一种 GMM-SVM混合说话人辨认模型*

冷自强[1], 王金明[2], 林大会[3]
(1.解放军理工大学通信工程学院研究生 1队, 江苏 南京 210007;
2.解放军理工大学通信工程学院电子信息工程系; 3.解放军理工大学通信工程学院研究生 3队)

摘 要: 文中提出了一种 GMM和 SVM混合说话人识别模型,在特征参数域和概率得分域对两种模型进行了融合。混合模型结合了 GMM和 SVM各自的优势,使 SVM的概率输出兼顾各说话人模型内部和模型之间的信息,并有效解决了 SVM训练算法复杂,难以处理大量样本的问题。采用 TIMIT数据库进行了说话人辨认实验,结果证明本文提出的 GMM-SVM模型比传统的 GMM模型和 SVM模型具有更好的辨识性能。

关键词: 说话人辨认; 支持向量机; 高斯混合模型
中图分类号: TN912 文献标识码: B 文章编号: CN32-1289(2009)01-0086-04

Speaker Identification Model Based on GMM-SVM

LENG Zi-qiang[1], *WANG Jin-ming*[2], *LIN Da-hui*[3]
(1. Postgraduate Team 1 ICE, PLAUST, Nanjing 210007, China;
2. Department of Electronic Information Engineering ICE, PLAUST; 3. Postgraduate Team 3 ICE, PLAUST)

Abstract: A hybrid speaker recognition model based on GMM and SVM was presented. GMM and SVM were mixed in both feature parameter and likelihood score domain. The new model combined the advantages of GMM and SVM, making the SVM output probabilities contain both the information inside and between the speaker models. The problem that SVM training algorithm is too complex to deal with large number of training data was resolved. The GMM-SVM model was tested on the TIMIT database and showed better performance than GMM and SVM.

Key words: speaker identification; support vector machine; Gaussian mixture model

说话人识别是利用语音信号中包含的说话人个性信息,对其身份进行识别的技术。说话人识别可分为辨认和确认,辨认是对一段未知出处的语音,从已有说话人模型集合中找出与之匹配最佳的模型,是多选一问题;而后者是针对未知出处的语音,与其声称的说话人模型进行匹配,做出"是"或"不是"的判断,是二选一问题。

目前有两种比较流行的说话人识别模型,一种是概率统计模型,如高斯混合模型(GMM)和隐马尔可夫模型(HMM);另一种是基于判决的模型,如人工神经网络(ANN)和支持向量机(SVM)。两种模型各有特点,概率统计模型从统计的角度充分表示了数据的分布情况,反映的是同类数据本身的相似度特性;而判决模型利用训练数据的类别标识信息,反映的是不同类数据之间的差异。判决模型往往要比概率统计模型的性能稍好一些,但其缺点是不能反映训练数据本身的特性。文献[1]中的研究表明 GMM和 SVM在同样的训练数据上识别错误有很大不同,这说明二者有互补的方面。因此若说话人识别系统能结合二者的优点,将会在一定程度上提高识别率。

实际应用时,支持向量机[2]的训练算法复杂,计算量大,难以处理大量样本数据。有实验表明,当训练样

* 收稿日期: 2008-04-30;修回日期: 2008-10-28
作者简介: 冷自强(1984-),男,硕士生。

本数据个数为 4000时,存储核函数矩阵所需的内存可能多达 128兆。而在文本无关的说话人识别中,往往需要较长的训练和测试语音才能够达到较好的识别性能,一段训练语音的样本数常常会达到数千甚至上万个,并且训练 SVM 模型往往需要多段语音。显然,如此数量巨大的样本若不经处理直接训练,算法对硬件的要求和训练的效率都无法满足实际的需要。并且样本数量巨大,各个说话人在特征空间上的混叠程度也很大,将导致支持向量的数目过多,收敛速度变慢,识别阶段的计算量也会增加。

