普通话孤立字的声调识别

李伟浩　王岚

摘　要　声调识别是汉语语音识别的重要部分。本文是基于支持向量机(SVM)的孤立字声调识别，在自相关域采用一种鲁棒性算法提取基音频率[1]，再根据语音声调的基音频率特征提取了6个特征作为识别的特征向量。在没有对说话人训练的条件下，声调识别率达到95.97%。

1　引言

汉语是一种有声调的语言，有四个声调(阴平、阳平、上声、去声)，即我们常说的一二三四声。它表现在音节高低升降，主要由音高决定。对于初学汉语者来说，声调是一个难点。因为声调包含了非常重要的信息，起到辩义的作用；对于同样的音节，会因为声调的不同而具有完全不同的意义，例如：灰(hui1)、回(hui2)、悔(hui3)、会(hui4)。声调在语音识别与合成中也起到非常重要的作用。在语音识别中，声调信息可以提高语音识别的精度和速度；而在语音合成中，声调可以使合成语音更有韵律感。

根据目前国内外关于汉语声调识别的研究可知，汉语声调识别与语音信号的频率，时域振幅包络和精细结构，刺激速率，持续时间有着密切相关。而近年来，许多学者利用声调与基音频率之间的关系，提出了包括隐马尔可夫[2]、神经网络[3,4]、决策树[5]和支持向量机[6]等声调识别的方法。基本上这些方法提取的特征都是基于基音频率的轮廓信息，但到目前为止，还没有一种可靠的方法可以非常精确地提取出基音频率，这就使得声调识别变得困难；而且在汉语中一个音节对应着一定的音调，而音调的高低是没有绝对的，只是相对的，即对同一个音节，不同说话人的基频曲线是不一样的。而声调在基音频率轮廓的表现可以用赵元任先生独创的”五度制”标记法表示，如图1(左)所示，但在现实中受环境、情绪等因素的影响，实际提取的基音频率却有点差异，如图1(右)所示。

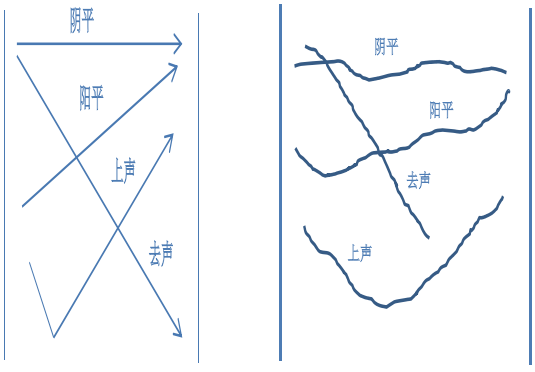

图1 四声调的”五度制”标记法(左)与现实中四声调基音频率曲线(右)

从图1可以看出四声调的基频曲线有如下特征：

1、 阴平的基频曲线较平坦，存在小扰动，且频率较高；

2、 阳平的基频曲线为上升型，前端有可能出现下降；

3、 上声的基频曲线为降升型的，其拐点大致居中。

4、 去声的基频曲线为下降型的，且其时长一般都比其他三个声调要短。

本文声调识别识别过程基本上就分为语音信号的预处理、基音频率检测、提取基频特征、分类识别。其结构框图如图2所示：

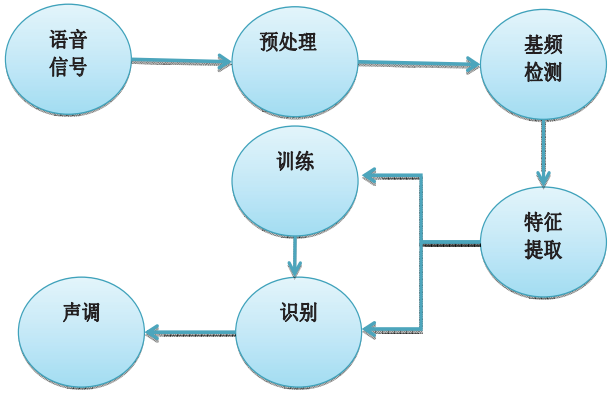

图2 声调识别的总体框图

2　信号预处理

在声调识别中，提取基音频是必不可少的，在提取之前，对语音信号进行预处理也是一项不可缺少的关键工作，它直接影响到提取基频的精度。其中语音信号的预处理包括预加重、分帧、加窗等处理。

