基于聚类优化GMM提高说话人识别性能的研究

吴庆棋,林江云
(厦门大学 计算机科学系,福建 厦门 361005)

摘 要:高斯混合模型(GMM)已广泛地应用于文本无关的说话人识别系统,该方法具有简单高效的特点。但如果GMM模型的高斯混合分量的数目比较多时,整个模型运算的复杂度会比较大。针对这个问题,提出将聚类算法和传统的高斯混合建模结合起来从而优化高斯混合模型,能够有效地提高说话人识别的速度。实验结果验证了这种算法的高效性。

关键词:说话人识别;高斯混合模型;聚类算法
中图分类号:TN92 **文献标识码**:A **文章编号**:1673-629X(2009)04-0035-03

A Study on GMM Optimization with Clustering for Improving Speaker Recognition

WU Qing-qi, LIN Jiang-yun
(Dept. of Computer Science, Xiamen University, Xiamen 361005, China)

Abstract: Gaussian mixture model (GMM) has been widely used for text-independent speaker recognition. This method has simple and efficient character. However, if it has a large number of Gaussians in GMM, it leads to a large complexity of computation. To solve this problem, proposes a new method which combines classical GMM with clustering algorithm to optimize the GMM for reducing the complexity of computation. Experimental results demonstrated that our approach was quite efficient to reduce the complexity of computation.

Key words: speaker recognition; Gaussian mixture model; clustering algorithm

0 引 言

语言是人的自然属性之一,由于说话人发音器官的生理差异以及后天形成的行为差异,每个人的语音都带有强烈的个人色彩,这使得通过分析语音信号来识别说话人成为可能。说话人识别[1,2]按识别任务可分成说话人辨认和说话人确认,按识别对象可分为文本无关的说话人识别和文本相关的说话人识别。在说话人识别的各种算法中,高斯混合模型(GMM)[2]性能较好,方法简单,是目前最好的说话人识别算法之一。

1 说话人识别系统

高斯混合模型(GMM)具有简单高效的特点,可以很好地描述说话人在不同环境和生理条件下的声音特征,已广泛地应用于文本无关的说话人识别系统。在GMM里,从说话人语音(以下简称话语)抽取出来的特征矢量 x_t 对应的似然率可以用 K 个高斯分量表示:

$$p(x_t|\lambda)=\sum_{k=1}^{K}c_kN(x_t,\mu_k,\Sigma_k) \tag{1}$$

其中 c_k 是第 k 个高斯分量的权重,这些权重满足关系式:

$$\sum_{k=1}^{K}c_k=1 \tag{2}$$

$N(x_t,\mu_k,\Sigma_k)$ 为高斯混合密度函数:

$$N(x_t,\mu_k,\Sigma_k)=\frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}}\exp[-\frac{1}{2}(x_t-\mu_k)\Sigma_t^{-1}(x_t-\mu_k)] \tag{3}$$

根据Eq. (1)和Eq. (3),话语 $X=\{x_t,1\leqslant t\leqslant T\}$ 的似然率可计算如下:

$$L(X|\lambda)=\prod_{t=1}^{T}p(x_t|\lambda)=\prod_{t=1}^{T}\sum_{k=1}^{k}c_kN(X_t,\mu_k,\Sigma_k)$$
$$=\prod_{t=1}^{T}\sum_{k=1}^{K}c_t\frac{1}{(2\pi)^{D/2}\prod_{l=1}^{D}\sigma_{kl}}\exp[-\frac{1}{2}\sum_{l=1}^{D}\frac{(x_t-\mu_{kl})^2}{\sigma_{kl}^2}] \tag{4}$$

其中,$\mu_k=[\mu_{kl}]_{l=1}^{D}$ 和 $\Sigma_k=[\sigma_{kl}^2]_{l=1}^{D}$ 为高斯对角矩阵的均值和方差参数,D 是特征矢量的维数。说话人辨认系统[2],设 S 个说话人,对应的GMM模型分别为 λ_1, $\lambda_2,\cdots,\lambda_S$,目标则是对一个观测话语序列 $X=\{x_t,1\leqslant t\leqslant T\}$,找到使之有最大后验概率的模型所对应的说