针对上述问题,本文提出了一种 GMM-SVM 混合识别模型,分别在特征参数域和概率得分域对支持向量机和高斯混合模型进行了混合,有效提高了说话人辨认系统的性能。

1　GMM-SVM模型

1.1　支持向量机多类分类

将支持向量机应用在说话人辨认方面,首先要解决多类分类问题。因为支持向量机是二类分类器,其训练时的输入是标识为"+ 1"和"- 1"的两类样本,识别时的标准输出也是测试样本属于"+ 1"或"- 1"中的哪一类,而说话人辨认的目的是判断一段语音是若干人中的哪一个所说,是多选一问题。支持向量机的多类分类方式[3]有一对一,一对余类,二叉树,纠错输出编码等方式。本文采用较为灵活的一对余类方式,即训练某个说话人模型时,将此说话人的特征向量标识为"+ 1"的训练样本,训练集中其余说话人的特征向量标识为"- 1"的训练样本,为每个说话人分别训练一个支持向量机模型。测试时,将测试语音的特征向量序列分别输入到每个说话人模型中,输出得分最高的模型对应的说话人被认为是真实说话人。

1.2　SVM与 GMM在特征参数域融合

由于计算量问题,支持向量机难以处理大量输入样本,而在与文本无关的说话人识别中,往往需要长时语音才能得到较为准确的说话人个性特征,相应的特征向量数目也是非常巨大,若直接将这些特征向量用于训练 SVM 模型,将导致难以接受的计算量。自然我们会想到从语音的大量样本中选取出具有代表性的样本作为支持向量机的输入。选取代表性的样本的方法很多,例如,对于 MFCC或 LPCC特征向量序列可以通过随机方式、矢量量化等方法选取。但这些方法具有很明显的缺点,随机选取的样本由于很强的偶然性,难以表示大量样本的分布情况。矢量量化方法虽然能较好的表示样本的分布中心,但仍包含有很多冗余信息(如语义信息、语种信息),并且鲁棒性较差。

GMM 模型作为一种统计模型,利用若干高斯概率密度函数的加权和来表示特征向量在概率空间的分布情况。GMM 模型使用较少的参数很好的描述了说话人的个性特征,在文本无关说话人识别方面得到了广泛应用。GMM模型由 EM 算法训练得到,其均值向量不但反映了各说话人在特征空间的分布,而且也较好的反映了说话人个性信息。因而可考虑采用 GMM 模型的均值向量作为 SVM 的训练样本。

1.3　SVM与 GMM在概率得分域融合

支持向量机作为二类分类器,标准输出方式是"+ 1"和"- 1",也就是测试向量属于两类中的哪一类。但是一帧语音的特征向量难免会包含冗余信息,硬性地将其分为某类显然并不合理,因此得到支持向量机的概率输出很重要。

支持向量机的软输出[4]是一个距离测度,表示测试向量到最优分类超平面的距离, Wahba把支持向量机的软输出通过 Sigmoid函数映射到后验概率上去,得到了支持向量机的一种概率输出形式:

$$P(y=1\mid \boldsymbol{x})=\frac{1}{1+\exp(-f(\boldsymbol{x}))} \tag{1}$$

其中 $\boldsymbol{x}$ 表示测试向量, $f(\boldsymbol{x})$ 表示支持向量机的软输出, y 为类别标识,取"+ 1"或"- 1"。

支持向量机依靠最优分类面分类,反映的是训练样本"类间"的关系;而高斯混合模型由数据本身特性生

成,反映了样本“类内”的关系,因此可以将两类输出进行融合。本文采用高斯混合模型的概率输出对支持向量机的概率输出进行调整,使支持向量机的输出兼顾了“类间”和“类内”的信息。考虑到高斯混合模型的概率输出动态范围较大,所以将高斯混合模型的输出取对数后进行归一化作为加权系数。一个具有 M 个混合数的 d 维 GMM[5]可表示为:

$$P_{\mathrm{GMM}}(\boldsymbol{x}\mid C)=\sum_{j=1}^{M}p_jN(\boldsymbol{x},\__j,\mathrm{E}_j) \tag{2}$$

$$N(\boldsymbol{x},\__j,\mathrm{E}_j)=(2\pi)^{-\frac{d}{2}}\mid\mathrm{E}\mid^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\__j)^T\mathrm{E}_j^{-1}(\boldsymbol{x}-\__j)\right) \tag{3}$$

