Software Engineering and Applications 软件工程与应用, 2013, 2, 1-5
http://dx.doi.org/10.12677/sea.2013.21001 Published Online February 2013 (http://www.hanspub.org/journal/sea.html)

Gaussian Mixture Model Training Method Based on Particle Swarm Optimizer for Speaker Recognition*

Liping Xue, Yinglong Yao, Zhiqiang Wang, Hong Zhou

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen
Email: xuelp@szu.edu.cn

Received: Nov. 14th, 2012; revised: Nov. 30th, 2012; accepted: Dec. 7th, 2012

Abstract: Expectation-Maximization (EM) algorithm is usually used to estimate parameters of Gaussian mixture model. Due to the hill-climbing characteristic of EM, any arbitrary estimation of the initial model parameters will usually lead to a sub-optimal model in practice. To resolve this problem, a hybrid training method based on Particle Swarm Optimization (PSO) is proposed. It utilizes the global searching capability of PSO and combines the effectiveness of EM. The particles perform basic operations of PSO (velocity updating and position updating) and EM algorithm, which can explore the training speech space to move toward the global optimum. The dependence of the final model parameters on the selection of the initial model parameters is also reduced. Experimental results have showed that this method can obtain more optimized GMM parameters and has better capability than EM in speaker recognition.

Keywords: Speaker Recognition; Particle Swarm Optimization (PSO); Gaussian Mixture Model (GMM)

说话人识别中基于粒子群优化的 GMM 训练方法*

薛丽萍，姚应龙，王志强，周　虹

深圳大学计算机与软件学院，深圳
Email: xuelp@szu.edu.cn

收稿日期：2012 年 11 月 14 日；修回日期：2012 年 11 月 30 日；录用日期：2012 年 12 月 7 日

摘　要：针对高斯混合模型(Gaussian Mixture Model, GMM)参数最优估计问题，常用的最大期望(Expectation-Maximization, EM)算法对初值敏感，在实际训练中极易得到局部最优参数，本文提出了一种 GMM 参数优化的新方法。将 EM 算法融入到粒子群优化(Particle Swarm Optimization, PSO)训练过程，形成了一种新的混合算法，利用 PSO 的全局探索和 EM 算法的局部深度搜索的混合策略，粒子在每次迭代中执行 PSO 速度位置更新和标准 EM 算法的混合更新操作，在训练语音矢量空间搜索最优高斯混合模型参数。从而避免传统 EM 算法陷入局部最优的缺点。说话人辨认实验表明，与 EM 算法相比，本文方法可以得到更优的模型参数，能有效提高系统的识别率。

关键词：说话人识别；高斯混合模型；粒子群优化

1. 引言

高斯混合模型(Gaussian Mixture Model, GMM)是近年来应用最广泛的说话人统计模型[1,2]，利用高斯密度函数的线性组合来表示每个说话人的训练语音在声学空间的分布，其模型的参数的估计一般利用最大期望(Expectation-Maximization, EM)算法[2]。EM 算法采用了最大似然为训练准则，具有很好的收敛性，然而 EM 算法是一种局部搜索算法，本质上采用爬山

*资助信息：深圳大学科学研究基金(200637)资助课题。

(Hill-Climbing)技术来寻找最优解，对初值十分敏感，容易陷入局部极优值，不能保证得到全局最优解。GMM 参数与结构的联合空间具有不可微、多峰值和欺骗性等特点，这使得近十多年来人们模拟自然界的一些自然现象而发展起了一系列智能优化算法成为 GMM 较好的优化途径。Hong 等提出了一种基于遗传分类的高斯混合模型训练方法[3]，林琳等提出了自适应小生境混合遗传算法对 GMM 模型参数进行优化[4]，王金明等结合模糊聚类和 Tabu 搜索算法优化 GMM 模型[5]，这些进化算法对 GMM 模型参数优化质量有一定的改善，但算法复杂性增加，求优过程耗时，对参数的设置敏感，收敛速度慢。在 GMM 参数训练的过程中，如何高效地得到 GMM 参数估计最优解是关键问题。

Eberhart 博士和 Kennedy 博士基于鸟群觅食行为提出了粒子群优化算法(Particle Swarm Optimization, PSO)[6]，由于该算法概念简明、实现方便、收敛速度快、参数设置少，是一种高效的搜索算法，近年来受到学术界的广泛重视，成为最受欢迎的优化算法。本文结合 PSO 和 EM 算法的特点，提出一种新的混合 GMM 训练方法(Hybrid Particle Swarm Optimization, HPSO)，以解决 GMM 参数估计问题，使其更准确地描述说话人的特征。利用粒子群优化算法的全局探索和 EM 算法的局部深度搜索的混合策略，粒子在每次迭代中执行 PSO 速度位置更新和标准 EM 算法的混合更新操作，在训练语音矢量空间搜索最优 GMM 参数。本文方法参数设置简单，不增加算法的复杂性，寻优能力较强。通过说话人识别实验证明了本文方法是一种有效的 GMM 优化方法，其识别性能优于 EM 算法。

