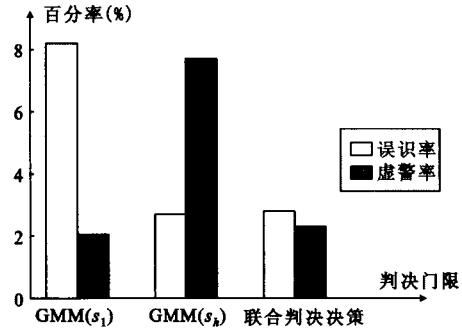

图 5 不同判决门限时的误识率和虚警率

Fig. 5 False positive rate and false alarm rate histogram of different decision threshold

根据实验结果可知，当采用 GMM 模型作为判决门限时，随着门限值的变化误识率和虚警率成反比，这个矛盾在单一门限中是固然存在的，而通过使用联合判决门限则使得误识率和虚警率都被降低到对识别率影响最小的状态。此外，所有样本中仅通过第一级识别就被拒绝的有 38.3%，而由其造成的虚警率仅为 0.3%，有效提高了识别效率。

5 结 语

本文提出了在使用 LBG 算法求取初始点的前提下，结合 EM 算法和 MAP 准则完成声纹识别的训练和识别过程，并在统一基准条件下研究了不同的求取初始点的算法和高斯混合数对识别率的影响。优化了判决门限的设定，从实验结果来看，本文提出的联合判决门限决策在没有增加计算量的情况下有效地克服了传统声纹识别的识别性能矛盾，误识率和虚警率明显低于传统的 GMM 模型识别方法，说明本文提出的方法是有效的，此联合判决门限决策对基于其他方法的语音识别都具有参考价值。但必须指出的是虽然 GMM 模型的混合数越多，识别的结果就会越接近测试样本的分布情况，但相对的，所花费训练和识别的时间也会随之增加。且上述识别结果都是在实验室良好的环境下取得的，在现实环境中，由于噪声和信道的干扰，严重影响了识别率。在今后的工作中，将针对如何高效地提高 GMM 模

型混合数和声纹识别的鲁棒性做进一步研究。

参考文献:

[1] SLEIT A, SERHAN S, and NEMIR L. A Histogram-BasedSpeaker Identification Technique[C]//International Conference on ICADIWT. Piscataway:IEEE Pr-ess, 2008:384-388.

[2] 吴朝晖,杨莹春. 说话人识别模型与方法[M]. 北京:清华大学出版社,2009:26-31.
WU Chao-hui, YANG Ying-chun. Speaker Recognition-Model and Method[M]. Beijing: Tsinghua Univers-ity Press,2009:26-31.

[3] 王韵琪,俞一彪. 自适应高斯混合模型及说话人识别应用[J]. 通信技术,2014,47(7):738-739.
WANG Yun-qi, YU Yi-biao. Adaptive Gaussian Mixture Model and Its Application in Speaker Recognition[J]. Communications Technology,2014,47(7):738-739.

[4] DUDA R O, HART P E, STORK DG. Pattern Classific-ation[M]. Second Edition. New York: Wiley Inters-cience,2000:108-112.

[5] GAUVAIN J, LEE C. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains[J]. Speech and Audio Processing, IEEE Transactions on,1994,2(2):291-298.

[6] YOSEPH L, Buzo A, GRAY R M. An Algorithm for Vec-tor Quantizer Design[J]. IEEE Tran Commun,1980,28(1): 84-95.

[7] 王炳锡. 语音编码[M]. 西安: 西安电子科技大学出版社, 2002.
WANG Bing-xi. Speech Coding[M]. Xi'an: Xidian U-niversity Press,2002.

[8] 王秋雯. 基于 GMM-UBM 的快速说话人识别方法[D]. 哈乐滨:哈尔滨工业大学, 2011.
WANG Qiu-wen. Rapid Speaker Recognition Based on GMM-UBM[D]. Harbin: Harbin Institute of Tech-nology,2011.

作者简介:

于　娴(1989—),女,硕士研究生,主要研究方向为声纹识别、语音信号处理;

YU Xian (1989-), female, graduate student, majoring in voiceprint recognition and speech signal processing.

贺　松(1974—),男,硕士,副教授,主要研究方向为信号处理;

HE Song (1974-), male, M. Sci., associate professor, mainly working at signal processing.

彭亚雄(1963—),男,副教授,主要研究方向为信号处理;

PENG Ya-xiong (1963-), male, associate professor, mainly working at signal processing.

周　晚(1987—),女,硕士,助讲,主要研究方向为电子通信。

ZHOU Wan(1987-), female, M. Sci., assistant lecturer, mainly working at electronic communication.