文章编号:1002-8684(2005)07-0060-03

一种基于多特征的带噪语音信号端点检测与音节分割算法

·论文·

卢艳玲，侯榆青，王　宾，唐　升，史　燕
（西北大学 电子科学系，陕西 西安 710069）

【摘　要】语音信号的端点检测和音节分割直接决定语音识别率。在传统方法的基础上提取语音信号的多个特征参数，并综合利用各个参数的特性进行检测和分割，提高了端点检测和音节分割的准确度。

【关键词】端点检测；音节分割；自相关函数
【中图分类号】TN912　　【文献标识码】A

An Endpoint Detection and Syllable Separation Algorithm Based on Multi-characteristic for Noisy Speech Signal

LU Yan-ling, HOU Yu-qing, WANG Bin, TANG Sheng, SHI Yan
(Department of Electronic Science, Northwest University, Xi'an 710069, China)

【Abstract】The endpoint detection of speech signal is an important step in speech recognition. It determines directly the recognition rate. Several characteristic parameters of speech signal are extracted based on the traditional methods for endpoint detection and syllable separation, and the accuracy of endpoint detection and syllable separation is improved.

【Key words】endpoint detection; syllable separation; autocorrelation function

1　引言

目前，语音信号端点检测的算法有很多种，基本思想是通过计算语音信号的某些特征参数，使语音信号与背景噪声区分开来。应用较多的有基于短时过零率与短时能量的算法[1]、基于子相关相似距离的算法、基于倒谱特征的算法等。这些算法在安静环境下基本可得到比较理想的检测效果，但在背景噪声较大或较复杂时，检测效果就会显著下降。产生这种现象的原因是检测算法只利用了语音信号的某一个特征，而没有充分利用其它特征信息[2]。笔者提出了一种利用语音信号多个特征进行端点检测和分割的算法，经过仿真得到了较好的效果。

2　特征参数的选取及其定义

2.1　特征参数的选取

对语音信号波形的分析可知，语音信号可分为无声段、清音段和浊音段，无声段的能量最低，清音段其次，浊音段最高。在噪声较低的环境下，清音段的能量一般比无声段的能量高出几到几十倍，而浊音段的能量比无声段的能量高出几十至上百倍，应用能量参数在背景噪声较低的情况下基本上可将它们区分开。但有些清音能量较低，使其很难与背景噪声区分开，这时需要用到另一个同等重要的特征参数——过零率。清音段的过零率大多数情况下最高，无声段的过零率变化较大，一般情况下比浊音段低一些，有时会比浊音段稍高或差不多[3]。利用上述 2 个参数在背景噪声较大环境下的检测效果仍不是十分好，如果能充分利用相对能频积、相对能频比和自相关函数这 3 个特征参数，则检测效果明显提高。

能频积是指能量与过零率的乘积，用于检测语音的端点。汉语语音的起点一般是无声段与清音段的分界点，无声段的能量和过零率都较低，而清音段相对于无声段来说能量和过零率都较高，因此能频积在 2 段的差别加大，可更好地检测语音的起点。汉语语音的末尾一般是浊音，只用短时能量就能较好地判断一个词语的末尾，一般只要短时幅度降低到该音节最大短时幅度的 1/16 左右以后就认为该音节已经结束[4]。汉语语音信号的声母和韵母一般对应清音和浊音，由前面

[基金项目] 陕西省自然科学基础研究计划项目(2003F23).

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

分析可知浊音段能量较高而过零率较低，因此能频比可以更好地区分音节的声母和韵母。

当背景噪声较大时，可能会淹没掉一部分语音信号，用短时能量和过零率的检测效果会明显下降，这时最好用短时自相关来检测。关于自相关的分析在其定义部分有详细说明。

2.2 特征参数的定义

(1)短时能量

短时能量的定义一般有 2 种，一种用振幅的平方表示，另一种用振幅的绝对值表示。由于某些声母发声短促，用振幅的平方表示能量时数值过大，所以笔者选用后者。定义为

$$E=\sum_{n=0}^{N-1}s(n) \tag{1}$$

其中，N 为一帧中的采样点数。

(2)短时过零率

$$Z=\frac{1}{2}\left\{\sum_{n=0}^{N-1}\left|sgn[s(n)]-sgn[s(n-1)]\right|\right\} \tag{2}$$