收稿日期:2008-07-02
基金项目:"985工程"二期"信息技术"创新平台资助项目(0000-X07204)
作者简介:吴庆棋(1982-),男,研究方向为声音识别;林江云,硕士,研究方向为说话人识别。

话人 λ_s，即：

$$\hat{S} = \arg\max_{1\leqslant k\leqslant S} P(\lambda_k \mid X) = \arg\max_{1\leqslant k\leqslant S} \frac{P(X\mid\lambda_k)P(\lambda_k)}{P(X)} \quad (5)$$

假定 $P(\lambda_k)=1/S$，即每个说话人出现为等概率，且因 $P(X)$对每个说话人是相同的，上式可简化为：

$$\hat{S} = \arg\max_{1\leqslant k\leqslant S} P(X\mid\lambda_k) \quad (6)$$

如果使用对数得分，说话人辨认的任务就是计算：

$$\hat{S} = \arg\max_{1\leqslant k\leqslant S} \sum_{t=1}^{T} \log\ p(x_t\mid\lambda_k) \quad (7)$$

说话人辨认系统可表示如图 1 所示。

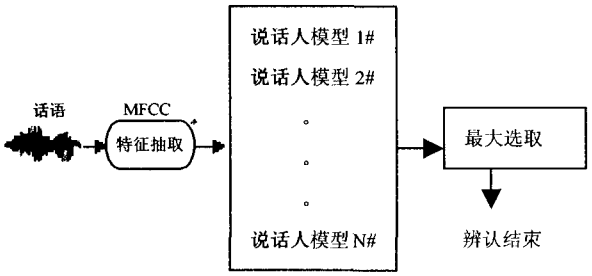

图 1 说话人辨认系统

说话人确认系统通过与一个预设阈值的比较，来做出接受或拒绝的判定。假设 H_0 表示输入话语 X 是来自目标说话人，H_1 表示来自冒充者，一个对数似然比分数定义如下[2]：

$$S(X) = \log\{p(H_0)/p(H_1)\} \quad (8)$$

为了做出确认决策，对数似然比分数 $S(X)$ 的值将与预设阈值 η 进行比较：

* $S(X) > \eta$：H_0 是真的，说话人被接受；

* $S(X) \leqslant \eta$：H_1 是真的，说话人被拒绝。

说话人确认系统可表示如图 2 所示。

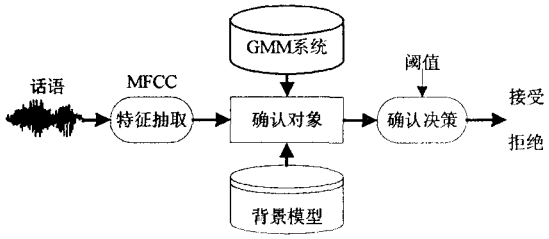

图 2 说话人确认系统

确认过程中，真实说话人的错误率叫错误拒绝率(FRR)或遗漏率(miss probability)，而冒充者的错误率叫错误接受率(FAR)或错误报警率(false alarm probability)[3]。两种类型的错误可以通过阈值 η 来调整。在实际操作过程中，更通常的是采用等错率(EER)，也就是 FAR 跟 FRR 相等的值，来评价说话人确认的性能。

为了计算似然分数，Eq. (8)以进一步表示为：

$$S(X) = \log\ p(H_0) - \log\ p(H_1) \quad (9)$$

其中，第一项是目标说话人的平均对数似然率，需要用话语的帧长 T 来归一化：

$$\log\ p(H_0) = \frac{1}{T}\log\ L(X\mid\lambda_v) \quad (10)$$

该项的值可能被许多因素如噪声、话筒等所影响。因此，第二项 $\log\ p(H_1)$ 的作用是使跟说话人无关的变化因素最小化，其计算一般是基于背景模型[2]，背景模型的选择标准是尽可能体现冒充者声音的统计分布。通常有三种方法：均值法、最大值法和算术平均法。由于最大值法有比较好的性能，文中选取对给定话语有最大对数似然率的说话人将被选为背景模型，即：

$$\log\ p(H_1) = \max_{\theta\neq v}[\log\ p(X\mid\lambda_\theta)] \quad (11)$$

对说话人 θ 的对数似然率也必须用话语的帧长来归一化：

$$\log\ p(X\mid\lambda_\theta) = \frac{1}{T}\log\ L(X\mid\lambda_\theta) \quad (12)$$

