基于最大似然聚类的 GMM 优化方法及其在说话人辨认中的应用

胡 婕 周 琳

（东南大学 信息科学与工程学院，江苏 南京 210096）

摘 要：模式识别中基于高斯混合模型 GMM(Gaussian Mixed Model)的说话人辨认系统在训练样本充分的条件下获得了较高的识别率，但其计算复杂度往往限制了系统应用。从提高系统实用性的角度，介绍了一种基于最大似然 ML(Maximum Likelihood)聚类的简化 GMM 算法。仿真结果表明在保证系统识别性能的前提下，简化算法有效降低了计算开销。

关键词：模式识别；高斯混合模型；最大似然；说话人辨认

GMM Optimization Based on ML Clustering and Its Application in Speaker Identification

Hu Jie Zhou Lin

(Department of Information Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: In the domain of pattern recognition, although speaker identification system based on GMM (Gaussian Mixed Model) has achieved high performance with sufficient training data, its application is restricted by a large complexity of computation. This paper presents a simplied GMM algorithm based on ML (Maximum Likelihood) cluster for practical application. The simulation result indicates that the method reduces the computational complexity, while keeps a high identification performance.

Keywords: pattern recognition; GMM; ML; speaker identification

说话人辨认是根据语音信号中的个人特征决定说话人身份的过程。在与文本无关的辨认方式下，高斯混合模型以及由其衍生出来的统计学模型是目前较为有效的方法。通常，GMM 方法对所有用户使用统一的模型结构，通过期望最大化 EM(Expectation Maximization)算法进行参数估计，从而建立说话人的概率模型。然而，EM 算法的计算复杂度较大、收敛速度慢，使得传统 GMM 算法很难满足大数据集应用中的实时性要求。

在实际应用中，运算速度和识别率是同时需要考虑的问题。识别率要求一定数量的训练语音样本，而运算速度则随训练样本数的增加而减慢。在使用 DSP 实时实现说话人辨认系统时，算法计算量是需要解决的关键问题。

1 基于 GMM 的说话人辨认系统

混合阶数为 M 的 GMM 可以表示为：

$$p(x|\lambda)=\sum_{i=1}^{M} w_i \xi_i(x) \tag{1}$$

其中，x 为 D 维观测矢量，$\xi_i(x)$ 为 GMM 的第 i 个高斯分量，w_i 为混合权重，满足 $\sum_{i=1}^{M} w_i = 1$。D 维高斯函数可以表示为：

$$\xi_i(x)=\frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\} \tag{2}$$

其中，μ_i 为均值矢量，Σ_i 为协方差矩阵。整个 GMM 由各混合分量的均值矢量、协方差矩阵及混合权重描述，可用模型参数 $\lambda=\{w_i,\mu_i,\Sigma_i,i=1,\ldots,M\}$ 表示。

基于最大似然（ML）准则的 GMM 参数估计问题可以描述为：

$$\hat{\lambda}_{ML}=\arg\max_{\lambda} p(\mathbf{X}|\lambda) \tag{3}$$

其中，观测矢量序列 $X=\{x_1, x_2, \ldots, x_N\}$ 的似然函数为：

$$p(X|\lambda)=\prod_{n=1}^{N} p(x_n|\lambda) \tag{4}$$

式（3）的最大化过程可以通过 EM 算法来进行[1]。该算法迭代求解 GMM 参数，每步迭代可以使估计模型的似然函数值单调递增。当前后两次迭代所得的似然函数值之差小于预先设定的门限时，即输出估计结果。实验表明，GMM 的协方差阵可以用高阶对角阵近似，这种近似在保证识别性能的同时可有效降低系统运算量[2]。此时，协方差阵 Σ_i 的估计简化为方差 σ_i 的估计。可以证明[1]，下面的重估公式保证模型收敛到局部最优。

$$\hat{w}_i=\frac{1}{N}\sum_{n=1}^{N} p(i|x_n,\lambda) \tag{5}$$

$$\hat{\mu}_i=\frac{\sum_{n=1}^{N} p(i|x_n,\lambda)x_n}{\sum_{n=1}^{N} p(i|x_n,\lambda)} \tag{6}$$

$$\hat{\sigma}_i=\frac{\sum_{n=1}^{N} p(i|x_n,\lambda)x_n^2}{\sum_{n=1}^{N} p(i|x_n,\lambda)}-\mu_i^2 \tag{7}$$

