第 37 卷 第 14 期
Vol.37 No.14

计 算 机 工 程
Computer Engineering

2011 年 7 月
July 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)14—0172—03 文献标识码：A 中图分类号：TP18

基于 SVM-GMM 的开集说话人识别方法

陈 黎，徐东平

(武汉理工大学计算机科学与技术学院，武汉 430063)

摘 要：建立一种支持向量机-高斯混合模型(SVM-GMM)，用以提高开集说话人识别的识别率。该模型的基本思想是将 SVM 的分类结果用 GMM 模型进行确认。由于 SVM 模型具有较好的分类性能，而 GMM 模型能够较好地描述类别内部的相似性，因此这 2 个模型的组合能够优势互补，从而获得较好的识别效果。实验结果表明，使用 SVM-GMM 模型能有效地提高开集说话人识别的识别率。

关键词：支持向量机；高斯混合模型；开集说话人识别；等误识率

Method of Open-set Speaker Recognition Based on SVM-GMM

CHEN Li, XU Dong-ping

(School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China)

【Abstract】This paper sets up a new Support Vector Machine-Gussian Mixture Model(SVM-GMM) to improve speaker recognition rate based on open-set. The basic idea of the new model is that the classification results of the SVM are confirmed with GMM. Due to the good classification performance of SVM and the good description of the internal similarity of some category of GMM, the good recognition effect can be obtained by combining the two models. Experimental results show that using SVM-GMM model can improve the open-set speaker recognition rate effectively.

【Key words】Support Vector Machine(SVM); Gussian Mixture Model(GMM); open-set speaker recognition; Equal Error Rate(EER)

DOI: 10.3969/j.issn.1000-3428.2011.14.057

1 概述

说话人识别，是指从说话人的一段语音中提取说话人的个性特征，通过对这些个性特征的分析和识别，达到对说话人辨识或确认的目的[1]。说话人识别包括说话人辨认和说话人确认。按照测试集来分，说话人识别又可分为开集说话人识别和闭集说话人识别。

由于传统方法对集内外话者的区分能力不够，会产生大量的误识。之所以会产生大量的误识，是因为在正确的特征选择前提下，主要瓶颈之一在于阈值 η 的选取。

为了改善传统方法的不足，从阈值 η 的选取方法上着手，本文提出一种基于支持向量机(Support Vector Machine, SVM)和高斯混合模型(Gaussian Mixture Model, GMM)的开集说话人识别方法。

2 传统说话人识别性能

说话人辨认是指，将待测语音与某个集合相比较，识别出说话人是集合中的哪一个人。说话人确认是指，将待测语音与某一模型相比较，系统做出“是”或“不是”的二元判决。开集说话人识别是指，待识别说话人可能是训练集之外某个说话人。闭集说话人识别是指，待识别说话人必定是训练集之内的某个说话人[2]。

目前，闭集说话人识别已经取得了较好的识别性能。但开集说话人识别的识别性能还有待提高。传统的开集说话人识别方法如图 1 所示[3]，在训练阶段为集内的每个话者建立一个模型；在识别阶段，待测语音经过特征提取后，由集内的每个话者对其进行打分，选取最高得分 S，将 S 与阈值 η 进行比较，若最高得分大于阈值，则打 S 分的话者 n 就是与待测语音相匹配的话者；若最高得分小于阈值，则将待测语音判定为训练集之外的话者。

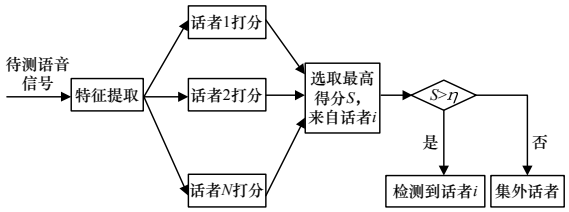

图 1 传统开集说话人识别模型

误识的多少严重影响说话人辨识系统的实用性，因此，需要研究减少说话人误识的方法，即减少集外话者辨识结构风险的方法，也就是集外话者的拒识方法[4]。

支持向量机模型有着优秀的分类性能[5]，高斯混合模型能够很好地描述类别内部的相似性。本文研究的辨识方法就是结合这 2 种模型的优点提出的。

3 支持向量机

支持向量机的基本思想是要求分类面不仅能将 2 类样本无错误地分开，而且要使 2 类样本的分类间隔最大。分类面是指能将 2 类样本分隔开来的平面。也就是说支持向量机要解决的问题是找到一个最优分类面，将 2 类数据分离开来。

