支持矢量机的汉语声调识别

顾明亮[1,2], 夏玉果[2], 杨亦鸣[2]
(1. 徐州师范大学物理与电子工程学院智能信息处理研究室, 江苏徐州 221116;
2. 江苏省语言科学与神经认知工程重点实验室, 江苏徐州 221116)

摘要: 提出了一种支持矢量机的汉语声调识别新方法。论文首先在基频和对数能量的基础上,建立了一个适合于支持矢量机分类的等维声调特征。然后对支持矢量机的多分类策略和不同核函数对声调识别的影响进行了实验研究。与 BP 神经网络相比,支持矢量机具有更高的识别率和更强的推广能力。

关键词: 基音检测;支持矢量机;声调识别
中图分类号: TP391.4 **文献标识码**: A **文章编号**: 1000-3630(2007)-06-1186-05

Support vector machine based Chinese tone recognition

GU Ming-liang[1,2], XIA Yu-guo[2], YANG Yi-ming[2]
(1. Physics Department Xuzhou Normal University Engineering, Xuzhou 221116, Jiangsu China;
2. Jiangsu Key Laboratory of Language Science and Neural Cognition, Xuzhou 221116, Jiangsu China)

Abstract: This paper presents a novel support vector machine based Chinese tone recognition method. A new tone recognition feature is first extracted using the fundamental frequency (F0) and logarithmic energy. And how to select the method of SVM multi-class classification and kernel function is also discussed by experiments. Compared with BP neural network, SVM has higher recognition rates and more strong generalization.

Key words: pitch detection; support vector machine; chinese tone recognition

1 引 言

汉语是一种典型的有调语言, 声调在汉语语音识别、语音合成等言语信息处理中具有十分重要的应用价值。早在 1000 多年前的隋朝,古代先贤就凭借自身的感知器官对汉语声调进行了大量的定性研究[1]。用科学仪器进行声调的定量研究则是 20 世纪初赵元任先生、刘复先生等留学海外的学者回国后才真正开始。其中,赵先生独创的五度声调表示法已得到广泛认可, 有关基频包络与汉语四声的对应关系更是现代汉语语音工程研究的重要基础。六十年代,国内外学者开始讨论语音信号的基频提取问题。七十年代,此项研究达到高潮,人们提出了很多至今仍在使用的经典算法[2]。八十年代以后,汉语等有调语言国家的学者开始将基频提取算法应用到声调识别研究中,并提出了包括隐马尔可夫模型[3]和神经网络等在内的许多声调识别方法[4]。归纳起来,声调识别主要由四个部分组成:(1) 语音信号预处理,(2) 基音频率的提取,(3) 基音频率曲线的平滑和声调特征的提取,(4) 分类器设计。

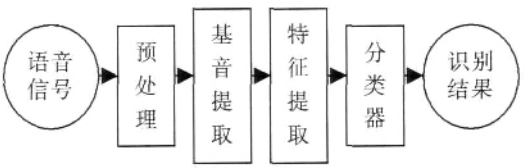

图 1 汉语声调识别框图
Fig.1 Chinese tone recognition framework map

收稿时期: 2006-11-03, 修回时期: 2007-02-01
基金项目: 江苏省 "十五"社科基金项目(K3-013), 江苏省高校自然科学基金(99KJB510002)。
作者简介: 顾明亮(1963-), 男, 江苏人, 博士, 副教授, 研究方向: 语音信号处理、模式识别、机器学习等。
通信作者: 夏玉果, E-mail:xyg810824@163.com

随着人们对声调本质属性认识的不断化，以及现代模式识别新方法的不断推出，一些好的基音提取方法、好的声调特征和好的模式分类方法仍在不断涌现。其中，支持矢量机就是最近几年出现的一种优秀的模式分类器[5]，它在解决非线性分类问题，尤其对解决小样本情况下的非线性分类问题表现出独特的优势，目前该方法已在邮政手写体数字识别、文本分类、人脸识别、基因检测等领域得到广泛应用[6]。由于语音信号是典型的动态模式，每个音节包含的特征矢量的总维数是可变的，而支持矢量机是典型的静态分类器，它要求模式的特征维数必须一致。为此，本文首先提出了一种新的齐次化声调特征，然后结合四声分类的特点，对支持矢量机多分类策略和核函数进行了实验研究，找到适合四声分类的最佳模式。