2.1 预加重

由于语音信号的平均功率受声门激励和口鼻辐射影响，信号频谱的高频部分的能量按-6db/倍频程衰减。因此，在语音信号预处理中必须先进行预加重处理，以补偿语音信号的高频，使信号频谱变得平坦。一般采用一阶数字滤波器对语音信号进行预加重处理，滤波器的传递函数可表示为：

$$H(z) = 1 - cz^{-1} \quad (1)$$

其中c为接近1的常量，一般取0.97。

2.2 分帧与加窗

语音信号是一种典型的非平稳时变信号，其信号波形随着时间不断变化，但在一个较短的时间内，一般认为10ms-30ms之间，可以认为信号是平稳的。利用信号的短时平稳性，对信号进行分帧，帧与帧之间存在交叠。为减小语音帧间的截断效应，使语音帧两端能够平滑过渡，分帧后需要对语音信号进行加窗处理，一般采用汉宁窗，其窗函数可以为：

$$w(n) = \begin{cases} 0.5 - 0.5 * \cos(2 * \pi * \frac{n}{N-1}; 0 \le n \le N-1 \\ 0; 其他 \end{cases} \quad (2)$$

3 特征提取

3.1 基音频率

声调识别主要是依据基音频率轮廓曲线，因此能否准确地检测出基音频率是关键。本文采用鲁棒性的基频检测方法[1]。其算法概述如下：

1、语音信号短时分析。因为人的语音基频范围是在75-600HZ，每帧最大的基频候选数为4，帧长设为10ms，而窗长为40ms。为了消除信号直流分量，采用去均值法。求全局峰值来确定静音与有声的阀值

2、自相关处理。对信号进行汉宁窗处理，及快速傅立叶变换，对频率域的能量进行傅立叶反变换，再除以窗函数的自相关函数。其相关的计算公式如下：

$$a(t) = s(t) * w(t) \quad (3)$$

$$A(w) = \int a(t) e^{-iwt} dt \quad (4)$$

$$r_a(\tau) = \int |A(w)|^2 e^{i\omega\tau} \frac{dw}{2\pi} \quad (5)$$

$$r_x(\tau) \approx \frac{r_a(\tau)}{r_w(\tau)} \quad (6)$$

3、确定基频。通过找到自相关函数最大的位置作为候选值，再通过Viterbi算法求得最佳的路径。

此算法正是praat软件所使用的算法，它能够比较准确地提取出语音的基音频率，如下图所示分别为巴(ba1)、拔(ba2)、把(ba3)、爸(ba4)四个字基频曲线图。

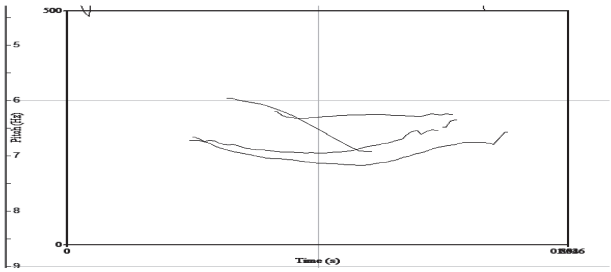

图3 基频曲线图

3.2 特征提取

为了消除不同说活人的差异，首先要对基频进行归一化[3]，归一化公式如下：

$$f' = \frac{f}{f_{mean}} \quad (7)$$

为了提取特征，把基频分成3段，提取6个特征，包括每一段基频的均值及基频曲线的一次斜率[3,4,5]。其中基频曲线斜率是通过最小二乘法拟合得到[7,8,9]。如图5所示，6个特征从基频的曲线中提取出来。

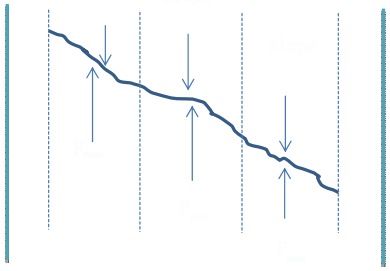

图4 6个基频特征的图

4 支持向量机的分类器设计

支持向量机(SVM)的分类是基于统计学习的理论基础，其核心内容是Vapnik提出的结构风险最小化原则[10,11]。该算法在文本分类、手写识别、生物信息等领域中获得了广泛应用。

对于类一个两类线性可分问题，如图5所示，图中实心点与空心点分别代表两类样本，H2为分类线，H1与H3分别为过各类中离分类线最近的样本且平行于分类线的直线，其中分类线H2使得分类间隔最大。