对于线性可分的情况，假定最优分类面为 $\omega \cdot x + b = 0$。那么，$f(x) = \omega \cdot x + b$ 即为分类函数。在进行分类时，待分类样本为 X，若 $f(X) > 0$，则待测样本归为一类；若 $f(X) < 0$，则待测样本归为另一类。所以问题的关键是要通过已知样本集求出 ω 和 b，这即是在训练阶段需要求得的参数。

设已知训练集 $T = \{(x_1, y_1), (x_2, y_2), \mathrm{L}, (x_l, y_l)\}$，其中，

作者简介：陈 黎(1986－)，女，硕士研究生，主研方向：模式识别，数据挖掘；徐东平，教授

收稿日期：2010-12-30 **E-mail：**lizi_sky1000@yahoo.com.cn

$x_i \in R^n$，$y_i \in \{1,-1\}, i=1,2,\mathrm{L}\ ,l$，即样本被分为 1 和－1 两类。目标是寻找一个分类超平面 $\omega \cdot x + b = 0 (\omega \in R^n, b \in R)$，将样本输入空间划分为 2 个子空间，不同的模式样本属于不同的子空间。通过求解一个带约束的二次规划问题，可以求得：

$$\omega = \sum_{i=1}^{l} \alpha_i^* y_i x_i, b = y_j - \sum_{i=1}^{l} y_i \alpha_i^* (x_i \cdot x_j) \tag{1}$$

其中，下标 $j \in \{j \mid \alpha_j^* > 0\}$；$\alpha^*$ 为在二次规划问题中求解出的最优解对应的拉格朗日乘子。那么，对应的分类函数为：

$$f(x) = \sum_{i=1}^{l} \alpha_i^* y_i (x_i \cdot x) + y_j - \sum_{i=1}^{l} y_i \alpha_i^* (x_i \cdot x_j) \tag{2}$$

由于只需要判断 $f(x)$ 的符号，因此通常情况下是将 $f(x)$ 用符号函数的形式表示出来，即：

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{l} \alpha_i^* y_i (x_i \cdot x) + y_j - \sum y_i \alpha_i^* (x_i \cdot x_j))$$

在实际问题中，通常是非线性的问题较多。那么对于非线性的问题，通常可以通过非线性变换，即通过核函数 $K(x,x')$ 将原始集合映射高维特征空间，转化为某个高维空间中的线性问题，在变换空间求最优分类面。那么，在线性不可分的情况下，分类函数为：

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{l} \alpha_i^* y_i K(x_i \cdot x) + y_j - \sum y_i \alpha_i^* K(x_i \cdot x_j)) \tag{3}$$

目前，研究最多的核函数有 3 类：

(1)多项式核函数 $K(x,x') = (< x \cdot x' > +1)^4$；

(2)径向基函数 $K(x,x') = \exp(-|x-x'|^2 / \delta^2)$；

(3)Sigmoid 核函数 $K(x,x') = \tanh(v_k < x \cdot x' > + C_k)$。

4　GMM 模型

高斯混合模型(Gusssian Mixture Model, GMM)是分别对不同说话人的不同语音部分求取高斯分布，再用一组总和为一的权重系数，将之前求取的高斯分布组合成一个状态，用这个状态表示一个说话人的模型[6]。

高斯混合模型本质上是一种多维概率分布函数。一个具有 M 个混合成分的 D 维 GMM，可以用 M 个高斯成员的线性加权和来表示，即：

$$f(\boldsymbol{x}_t \mid \lambda) = \sum_{k=1}^{M} \omega_k N(\boldsymbol{x}_t \mid \boldsymbol{m}_k, \boldsymbol{\Sigma}_k) \tag{4}$$