2 聚类优化 GMM 算法

传统的高斯混合模型是在混合数不变情况下的一种建模方法。而基于聚类优化的高斯混合模型初始状态混合数与最终的混合数是不一致的。在优化高斯混合模型过程中可将相近的几个高斯分量聚类在一起，然后再用 EM 算法进行训练直至收敛，使得整个 GMM 的混合数减少，在保证识别率不明显降低情况下，有效地提高了识别速度。文中是采用将两个相近高斯分量合并成一个，然后再用 EM 算法重新训练直至收敛，以此往复，最终优化整个 GMM 模型。方法如下：

设一个观测话语序列 $X = \{x_t, 1\leqslant t\leqslant T\}$。

1. 设置初始混合数 $K = k_0$；

2. 根据 Eq. (1) 和 Eq. (3) 利用 K - Means 算法初始化 GMM 参数 λ；

3. 运用 EM 算法训练直至收敛；

4. 在 GMM 模型中寻找距离相近的两个高斯分量，若它们的距离小于预设的距离阈值 α，将它们合并成一个新的高斯分量，并且将混合数 K 减少 1。

两个高斯分量距离计算公式采用 kullback - leibler[4] 距离计算方法：

如果两个高斯分量分别为：

$N_i\{c_i,\mu_i,v_i\}, N_j\{c_j,\mu_j,v_j\}$

$$D(N_i,N_j) = \sum_{k=1}^{l} \frac{v_i^k}{v_j^k} + \frac{v_j^k}{v_i^k} + \frac{(\mu_i^k-\mu_j^k)^2}{v_i^k+v_j^k} \quad (13)$$

其中 l 为语言特征向量的维数。

合并两个高斯分量的规则如下：

设两个高斯分量的参数分别为：$\{c_i,\mu_i,v_i\}$ 和 $\{c_j,\mu_j,v_j\}$，合并后新的高斯分量的参数为：$\{c_l,\mu_l,v_l\}$，

则:

$$p_1 = c_i / (c_i + c_j) \tag{14}$$

$$p_2 = c_j / (c_i + c_j) \tag{15}$$

$$c_l = c_i + c_j \tag{16}$$

$$\mu_l = p_1 * \mu_i + p_2 * \mu_j \tag{17}$$

$$v_l = (p_1 * (\mu_i * \mu_i^{\mathrm{T}} + v_i) + p_2 * (\mu_j * \mu_j^{\mathrm{T}} + v_j)) - \mu_l * \mu_l^{\mathrm{T}} \tag{18}$$

5. 返回步骤 3,直至整个 GMM 模型中没有两个高斯分量的距离小于预设的距离阈值 α 即可。

3　实验结果

实验采用 TIMIT 数据库[5]进行说话人辨认和说话人确认实验。TIMIT 数据库共包含 630 个说话人,它包含两个目录 TRAIN 目录和 TEST 目录,每个目录下又分别包含 8 个文件夹从 DR1 到 DR8。每个说话人 10 个句子,每个句子大约有 3 秒时间。进行说话人识别实验对语音数据,以帧长 25 毫秒、帧移 10 毫秒对语音分帧,提取 12 维 MFCC 特征和 12 维一阶差分 MFCC 特征。

3.1　说话人辨认实验

从 TIMIT 数据库的 TEST 目录中取 168 个人,每个说话人模型有 5 个训练句子(2SA + 3SI),5 个测试句子(5SX),用滤波器 $H(z) = 1 - 0.97z^{-1}$ 对语音预处理,以帧长 25 毫秒、帧移 10 毫秒对语音分帧,提取 12 维 MFCC 特征和 12 维一阶差分 MFCC 特征,模型的混合数目取 32 个。在聚类优化 GMM 过程中,距离阈值 α 取不同的值,对识别率和所有的 GMMS 高斯计算次数影响很大,实验结果如图 3、图 4 所示。