其中 p_j 为权重,$\__j$ 和 E_j 分别表示多维高斯概率密度函数的均值和协方差矩阵。

设共有 N 个说话人,对应的模型分别为 $C_1, C_2,\cdots, C_N$,输入测试语句 $\mathbf{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_k,\cdots,\boldsymbol{x}_T\}$,$1\leqslant k\leqslant T$,则对说话人辨认来说,就是找到 $\mathbf{X}$ 对应的具有最大后验概率的模型。

$$\hat{S}=\arg\max_{1\leqslant i\leqslant N}P(C_i\mid\mathbf{X})=\arg\max_{1\leqslant i\leqslant N}\frac{P(\mathbf{X}\mid C_i)P(C_i)}{P(\mathbf{X})} \tag{4}$$

其中,$\hat{S}$ 表示目标说话人模型的序号。通常认为每个说话人出现的概率是相等的,即 $P(C_i)=1/N$,且 $P(\mathbf{X})$ 对每个说话人模型都是相等的,因此上式可简化为:

$$\hat{S}=\arg\max_{1\leqslant i\leqslant N}P(\mathbf{X}\mid C_i) \tag{5}$$

通常还认为特征向量之间是相互独立的,因此判别公式为:

$$\hat{S}=\arg\max_{1\leqslant i\leqslant N}\prod_{k=1}^{T}a_{ik}P_{\mathrm{svm}}(\boldsymbol{x}_k\mid C_i) \tag{6}$$

其中 a_{ik} 是 $\boldsymbol{x}_k$ 对第 i 个 GMM 模型输出的归一化对数得分,$P_{\mathrm{svm}}(\boldsymbol{x}_k\mid C_i)$ 表示 $\boldsymbol{x}_k$ 对第 i 个 SVM 模型的概率输出。计算时对上式等号两边取对数,求平均值,得:

$$\hat{S}=\arg\max_{1\leqslant i\leqslant N}\frac{1}{T}\sum_{k=1}^{T}\log(a_{ik}P_{\mathrm{svm}}(\boldsymbol{x}_k\mid C_i)) \tag{7}$$

1.4 GMM-SVM混合识别模型

图 1和图 2分别是 GMM-SVM混合模型训练和识别的流程框图。假设共有 N 个说话人,则第 k 个说话人模型的训练流程如图 1所示,训练语音经预处理后,提取特征参数,用 EM算法训练 GMM模型,用此 GMM 模型的均值作为“+1”的训练样本,其余 $N-1$ 个 GMM 模型的均值作为“-1”的训练样本,训练 SVM 模型。识别时,测试语音经预处理,提取特征参数,而后分别输入到每一个 GMM-SVM 混合模型,得到概率得分,经判决后输出识别结果,识别流程如图 2所示。

2 实验结果及分析

首先对代表性训练样本的选取方法进行了对比实验。实验采用了 TIMIT数据库,从库中随机选取了 34人,每名说话人的训练语音长度为 8至 12秒,测试语音长度为 2秒,对每名说话人进行 3次测试。

在预处理阶段,首先去除语音中的无声段,而后经过 $H(z)=1-0.96z^{-1}$ 高通滤波器进行预加重,使用汉明窗进行分帧处理,帧长 20 ms,帧移 10 ms。特征参数使用美尔频率倒谱系数(MFCC),美尔滤波器个数为

