基于SVM的非特定人声调识别的研究

肖汉光[1],蔡从中[2]
XIAO Han-guang[1],CAI Cong-zhong[2]

1.重庆工学院 数理学院,重庆 400054
2.重庆大学 数理学院,重庆 400044
1.School of Mathematics and Physics,Chongqing Institute of Technology,Chongqing 400054,China
2.School of Mathematics and Physics,Chongqing University,Chongqing 400044,China
E-mail:simenxiao1211@163.com

XIAO Han-guang,CAI Cong-zhong.Study of speaker-independent tone recognition based on support vector machine. Computer Engineering and Applications,2009,45(9):174-176.

Abstract: A speaker-independent tone database of Chinese speech(putonghua) is established.The Mel-frequency cepstrum coefficients(MFCCs) are used for extraction of the tone feature parameters.The four recognizing models of four tones are trained by using support vector machine(SVM),and are tested by using the testing tone data.The results show that a recognition accuracy can reach 97.6% by combining MFCCs and SVM.

Key words: tone recognition;feature extraction;Mel-Frequency Cepstrum Coefficients(MFCCs);Support Vector Machine(SVM)

摘 要:在建立非特定人普通话四声语调语音数据库的基础上,采用Mel频率倒谱系数(MFCCs)对语音数据进行特征参数的提取,并利用支持向量机(SVM)对语音中的四种声调进行了训练和识别研究。实验结果表明MFCCs和SVM的结合得到的平均识别率达到了97.6%。

关键词:声调识别;特征提取;Mel频率倒谱系数(MFCC);支持向量机
DOI:10.3778/j.issn.1002-8331.2009.09.050 **文章编号**:1002-8331(2009)09-0174-03 **文献标识码**:A **中图分类号**:TP391.4

1 前言

声调是汉语语音的重要特征。对于同一音节,由于声调不同,其含意有很大的不同。所以声调的识别在汉语语音识别中占有重要地位,其在汉语语音合成、汉语方言辨识中得到了广泛的应用[1-3]。

普通话单音节声调的识别中最常用的方法是:首先提取语音数据的基音频率参数,然后在实验观察的基础上定义一定的规则,当基音频率轨迹的某一参数超过规则中预先设定好的某一阈值时,则判定为某一声调[2-3]。此方法有两方面的缺点:一方面,由于不同说话人的基音有很大的差别,仅提取基音作为特征参数是不够的,特别是当识别的非特定人数很大时,识别率会有明显的下降;另一方面,特定的规则需要预先设定,不能达到模型的自动建立。为达到模型的自动建立,神经网络被逐渐应用于语调识别[4-5]。

MFCC是一种符合人耳听觉特性的参数,能很好地表达声调等语音信息,且在有信道噪声和频谱失真情况时表现稳健。支持向量机(Support Vector Machine,SVM)建立在统计学习理论的VC维(Vapnik Chervonenks Dimension)理论和结构风险最小原理(Structural Risk Minimization)基础上,根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷,以期获得最好的推广能力[6]。支持向量机能较好地解决小样本、非线性、高维数和局部极小点等实际问题,在很大程度上解决了模型选择与过学习问题、非线性和维数灾难问题以及局部极小点等问题,因此目前已成为机器学习界研究的热点,并应用于诸多领域[7-10]。

笔者提出利用MFCC提取语音声调特征参数,并采用一对多的分类策略对支持向量机进行训练学习并自动建立识别模型。

2 SVM的分类原理

以两类(正样本和负样本)分类问题为例,在线性可分的情

基金项目:国家教育部新世纪人才支持计划(the New Century Excellent Talent Foundation from MOE of China under Grant No.NCET-07-0903),重庆市自然科学基金(the Natural Science Foundation of Chongqing city of China under Grant No.CSTC,2006BB5240);重庆工学院青年教师科研基金(the Young Teacher Scientific Research Foundation of Chongqing Institute of Technology under Grant No.20062D39)。

作者简介:肖汉光(1980-),男,硕士,新加坡国立大学访问学者,主要研究方向:机器学习,模式识别等;蔡从中(1966-),男,博士,研究员,博士生导师,主要研究方向:人工智能和机器学习,计算物理学,计算生物信息学等。

