文章编号:1673-159X(2010)01-0058-04

基于 SVM-GMM混合模型的说话人辨认研究

崔 宣[1],孙 华[1],刘 浏[2]

(1.西华大学机械工程与自动化学院,四川 成都 610039;
2.四川外语学院成都学院 计算机教研室,四川 成都 610039)

摘 要:通过建立一种新的混合模型——SVM-GMM模型,用以提高说话人辨认的识别率。其中介绍了高斯混合模型(GMM)和支持向量机(SVM)建立的基本原理,指出了高斯混合模型和支持向量机在实际应用中的不足之处,并针对这两种模型各自的特点,提出了将 GMM模型的输出机制引入到 SVM模型中,以便于调整支持向量机(SVM)模型的概率输出,并建立了 SVM-GMM混合模型。通过实验对比,验证了使用 SVM-GMM模型能有效的提高系统识别率。

关键词:说话人识别;高斯混合模型(GMM);支持向量机(SVM);SVM-GMM混合模型
中图分类号:TN912.34 **文献标识码**:A

Research of the speaker verification based on the SVM-GMM mixed model

Cui Xuan[1], iu liu[1], un hua[2]

(1. School of Mechanical Engineering and Automation, Xihua University, Chengdou 610039;
2. Chengdu Institute Sichuan International Studies University, Chengdu 610039 China)

Abstract: The authors put forward a new SVM-GMM mixed model to improve the recognition rate of a speaker verification system in the paper. Support vector machines (SVM) and Gaussian mixture model (GMM) are widely applied to speaker verification, but they both have disadvantages. The new approach for speaker verification presented is based on their features. The novel model introduces the output of Gaussian mixture model to support vector machines in order to adjust the probabilistic output of the support vector of machines. It can complement support vector machines with probabilistic information. The experiments show that SVM-GMM mixture model can effectively enhance the recognition rate of a speaker verification system.

Key words: speaker recognition; gsuss mixed model(GMM); support vector machines(SVM); SVM-GMM mixed model

0 引言

随着社会发展和科技进步,信息服务正走向智能化,而具有高度智能化的说话人识别(Speaker Recognition)服务已勃然兴起并广泛的应用于刑侦破案、国防监听、证券交易、声控、身份认证等等[1]。自 20世纪 70年代末至今,说话人识别的研究重点已转向对各种声学参数的线性或非线性处理以及新的模式匹配方法上[2],如动态时间规整、主成分分析、隐马尔可夫模型、高斯模型、支持向量机、神经网络和多特征组合等技术。

说话人识别可分为说话人辨认和说话人确认两类[2]。而对这两类都需要进行两个最基本问题的研究:第一是说话者声音特性特征参数选取;第二是说话人模型的建立。本文针对第二个基本问题进行研究:首先介绍 GMM模型和 SVM模型建立理论原理,接着讨论如何建立 SVM-GMM混合模型,最后对实验结果进行分析和总结。

1 背景知识

1.1 高斯混合模型

高斯混合模型[2-3,5](GMM)由于具有良好的识别性能,目前常被用于说话人识别,其本质是一种多维概率密度函数,一个具有 M个混合数的 D维 GMM,用 M个高斯分量的加权和来表示。

设 S个说话人,对应的 GMM模型分别为 λ_1,

收稿日期:2009-07-10
基金项目:西华大学青年基金(Q0920211)
作者简介:崔 宣(1982-),女,四川省绵阳市人,硕士,助教,主要研究方向为模式识别。

$\lambda_2, \cdots, \lambda_s$。目标则是对一个观测序列 X,找到最大后验概率的模型所对应的说话人 λ_k,即

$$\hat{S} = \arg\max_{1\leqslant k\leqslant s} p_r(X|\lambda) = \arg\max_{1\leqslant k\leqslant s} \frac{p(X|\lambda_k)p_r(x|\lambda_k)}{p(X)} \quad (1)$$

假定 $p_r(\lambda_k)=1/S$ 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:

$$\hat{S} = \arg\max_{1\leqslant k\leqslant s} p(X|\lambda) \quad (2)$$

1.2 支持向量机模型

统计学习理论指出:最优分类面具有最好的推广性能,这样求最优分类面的问题即转化为求优化问题

$$\min\varphi(w) = \frac{1}{2}\|w\|^2 = \frac{1}{2}<w\cdot w> \quad (3)$$

服从　$y_i[<w\cdot x_i>+b]-1\geqslant 0$　(4)