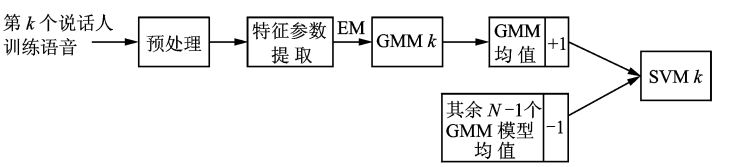

图 1 训练流程图

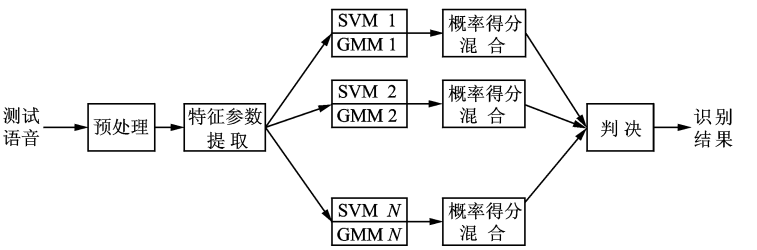

图 2 识别流程图

24,去除第 1阶分量,取第 2到第 17阶分量得到 16阶的 MFCC. 分别采用随机选取、LBG方法选取 64个 MFCC向量, GMM均值通过训练一个混合数为 64的 GMM模型得到. 然后用选取的特征向量作为训练样本输入支持向量机.

支持向量机采用径向基核函数, V设置为 0.005,惩罚系数 C设置为 1000. 输出得分采用正确分类的测试向量占总测试向量的百分比. 表 1列出了应用各种方法的辨认性能,以误识率作为衡量指标.

表 1　各种选取代表性样本方法的辨认性能

方法	误识率(%)
随机选取	22.06
LBG聚类	10.29
GMM均值	8.82

由实验结果可以看出在同样数量的输入样本情况下,采用 GMM均值向量作为 SVM训练样本得到了最好的识别性能. 这主要是因为 GMM模型较好的去除了语音中的冗余信息,如语义信息,而保留了说话人的个性信息,更好的描述了不同说话人在特征空间的分布. 所以本文采用 EM算法处理大量的语音特征向量,以 GMM均值作为支持向量机的输入,较好的解决了支持向量机的大量训练样本问题.

而后,对 GMM模型、采用概率输出的 SVM模型以及本文提出的 GMM-SVM模型的说话人辨认性能进行了实验对比,实验条件同前面一样.

提取 16维的 MFCC特征向量,利用 EM算法训练混合数为 64的 GMM模型, GMM模型的均值向量作为支持向量机的输入. 实验结果如表 2所示.

表 2　各种模型的辨认性能比较

模型	误识率(%)
GMM	9.81
SVM概率输出	5.88
GMM-SVM	4.90

从表 2中看出,本文提出的 GMM-SVM模型性能优于 GMM模型和 SVM模型. 这是因为 GMM的输出对 SVM进行了调整,使得 GMM的分类结果可以反应到 SVM的输出中. GMM-SVM兼顾了各说话人模型之间和模型内部的信息,也就是说 GMM-SVM混合模型既保留了 SVM判决能力强的优点,又体现了 GMM表征数据统计特性强的优势,所以取得了比 GMM和 SVM分立模型更低的误识率.

3　结　论

支持向量机最为一种新的模式识别方法具有很好的分类性能,但其不能反映训练数据本身的特性,且训练算法复杂,难于处理大量样本数据,限制了其在说话人识别领域的应用. 本文提出了一种 GMM-SVM混合说话人辨认模型,在特征参数域,采用高斯混合模型的统计参数训练支持向量机,解决了支持向量机难以处理大量训练样本的问题;并且在概率得分域对 GMM和 SVM的输出进行了融合,有效提高了说话人辨认系统的性能. 下一步研究的重点是进一步降低 GMM-SVM混合模型的复杂度,实现基于 DSP芯片的独立模块系统.

参考文献:

[1] Fine S, Navratil J, Gopinath R. A hybrid GMM/SVM approach to speaker identification [C]// ICASSP'01. US: IEEE Press, 2001: 417-420.

[2] Vapnik V N. An overview of statistical learning theory [J]. IEEE Trans on NN, 1999, 10(5): 988-999.

[3] 苟　博,黄贤武. 支持向量机多类分类方法 [J]. 数据采集与处理, 2006, 21(3): 332-339.

[4] Wan V, Renals S. Speaker verification using sequence discriminant support vector machines [J]. IEEE Trans on Speech and Audio Processing, 2005, 13(2): 203-210.

[5] Reynolds D A, Rose R C. Robust text-independent speaker identification using Gaussian mixture speaker models [J]. IEEE Transactions on Speech and Audio Processing, 1995, 3(1): 72-82.