2004 年 4 月
第 20 卷第 2 期

皖西学院学报
Journal of West Anhui University

Apr.,2004
Vol.20 NO.2

文本文件的语音识别中音节的自动切分

张晓东,崔仁涛

(皖西学院 物理系,安徽 六安 237012)

摘 要:汉语语音识别中对孤立词、小词汇特定人的语音识别率较高,但对于连续的大词汇量语音识别率较差。把连续的大词汇语音实时自动地切分为单个音节,可以提高其系统的识别率。本文根据汉语语音在能量和频率等方面的特征,找到了短时平均幅度和短时平均过零率的方法来检测音节的端点,从而得到对文本文件中汉语语音的音节自动切分算法。

关键词:短时平均能量;短时平均过零率;短时平均幅度

中图分类号:O429 **文献标识码**:A **文章编号**:1009-9735(2004)02-0018-03

随着语音处理技术的进步,语音人机交互技术在过去十多年里取得了很大进展,并在某些方面有所突破。国际上,孤立单词识别系统的词汇已经扩大到几万[1,2],对孤立词,小词汇特定人的语音识别率达 90% 以上[3,4]。但对连续汉语语音识别还存在很多的困难。

本文综合利用短时平均幅度和短时平均过零率对语音信号作分帧处理,计算每帧的短时平均幅度和平均过门限率,确定语音的音节端点,实现实时自动切分音节的目的。

1 短时平均幅度和短时平均过零率的原理

1.1 短时平均能量和短时平均幅度

由于语音信号随时间而变化,清音和浊音之间的能量差别相当显著,因此对其短时能量和平均幅度进行分析,可以描述语音的这种特征变化情况。

定义短时能量为:

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m) = \sum_{m=n-N+1}^{n} x^2(m)h(n-m) \quad (1)$$

从式(1)中可以看出,语音信号各样点值先平方,然后通过一个冲激响应为 h(n)的滤波器,输出为由短时能量构成的时间序列。若 h(n)幅度恒定且序列长度 N(即窗长)很长,这样的窗函数等效为很窄的低通滤波器,此时 h(n)对 $x^2(m)$的平滑作用非常显著,使得短时能量几乎没有多大变化,无法反映语音的时变特性。反之,若 h(n)序列长度 N 过小,那么等效窗又不能提供足够的平滑,以至于语音振幅瞬时变化的许多细节仍然被保留了下来,从而看不出振幅包络的变化规律。

图 1 画出了一段实际语音(女声"他去无锡市")的短时能量函数随矩形窗长的变化曲线,帧之间无交叠[5]。选择帧长为 10-20ms。

短时平均能量特征可以作为区分清音段和浊音段的特征参数。实验结果表明浊音段的能量 E_n 明显高于清音段。通过设置门限值,可以判定浊音变为清音的时刻,同时也可以判定清音区间和浊音区间。短时能量函数的一个主要问题是 E_n 对信号电平值过于敏感。由于需要计算信号样值的平均和,在定点容易溢出。为了避

* **收稿日期**:2003-12-12
基金项目:安徽省教育厅自然科学基金项目(03kj324)。
作者简介:张晓东,男,皖西学院物理系副教授。

免这种情况，我们引入平均幅度 M 的概念来衡量语音的变化：

$$M_n = \sum_{m=-\infty}^{\infty} |x(n)| \omega(n-m) = \sum_{m=n-N+1}^{n} |x(n)| \omega(n-m) \tag{2}$$