(其中,x_i是输入向量,w为权系向量,b为分类阈值)

该优化问题可以转化为对偶问题

$$\max W(a) = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i\ j=1}^{n} a_i a_j y_i y_j <x_i y_j> \quad (5)$$

服从 $a_i\geqslant 0(i=1,2,\cdots n)$ 和 $\sum_{i=1}^{l} y_i a_i = 0$　(6)

对于线性不可分的两类模式分类问题,可以通过选择适当的非线性变换映射到某个高维的特征空间,使得在目标高维空间中线性可分,进而转化成线性可分问题.高维空间的维数可能是非常高的,但 SVM巧妙地应用了核函数,不仅将样本映射到了高维空间,而且计算的复杂度也没有增加。核是一个函数 K,对所有 $x_s, x\in X$,满足 $K(x_s, z)=<\varphi(x)\cdot\varphi(z)>$,这里 φ是从 X到(内积)特征空间 F的映射.引入核函数后,其优化方程为:

$$\max W(a) = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i\ j=1}^{n} a_i a_j y_i y_j K(x_i, x_j) \quad (7)$$

服从 $a_i\geqslant 0(i=1,2,\cdots l)$ 和 $\sum_{i=1}^{l} y_i a_i = 0$　(8)

引入核函数后,最优超平面的分类函数为

$$f(x) = \mathrm{sgn}\{\sum_{i=1}^{n} a_i y_i K(x_i, x) + b\} \quad (9)$$

目前,应用较多的有线性核函数、多项式核函数和高斯核函数。支持向量机的特点是寻找不同类别之间的最优化分类面,反映的是异类数据之间的差异,但是其训练时间长,并且不能反映训练数据本身的特性[4,7-8]。

2 SVM-GMM混合模型的建立与分析

由于 SVM模型和 GMM模型各具优点和缺点,因此本文将高斯混合模型和支持向量机模型结合起来,建立 SVM和 GMM的混合模型,结合 GMM对数据表征能力强和 SVM对数据区分能力强的特点,利用 GMM对 SVM的输出进行调整,实现 SVM的概率输出[8,10]。

2.1 引入 SVM模型

假设支持向量机的输出格式是:

$$y = \mathrm{sgn}(f(x)) \quad (10)$$

其中　$f(x) = (w\cdot x)+b$　(11)

式中,x是输入向量,w为权系向量,b为分类阈值。

在计算过程中,对训练样本进行归一化,即对于离分类面最近的点(支持向量)满足[8,10]:

$|f(x)|=1$　即要求:

$$\min|(w\cdot x_i)+b|=1 \quad (12)$$

其中 x_1, $x_2\cdots x_n$是训练集中的数据。

实际上 SVM输出的 f(x)是一个距离测度[2],使用一个映射函数将 SVM的距离映射到后验概率上去。通过 Sigmoid函数可以给出支持向量机的概率输出形式:

$$p(\lambda|x) = p(y=1|x) = p(x) = \frac{1}{1+\exp(-f(x))} \quad (13)$$

即:$p(C_{+1}|x) = \frac{1}{1+\exp(-f(x))}$　(14)

以及　$p(C_{-1}|x) = \frac{1}{1+\exp(f(x))}$　(15)

式中,C_{+1}和 C_{-1}分别表示对应 +1和 -1类的样本集合,显然,处在分类面上的点对应的 +1类和 -1类的概率都是 0.5。

2.2 引入 GMM模型

下面介绍如何引入 GMM模型和如何应用 GMM模型对 SVM模型的输出进行控制。

首先,将 GMM模型引入 SVM模型,为每一类建立高斯混合模型。对于给定的向量,第 i个 GMM的概率输出为[2]

$$P_{GMM}(x_i|C_k) = \sum c_{km} N(x_i, \mu_{km}, \sum_{km}) \quad (16)$$

其中　$N(x_i, \mu_{km}, \sum_{km}) = \frac{\exp[-\frac{1}{2}(x_i-\mu_{km})^T\sum_{km}^{-1}(x_i-\mu_{km})]}{(2\pi^{d/2}|\sum_{km}|)^{1/2}}$　(17)