本文第二部分主要说明如何解决声调特征的选择、特征齐次化问题，第三部分则介绍支持矢量机多分类策略和核函数的选择问题，第四部分给出相关的实验及结果，最后进行了简要的总结。

2　齐次化声调识别特征的提取

2.1　特征参数的计算

特征是影响分类效果的重要因素，基音频率无疑是公认的声调识别特征参数，此外，基频差分、能量及能量差分也是声调识别的重要参数。其中，基音频率的计算最为关键。目前的计算方法大致可以分成三类：(1) 时域法：如自相关方法、平均幅度差法；(2) 频域法：如倒谱提取法；(3) 时频结合法：如简单逆滤波法、小波变换法。有关基频提取算法的综述可以参考文献[2]。由于基音不是一个完全周期信号，它既要受声门开闭变化的影响、又要受讲话者、噪声等因素的影响，因此，要想正确提取语音信号的基音频率仍然不是一件容易的事情。考虑到基音周期计算应尽可能简单并具有一定的容错性，这里采用加权自相关法来计算各帧的基音频率[7]。设分帧后的语音信号为 $x(n)$，帧长为 M，延迟点数为 τ。语音信号的自相关函数定义为：

$$\phi(\tau)=\sum_{n=0}^{M-1}x(n)x(n+\tau) \tag{1}$$

平均幅度差函数定义为：

$$\phi(\tau)=\frac{1}{M}\sum_{n=0}^{M-1}|x(n)-x(n+\tau)| \tag{2}$$

则改进的加权相关函数定义为：

$$\mu(\tau)=\phi(\tau)/(\varphi(\tau)+k) \tag{3}$$

其中，k 是大于零的某个常数。利用该加权函数，可以较好地得到浊音语音信号各帧的基音频率。再通过中值平滑处理基频轨迹上的随机错误点，即可得到一组基频特征参数：$F_0=\{f_0^1, f_0^2, \cdots, f_0^N\}$，N 为总的语音帧数，其对应的差分基频参数记为：$\Delta F_0=\{\Delta f_0^1, \Delta f_0^2, \cdots, \Delta f_0^{N-1}\}$。各帧的能量参数记为：$E=\{e_1, e_2, \cdots, e_N\}$，其中，$e_i$ 可用公式(4)得到：

$$e_i=\sum_{n=0}^{M-1}|x(n)|^2 \tag{4}$$

考虑人耳的听觉感知特点，取对数值作为特征参数，即：$\lg E=\{\lg(e_1), \lg(e_2), \cdots, \lg(e_N)\}$，其对应的差分能量记为：$\Delta\lg E=\{\Delta\lg(e_1), \Delta\lg(e_2), \cdots, \Delta\lg(e_{N-1})\}$。由此可见，每一帧的特征矢量可选择四个参数组成：$\{f_0^i, \Delta f_0^i, \lg(e_i), \Delta\lg(e_i)\}$，其中 $1\le i\le N-1$，将各帧特征矢量参数组合起来即可得到一个描述该语音的声调特征参数矩阵 X，见式(5)。将矩阵中的某几个列矢量串接起来即可得到不同维数的音节特征矢量。

$$X=\begin{bmatrix} f_0^1 & \Delta f_0^1 & \lg(e_1) & \Delta\lg(e_1) \\ f_0^2 & \Delta f_0^2 & \lg(e_2) & \Delta\lg(e_2) \\ \vdots & \vdots & \vdots & \vdots \\ f_0^{N-1} & \Delta f_0^{N-1} & \lg(e_{N-1}) & \Delta\lg(e_{N-1}) \end{bmatrix} \tag{5}$$

2.2　特征的齐次化

由于音节发音长度不可能相同，则每个音节包含的帧数 N 也不相同，因此，音节特征矢量的维数必然不同，而支持矢量机只能对相同特征维数的模式矢量进行分类，因此，必须将不同维数的特征矢量转化为具有相同特征维数的新的特征矢量。此项工作称为特征的齐次化，这里采用两种齐次化策略。

(1) 插值法，即将矩阵 X 中的每一列特征参数值看成是函数 $f(x_k)$ 在自变量 $(x_1, x_2, \cdots, x_k, \cdots, x_{N-1})$ 各点的函数值。将自变量取值区间进行 P+1 等分，然后计算中间 P 个等分点上的函数值作为新的特征参数，中间各等分点上的函数值可按公式(6)计算得到。

$$c_j^i=\frac{z_j-x_{k+1}}{x_k-x_{k+1}}f(x_k)+\frac{z_j-x_k}{x_{k+1}-x_k}f(x_{k+1}) \tag{6}$$