收稿日期:2008-01-11 **修回日期**:2008-04-09

况下,SVM 构建一个超平面 H:

$$\boldsymbol{W}\cdot\boldsymbol{P}+b=0 \tag{1}$$

式中,$\boldsymbol{W}$ 为权重向量,$\boldsymbol{P}$ 为特征向量,b 为一参数。该超平面以最大边界的形式将正负样本区分开。该超平面的构建是通过寻找向量 $\boldsymbol{W}$ 和参数 b,使其在满足条件

$$\boldsymbol{W}\cdot\boldsymbol{P}_i+b\geqslant 0,(\text{对正样本},y=+1) \tag{2}$$

$$\boldsymbol{W}\cdot\boldsymbol{P}_i+b<0,(\text{对负样本},y=-1) \tag{3}$$

时,$\|\boldsymbol{W}\|^2$ 达到最小。式中 $\boldsymbol{P}_i$ 代表第 i 个训练样本的特征向量,$\|\boldsymbol{W}\|^2$ 代表权重向量 $\boldsymbol{W}$ 的欧几里德范数,y 为样本类别标记。在求出 $\boldsymbol{W}$ 和 b 后,通过决策函数

$$y_i=sign[\boldsymbol{W}\cdot\boldsymbol{P}_i+b] \tag{4}$$

判断向量 $\boldsymbol{P}_i$ 所对应测试样本的类别。若决策函数值为+1,该样本属于正样本;否则,属于负样本。

在线性不可分的情况下,SVM 利用核函数 $K(\boldsymbol{P}_i,\boldsymbol{P}_j)$ 将特征向量映射到一个高维空间。在此高维空间中,线性不可分问题被转化为线性可分问题,其决策函数为:

$$y_j=sign[\sum_{i=1}^{l}\alpha_i y_i K(\boldsymbol{P}_i,\boldsymbol{P}_j)+b] \tag{5}$$

上式中,l 为训练样本数,系数 α_i 和 b 应使拉格朗日表达式:

$$\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}a_i a_j y_i y_j K(\boldsymbol{P}_i,\boldsymbol{P}_j) \tag{6}$$

达到最大值,且应满足:

$$C>a_i\geqslant 0 \text{ 和 } \sum_{i=1}^{l}a_i y_i=0 \tag{7}$$

其中,C 为错误惩罚参数,它控制对错误分类样本的惩罚程度,C 越大支持向量的个数越多,最优超平面越复杂。

核函数 $K(\boldsymbol{P}_i,\boldsymbol{P}_j)$ 一般取径向基函数:

$$K(\boldsymbol{P}_i,\boldsymbol{P}_j)=e^{-\|\boldsymbol{p}_i-\boldsymbol{p}_j\|^2/(2\sigma^2)} \tag{8}$$

一般训练过程中需要对径向基函数中的参数 $g=-\frac{1}{2\sigma^2}$ 进行优化,大多采用的方法为网格搜索法。

3 实验及分析

声调识别本质上是一种模式识别的过程,主要包括信号的采集、预处理、特征提取、模型训练、参数优化和声调识别等过程。

3.1 语音数据库的建立

采用专业录音笔采集语音数据。采集对象为男女生各 7 名。采集内容为 120 个汉字,即四种声调各 30 个字,14 名采集对象对 120 个汉字各读三遍,录音笔的采样频率为 44 KHz,精度为 8 bit。在预处理前将录好的语音文件按四种声调分别放进 4 个文件夹。将所有语音文件从 *.mp3 格式转换为 *.wav 格式,然后进行预处理。

3.2 预处理

由于录制的语音数据中存在非语音段,所以需要将非语音段剔除。一般采用的方法为平均能量和平均过零率检测。平均能量和平均过零率是最基本的时域特征,其定义分别是:

平均能量:

$$RMS=\frac{1}{L}\sum_{n=1}^{L}s^2(n) \tag{9}$$

平均过零率:

$$ZCS=\frac{1}{2L-1}\sum_{n=1}^{L-1}|\text{sgn}(s(n+1))-\text{sgn}(s(n))| \tag{10}$$

其中 L 为一段语音的采样点数,n 为 L 个采样点中的任意一点。根据不同的采样频率可以选择不同大小的 L,一般选择为加窗处理中的窗口大小(一般为 256 或 512,本文选择 512)。