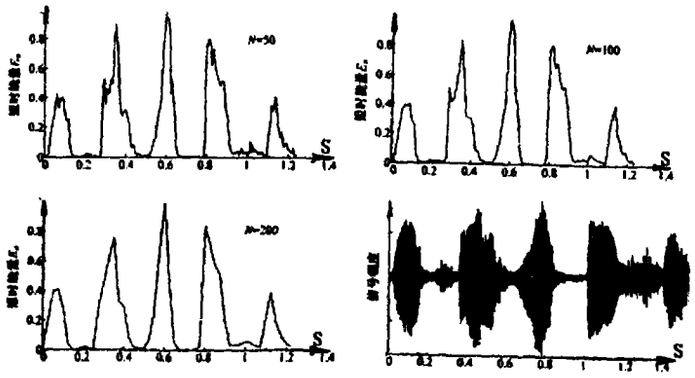

图 1　不同矩形窗长 N 时的短时能量函数

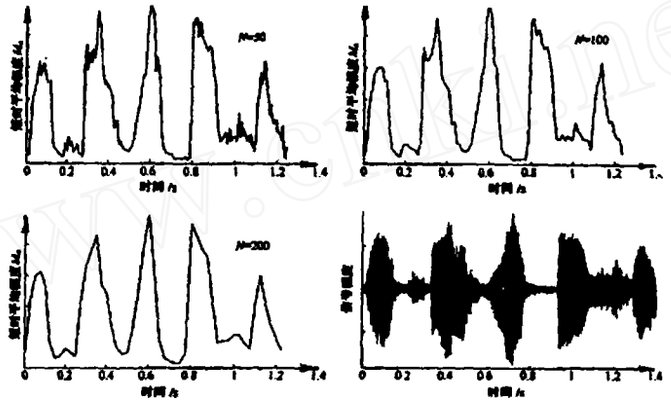

图 2　不同矩形窗长 N 时的短时平均幅度函数

图 2 画出了短时平均幅度随矩形窗长 N 的变化情况，帧之间无交叠。我们比较图 1 和图 2 可以看出对短时平均能量和短时平均幅度的分析结果的结论是完全一致的，但短时平均能量的变化情况比短时平均幅度明显。由于平均幅度函数没有平方运算，所以动态变化范围要比平均能量小。区分清音和浊音时，短时平均能量要比短时平均幅度明显[5]。

1.2　短时平均过零率

过零率反映信号的频谱特性。当相邻的两个样点的正负号异号时，我们称之为“过零”，即此时信号的时间波形穿过零电平的横轴。把单位时间内的样点值改变符号的次数叫平均过零率。对于窄信号来说，用平均过零率衡量相当精确。

由于语音是一种短时平稳的宽带信号，因此在统计平均过零率时通常按帧来计算，这样得出短时平均过零率，定义为：

$$Z_n = \sum_{-\infty}^{\infty} |\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]| \omega(n-m) \tag{3}$$

$\omega_{(n)}$为窗函数，计算时常采用矩形窗，窗长为 N。当相邻两个样点符号相同时，$|\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]| = 0$，没有产生过零；而当相邻两个样点符号相反时，$|\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]| = 2$，为过零次数的 2 倍。因此在统计一帧(N 点)的短时平均过零率，求和后必须要除以 2N。在矩形窗条件下，式(7)还可以简化为下式表示：

$$Z_n=\frac{1}{2N}\sum_{m=n-M+1}^{n}|\mathrm{sgn}[x(m)]-\mathrm{sgn}[x(m-1)]| \tag{4}$$