这里，c_j^i 表示矩阵(5)中第 i 列中的第 j 个元素变换后的结果。$z_j=x_1+j\cdot\frac{x_{N-1}-x_1}{P+1}$，$z_j\in[x_k, x_{k+1}]$，$i=1, 2, \cdots, 4$；$j=1, 2, \cdots, P$；$k=1, 2, \cdots, N-1$。上述参数最终形成一个新的特征参数矩阵 C，见式(7)，将各列串起来即可得到维数为 4*P 维的音节特征矢量。

$$C=\begin{bmatrix} c_1^1 & c_1^2 & c_1^3 & c_1^4 \\ c_2^1 & c_2^2 & c_2^3 & c_2^4 \\ \vdots & \vdots & \vdots & \vdots \\ c_P^1 & c_P^2 & c_P^3 & c_P^4 \end{bmatrix} \tag{7}$$

(2) 曲线拟合法，即矩阵 X 中每一列的特征参数看成是某一拟合曲线上的采样点，然后利用这些采样点求出该曲线的拟合系数，将这些拟合系数组合在一起即可得到一个等维的特征矢量。这里采用以勒让德多项式为基函数的最小二乘法进行拟合。如公式(8)所示

$$g_0^i(x)=\sum_{j=1}^{M-1} c_j^i e_j(x) \tag{8}$$

其中 $g_0^i(x)$ 为矩阵 X 中第 i 列语音特征参数构成的曲线，$e_j(x)$ 为第 j 阶勒让德多项式。c_j^i 为曲线拟合系数，M 为勒让德多项式的阶数，这里取 M=4，这样就可以用 4 个曲线拟合系数来表示每个语音特征参数。由于勒让德多项式在[-1,1]上带权 $\rho(x)=1$ 的正交多项式，所以需要将语音特征参数归一化处理。归一化公式见(9)。

$$\tilde{x}=2.0\times\frac{x-x_{min}}{x_{max}-x_{min}}-1.0 \tag{9}$$

这样拟合系数 c_j^i 由公式(10)直接求得。有关此项工作的具体方法可参考文献[8]。

$$c_j^i=\frac{\langle e_j(x), g_0^i(x)\rangle}{\langle e_j(x), e_j(x)\rangle} \tag{10}$$

这时齐次化的语音声调特征参数可用矩阵 C1 来表示，见式(11)，将各列串起来即可得到维数为 4*M 维的特征矢量。

$$C1=\begin{bmatrix} c_0^1 & c_0^2 & c_0^3 & c_0^4 \\ c_1^1 & c_1^2 & c_1^3 & c_1^4 \\ \vdots & \vdots & \vdots & \vdots \\ c_M^1 & c_M^2 & c_M^3 & c_M^4 \end{bmatrix} \tag{11}$$

3 支持矢量机分类器设计

支持矢量机(Support Vector Machines，简称 SVM)是建立在统计学习理论基础上的一种新型模式分类器。它的主要特点是：(1)利用那些对分类有较好区分能力的支持向量，构造分类空隙最大的有较好推广性能和较高分类准确率的最优超平面，图 2 中的黑点表示支持向量；(2)通过核函数变换的方法，将低维空间非线性分类问题转换为高维空间线性可分的问题，同时很好地解决了高维空间中的计算复杂性问题；(3) SVM 分类器不存在过学习问题。在二类模式分类问题中，支持矢量机的最优超平面可用图 2 表示，其结构如图 3 所示，其中，最优超平面的分类函数可以表示为：

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{s} a_i y_i K(x_i, x)+b\right) \tag{12}$$

$K(x_i, x)$ 表示内积函数。选择不同形式的内积函数，支持矢量机就会形成不同的算法，常用的内积函数主要有以下三种：多项式核函数，径向基核函数和 sigmoid 型核函数。本文选择了前两种形式的内积函数。