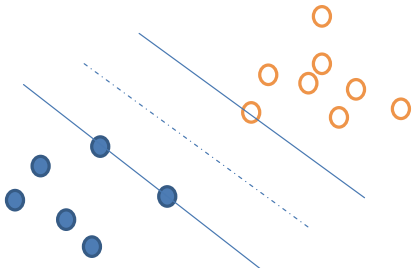

图5 线性分类示意图

对于非线性分类问题，SVM能够将输入向量x映射到高维特征空间中，这样非线性可分类问题就会转变为线性可分类问题，即最大化函数：

$$W(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \quad (8)$$

其决策函数可以表示为：

$$y_j=\sum_{i=1}^{N}\alpha_i y_i K(x,x_i)+b \quad (9)$$

且满足

$$\sum_{i=1}^{N}\alpha_i y_i=0\ ,\ \alpha_i\geq 0 \quad (10)$$

其中K(，)为核函数。SVM的核函数主要包括多项式核函数、径向基核函数和sigmoid型核函数。

5 实验与分析

本文是按照图3所示流程进行实验研究的，先建立声调训练与测试的语音数据库。本实验所用的语音数据库是在安静环境下采集了10人(4男6女，4-6岁小孩与20-25岁的学生组成)，其中语音的采样频率为16KHz，单声道，16bit。数据采集的内容包括6个单韵母及涵盖了20个声母的单音节(a,o,e,l,u,v,ba,po,ma,fang,da,ti,nan,liu,ji,qi,xie,ge,ke,hui,zang,can,zhai,chi,shi)。每个音节分别读作4种声调，每个人的发音为一个音频文件，即实验数据可视为1040(10x26x4)个单音节音频信号，除去一些不可用数据，只有812个音频文件。将4人(2男2女)总共365条语音作为训练样本，6人(2男4女)共447条语音作为测试样本。实验结果如下表所示。

从表一可以看出，SVM对孤立字的四种声调的识别别率为95.97%。其中一声的识别率是最高的，而三声的识别率最低，这是因为一声的基频曲线较平坦，波动性小；而三声的基频曲线波动性较大，且轨迹图与二声的基频轨迹图有点相似，容易把三声误判为二声。

6 结论

本文讨论了基于SVM的孤立字声调识别，及在praat软件所使用的基频提取算法，具有较强的抗干扰力，鲁棒性强。同时在提取特征时，利用基频曲线的规律，对基频进行分段拟合；使得特征向量不容易受两端基频曲线不准确的影响，最后的实验取得了很好的结果。但是本文的方法不能简单地推广到连续语音的声调识别，在连续语音中，由于存在前后音节相互影响，声调基音频率特性曲线变得更为复杂，这也是未来工作中要去研究的。

表1 实验结果

	训练样本数(条)(2男2女)	测试(条)(2男4女)	正确识别数(条)	识别率(%)
一声	100	139	136	97.84
二声	93	143	138	96.5
三声	74	51	46	90.2
四声	98	114	109	95.61
总计	365	447	429	95.97

参考文献

[1] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to noise ratio of a sampled sound. Institute of Phonetic Sciences, University of Amsterdam, Proceedings 17(1993),97-110.

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

[3] P. C. Chang, San Wei Sun, and Sin Horng Chen. Mandarin tone recognition by muliti-layer perceptron, Acoustics, Speech, and Signal Processing, 1990.

[4] T. Lee, P. C. Ching, L. W. Chan, Y. H. Chen, and Brian Mak. Tone Recognition of Isolated Cantonese Syllables. IEEE Transactions on Speech and Audio Processing, VOL. MAY 1995.

[5] C. Yang, etal. Tone Recognition Mandarin Using Focus, Interspeech, 2005.

[6] G. Peng, William S. Y. Wang. Tone recognition of continuous Cantonese Speech based on support vector machines. Speech Communication 45 (2005) 49-62.

[7] 顾明亮，夏玉果，杨亦鸣. 支持矢量机的汉语声调识别，声学技术，2007.

[8] 傅德胜，李仕强，王水平. 支持向量机的汉语连续语音声调识别方法，计算机科学，2010.

[9] 宋刚，姚艳红. 用于汉语单音节声调识别的基频轨迹拟合方法，计算机工程与应用，2008.

[10] B. Boser, Guyon I, Vapnik V. A Training Algorithm for Optimal Margin Classifiers, Fifth Annual Workshop on Computational Learning Theory. Pittsburgh:ACM Press,1992.

[11] C. Cortes, V. Cortes. Support Vector Networks. Machine Learning, 1995.

作者简介

李伟浩 男，毕业于深圳大学，现工作于环绕智能中心，研究方向为语音识别。

王 岚 作者简介详见本期封2页。