支持向量机的汉语连续语音声调识别方法

傅德胜[1] 李仕强[2] 王水平[1]
(南京信息工程大学计算机与软件学院 南京 210044)[1]
(南京信息工程大学信息与控制学院 南京 210044)[2]

摘 要 声调信息在汉语语音识别中具有非常重要的意义。采用支持向量机对连续汉语连续语音进行声调识别实验,首先采用基于Teager能量算子和过零率的两级判别策略对连续语音进行浊音段提取,然后建立了适合于支持向量机分类模型的等维声调特征向量。使用6个二类SVM模型对非特定人汉语普通话的4种声调进行分类识别,与BP神经网络相比,支持向量机具有更高的识别率。

关键词 声调识别,基音频率,支持向量机
中图法分类号 TP391.4 **文献标识码** A

Tone Recognition Based on Support Vector Machine in Continuous Mandarin Chinese

FU De-sheng[1] LI Shi-qiang[2] WANG Shui-ping[1]
(Computer and Software Institute, Nanjing University of Information Science and Technology, Nanjing 210044, China)[1]
(School of Information and Control, Nanjing University of Information Science and Technology, Nanjing 210044, China)[2]

Abstract Tone is an essential component for word formation in Chinese languages. It plays a very important role in the transmission of information in speech communication. We looked at using support vector machines (SVMs) for automatic tone recognition in continuously spoken Mandarin. The voiced segments were detected based on Teager Energy Operation and ZCR. Compared with BP neural network, considerable improvement was achieved by adopting 6 binary-SVMs scheme in a speaker-independent Mandarin tone recognition system.

Keywords Tone recognition, Fundamental frequency, Support vector machine

1 引言

声母、韵母和声调是汉语语音音节的三要素,相同声母和韵母构成的音节会随着声调的不同而具有不同的读音和含义。汉语普通话中把声调分成4种:一声(阴平)、二声(阳平)、三声(上声)及四声(去声)。在语音信号处理的各个领域,如编码、识别及合成中,声调的识别是一项至关重要的任务。

近几十年来,许多有调语言国家的学者利用声调与基因频率之间存在的因果关系,提出了包括隐马尔科夫模型[1,2]、神经网络[3]、决策树[4]及支持向量机[5]等音调识别的方法。其中,支持向量机(SVM)是一种比较新颖的统计模式识别的方法,它在解决非线性小样本模式分类问题中具有极大的优势[5]。由于支持向量机属于静态分类器,它要求特征向量的维数必须一致,而语音信号是一种典型的动态模型,每个音节的特征向量的维数不一定相同,本文在提取出各音节的特征向量后,首先采用基于勒让德多项式的曲线拟合方法对特征进行齐次化,然后采用支持向量机及神经网络等方法分别进行声调识别实验,在支持向量机实验中重点对如何利用多个SVM二分类器实现多分类问题进行研究。

本文设计的连续语音声调识别系统由3部分组成:(1)清浊音的判断,(2)特征提取,(3)分类器设计。

2 浊音段提取

中文汉字有几万个,但其读音仅由22个声母及35个韵母组合而成,普通话声韵母组合有408个[6]。汉语声调主要体现在基音轮廓曲线上,基音频率都是随时间变化的,根据其不同的变化趋势形成了4种不同的声调。就浊声母和零声母汉字语音而言,基音轮廓曲线是跟字音同时开始的,而清声母汉字语音的声调是从韵母开始的。在普通话语音中,清声母可视为噪音,所以本文首先对语音信号进行清浊音的判别,再对浊音段进行下一步的处理。清音和浊音在平均能量频带分布和过零率这两个特征参数上差别较大,因此常被用来作为区分清浊音的依据。本文运用Teager能量算子(TEO)来计算每帧语音信号的能量,并提出一种基于TEO能量和过零率(ZCR)两个特征参数组合的两级判决算法。

(1)TEO能量

Teager能量算子是由Kaiser等人提出的一种非线性算子,它能有效地提取出信号的"能量"[7],并且已经成功地应用于语音信号处理中。TEO的计算公式为:

到稿日期:2009-07-03 返修日期:2009-09-25
傅德胜(1950-),男,教授,博士生导师,主要研究方向为图像处理与模式识别、信息安全,E-mail:dsfu@vip.sina.com;李仕强(1979-),男,硕士生,讲师,主要研究方向为信号处理、模式识别;王水平(1977-),女,博士生,讲师,主要研究方向为信息处理、模式识别。

$$\psi[x(n)]=[x(n)]^2-x(n+1)x(n-1) \quad (1)$$

为了克服噪声对能量计算带来的影响，首先对原始语音信号进行小波分解，图 1(b)为 db2 第 4 层分解系数 W，图 1(c)为对小波分解系数 W 用 TEO 算子计算并取模后的“能量”E。