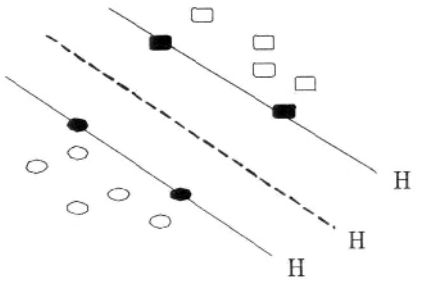

图 2 最优超平面示意图

Fig.2 The optimal separating hyperplane map

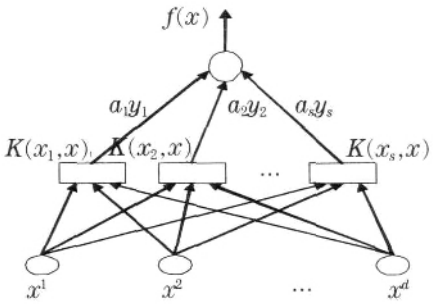

图 3 支持矢量机示意图

Fig.3 Support vector machine map

对于多类模式的分类问题，目前主要有“一对一”、“一对其余”和决策导向无环图三种多类分类策略。考虑到“一对一”分类方法精度高、容错性好，而待区分的类别数(四类)较少的实际情况，本文采用“一对一”多类分类支持矢量机。它的原理是：首先利用已知的 k 类训练样本训练 k(k-1)/2 个二类支持矢量机，然后，用这些二类支持矢量机对测试样本进行分类，最后将判入次数最多的类别作为测试样本的分类结果。

4 实验与分析

为了检验特征的可行性和支持矢量机的分类效果，经与本校语言研究所有关专家合作，首先建立了一个专门进行声调训练和测试的语音数据库，然后按照图 1 所示的流程进行了系列实验研究。图 1中

的语音信号预处理主要包括滤波、分帧、加窗、清浊音区分等多个环节。实验研究主要包括特征的齐次化方式、特征的组合方式、核函数对 SVM 分类效果的影响和不同分类器对声调识别效果的影响等。

4.1　语音数据库的建立

实验所用的发音语料选自《普通话水平测试实用手册》中的 100 个单音节字，这些单字包括单元音和复元音，每个音节大部分都有四声读音。录音由声卡和 Cool Edit 软件共同完成，其中语音的采样频率为 11025Hz、量化级长为 16bit。发音人是 10 位在校大学生（男生和女生各 5 人），年龄介于 18~23 岁，都来自于徐州市区。每个发音人每个单字读 5遍共得到 5 000 个单字的读音，发音不准时要求重录。将其中 3 男 3 女的前三次发音共计 1 800 个的读音作为训练集（记作 TRAIN），将他们后两次发音共计 1,200 个读音组成测试集 1（记作 TEST1），将另外 2 男 2 女的所有发音共计 2,000 个读音组成测试集 2（记作 TEST2）。

4.2　特征齐次化方式对声调识别的影响

不同的特征齐次化方式对声调识别的影响是不相同的。表 1 给出了两种不同齐次化方式下的声调识别结果。该结果是采用混合特征参数：基频及其差分、对数能量及其差分的情况下得到的。有此可见，曲线拟合法比插值法识别效果要更好一点。原因是插值法拟合曲线必需经过节点，而节点值并非精确值，它存在一定的误差。而曲线拟合时只考虑所有节点处的总体误差小，因此，某个局部的误差不会影响整体拟合的效果。

表 1　两种特征齐次化处理的平均识别率

Table 1　The mean recognition rate of two uniformed features

齐次化方式	训练集 2	测试集 1	测试集 2
插值法	100%	98.2%	95.5%
曲线拟合法	99.9%	98.9%	96.6%

4.3　不同特征对声调识别的影响

表 2 给出了不同特征组合下，声调识别的结果表明：(1) F_0 的识别率明显高于 lgE 的识别率，说明基音频率对声调识别起主导作用；(2) 组合特征参数的识别率高于单一特征参数的识别率，说明特征组合可以提高声调识别的效果。

表中 $F_0=[f_0^1, f_0^2, \cdots, f_0^{N-1}]$，$\Delta F_0=[\Delta f_0^1, \Delta f_0^2, \cdots, \Delta f_0^{N-1}]$，$lgE=[lge_1, lge_2, \cdots, lge_{N-1}]$，$\Delta lgE=[\Delta lge_1, \Delta lge_2, \cdots, \Delta lge_{N-1}]$

4.4　不同核函数对声调识别实验

表 2　不同特征和特征组合对声调识别的影响

Table 2　The influence of different feature and assembled feature on tone recognition