其中，后验概率为：

$$p(i|x_n,\lambda)=\frac{w_i\xi_i(x_n)}{\sum_{j=1}^{M} w_j\xi_j(x_n)} \tag{8}$$

在说话人辨认系统中，每个说话人可以由一个参数为 $\lambda^{(k)}$（$k = 1,\ldots,K$,K 为闭集人数）的 GMM 代表。辨认的任务是找到一个说话者 k^*，其对应的模型参数 $\lambda^{(k^*)}$ 使得待识别语音的特征矢量序列 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ 具有最大似然度，即：

$$k^* = \arg\max_{1\le k\le K} p\left(\boldsymbol{X}|\lambda^{(k)}\right) = \arg\max_{1\le k\le K} \ln p\left(\boldsymbol{X}|\lambda^{(k)}\right) \tag{9}$$

2 改进 GMM 算法

式（1）中，观测矢量 $\boldsymbol{x}_n$ 的似然函数是由 M 个高斯概率密度加权求和得到的，该步计算可以通过维特比算法简化。维特比算法解决了给定观测矢量序列 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ 以及模型参数 λ 时，如何在最大似然的意义上确定一个最佳状态序列 $S = \{s_1, s_2, \ldots, s_N\}$ 的问题。

实验表明[3]，在 GMM 中，对于单个观测矢量 $\boldsymbol{x}_n$，只有一个高斯分量 i_n 对应的概率值较大，即单个观测矢量落入混合模型中某个高斯分布的可能性较大，因此其似然函数可以用该分布的概率密度函数 $\xi_{i_n}(\boldsymbol{x}_n)$ 近似，而 i_n 即代表了观测矢量所对应的状态。$\boldsymbol{x}_n$ 对应的高斯分量可以通过下式求解：

$$\begin{aligned} i_n &= \arg\max_{1\le i\le M} \xi_i(\boldsymbol{x}_n) \\ &= \arg\max_{1\le i\le M} \{\ln \xi_i(\boldsymbol{x}_n)\} \\ &= \arg\min_{1\le i\le M} \left\{(\boldsymbol{x}_n - \boldsymbol{\mu}_i)' \Sigma_i^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_i) + \ln|\Sigma_i|\right\} \end{aligned} \tag{10}$$

此时，观测矢量 $\boldsymbol{x}_n$ 的似然函数简化为：

$$p(\boldsymbol{x}_n|\lambda) \approx \xi_{i_n}(\boldsymbol{x}_n) \tag{11}$$

第 i 分量的后验概率简化为：

$$p(i|\boldsymbol{x}_n, \lambda) \approx \begin{cases} 1 & i = i_n \\ 0 & i \ne i_n \end{cases} \tag{12}$$

将式(11)和式(12)代入式(4)至(9)得改进算法流程如下。

训练流程：

1）输入训练样本序列 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$，设置最大迭代次数为 MaxIter；

2）初始化[4]：从训练样本集中均匀抽取 M 个样本作为均值向量的初始值 $\boldsymbol{\mu}_i^{(0)}$

协方差阵 $\Sigma_i^{(0)} = \mathrm{I}$

迭代次数 $m = 0$

3）找到每个样本点 $\boldsymbol{x}_n$ 对应的高斯分量：

$$i_n = \arg\min_{1\le i\le M} \left\{\left(\boldsymbol{x}_n - \boldsymbol{\mu}_i^{(m)}\right)' \left(\Sigma_i^{(m)}\right)^{-1} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_i^{(m)}\right) + \ln|\Sigma_i^{(m)}|\right\} \tag{13}$$

4）定义 $S_i = \{\boldsymbol{x}_n \in \boldsymbol{X} \mid i_n = i\}$ 为每个状态对应的样本集，更新模型参数：

$$\boldsymbol{\mu}_i^{(m+1)} = \frac{1}{N_i} \sum_{\boldsymbol{x}_n \in S_i} \boldsymbol{x}_n \tag{14}$$

$$\sigma_i^{(m+1)} = \frac{1}{N_i} \sum_{\boldsymbol{x}_n \in S_i} \boldsymbol{x}_n^2 - \left(\boldsymbol{\mu}_i^{(m+1)}\right)^2 \tag{15}$$

其中 N_i 为 S_i 中包含的样本点数；

1）判断 $m <$ MaxIter? 若否，转入 6）执行，否则，令 $m = m+1$，转入 2）执行；

2）迭代终止，输出模型参数：

$$w_i = \frac{N_i}{N}$$
$$\mu_i = \mu_i^{(m)}$$
$$\sum_i = \mathrm{diag}\left\{\sigma_i^{(m)}\right\}$$