2. 高斯混合模型及 EM 算法

高斯混合模型利用高斯密度函数的线性组合来描述每个说话人的训练语音在声学空间的分布。设某个说话人的语音特征矢量序列为 $\boldsymbol{X}=\{\boldsymbol{x}_t,1\le t\le T\}$，则其对应的混和数为 M 的高斯混合模型 $\lambda=\{C_i,\boldsymbol{\mu},\boldsymbol{\Sigma}_i\}$ $1\le i\le M$ 可以表示为：

$$P(\boldsymbol{x}_t|\lambda)=\sum_{i=1}^{M}C_iN(C_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \tag{1}$$

其中，C_i 表示混和权重，$N(C_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ 表示期望矢量为 $\boldsymbol{\mu}_i$，协方差矩阵为 $\boldsymbol{\Sigma}_i$ 的高斯分布密度函数。

$$N(\boldsymbol{x}_t,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)=\frac{1}{(2\pi)^{L/2}|\boldsymbol{\Sigma}_i|^{1/2}}\times\exp\left(-\frac{1}{2}(\boldsymbol{x}_t-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_t-\boldsymbol{\mu}_i)\right) \tag{2}$$

(2)式中 L 表示语音特征矢量维数。通常在说话人识别中，高斯密度函数的协方差矩阵采用对角型。

在说话人识别时，计算每一个说话人模型的后验概率，选择概率最大的说话人 k^* 为识别结果。其判决准则为：

$$k^*=\arg\max_{1\le k\le N}\sum_{t=1}^{T}\log P(x_t|\lambda_k) \tag{3}$$

GMM 的模型训练是一个有监督的训练过程。对于给定的训练集，GMM 训练的“好坏”需要一个评价标准，一般采用最大似然(Maximum Likelihood, ML)准则，由模型 λ 产生 $\boldsymbol{X}$ 的似然概率通常由对数似然度表示：

$$L(\boldsymbol{X}|\lambda)=\frac{1}{T}\sum_{t=1}^{T}\log P(\boldsymbol{x}_t|\lambda) \tag{4}$$

EM 算法一般由 K-means 聚类算法产生 λ 的初始值，然后估计出新的参数集 $\bar{\lambda}$ 使得 $P(\boldsymbol{X}|\bar{\lambda})\ge P(\boldsymbol{X}|\lambda)$，即最优模型参数能够使得训练集与模型匹配似然度达到最大。新模型参数 $\bar{\lambda}$ 再作为当前参数进行训练，这样迭代运算直到模型收敛。每一次迭代运算，都需要使用重估公式保证模型似然度的单调递增。各参数的重估公式为：

$$\bar{C}_i=\frac{1}{T}\sum_{t=1}^{T}P(i|\boldsymbol{x}_t,\lambda) \tag{5}$$

$$\bar{\boldsymbol{\mu}}_i=\frac{\sum_{t=1}^{T}P(i|\boldsymbol{x}_t,\lambda)\boldsymbol{x}_t}{\sum_{t=1}^{T}P(i|\boldsymbol{x}_t,\lambda)} \tag{6}$$

$$\bar{\boldsymbol{\Sigma}}_i=\frac{\sum_{t=1}^{T}P(i|\boldsymbol{x}_t,\lambda)(\boldsymbol{x}_t-\boldsymbol{\mu}_i)(\boldsymbol{x}_t-\boldsymbol{\mu}_i)'}{\sum_{t=1}^{T}P(i|\boldsymbol{x}_t,\lambda)} \tag{7}$$