式中，α_{km}，μ_{km}和$\sum_{km}$分别表示第 i个 GMM的第 m个高斯模型的权值、均值和协方差。

接着建立 GMM-SVM混合模型，为使 GMM能更好的调整 SVM模型的输出，我们引入了两个调整因子 $S(x, C_{+1})$和 $S(x, C_{-1})$其形式如下：

$$S(x, C_{+1}) = \begin{cases} P_{GMM}(C_{+1}|x) + 0.5 & f(x) \geqslant 0 \\ 1.5 - P_{GMM}(C_{+1}|x) & f(x) < 0 \end{cases} \quad (18)$$

$$S(x, C_{-1}) = \begin{cases} 1.5 - P_{GMM}(C_{-1}|x) & f(x) \geqslant 0 \\ P_{GMM}(C_{-1}|x) + 0.5 & f(x) < 0 \end{cases} \quad (19)$$

其中

$$P_{GMM}(C_{+1}|x) = \frac{P_{GMM}(x|C_{+1})}{P_{GMM}(x|C_{+1}) + P_{GMM}(x|C_{-1})} \quad (20)$$

$$P_{GMM}(C_{-1}|x) = \frac{P_{GMM}(x|C_{-1})}{P_{GMM}(x|C_{+1}) + P_{GMM}(x|C_{-1})} \quad (21)$$

这里引入调整因子 $S(x, C_{+1})$和 $S(x, C_{-1})$的目的是根据高斯混合模型的结果对支持向量机的输出进行调整。再将 $S(x, C_{+1})$和 $S(x, C_{-1})$融合到前面的(14)、(15)式中，就得到了我们所建立的 SVM-GMM混合模型。对于给定输入矢量 x两类后验概率为：

$$p(C_{+1}|x) = \frac{1}{1 + \exp[-f(x)S(x, C_{+1})]} \quad (22)$$

$$p(C_{-1}|x) = \frac{1}{1 + \exp[f(x)S(x, C_{-1})]} \quad (23)$$

显然，文中新建模型很好的将 GMM模型和 SVM模型融合在一起，并利用调整因子的变化去增强 SVM模型的概率输出。具体调整过程如下：

利用 GMM模型的结果对 SVM进行调整，当 GMM的结果与 SVM的结果一致的时候，SVM的输出相应地增加，反之减少，也就是说，在 SVM的概率输出中，引入 GMM的分类结果可以反映到 SVM的输出中，这样按照实际情况起到了调整作用[8-9]。使得类间信息和类内信息能同时被反映在这个新的 SVM-GMM模型中。

3 实验及结果

实验中采用的语音数据为 25名录音人员，包含 10名女性，15名男性，在相距 30天的时段录了两次语音。录音在实验室环境下进行，都是采用 8K采样率。录音人员全部为学生，每人朗读随机 1～9的数字。语句长度为 10～30s，每个语句存为一个文件，每次录取 5段语句。特征参数提取前进行的预加重 $1-0.9375\,Z^{-1}$，分词处理，每段语音分割为帧长 30 ms 240个采样点，帧移设为 0，帧数共 48，加 Hamming窗平滑。在实验中的说话人系统，采用两种方式：

(1)采用 20阶 MFCC特征参数，采用高斯混合模型(GMM)作为分类器模型训练模板，建立说话人辨认系统。

(2)采用 20阶 MFCC特征参数，采用支持向量机(SVM)和高斯混合模型(GMM)混合模型作为分类器模型训练模板，建立说话人辨认系统。

使用[y, s]=wavread('train.wav')可得到原始语音 train1.wav的语音数据值为[84260* 2, 22050]。由此可以看出，对于每一个语音信号，都包含了大量的数据，这些数据就代表了每个人的说话内容，其原始语音信号如图 1所示[10]：