识别流程：

1）输入待识别的特征序列 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$；

2）找到样本 $\boldsymbol{x}_n$ 在模型 $\lambda^{(k)}$ 下对应的高斯分量：

$$i_n^{(k)} = \arg\min_{1\le i\le M}\left\{\left(\boldsymbol{x}_n - \mu_i^{(k)}\right)'\left(\sum_i^{(k)}\right)^{-1}\left(\boldsymbol{x}_n - \mu_i^{(k)}\right) + \ln\left|\sum_i^{(k)}\right|\right\} \tag{16}$$

则该模型下，特征矢量序列的似然函数：

$$p\left(\boldsymbol{X}\mid\lambda^{(k)}\right) \approx \prod_{n=1}^{N} \xi_{i_n}^{(k)}(\boldsymbol{x}_n) \tag{17}$$

其中 $k = 1,\ldots,\mathrm{K}$ 为对应的说话人；

3）输出识别结果：

$$\begin{aligned} k^* &= \arg\max_{1\le k\le \mathrm{K}}\left\{\sum_{n=1}^{N} \ln \xi_{i_n}^{(k)}(\boldsymbol{x}_n)\right\} \\ &= \arg\min_{1\le k\le \mathrm{K}}\left\{\sum_{n=1}^{N}\left[\left(\boldsymbol{x}_n - \mu_{i_n}^{(k)}\right)'\left(\sum_{i_n}^{(k)}\right)^{-1}\left(\boldsymbol{x}_n - \mu_{i_n}^{(k)}\right) + \ln\left|\sum_{i_n}^{(k)}\right|\right]\right\} \end{aligned} \tag{18}$$

上述改进算法避免了高斯分布函数中的指数和除法运算，可通过乘累加和比较逻辑实现，在 DSP 上实现时，可有效提高识别系统的实时性。

3 实验结果

3.1 语音库描述

实验采用 TIMIT 语音库，该语音库为纯净的阅读语音，16kHz 采样，16bit 线性量化。实验中选取了 100 名说话人的“sx”语音，该类语音对于所有说话人的语料内容相同。每个人 5 句话，每句话大约持续 3 秒，其中 3 句用于训练，2 句用于识别。由于整句话较短，因此没有采用分段测试，而是采用整句测试。整句测试时识别率的计算，即用所有正确识别的人数除以总人数。

3.2 特征提取

语音特征采用 12 维 MFCC 参数，特征提取的其它参数设置为：

1）语音信号 16kHz 采样，帧长 256 个采样点，帧移 128 个采样点；

2）预加重系数 $\alpha = 0.97$；

3）数据窗采用 Hamming 窗，256 点 FFT；

4）Mel 滤波器的个数为 24；

5）由于使用的是 GMM 模型，因此不对特征进行倒谱加权。

3.3 收敛性能

图 1 比较了原始 GMM 算法和改进算法的收敛性能。其中，两种算法的模型阶数均取为 16，纵坐标为所有训练语音帧对数似然函数求和所得。可以看出，改进算法的收敛性能较原始算法有所降低，但仍可保证训练模型以较快速度收敛到局部最优。

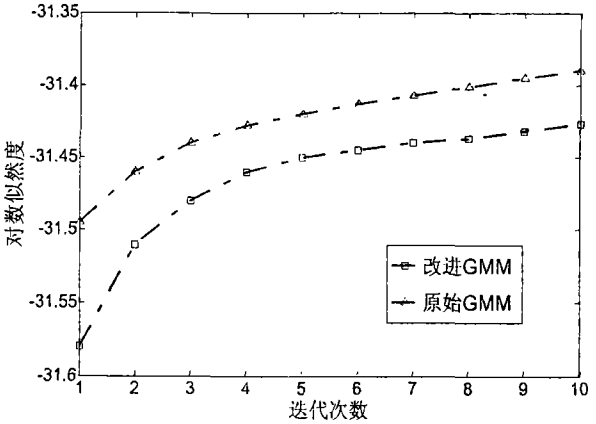

图 1 收敛性能比较

3.4 识别性能

表 1 比较了原始 GMM 算法和改进算法的 CPU 运行时间。该仿真在 CPU 主频为 2.66GHz 的 Pentium(R) 4 机器上进行。其中，两种算法的迭代次数均设为 13 次，训练和识别时间均为一个人的平均时间。可以看出，在相同模型阶数条件下，改进算法提高了系统的运行速度。