EM 算法本质上是一种局部搜索技术，能最终收敛到一个局部极值点。

3. 粒子群优化的 GMM 训练方法

3.1. 粒子群优化算法

在一个 D 维的目标搜索空间中，随机生成 P 个粒

子，第 i 个粒子的位置可表示为 $\boldsymbol{z}_i=(z_{i1},z_{i2},\cdots,z_{iD})$，速度为 $\boldsymbol{v}_i=(v_{i1},v_{i2},\cdots,v_{iD})$，根据适应度函数计算 $\boldsymbol{z}_i$ 当前的适应值，来衡量粒子位置的优劣。粒子 i 迄今为止搜索到的最优位置为 $\boldsymbol{p}_i=(p_{i1},p_{i2},\cdots,p_{iD})$，整个粒子群迄今为止搜索到的最优位置为 $\boldsymbol{p}_g=(p_{g1},p_{g2},\cdots,p_{gD})$。每次迭代中粒子 i 根据以下公式更新速度和位置[7]。

$$v_{id}^{k+1}=wv_{id}^{k}+c_1r_1\left(p_{id}-z_{id}^{k}\right)+c_2r_2\left(p_{gd}-z_{id}^{k}\right) \quad (8)$$

$$z_{id}^{k+1}=z_{id}^{k}+v_{id}^{k+1} \quad (9)$$

其中，$d=1,2,\cdots,D$；k 是迭代次数；r_1 和 r_2 为均匀分布在[0,1]之间的随机数；w 为惯性权重；c_1，c_2 为学习因子。粒子在目标搜索空间中不断跟踪 $\boldsymbol{p}_i$ 和 $\boldsymbol{p}_g$ 进行搜索，直到达到预定的迭代次数为止。速度 v_{id} 取值范围为 $[v_{\min},v_{\max}]$，位置 z_{id} 的取值范围为 $[z_{\min},z_{\max}]$。

3.2. 粒子结构和适应度函数

本文算法中粒子结构的设计是基于 GMM 的参数。每个粒子代表着一个 λ。粒子结构如图 1 所示。

粒子的维数为 D 维，其中 $D=M\times(2\times L+1)$。第 i 个粒子的位置由图 1 表示。

粒子的最优位置由适应度函数值决定，适应度函数的选取要体现 GMM 设计质量，因此 HPSO 选取由模型 λ 产生 $\boldsymbol{X}$ 的对数似然度表示，利用公式(4)计算。

3.3. 粒子群初始化

本文算法群体初始化分为两步，首先，同 EM 算法一样，利用 K-means 聚类算法产生初始模型 λ_1，作为第 1 个粒子的位置。其次，其它粒子的初始化在初始模型 λ_1 的基础上加随机扰动而成。初始化步骤如下：

Step1：由 K-means 聚类算法产生初始模型 λ_1；

C_1	μ_{11}	…	μ_{1D}	Σ_{11}	…	Σ_{1D}	…

…	C_M	μ_{M1}	…	μ_{MD}	Σ_{M1}	…	Σ_{MD}

Figure1. Configuration of particle
图 1. 粒子结构

Step2：FOR $n=2$ to *Popsize*

FOR $k=1$ to M

$\lambda_n\cdot C_k=\lambda_1\cdot C_k*G(1.0,0.2)$

FOR $d=1$ to L

$\lambda_n\cdot\mu_{kd}=\lambda_1\cdot\mu_{kd}*G(1.0,0.2)$

$\lambda_n\cdot\Sigma_{kd}=\lambda_1\cdot\Sigma_{kd}*G(1.0,0.3)$

END

END

END

其中，*Popsize* 为群体粒子总数；M 为高斯混合数；L 为语音特征矢量的维数；$G(1.0,0.2)$ 是均值为 1，方差为 0.2 的高斯随机数；$G(1.0,0.3)$ 是均值为 1，方差为 0.3 的高斯随机数。

3.4. 粒子的混合更新策略

本文算法利用 EM 算法作为一个局部寻优操作子，与 PSO 的粒子速度和位置更新操作结合，形成一种新的混合更新操作策略，从而改善 PSO 算法的收敛速度和求解精度。粒子完成速度和位置更新之后，在进入下一次迭代之前，进行 EM 局部优化，使粒子能够尽快移动到最优点上，加快 HPSO 的收敛速度，取得计算代价与求解质量之间的较好平衡，进一步优化 GMM 的参数。

3.4.1. 粒子的 PSO 更新

利用更新公式(8)、(9)对粒子进行更新。每个粒子有自我学习和向优秀粒子学习的能力，根据自身最优和群体最优估计和调整每个 GMM 参数的最佳移动方向，向自己的历史最优点以及群体内历史最优点靠近，使 GMM 参数趋于最优。