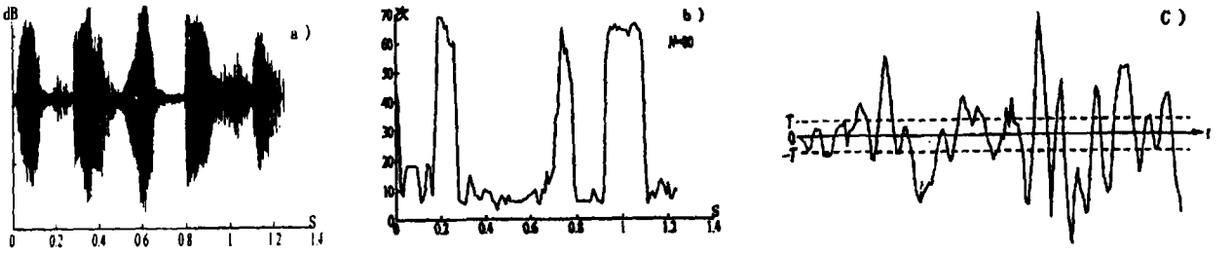

图3 一段语音的短时平均过零曲线及过零门限率

a)语音时域波形 b)短时平均过零次数 c)过门限率

短时平均过零率可以估计出语音的频谱特性。根据式(4)可知高频率对应着高过零率,低频率对应低过零率,那么过零率与语音的清浊音特性就存在着对应的关系。一般经验结论是,清音的过零率分布大致为高斯分布,清音每10ms的短时平均过零次数的均值为49次,浊音每10ms的短时过零次数的均值为14次。两种分布之间相互交叠的区域,它们在过零次数为24左右的概率基本相等。因此,单纯依靠短时平均过零率来准确判断清浊音是不可能的,在实际运用中往往是采用语音的多个特征参数来综合判断。

在实际应用中,过零率容易受到A/D转换时的直流偏移以及噪音的影响。为了减少影响我们采用了过门限频率来修改过零率。即在零电平附近设置正负门限±T,与平均过零率定义相似,短时平均过门限率为式(5),过门限率反映了穿过正负门限的次数,用它来修改过零率就具有一定的抗噪声干扰,只要信号没有超过[-T,T]的范围,就不可能产生虚假的过零数.在语音的音节端点检测时,经常采用类似的多门限过零率特征来检测语音音节的起始位置和结束位置。

$$Z_n=\sum_{m=-\infty}^{\infty}\{|\mathrm{sgn}[x(n)-T]-\mathrm{sgn}[x(n-1)-T]|+|\mathrm{sgn}[x(n)+T]-\mathrm{sgn}[x(n-1)+T]|\}\omega(n-m) \tag{5}$$

2 用短时平均幅度和短时平均过零率对音节自动切分

假设要检测的语音进行分帧处理,每帧10ms。若抽样频率为8kHz。即窗函数长度N=80。分别采用式(5)和式(11)计算每帧语音的短时平均幅度和平均过门限率。由于语音一般都存在能量较高浊音段,因此考察语音的平均幅度轮廓可以设定一个较高的门限T_1,使语音的起点和终点落在T_1和M_n所确定的时间间隔AB之外。然后根据背景噪音的平均幅度确定一个门限较低的T_2,并从A点往起点方向、从B点往终点方向搜索,分别找到与门限T_2相交的两点C、D。这样我们就用双门限完成了第一级粗判。第二级判决要利用短时平均过门限率。同样根据背景噪声的Z_n可以设定一个较低的门限T_3,从C和D点分别向起点和终点方向搜索,可以找到Z_n与门限T_3相交的两个点E、F。这样就确定出了语音音节的起始的端点E、F。

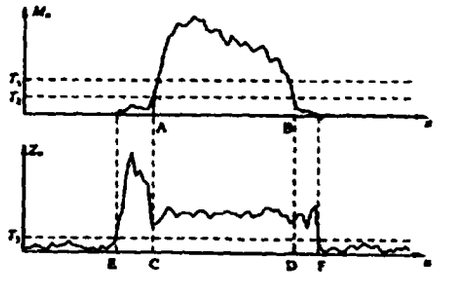

图4 利用能量和过零率的语音端点检测

参考文献:

[1]Averbuch,A.et al.Experiments with the TANGORA 20,000 word speech recognizer,proc.ICASSP(Dallas Texas),1987.

[2]Penny,P.et al.Experiments in continuous Speech recognition with a 60,000 word Vocabulary,proc.ICSP'92 (Banff, Canada),Vol.1,1992.

[3]Bates,M.et al.The BBN/HARC Spoken Language Understanding System,Proc.ICASSP'93,1993.

[4]杨家沅.语音识别与合成[M].成都,四川科学出版社,1994.

[5]张雄伟,等.现代语言处理技术及应用[M].北京:机械工业出版社,2003.