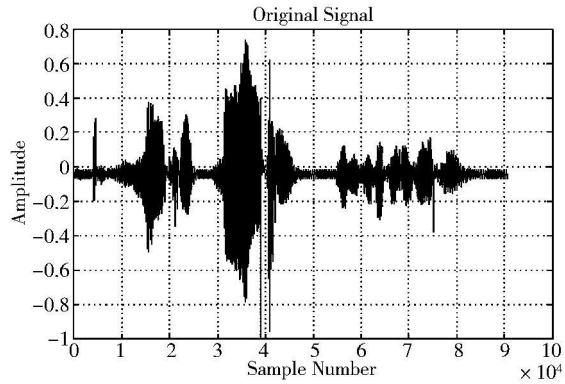

图 1 原始语音的时域波形图

此语音信号经过预加重、分帧、加汉明窗平滑后得到语音信号波形如图 2所示：

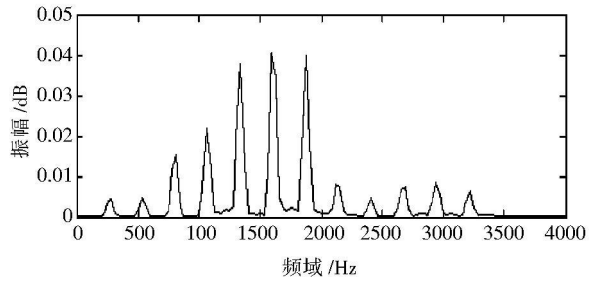

图 2 加汉明窗后的语音信号波形图

经过以上预处理后，提取 20阶 MFCC特征参数，得到了 20* 841的 MFCC数据，此时相对于原始的信号数据已经大大的压缩了原始数据量。由于所得数据是在 20维空间上的图形，不利于观察，把训练特征向量转换到 2维空间，得出图 3所示图形(给出两个语音信号 MFCC加以对比)。