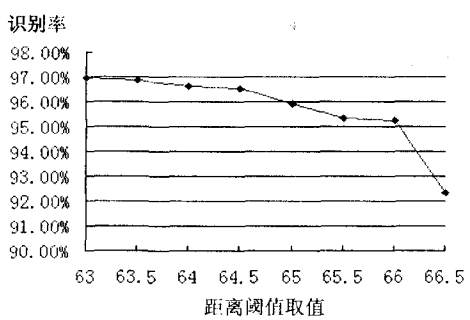

图 3　α 取不同值的识别率

从图 3、图 4 中可以看到,当 α 取值从 63 增大到 66.5,识别率就从 97.95% 降低到 92.34%,高斯计算次数也相应地从 4826 减少到 3603。而且可以看到,当 α 增大到一定值时,识别率会急速降低。这主要是因为 α 取值太大会把不大相近的高斯分量也会聚类在一起,从而造成了 GMM 对数据分布的描述不精确,降低系统的识别率。因此,距离阈值 α 取一个合理的值相当重要,不仅可以保证识别率不明显降低,而且可以有效减少系统高斯计算次数。接下来,比较非优化 GMM 系统和聚类优化 GMM 系统,文中 α 取 64.5。实验结果比较如表 1 所示。

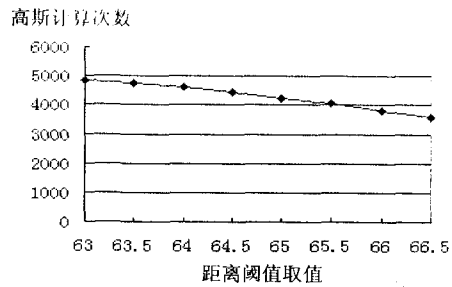

图 4　α 取不同值需要高斯计算次数

表 1　两种方法结果比较

	识别率	高斯计算次数
非优化 GMM 系统	97.33%	5376
聚类优化 GMM 系统	96.51%	4452

从表 1 可以看出,聚类优化 GMM 方法相对传统的 GMM 方法,识别率降低 0.82%,而速度却提高了 17.19%,有效地提高说话人辨认识别速度。

3.2　说话人确认实验

实验数据为 TIMIT 数据库中分别选取 TRAIN 目录和 TEST 目录中的 DR4 文件夹中的人,合起来总共 100 人,每个人选取 5 句语音(2SA 和 3SI)作为训练数据,另外 5 句语音(5SX)作为测试数据,MFCC 维数为 24 维,GMM 的混合度为 32。模型训练分别采用传统的 GMM 方法和聚类优化 GMM 方法,距离阈值 α 取 63。实验结果比较如表 2 所示。

表 2　两种方法结果比较

	等错误率	高斯计算次数
非优化 GMM 系统	1.36%	3200
聚类优化 GMM 系统	1.91%	2815

从表 2 可以看出,聚类优化 GMM 方法相对传统的 GMM 方法,EER 降低 0.55%,而速度却提高了 12.03%,有效地提高说话人确认识别速度。

4　结束语

采用的基于聚类优化 GMM 的方法相对传统的 GMM 方法,在说话人辨认中识别率降低 0.82%,而速度却提高了 17.19%,在说话人确认中 EER 降低 0.55%,而速度却提高了 12.03%。实验结果表明,这种方法采用聚类算法,将相近的高斯分量合并起来,从而达到优化模型的效果,有效地提高了说话人识别速度,从而证明了该算法的高效性。

(下转第 40 页)

$d(H(I_1),H(Q_1))$ $d(H(I_1),H(Q_2))$ $d(H(I_1),H(Q_3))$
$d(H(I_2),H(Q_1))$ $d(H(I_2),H(Q_2))$ $d(H(I_2),H(Q_3))$
$d(H(I_3),H(Q_1))$ $d(H(I_3),H(Q_2))$ $d(H(I_3),H(Q_3))$