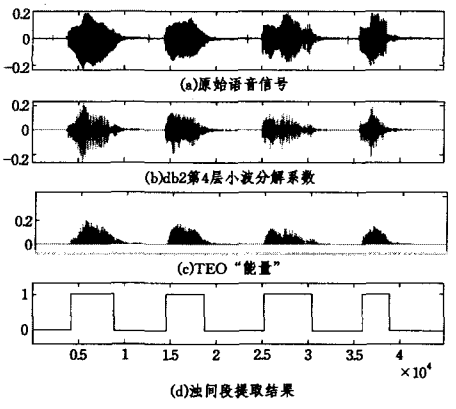

图 1　浊音段提取过程

(2)短时过零率

语音的过零率 ZCR 表示每帧内信号通过零值的次数，第 i 帧语音信号的过零率表示为

$$ZCR_i=\sum_{n=0}^{N-1}|\mathrm{sgn}[x_i(n)]-\mathrm{sgn}[x_i(n-1)]| \quad (2)$$

经研究发现，清音的过零率较高，浊音的过量率较低，可设定某一阈值，小于该阈值的帧视为浊音帧。但在实验中发现若使用固定阈值，算法的正确性不高，本文采用语音段中所有帧的过零率中值作为阈值，即

$$TH=\mathrm{median}(\{ZCR_m\}_{m=1}^{M}) \quad (3)$$

(3)两级判别算法

首先计算每帧语音信号的 E_i，若 $E_i\leqslant T_1$，则将该帧视为清音帧；若 $E_i\geqslant T_2$，则将该帧视为浊音帧；若 E_i 介于 T_1 和 T_2 之间，则根据短时过零率进行第二级判断，若 $ZCR_i\geqslant TH$，则该帧为清音帧，反之则为浊音帧。

(4)后处理

由于浊音段和清音段常常交叠在一起，上述算法有时会产生部分误差，需要做一些后处理，其遵循的原则为：短的浊音段不可能出现在连续的清音段中，反之亦然。图 1(d)为浊音段提取结果。

3　特征提取及归一化

汉语声调主要体现在基音轮廓曲线上，基音频率都是随时间变化的，根据其不同的变化趋势便形成了 4 种不同的声调，图 2 为声韵母组合“bao”在 4 种声调下的语音信号的语谱图。

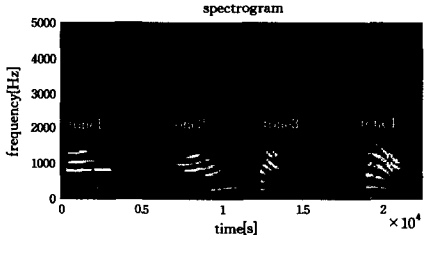

图 2　“bao”的 4 种声调语音语谱图

3.1　特征提取

基音频率无疑是公认的声调识别特征参数，此外，本文还选择基频差分、能量及能量差分等参数作为音频的特征向量。

平均幅度差函数可以用来计算音频的基因轨迹，与自相关函数法相比具有计算复杂度小的优点，因此在实时语音处理系统中经常被使用。短时平均幅度差函数(AMDF)定义为

$$AMDF(k)=\sum_{m=-\infty}^{\infty}|x(n+m)w(m)-x(n+m-k)w(m-k)| \quad (4)$$

从图 2 可以看出，基频的变化规律也是声调的重要特征之一，因此本文在提取基频特征参数 $F_0=\{f_0^1,f_0^2,f_0^3,\cdots,f_0^N\}$ 的基础上，进一步计算了差分基频参数

$$\Delta F_0=\{\Delta f_0^1,\Delta f_0^2,\Delta f_0^3,\cdots,\Delta f_0^{N-1}\} \quad (5)$$