由改进算法的训练、识别流程可以看出，计算开销的节省主要体现在以下几个方面：

1）简化算法计算每个特征向量的似然函数时，用单模高斯模型近似 GMM，这种近似避免了式（1）中的 M 次乘累加操作；

2）简化算法不需要计算似然函数的具体值，只需要比较对数似然函数的大小，从而避免了式（2）中的指数运算；

3）简化算法的模型更新公式只需乘上相应的比例因子，避免了原始公式（6）（7）中的除法操作；

4）简化算法不需要计算式（8）中的后验概率，只需根据比较逻辑的结果，对特征向量的后验概率进行 2 进制赋值操作。

表 1 CPU 时间比较（单位：秒）

		改进 GMM	原始 GMM
M = 16	训练	0.33	0.45
	识别	2.88	3.02
M = 32	训练	0.65	0.75
	识别	5.62	6.13
M = 64	训练	1.33	1.47
	识别	11.71	12.64

表 2 比较了原始 GMM 算法和改进算法的识别率性能，实验条件同上。该实验共进行了 3 次，每次从语音库中抽取 100 个说话人进行测试，相应的识别率为 3 次的平均值。可以看出，改进算法在提高系统运行速度的同时并未显著降低识别率性能。但该算法对模型阶数的选择更为敏感，当模型阶数选择的不合适时，会对识别率有较大影响。

表 2 识别率比较（单位：%）

	改进 GMM	原始 GMM
M = 16	88.7	91.0
M = 32	87.3	89.7
M = 64	86.3	87.0

3.5 模型阶数与迭代次数的选择

图 2 为模型阶数 M = 16、32、64 时识别率与迭代次数的关系曲线。所有实验均在相同的数据集中进行。由该图可知，对于实验中采用的语音库，当模型阶数取为 32 时可获得较好的识别性能。识别率随迭代次数递增，且当迭代次数大于 10 次时，模型趋于稳定。M = 64 时，模型出现欠拟合，单纯增加迭代次数对识别性能的提升影响不大。

由于改进算法对 GMM 参数估计问题进行了简化，每个特征向量的似然函数值仅用单模高斯模型近似，当模型阶数较小时，不能充分拟合特征空间的分布，增加了模型参数的估计误差；同时，GMM 参数的个数随模型阶数呈线性递增，当模型阶数取的较大时，对应于较大的参数集，参数估计误差对识别性能的影响也较大，但这种影响可以通过增加训练的数据量克服。

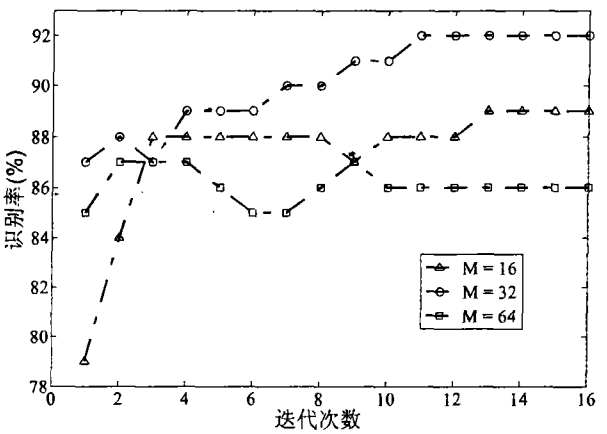

图 2 不同模型阶数的识别率与迭代次数关系曲线

4 结束语

本文从基于 GMM 说话人辨认系统的实用性角度出发，介绍了一种基于最大似然聚类的简化算法，并将该算法与原始 GMM 算法进行了比较。改进算法避免了计算高斯分布函数时的指数和除法运算，更适合 DSP 实时实现。实验表明，该算法在提高系统运行速度的同时不

会显著降低识别率性能，但对模型阶数的选择较为敏感，通常将模型阶数取为保证识别率性能的最小值，以节省系统开销。

参 考 文 献

[1] BISHOP C.M., JORDAN M., KLEINBERG J., Pattern Recognition and Machine Learning [M]. New York: Springer-Verlag, 2006.

[2] REYNOLDS D.A., DODDINGTON G., DUNN R.B., Automatic speaker recognition using Gaussian mixture speaker models [J]. Lincoln Lab, 1996, J. 8: 173-192.

[3] HAUTAMAKI V., KINNUNEN T., KARKKAINEN I., Maximum a Posteriori Adaptation of the Centroid Model for Speaker Verification [J]. IEEE Signal Processing Letters, 2008, Vol. 15: 162-165.

[4] REYNOLDS D.A., ROSE R.C., Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models [J]. IEEE Trans. on Speech and Audio Processing, 1995, 3(1): 72-83.

作者简介：

胡婕（1983-），女，江苏徐州，硕士研究生，东南大学，研究方向为语音信号处理；

周琳（1978-），女，博士，江苏镇江，东南大学，副教授，研究方向为语音信号处理。