在处理过程中,首先计算待处理语音的总平均能量和总平均过零率,然后用大小为 L 的窗口采用重叠式连续截取语音,重叠大小取窗口的一半或四分之一,并计算各段语音数据的平均能量和平均过零率,最后设定平均能量和平均过零率的门限值,将低于或高于该门限的语音段判断为非语音段,并予以剔除。在分析过程中,得到如图 1 所示的平均能量和平均过零率的数据。图 1 中,(a)为待处理的一段语音,(b)为各窗口的平均能量,(c)为各窗口的过零率,通过比较(a)和(b)图,发现平均能量较大时对应语音段,平均能量较小时对应非语音段,并且对应关系较为良好。而(c)和(a)图的对应关系不明显,这可能是背景噪音的不同引起的。所以只采用平均能量进行非语音段剔除,门限值设为待处理语音平均能量的十分之一,如果截取语音段的平均能量大于该值,即判断为语音段。图 1(d)为非语音段剔除后的语音序列。通过比较图(a)和(d),可以看出新的语音序列保存了原语音序列的语音段。

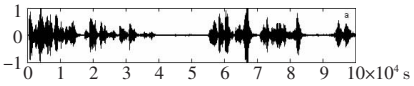

图 1(a)原始语音信号

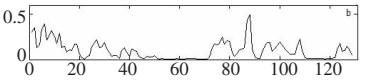

图 1(b)能量检测数据

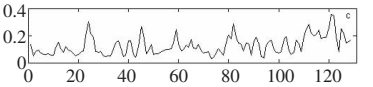

图 1(c)过零率检测数据

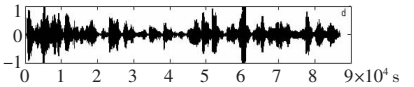

图 1(d)检测和剪切后的语音数据

语音信号是一种非平稳信号,但由于语音的形成过程与发音器官运动密切相关,这种物理运动比起声音的运动速度来要缓慢得多。因此语音信号可以假定为短时平稳的,则可对其分帧提取短时性。本文采用帧长约为 45 ms,帧移约为 30 ms,即重叠式截取短时语音序列,如图 2 所示。

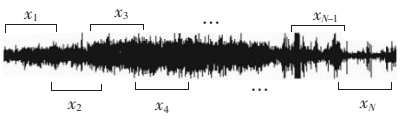

图 2 重叠式连续截取短时时间序列$\{x_1,x_2,\cdots,x_N\}$

为了避免进行 FFT 产生吉布斯效应,必须对短时语音序列$\{x_1,x_2,\cdots,x_N\}$进行平滑过滤。平滑过滤器一般选择汉明窗口(Hamming window)。其表达式为:

$$W_i=0.54-0.46\cos(2\pi\frac{i}{L}),i=0,1,\cdots,L-1 \quad (11)$$

$$\boldsymbol{X}_{ji}=\boldsymbol{X}_{ji}W_i,j=1,2,\cdots,N;i=0,1,\cdots,L-1 \quad (12)$$

图 3 为最大幅值为 1、窗口长度为 512 个采样点的汉明窗口的示意图。

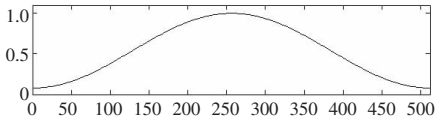

图 3 最大幅值为 1,窗口长度为 512 个采样点的汉明窗口

3.3 MFCC 的特征提取

设 $\boldsymbol{x}_j$ 代表某一语音信号经预处理后的其中一帧,对 $\boldsymbol{x}_j$ 进行快速傅立叶变换 $\boldsymbol{f}_j$,获取频谱系数并求出能谱系数,然后利用梅尔频谱特征标度的三角形滤波器组进行滤波处理。三角形滤波器组如图 4 所示。每个三角形可以作为一个中心频率和一个上下截止频率的带通滤波器。中心频率为人耳对某频段的感知中心,上下截止频率为人耳在该频段的感知范围。滤波器在低频段的个数比较均匀,随着频率的增加,滤波器的个数呈指数衰减。滤波器的形状可供选择,如三角形、汉明形和汉宁形。但使用最多的是三角形。滤波器的个数一般选择为 24。