其中，$\boldsymbol{x}_t$ 是一个 D 维观测矢量；$\omega_k, k=1,2,\mathrm{L}\ ,M$ 为混合权值，相当于每个高斯成员出现的概率，且 $\sum \omega_k = 1$；$N(\boldsymbol{x}_t \mid \boldsymbol{m}_k, \boldsymbol{\Sigma}_k)$ 为 D 维高斯混合函数，即：

$$N(\boldsymbol{x}_t \mid \boldsymbol{m}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{D/2} \left|\boldsymbol{\Sigma}_k\right|^{1/2}} \exp(-\frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{m}_k)'(\boldsymbol{\Sigma}_k^{-1})(\boldsymbol{x}_t - \boldsymbol{m}_k)) \tag{5}$$

其中，$\boldsymbol{m}_k$、$\boldsymbol{\Sigma}_k$ 分别是均值向量和协方差矩阵。

从上面的分析可以看出，$(\omega_k, \boldsymbol{m}_k, \boldsymbol{\Sigma}_k)$ 是说话人模型中很重要的部分，故使用 $\lambda = (\omega_k, \boldsymbol{m}_k, \boldsymbol{\Sigma}_k)$ 描述说话人的模型。对于一个说话人，$\lambda = (\omega_k, \boldsymbol{m}_k, \boldsymbol{\Sigma}_k)$ 是未知的，但是它可以通过 EM 算法从说话人的语音特征向量中估计得到。

EM 算法是期望最大化算法(Expectation Maximization)的简称。对于一组独立同分布的向量 $\boldsymbol{X} = (x_1, x_2, \mathrm{L}\ , x_T)$，EM 算法的主要思想就是采用最大似然概率估计的方式估计高斯混合模型中的参数 $\lambda = (\omega_k, \boldsymbol{m}_k, \boldsymbol{\Sigma}_k)$。即：

$$\lambda_{ML}^* = \arg\max_{\lambda} f(\lambda \mid x) = \arg\max_{\lambda} f(x \mid \lambda) \tag{6}$$

欲求得最佳的参数来描述所观测到的数据点，可由最大似然估计的概念来求得。在上述高斯混合密度函数的假设下，当 $x = x_i$ 时，其概率密度为 $f(x_i)$。若假设 $x_i, i=1,2,\mathrm{L}\ ,T$ 之间独立同分布，则发生 $\boldsymbol{X} = (x_1, x_2, \mathrm{L}\ , x_T)$ 的概率为 $f(X) = \prod_{i=1}^{T} f(x_i)$。由于 X 是已经发生过的事情(已有的训练语音)，因此希望找到 $\lambda = (w_k, \boldsymbol{m}_k, \boldsymbol{\Sigma}_k)$，使得 $f(X \mid \lambda)$ 能有最大的观测值。此种估计的方法称为最大似然估计(Maximum Likelihood Estimation, MLE)。

5　基于 SVM-GMM 的开集说话人识别

SVM-GMM 模型即是本文提出的一种改进的开集说话人识别方法。这种方法同样也分为训练阶段和识别阶段，训练阶段模型如图 2 所示，识别阶段模型如图 3 所示。

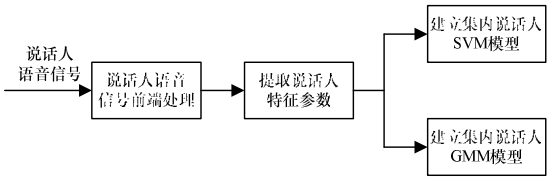

图 2　训练阶段模型示意图

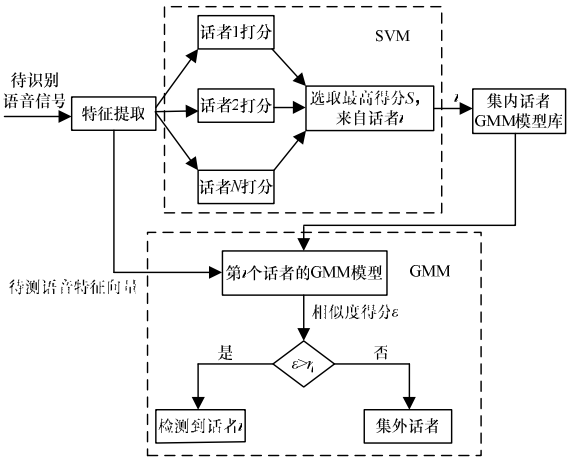

图 3　识别阶段模型示意图

在训练阶段，每个集内说话人不仅要建立 SVM 模型，还要建立 GMM 模型。而在识别阶段，SVM 模型主要是用于对待识别说话人进行分类，而 GMM 模型主要是用于对分类后的待识别说话人进行确认。