(4)D_{3*3} 的任意一行或列表示该行或列所在的小图与另外图像的三幅小图比较所得的距离值,取其最小值即可表示该小图与另外图像的三幅小图比较相似度最大值。

$$B(I_i) = \min\{D[i][j] \mid j = 1,2,3\}$$

$$C(Q_j) = \min\{D[i][j] \mid j = 1,2,3\}$$

(5)将小图结合成整图得到两个整图的相似度,为了防止个别小图对整个大图的影响过大,取最大值为图像之间最终相似度。

$$D(Q,I) = \max(\sum_{i=1}^{3} B(I_i), \sum_{i=1}^{3} C(Q_i))$$

对拍摄者表达思想的中心位置可以加以一个较大的权值,再采用加权欧氏距离法来描述两图像的相似度,这样产生的效果会更好。

4 实 验

风景图像分类问题一直是计算机视觉研究的开放性课题。风景图像的分类可以提供对图像内容的理解,因此是图像检索中一个非常值得研究的问题[8]。王小玲提出的两种方法提高了对图像检索的能力,而且平均面积直方图法在一般情况下比主要面积法检索效果更好。在此使用日出、瀑布、花草和山脉图像各 100 幅,随机从中各抽取一幅图像作为模糊实例查询对象,设定阈值 0.33,使用新方法与平均面积直方图进行实验对比。结果如表 1、表 2 所示。

表 1 平均面积直方图实验结果

模糊实例	日出(100)	瀑布(100)	花草(100)	山脉(100)
日出图	71	7	5	17
瀑布图	6	66	4	13
花草图	3	9	81	5
山脉图	16	5	8	76
检全率	0.71	0.66	0.81	0.76
检准率	0.71	0.74	0.83	0.72

表 2 图像分割主要面积直方图实验结果

结果	日出(100)	瀑布(100)	花草(100)	山脉(100)
日出图	90	5	5	9
瀑布图	3	88	2	9
花草图	3	3	85	6
山脉图	9	4	5	91
检全率	0.9	0.88	0.85	0.91
检准率	0.82	0.87	0.83	0.83

从表 1 和表 2 的对比可以看出,不论是从检全率还是从检准率的角度上,图像分割主要面积直方图比平均面积直方图的检索效果都要好。

5 结束语

无论从理论上还是在实践中,传统的基于颜色的 CBIR 效果没有王小玲提出的基于平均颜色直方图或是最大面积直方图显著,而文中提出的图像分割面积直方图克服了平均颜色直方图的某些缺点,又进一步提高了图像检索的效率。

参考文献:

[1] Swain M J, Ballard D H. Color indexing[J]. International Journal of Computer Vision, 1991, 7(1): 11 - 32.

[2] Moxon B. Defining Data Mining[M]. [s.l.]: Miller Freeman, Inc., 1996.

[3] 王小玲. 基于内容的图像检索技术研究[D]. 上海: 上海交通大学, 2006.

[4] 章毓晋. 图像处理和分析[M]. 北京: 清华大学出版社, 1999.

[5] 阮秋琦. 数字图像处理学[M]. 北京: 电子工业出版社, 2001.

[6] 涂占新. 数据挖掘方法及其应用展望[J]. 中南财经政法大学学报, 2003, 2(2): 117 - 120.

[7] 韩祯祥, 张 琦, 文福拴. 粗糙集理论及应用[J]. 信息与控制, 1998(2): 37 - 44.

[8] 王 珏, 苗夺谦, 周育健. 关于 Rough Set 理论与应用的综述[J]. 模式识别与人工智能, 1996, 9: 337 - 344.

[9] Lin T Y, Cercone N. Rough Sets and Data Mining[M]. Boston: Kluwer Academic Publisher, 1997: 47 - 76.

(上接第 37 页)

参考文献:

[1] Reynolds D A, Rose R C. Robust text - independent speaker identification using Gaussian mixture speaker models[J]. IEEE Trans. Speech Audio Processing, 1995, 3(1): 72 - 83.

[2] Reynolds D A. Speaker identification and verification using Gaussian mixture speaker models[J]. Speech Communication, 1995, 17: 91 - 108.

[3] Doddington G R, Przybocki M A, Martin A F, et al. The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective[J]. Speech Communication, 2000, 31: 225 - 254.

[4] Yared G F G, Violaro F, Sousa L C. Gaussian elimination algorithm for HMM complexity reduction in continuous speech recognition systems[C]//Ninth European Conference on Speech Communication and Technology. Brazil: ISCA, 2005: 377 - 380.

[5] Fisher W, Zue V, Bernstein J, et al. An acoustic - phonetic database[C]//JASA, suppl. A. [s.l.]: [s.n.], 1986.