3.4.2. 粒子的 EM 操作

粒子完成速度和位置更新之后，执行迭代次数为 5 的 EM 算法。

由于更新操作会破坏 GMM 参数的限制条件，因此在粒子执行完混合操作后会对混和权重 C_i 和协方差矩阵 $\boldsymbol{\Sigma}_i$ 的元素进行阈值限定，混和权重的阈值为 $C_{\min}=0.0001$，协方差的阈值为 $\sigma_{\min}=1/30\,M$。同时对混和权重进行归一化处理。

$$\overline{C_i}=C_i\Big/\sum_{i=1}^{M}C_i \quad (10)$$

3.5. 粒子群优化算法的 GMM 模型训练

粒子群优化算法的 GMM 模型训练流程图如图 2 所示。

算法主要步骤如下：

1) 粒子群初始化。设置最大迭代次数 $K_{\max}$，迭代次数计数器 $k=1$。

2) 对群体中的每个粒子：

i) 按适应度值更新确定 $\boldsymbol{p}_i$ 和群体的 $\boldsymbol{p}_g$；

ii) 粒子按公式(8)、(9)更新速度和位置；

iii) 粒子执行迭代次数为 5 的 EM 操作；

3) $k=k+1$，当 $k>K_{\max}$ 时，转向步骤 4)，否则转向步骤 2)。

4) 结束迭代，找出群体中的最优粒子，作为所训练的说话人的 GMM 模型。

4. 实验结果与分析

本文进行与文本无关的说话人辨认实验，实验中采用的说话人语音数据取自 TIMIT 语音数据库[8]。说话人语音数据分为 30 个女性说话人和 76 个说话人作为训练和辨认的集合，从 TIMIT 语音数据库的方言区 dr1、dr2、dr3 和 dr4 目录中随机抽取 30 个女性。选取 TIMIT 语音数据库的方言区 dr2 的 76 人，其中女性 23 人，男性 53 人。训练语句选择每个说话人 6 个时间较长的语句，训练语音时长约为 16 秒，测试语句为每个说话人的另外 4 个语句，以 1 句为一个测试语音，最短测试语音时长为 0.65 秒，最长测试语音时长为 5 秒。语音信号经过预加重系数为 0.95 的滤波，采用汉明窗进行分帧和加窗，提取 12 维 Mel 频率倒谱系数(Mel Fequency Cepstrum Coefficient, MFCC)及其 12 维一阶动态系数的组合作为说话人的特征矢量。

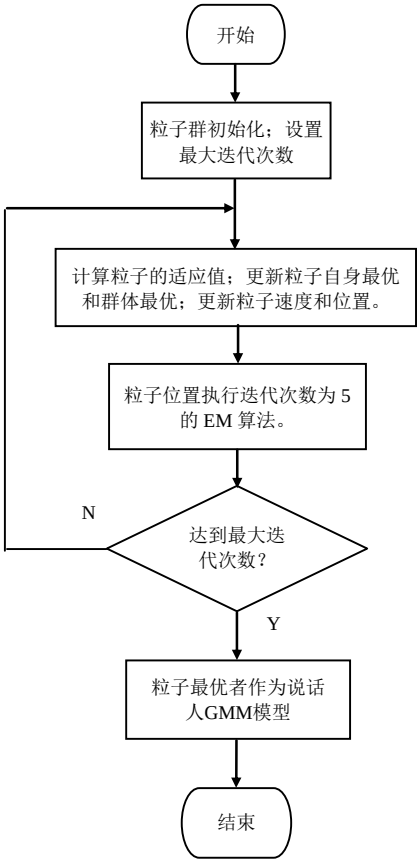

Figure 2. The flow chart of GMM training based on particle swarm optimizer
图 2. HPSO 的 GMM 模型训练流程图

实验中，对本文提出的算法与 EM 算法进行说话人辨认实验对比。

HPSO 算法的迭代终止条件设为最大迭代次数 $K_{\max}$，EM 算法的迭代终止条件设为 $L^k(X/\lambda)-L^{k-1}(X/\lambda)\le\varepsilon$。其中，$k$ 表示迭代次数。EM、HPSO 算法的参数设置如表 1 所示。

GMM 混合数 M 的选择与训练数据量密切相关，当训练数据量大时，较大的混合数 M 能提高辨认精度，而当训练数据有限时，M 值大反而会降低辨认率。由于 TIMIT 语音数据库每个说话人仅有 10 句，而且句子长短不一，每句话的有效语音时间较短平均约为 2.5 秒，因此本文实验混合数取值 $M=8$ 或 $M=16$。

在 GMM 混合数分别为 $M=8$、$M=16$ 情况下，HPSO、EM 算法的说话人辨认性能比较。表 2 给出了 10 次训练和辨认实验的平均误识率。

Table 1. The parameters for EM and HSPO
表 1. EM 和 HSPO 算法的参数设置

算法	ε	*Popsize*	$K_{\max}$	w	c_1	c_2
EM	0.001	-	-	-	-	-
HPSO	-	10	30	0.1	0.5	0.5

Table 2. Comparison of speaker recognition error rates
表 2. 说话人误识率比较

算法	30 人误识率(%)		76 人误识率(%)	
	$M=8$	$M=16$	$M=8$	$M=16$
EM	4.50	2.08	3.91	1.88
HPSO	3.42	1.83	3.23	1.65