特征组合	训练集	测试集 1	测试集 2
F_0	97.5%	95.0%	90.3%
$F_0, \Delta F_0$	99.5%	96.0%	91.1%
lgE	71.8%	69.6%	61.1%
lgE, ΔlgE	89.3%	75.8%	68.9%
F_0, lgE	97.4%	96.6%	94.5%
F_0, lgE, ΔlgE	98.1%	97.4%	95.5%
$F_0, \Delta F_0$, lgE	99.9%	99.3%	96.1%
$F_0, \Delta F_0$, lgE, ΔlgE	99.9%	98.9%	96.6%

对于不同的核函数，SVM 分类器的识别效果是不同的，这里对径向基函数和多项式核函数的声调识别情况进行了测试，结果表明两者都具有很高的声调识别率，但多项式核函数的推广能力比径向基函数要更强一些。

表 3　不同核函数对声调识别的影响

Table 3　The influence of different kernel function on tone recognition

核函数	训练集	测试集 1	测试集 2
径向基函数多	100%	98.2%	96.2%
项式核函数	100%	97.5%	95.4%

4.5　与 BP 神经网络的比较

神经网络也是应用非常广泛的一类非线性分类器，它是目前声调识别效果最好的少数几种方法之一，为此，这里将 SVM 与 BP 网络进行了对比实验，结果如图 4 所示。从实验结果可以看出：(1) SVM 分类器比 BP 神经网络分类器具有更高的识别率。(2) 从比较两者的训练集、测试集 1 和测试集 2 识别结果，可以发现 BP 网络的识别结果有明显下降，而SVM 分类器的测试结果下降较少，说明 SVM 的推广能力比 BP 网更强。

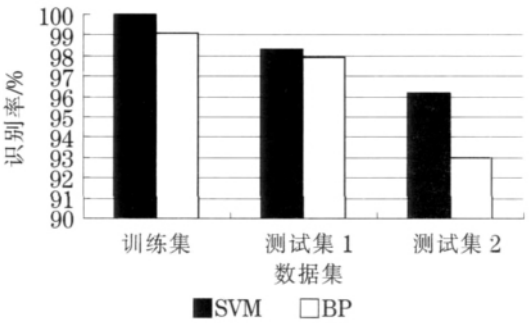

图 4　不同分类器对声调识别的影响

Fig.4　The influence of different classifier on tone recognition

5　结　论

本文讨论了支持矢量机在汉语声调识别中的应

用,实验结果表明,(1)基音频率是声调识别的主要特征,但增加能量及其差分参数可以进一步提高声调识别率;(2)曲线拟合得到的齐次声调特征是一种好的识别特征;(3)对于模式类别较少的声调识别来说,采用"一对一"的SVM多分类策略具有更好的识别效果;(4)与BP神经网络方法相比,SVM具有更高的识别率和更强的推广能力。

参 考 文 献

[1] 刘俐李. 二十世纪汉语声调理论的研究综述[J]. 当代语言学, 2004, 6(1): 45-56.
LIU Lili. Chinese tone theory research in the 20 century[J]. Contemporary linguistics, 2004, 6(1): 45-56.

[2] Rabiner L R, CHENG M J, Rosenberg A E. C A. McgonegalA comparative performance study of several pitch detection algorithm[J]. IEEE Trans. on ASSP, 1976, 24(5): 399-418.

[3] YANG W J, LEE J C, CHANG Y C, et al. Hidden markov model for mandarin lexical tone recognition[J]. IEEE Trans, on ASSP, 1988, 36(7): 988-992.

[4] CHEN S H, WANG Y R. Tone recognition of continuous mandarin speech based on neural network[J]. IEEE Trans. on SAP, 1995, 3(2): 146-150.

[5] Vapnik V N. The nature of statistical learning theory, New York: Springer Verlag, 1995.

[6] Burges C J C. A tutorial on support vector machines for pattern recognition[J]. Knowledge Discovery Data Mining, 1998, 2(2): 121-167.

[7] Shimamura T, Kobayashi H. Weighted autocorrelation for pitch extraction of noisy speech[J]. IEEE Trans. on SAP, 2001, 9(7): 727-730.

[8] 顾明亮, 王太君, 何振亚. 语音信号时间动态规正新方法[J]. 东南大学学报, 1998, 28(2): 10-14
GU Mingliang, WANG Taijun, HE Zhenya. Novel dynamic timealignment method for speech signals[J]. Journal of Southeast University, 1998, 28(2): 10-14.