比较本文识别方法与传统的识别方法。图 3 表示的是本文的识别方法，图 1 表示的是传统的识别方法。不难看出 2 个模型的不同之处在于，对待测模型进行分类后，传统识别方法仅仅是对分类得分与阈值进行比较，分类得分值本身就是一个比较粗糙的值，即使选取到了一个较好的阈值也得不到较好的识别效果。故本文中提出了一种新的思路，通过 SVM 对待测语音进行分类，判断其跟集内的哪个话者是属于同一个子空间，若待测语音被分到与集内的第 i 个话者属于同一个子空间，则将待测语音输入到集内第 i 个话者的 GMM 模型中计算相似性得分，这里的相似性得分指的是相似性的概率。将这个相似性概率与某个概率阈值进行比较，以确定最后接受还是拒绝该语音。

6　仿真实验及实验分析

6.1　实验过程

在 Matlab7.0 实验环境下，对基于 SVM-GMM 的开集说话人识别进行了仿真实验，并对仿真结果进行了分析。语音样本来自 6 个说话人，其中，男性 3 人，女性 3 人。这 6 个说话人的语音均是来时普通实验环境下的录音。每个说话人读一段文字，内容不固定，长度为 1 min，普通声卡采样，采样频率为 44.1 kHz，量化精度为 16 bit。实验过程如下：

(1)语音信号的预处理和特征提取

对采集到的语音样本进行预处理和特征提取。预处理包括预加重、加窗分帧和端点检测。特征提取，即是提取语音中表示个性特征的特征向量。本实验中提取的特征参数采用的是美尔倒谱系数(MFCC)和其一阶差分组成的矢量。

1)预加重：为了消除发生过程中声带和嘴唇的效应，补偿语音信号受到发音系统所压抑的高频部分，将原始的语音信号 $s(n)$通过一个高通滤波：$H(Z)=1-0.937\ 5\ Z^{-1}$。

2)加窗分帧：采用一个帧长为 256，帧移为 128 的 hamming 窗在语音序列上滑动，对语音进行加窗分帧操作。

3)端点检测：采用双门限比较法进行端点检测，其中，短时能量高低门限为(10,2)，过零率高低门限为(10,5)，最大静音长度为 30 ms，最短语音长度为 150 ms。

4)特征提取：选用美尔倒频谱系数(Mel Frequency Cepstrum Coefficient, MFCC)及其一阶差分组成的矢量作为特征参数。首先将端点检测之后的帧信号经过离散 FFT 变换以得到其频率谱上的分布。将得到的能量谱乘以一组 20 个的三角带通滤波器。接着求取每一个滤波器输出的对数能量。再将这 20 个对数能量 E_k 进行离散余弦变换，求取 L 阶的 MFCC。在实验中取 L=16，同时还加上了差量倒谱系数，以显示倒谱系数对时间的变化。从而每一帧语音信号的特征参数即是一个 32 维的特征向量。

(2)SVM 训练和识别

本实验的 SVM 算法采用的是训练软件 LibSVM，其多类识别采用的是 one-against-one 分类算法，将多类问题转化为 2 类分类问题进行求解，为任意 2 类构建分类超平面，将 N 个类别两两分开，那么将会得到 $N\times(N-1)/2$ 个两类分类器。

(3)GMM 训练和识别

对 GMM 模型的训练，特征参数同 SVM 的训练参量是相同的，均是采用 16 维 MFCC 特征和 16 维一阶差分 MFCC 特征，GMM 的混合数目取 32。