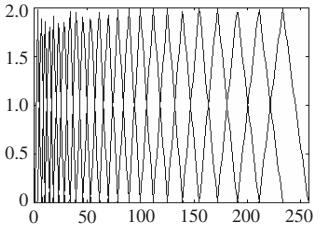

图 4 三角形滤波器组

利用滤波器组对频谱系数 $\boldsymbol{f}_j$ 进行加和滤波。即在同一三角形下,频谱与对应的三角形幅值相乘然后求和,得到 24 个系数。对这 24 个系数 $\boldsymbol{m}_j$ 取对数,并作 DCT 变换,得到 MFCC 系数,即:

$$C_i=\sqrt{\frac{2}{k}}\sum_{j=1}^{k}\ln(m_j)\cos|\pi\times i/k(j-0.5)| \quad (13)$$

其中 k 是三角滤波器个数,m_j 为第 j 个滤波器的输出,C_i 为 MFCC 的第 i 个分量,其中,$i=1,2,3,\cdots,k$。

3.4 SVM 的训练及识别

利用 MFCC 特征提取方法提取前 30 秒钟语音数据构成训练向量集,剩余的语音数据构成测试向量集。将声调的四种模式作为支持向量机的分类对象,采用一对多的训练策略对SVM 进行训练,即将一种声调的特征向量当作正样本,其他三种声调的特征向量当作负样本, 训练 SVM 构成一个 SVM 预测模型。依此类推得到 4 个 SVM 预测模型。四种语调共建立 4 个 SVM 预测模型,即 SVM1、SVM2、SVM3 和 SVM4。在训练过程中,各预测模型的训练准确率均达到了 100%。

测试时,将待测语音作为正样本,提取特征向量后输入到 4 个 SVM 预测模型中,得到 4 个准确率,准确率最高的 SVM预测模型的正样本声调即为待识别语音的声调。

测试准确率公式为:$Q_{svmi}=TP/(TP+FN)$ (14)

其中:TP(True Positive)代表在测试集中被正确地判断为正样本的个数;FN(False Negative)代表在测试集中被错判为负样本的样本个数。

SVM 总的准确率计算公式为:

$$Q_{all}=\frac{1}{4}\sum_{i=1}^{4}Q_{svmi} \quad (15)$$

其中 Q_{svmi} 代表声调 i 输入到 SVM_i 模型中的准确率。4 个 SVM 预测模型对不同声调的测试准确率如表 1 所示。

表 1 不同 SVM 模型对不同声调的识别率(%)

	Q_{svm1}	Q_{svm2}	Q_{svm3}	Q_{svm4}	Q_{all}
声调 1	**97.3**	7.3	12.7	9.2	97.6
声调 2	2.5	**97.1**	10.6	12.1	
声调 3	5.3	6.8	**97.7**	7.1	
声调 4	4.9	7.3	5.7	**98.2**	

从表 1 中可知:SVM 能有效地对四种声调进行有效的识别,总的识别率 Q_{all} 为 97.6%。

4 结论

本文建立了汉语声调数据库,并利用 MFCC 进行了声调特征向量的提取。利用一对多分类方法对 SVM 进行训练,得到了 4 个声调的预测模型。利用预测模型对待测声调进行识别,识别总准确率达到了 97.6%。该结果表明:利用 SVM 结合 MFCC 特征提取技术,能够对声调进行有效识别。

参考文献:

[1] 赵鹤鸣,周旭东,金延庆.基于小波变换的重叠语音基频提取及声调识别[J].声学学报,1999,24(1):87-93.

[2] 关存太,陈永彬.非特定人四声识别[J].声学学报,1993,18(5):379-385.

[3] 朱小燕,王昱,刘俊.汉语声调识别中的基音平滑新方法[J].计算机学报,2001,24(2):213-218.

[4] 汤霖,尹俊勋,粟志昂,等.基于两级 BP 模型的普通话声调识别系统[J].计算机工程与应用,2004,40(25):96-99.

[5] 孙放,胡光锐.一种新型前向神经网络用于汉语四声识别[J].上海交通大学学报,1997,31(5):36-38.

[6] Vapnik V.The nature of statistical learning theory[M].New York:Springer,1995.

[7] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000,26(1):32-42.

[8] 朱永生,张优云.支持向量机分类器中几个问题的研究[J].计算机工程与应用,2003,39(13):36-38.

[9] 肖汉光,蔡从中,廖克俊.利用声波和地震波识别军事车辆类型[J].系统工程理论与实践,2006,26(4):108-113.

[10] Cai C Z,Han L Y,Ji Z L,et al.SVM-Prot:Web-based support vector machine software for functional classification of a protein from its primary sequence[J].Nucleic Acids Research,2003,31(13):3692-3697.