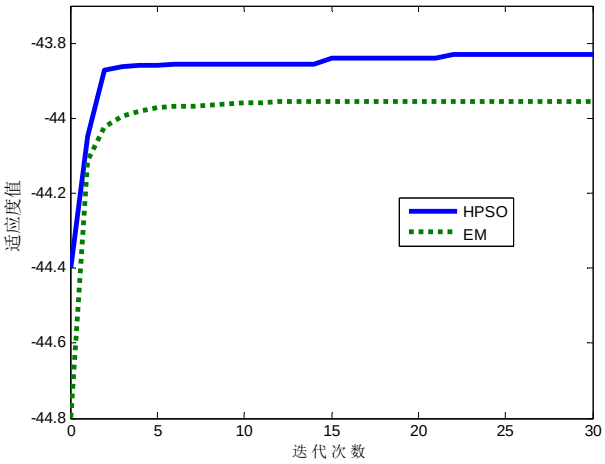

Figure 3. Comparison of converging procedure of the two algorithms
图 3. 两种算法的收敛特性比较

从表 2 可以看出，辨认 30 人时，HPSO 算法的误识率都低于 EM 算法，辨认集为同一方言区且人数增加到 76 人时，HPSO 算法误识率仍低于 EM 算法。可见 HPSO 算法有效地增强了优化 GMM 参数的能力。

为了进一步从收敛过程解释 HPSO 算法的有效性，图 3 给出了实验中两种算法对 1 个女性的 GMM 训练收敛过程。为了算法比较方便，在图 3 中纵坐标适应度均用对数似然度值，适应度值越大越好。HPSO 算法的对数似然度值取自于每次迭代中群体的最优值。

从图 3 整个收敛过程看，本文算法在粒子搜索前期收敛速度快，随着粒子位置与群体最优解的差值将会逐渐减少，粒子速度变小，EM 算法很快就收敛了到局部最优值。本文算法能够较成功地从局部最优值跳出，进一步接近全局的最优解，从而得到更优的模型参数。对其余说话人的 GMM 训练实验，也得到同样的趋势。

5. 结束语

本文提出了一种新的 GMM 训练方法应用于说话人识别，利用 EM 算法作为一个局部寻优操作子，与 PSO 的粒子速度和位置更新操作结合，形成一种新的混合更新操作策略，对 GMM 参数进行优化。算法采用群体规模较小的 10 个粒子，分别在目标搜索空间中的进行全局探索和局部精细搜索，从而较好地平衡全局优化和局部搜索，使得算法能够跳出局部最优。使用 TIMIT 语音数据库，进行了与文本无关的说话人辨认实验。取混合数分别为 M = 8、M = 16，比较了 EM 算法和本文算法的系统误识率，验证了混合更新策略可以提高系统的辨认性能。同时从收敛过程进一步证明了本文算法的有效性。实验结果表明新算法在与文本无关的说话人辨认性能方面优于 EM 算法，算法没有增加算法的复杂性，参数设置简单，寻优能力较强。在 GMM 训练方面，PSO 与其他群体智能优化算法对比研究或融合有待于进一步的研究。

参考文献 (References)

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

[2] D. A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication, 1995, 17(1): 91-108.

[3] Q. Y. Hong, S. Kwong. A genetic classification method for speaker recognition. Engineering Applications of Artificial Intelligence, 2005, 18(1): 13-19.

[4] 林琳, 王树勋. 基于自适应小生境混合遗传算法的说话人识别[J]. 电子学报, 2007, 35(1): 8-12.

[5] 王金明, 张雄伟. 一种模糊高斯混合说话人识别模型[J]. 解放军理工大学学报(自然科学版), 2006, 7(3): 214-219.

[6] J. Kennedy, R. Eberhart. Particle swarm optimization. Proceedings of the IEEE International Conference on Neural Networks 1995, Piscataway: IEEE Press, 1995: 1942-1948.

[7] Y. Shi, R. C. Eberhart. A modified particle swarm optimizer. IEEE International Conference on Evolutionary Computation Proceedings, Piscataway: IEEE, 1998: 69-73.

[8] J. S. Garofolo, L. F. Lamel. TIMIT acoustic-phonetic continuous speech corpus, 2012. http://www.ldc.upenn.edu/Catalog