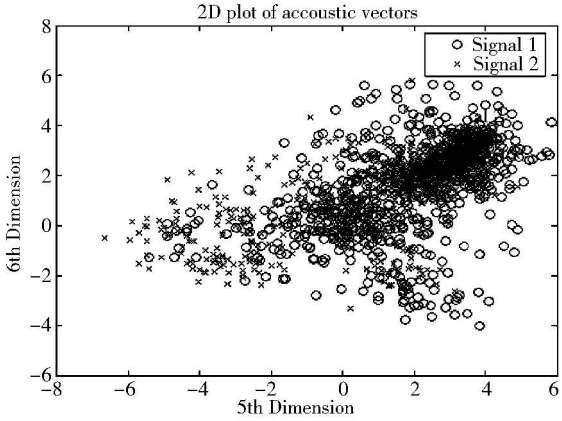

图 3　语音信号的 MFCC数据二维图

下面对基于新的 SVM-GMM混合模型的说话人辨认和基于当前流行的 SVM模型说话人辨认结果作比较。

根据训练时间的不同,两种模型的识别率也不同,可绘出说话人辨认实验中使用两种模型作为识别模型的误识率的比较图,以便分析。

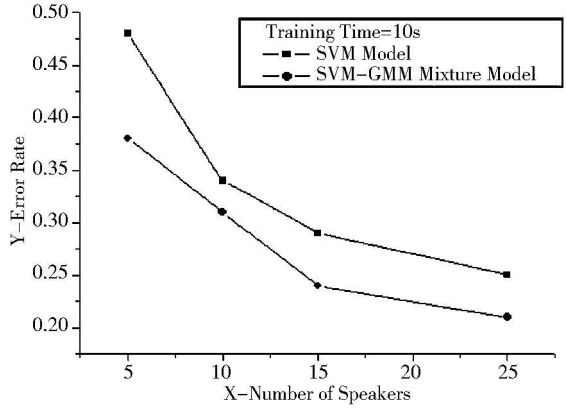

图 4　训练时间为 10S的误识别率对比图

从图 4中可以看出在训练时间为 10S时,两种模型的模式识别误识率都很高。相比较而言,SVM-GMM模型优于 SVM模型。

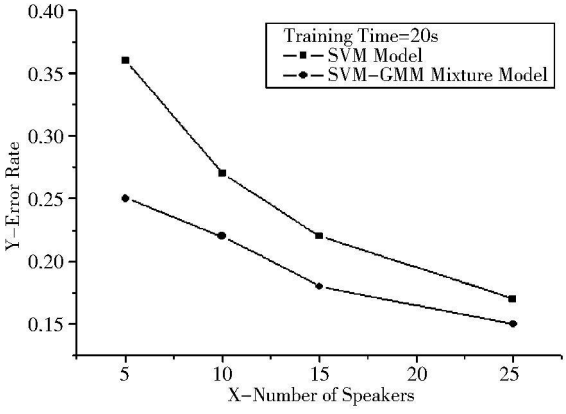

图 5　训练时间为 20S的误识别率对比图

如图 5所示,随着训练时间的增加,两种模型的误识率都在下降,但是基于 SVM模型的说话人辨认误识率始终高于基于新模型说话人辨认误识率。

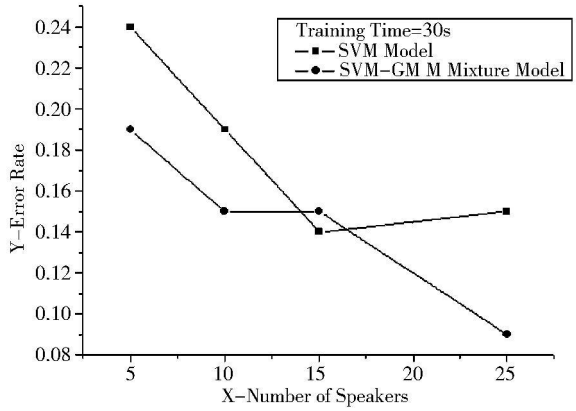

图 6　训练时间为 30S的误识别率对比图

从图 6中可以看出,当训练时间增加时新模型的误识率大幅度下降。并且参与识别的说话人人数越多,识别效果越显著。

基于 SVM模型的说话人辨认识别率以及基于 SVM-GMM混合模型说话人辨认的平均识别率如表 1和表 2:

表 1　使用 SVM模型的平均识别率

Trainnig time/s	10	20	30
	Average Rate of recognition (%)		
5(speakers)	0.52	0.64	0.76
10(speakers)	0.66	0.73	0.81
15(speakers)	0.71	0.78	0.86
25(speakers)	0.75	0.83	0.85

从表 1中数据结果可知,在 SVM模型中识别率随着说话人人数的增加而增高。但是随着训练时间的增长识别率有所降低。

表 2　使用 SVM-GMM模型的平均识别率

Trainnig time/s	10	20	30
	Average Rate of recognition (%)		
5(speakers)	0.62	0.75	0.81
10(speakers)	0.69	0.78	0.85
15(speakers)	0.76	0.82	0.85
25(speakers)	0.79	0.85	0.91

由表 2可知,在 SVM-GMM模型中识别率随着说话人人数的增加和训练时间增长而增高。

实验中,我们将基于新模型 SVM-GMM的说话人辨认和当前流行 SVM模型说话人辨认作了对比,实验数据结果证明,基于新模型的识别效果优于 SVM模型的识别效果。但是,通过对实验数据分析,随着说话人人数的增加,识别率 (下转第 88页)

成地震记录。

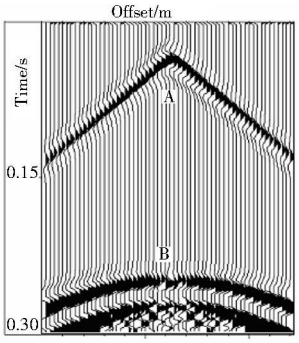

图 4 模型 I地震波方程模拟图

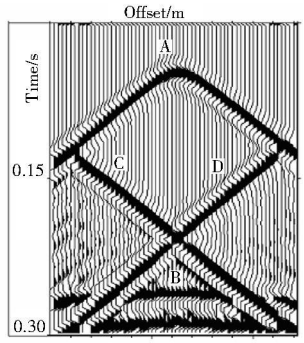

图 5 模型 II地震波方程模拟图

图 4中最上面的一组同相轴 A为直达波，另一组同相轴 B为来自于底部界面（y=0）的反射波。从图 5可以看出，在直达波 A和反射波 B中间又多出了两组同相轴 C、D。这是因为从震源发出的波到达地表曲面一侧后产生反射，该反射在接收时段内又传播到地表另一侧曲面上，被那里的检波器所接收，于是就形成了 C、D两组同相轴。

对于复杂地表或复杂界面情形的地球模型，用边界元法进行数值模拟非常方便，而且由曲面产生的反射，可以在合成地震记录剖面中得到正确的反映。

参 考 文 献

[1] Alterman Z. Karal F C. Propagation of Elastic Waves in Layered Media by Finite Difference Methods[J]. Bulletin of the Seismological Society of America, 1968, 58 (1): 367-398.