使用 GMM 模型进行确认识别时，采用传统的 GMM 方法。首先是通过实验的方法得到一个阈值 η。取一部分测试数据，根据等误识率(Equal Error Rate, EER)原则估计出阈值 η。等误识率是指错误拒绝率(False Rejection rate, FR)等于错误接受率(False Accept rate, FA)时的值。然后进行确认识别，即将待测的特征参数输入到 GMM 模型中得到待测特征参量与 GMM 模型的相似度得分 ε，再将相似度得分与阈值 η 进行比较，以确定最终结果。

6.2 实验结果和分析

6.2.1 阈值选择实验

图 4 中显示的 2 条曲线分别为错误拒绝率(FR)和错误接受率(FA)曲线，2 条曲线的交点即是等误识率(EER)点，此点所对应的阈值即为所要确定的阈值。

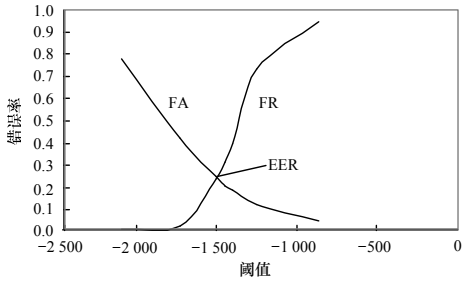

图 4 EER 阈值

训练集中的每一个话者均有一个阈值，图 4 显示的是话者 1 的 EER 阈值图示。话者 1 的阈值为－1 500，此时的错误拒绝率和错误接受率均为 24.50%。

6.2.2 传统方法与本文方法比较实验

将 6 个话者的语音中 4 个作为集内话者，另外 2 个作为集外话者。其中，集内的 4 个话者分别有 2 段 1 min 的录音，一个作为训练语音，另一个作为测试语音。每个测试语音分为 120 个样本，对这 120 个样本进行测试。

传统的开集说话人识别方法仅用 SVM 模型进行识别，集内话者的识别率和集外话者的拒识率如表 1 和表 2 所示。

表 1 SVM 模型对集内话者的识别 (%)

集内话者	错误拒识率	识别率
话者 1	16.83	83.17
话者 2	18.76	81.24
话者 3	16.50	83.50
话者 4	19.62	80.38

表 2 SVM 模型对集外话者的拒识 (%)

集外话者	错误接受率	拒识率
话者 1	30.67	69.33
话者 2	29.50	70.50

对本文提出的 SVM-GMM 模型的开集识别方法进行识别，集内话者的识别率和集外话者的拒识率如表 3 和表 4 所示。

表 3 SVM-GMM 模型对集内话者的识别 (%)

集内话者	阈值	错误拒识率	识别率
话者 1	−1 500	13.57	86.43
话者 2	−1 550	12.08	87.92
话者 3	−1 200	16.67	83.33
话者 4	−1 350	15.83	84.17

表 4 SVM-GMM 模型对集外话者的拒识 (%)

集外话者	错误接受率	拒识率
话者 1	20.54	79.46
话者 2	18.65	81.35

表 1~表 4 中的识别率是指对某话者的 N 个样本进行识别，其中，识别正确的样本所占的比例计算公式为：

$$\text{识别率}=\frac{\text{识别正确的样本数}}{\text{总的测试样本数}}$$

拒识率是指对集外某话者的 N 个样本进行识别，其中，拒识正确的样本所占的比例计算公式为：

$$\text{拒识率}=\frac{\text{成功拒识的样本数}}{\text{总的测试样本数}}$$

从表 1~表 4 的实验结果可以看出，本文提出的 SVM-GMM 模型进行开集说话人的识别，不管是对集内话者的识别性能还是对集外话者的拒识性能，都相对于传统开集说话人模型有了较大的提高。对集内话者的识别率，平均提高了约 3 个百分点；对集外话者的拒识率，平均提高了约 10 个百分点。

7 结束语

本文提出了一种结合 SVM 和 GMM 这 2 种识别方法的方法改进传统的开集说话人识别。实验结果表明，该方法有效地提高了对集外话者的拒识率。同时在实验中也发现阈值的选择也是影响识别结果的一个很关键的因素。在今后的工作中还会从阈值的选择、前端噪声的处理等方面进行研究，以期进一步提高开集说话人的识别效率。

(下转第 177 页)