另外，本文还计算了各帧的能量 $E=\{e_1,e_2,e_3,\cdots,e_N\}$，$e_i=\sum_{n=0}^{M-1}|x(n)|^2$ 及能量差分 $\Delta E=\{\Delta e_1,\Delta e_2,\Delta e_3,\cdots,\Delta e_{N-1}\}$。以帧为单位的特征向量由 4 个参数$\{f_0^i,\Delta f_0^i,e_i,\Delta e_i\}$组成，其中 $1\leqslant i\leqslant N-1$。以音节为单位的特征参数可以用如下的特征矩阵 T 表示

$$T=\begin{bmatrix} f_0^1 & \Delta f_0^1 & e_1 & \Delta e_1 \\ f_0^2 & \Delta f_0^2 & e_2 & \Delta e_2 \\ \vdots & \vdots & \vdots & \vdots \\ f_0^{N-1} & \Delta f_0^{N-1} & e_{N-1} & \Delta e_{N-1} \end{bmatrix} \quad (6)$$

3.2　归一化处理

由于每个音节的发音长度各不相同，从而其包含的帧数 N 也不相同，因此各音节的特征矩阵的维数必然不同。但是，支持向量机只能处理相同维数的模式分类问题，因此，必须对提取到的具有不同维数的各特征矩阵进行归一化处理，本文选用基于勒让德多项式系数的曲线拟合法，对不同维数的特征矩阵进行归一化。前 6 阶(n 从 0 到 5)勒让德多项式的曲线如图 3 所示。

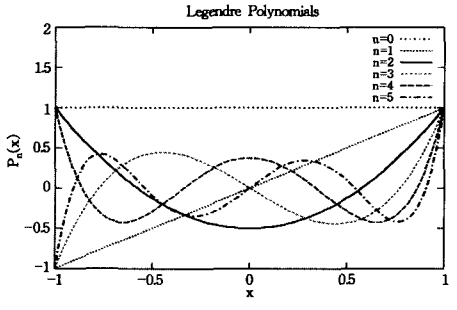

图 3　勒让德多项式前 6 阶曲线

设 $g_0^i(x)=\sum_{j=1}^{M}c_j^i e_j(x)$ 为特征矩阵 T 中第 i 列语音特征参数构成的曲线，其中，$e_j(x)$ 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，本文取 $M=4$。采用最小二乘法即可求出 4 个曲线拟合系数，将这 4 个曲线拟合系数进行组合，即可表示每个语音特征参数，如式(7)所示。

$$C=\begin{bmatrix} c_1^1 & c_1^2 & c_1^3 & c_1^4 \\ c_2^1 & c_2^2 & c_2^3 & c_2^4 \\ c_3^1 & c_3^2 & c_3^3 & c_3^4 \\ c_4^1 & c_4^2 & c_4^3 & c_4^4 \end{bmatrix} \quad (7)$$

4　分类器设计

支持向量机(SVM)是在统计学习理论的基础上发展起

来的新一代学习算法，该算法在文本分类、手写识别、图像分类、生物信息学等领域中获得了较好的应用。对于一个二分类问题，SVM 构建一个最优超平面 C_0，如图 4 所示，该超平面以最大边界的形式将正负样本区分开。在线性不可分的情况下，SVM 利用核函数 $K(P_i, P_j)$ 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。

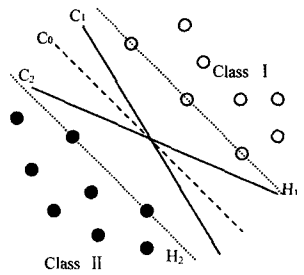

图 4　最优分类超平面示意图

本文将基本的二类 SVM 进行扩展，设计了 $6(C_4^2)$ 个如图 5 所示的二类支持向量机来解决 4 种声调的识别问题。其中 $K(x_i, x)$ 表示内积函数，常用的内积函数主要有以下 3 种：多项式函数、径向基函数和 sigmoid 型核函数，本文在实验环节选用了径向基函数。

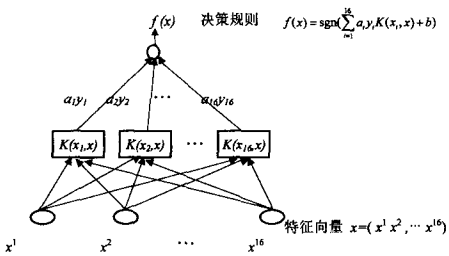

图 5　支持向量机示意图

表 1 列出了由各一对一声调识别模板结果组成的矩阵 R，其中 x/y 表示 Tone x-Tone y 分类器识别情况。若有一特征向量 $x=(x^1\ x^2,\cdots,x^{16})$，在通过 6 个二分类 SVM 分类器后所产生的 6 个输出结果表示为向量 $h(h_1, h_2,\cdots, h_6)$，采用最邻近原则，将 h 向量与矩阵 R 中的各行进行比较，即可得出特征向量 x 所属的声调类别。