[2] 张关泉. 利用低阶方程组的大倾角差分偏移 [J]. 地球物理学报, 1986, 29(3): 273-282.

[3] 张关泉. 波动方程的上行波方程和下行波的耦合方程组 [J]. 应用数学学报, 1993, 16 (2): 251-263.

[4] 崔兴福，张关泉，吴雅丽. 三维非均匀介质中真振幅地震偏移算子研究 [J]. 地球物理学报, 2004, 47(3): 509-513.

[5] 孙卫涛，杨慧珠. 地形构造中地震波传播的非对称交错网格模拟 [J]. 应用数学和力学, 2004, 25(7): 686-694.

[6] 王忠仁，何樵登. 叠前共炮点道集的奇性反演研究 [J]. 地球物理学报, 1998, 41(增刊): 326-336.

[7] 符力耘，牟永光. 弹性波边界元法正演模拟 [J]. 地球物理学报, 1994, 37(4): 521-529.

[8] Li-Yun Fu and Ru-Shan Wu. Infinite Boundary Element Absorbing Boundary for Wave Propagation Simulations [J]. Geophysics 2000, 65(2): 596-602.

(编校：叶 超)

(上接第 61页)

上升的趋势在减缓，这是由于人数增加的同时增加了新模型的复杂度。因此，我们所得出的结论还具有一定的局限性，将来还要做大量的试验工作去证明新模型的可用性。

4 结论

SVM-GMM模型的优点是将 GMM的概率与 SVM的结果结合起来，既保留了 SVM判决能力强的优点，又体现了 GMM表征数据本身统计特性的能力[2][7]，通过 GMM的训练后所得到的先验概率知识，对 SVM的输出进行调整，这样使输出不仅考虑到样本的距离，而且考虑到样本的分布情况，使得到的后验概率输出更能反映实际的情况，从而使得整个说话人辨认系统的识别率得到提高。

参 考 文 献

[1] 崔宣. 基于语音混合特征说话人识别的研究 [D]. 成都：西华大学, 2008(6).

[2] Fine S, J Navratil R. A. Gopinath. A Hybrid GMM/SVM Approach to Speaker Indentification [J]. Speech and Signal Processing 2001(2): 568-572.

[3] Malayath N., Hermansky H., Kajarekar S, et al. Data-driven Temporal Filters and Alternatives to GMM in Speaker Verification [J]. Digital Signal Processing 2000: 55-74.

[4] Vapnik V. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlag 2001: 39-43.

[5] Daniel Garcia-Romero Julian Fierrez-Aguilar. Using Quality Measures for Multilevel Speaker Recognition [J]. Computer Speech and Language 2006(20): 192-209.

[6] Reynolds D. A., Rose R. C. Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models [J]. IEEE Trans Speech Audio Process 1995, 3 (1): 72-83.

[7] D. Xin Z Wu. Speaker Recognition Usingcontinuous Density Support Vector Machines [J]. Electronics Letters 2001 (37): 1099-1101.

[8] Daniel Garcia-Romero Julian Fierrez-Aguilar. Using Quality Measures for Multilevel Speaker Recognition [J]. Computer Speech and Language 2006(20): 192-209.

[9] Nakagawa S, Zhang W., Takahashi M. Text-independent/text-prompted Speaker Recognition by Combining Speaker-specific GMM with Speaker Adapted Syllable-based HMM [J]. IEICE Trans Inform. Syst 2006(3): 1058-1064.

[10] W. M. Campbell J P. Campbell D. A. Reynolds E. Singer P. A., et al. Support Vector Machines for Speaker and Language Recognition [J]. Computer Speech and Language 2006 (20): 210-229.

(编校：邓 英)