表 1　声调识别调节矩阵

	One-versus-one					
	1/2	1/3	1/4	2/3	2/4	3/4
Tone 1	1	1	1	0	0	0
Tone 2	−1	0	0	1	1	0
Tone 3	0	−1	0	−1	0	1
Tone 4	0	0	−1	0	−1	−1

5　仿真实验及结果

为了检验算法的可行性，首先建立了用于声调训练及测试的语音数据库。采集对象为 12 名在校大学生(男生女生各 6 名)，采样频率为 22.05kHz，精度为 16 位，内容为普通话发音中选出的 45 个单音节(*a*, *ai*, *bao*, *bo*, *can*, *chi*, *du*, *duo*, *fa*, *fu*, *ge*, *hu*, *ji*, *jie*, *ke*, *la*, *ma*, *na*, *pao*, *pi*, *qi*, *qie*, *sha*, *shi*, *shu*, *tu*, *tuo*, *wan*, *wen*, *wu*, *xia*, *xian*, *xu*, *ya*, *yan*, *yang*, *yao*, *yi*, *ying*, *yu*, *yuan*, *yun*, *zan*, *zhi*, *zi*)，每个音节分别读作 4 种声调，每名学生的发音为一个连续音频文件，即实验数据可视为 2160(12×45×4)个单音节音频信号，其中 1/2(3 名男生和 3 名女生)的发音被作为训练集，剩余的作为测试集。

实验中的 6 种错误识别分别为 Tone 1-Tone 2，Tone 1-Tone 3，Tone 1-Tone 4，Tone 2-Tone 3，Tone 2-Tone 4 及 Tone 3-Tone 4，测试结果如表 2 所列。测试集中，每种声调均有 270 个样本，总体识别率为 93.52%。

表 2　声调识别结果表

		heard				
		Tone 1	Tone 2	Tone 3	Tone 4	Accuracy(%)
spoken	Tone 1	252	6	3	9	93.33%
	Tone 2	9	250	4	7	92.59%
	Tone 3	2	3	259	6	95.93%
	Tone 4	10	6	5	249	92.22%
Overall						93.52%

神经网络也是应用比较广泛的一类非线性分类器，是公认的识别效果较好的分类器之一。为此，本文将实验数据在 SVM 及 BP 网络上进行了比较试验，结果分别由训练集实验、测试集实验及合集实验组成，如图 6 所示。

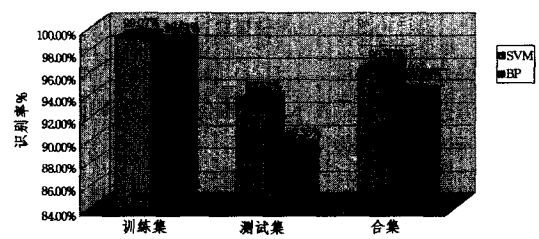

图 6　SVM 与 BP 网络识别算法结果比较

结束语　本文建立了普通话声调数据库，并利用基频、基频差分、能量、能量差分作为声调的基本特征，利用 6 个二类 SVM 分类器对待测声调进行识别，总识别率达到 93.52%。实验结果表明：(1)增加能量及差分作为声调的特征可以进一步提高声调识别率；(2)基于勒让德多项式系数的曲线拟合法较好地保证了原始特征的有效性；(3)与 BP 神经网络方法相比，SVM 具有更高的识别率和更强的实用性。

参 考 文 献

[1] Chen X-X, Cai C-N, Guo P, et al. A Hidden Mardov model applied to Chinese four-tone recognition[C]// Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 1987:797-800

[2] Yang W-J. Hidden Markov Model for Mandarin lexical tone recognition[J]. IEEE Trans. Acoust. Speech Signal Process, 1988, 36:988-992

[3] Emonts M, Lonsdale D. A memory-based approach to Cantonese tone recognition[C]// Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH). 2003:2305-2308

[4] Cao Yang, et al. Tone Recognition in Mandarin using Focus[C]// INTERSPEECH. 2005:3301-3304

[5] Gang Peng, William S-Y, Wang. Tone recognition of continuous Cantonese speech based on support vector machines[J]. Speech Communication, 2005, 45:49-62

[6] 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，2004，25：96-99

[7] Kaiser J F. On a simple algorithm to calculate the energy of a signal [C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP' 90). Albuquerque, USA, 1990

[8] 魏延，石磊，陈琳琳. 基于后验概率加权的模糊支持向量机[J]. 重庆工学院学报：自然科学版，2009，23(8)：80-84