其中，$sgn(\cdot)$为符号函数。

用以上方法计算短时过零率容易受低频干扰，尤其是 50 Hz 交流干扰。解决的办法就是设置一个门限 T_0，将过零的含义修改为跨过正负门限。这样过零率就有了一定的抗干扰能力，即使存在小随机噪声，只要选择合适的门限，使噪声信号不越过正负门限构成的带就不会产生虚假的过零数。定义为[4]

$$Z=\sum_{n=0}^{N-1}\left\{\left|sgn[s(n)-T_0]-sgn[s(n-1)-T_0]\right|+\left|sgn[s(n)+T_0]-sgn[s(n-1)+T_0]\right|\right\} \tag{3}$$

(3)短时能频积

$$A=EZ \tag{4}$$

(4)短时能频比

$$B=E/Z \tag{5}$$

(5)短时自相关函数

$$R(m)=\sum_{n=0}^{N-1-m}s(n)s(n+m) \tag{6}$$

噪声的自相关函数除了在 $m=0$ 时刻外都很小，而语音信号中的浊音由于有比较明显的周期性，所以其相关性较高，除了主峰外还有较高的副峰，采用求主副峰比的方法可将噪声与语音信号区分开。定义主副峰比

$$T=R(0)/R(k) \tag{7}$$

其中，$R(0)$为主峰幅度，$R(k)$为最近副峰幅度。人的基音频率为 75~300 Hz，在 8 kHz 采样频率下 m 在 25~108 范围时会有至少一个副峰，在这个范围内取最近的副峰来求主副峰比，其值很大时认为是噪声，较小时认为是语音。

利用短时自相关法检测存在的较大问题是求自相关函数的计算量比较大，但目前高速数字信号处理器(DSP)可在很短的一个指令周期内做一次乘加运算，而且专为卷积运算和递归运算设计了一些效率很高的指令，使利用定义直接求自相关变得简单有效。

3 端点检测与音节分割算法

3.1 对语音信号的预处理

在计算语音信号各个参数之前应该对语音信号进行预处理，将语音信号通过一个一阶高通滤波器(一般称为预加重滤波器)，该滤波器的数学表达式为 $1-\mu z^{-1}$，一般 $0.93<\mu<0.98$，典型值为 0.94。它的目的在于消除低频干扰，尤其是 50 Hz 的工频干扰，将对语音识别更有用的高频部分的频谱进行提升[5]。

3.2 参数阈值的选定

对连续语音信号经过 8 kHz 采样，16 bits 量化得到数字语音信号。由于录音和发声有间隔，一般情况下语音信号前 100 ms 是无声段，可提取这一段的特征参数作为噪声段的参数值，再将噪声段的参数值乘以相应的系数作为判定的阈值，系数的大小需要通过大量的实验得到。首先计算起始 100 ms 平均每帧的上述 5 个参数值，分别记为：短时能量 E_{100}，短时过零率 Z_{100}，短时能频积 A_{100}，短时能频比 B_{100} 和自相关主副峰比 T_{100}。判断语音开始各阈值的选择如下：

(1)短时能量阈值

$$E_{th}=(E_{max}-E_{100})\times 2 \tag{8}$$

其中，E_{max}为前 100 ms 短时能量的最大值。

(2)短时过零率阈值

$$Z_{th}=\min(25,Z_{100}+2\sigma_z) \tag{9}$$

其中，σ_z为前 100 ms 短时过零率的标准差。

(3)短时能频积阈值

$$A_{th}=A_{100}\times 2 \tag{10}$$

(4)短时能频比阈值

$$B_{th}=B_{100}\times 1.5 \tag{11}$$

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(5)自相关主副峰比阈值

$$T_{th}=(T_{max}-T_{100})\times 2.8 \quad (12)$$

其中,T_{max}为前 100 ms 主副峰比的最大值。

判定清浊音时,各系数要做进一步修正,短时能量系数为 4,短时过零率的系数为 1.6,短时能频积的系数为 3.4,短时能频比的系数为 1.3。利用这种修改简单方便,几乎不增加额外的计算量。

3.3 端点检测与音节分割算法

①根据 E_{100} 值确定选用短时能量还是自相关主副峰比作为检测的一个参数,如果 E_{100} 值较小用短时能量,如果 E_{100} 值较大则用自相关主副峰比。

②自适应电平均衡。$E_0(k)=E(k)-E_{100}$,$E_0(k)$表示均衡后各帧的能量,静态时在 0 附近摆动,有语音时值较大。

③语音起点的判定。如果连续 3 帧的各参数值超过各自对应的起点判定阈值,则 3 帧中的第 1 帧判为语音的起点。

④清浊音的判定。如果连续 8 帧的各参数值超过各自对应的清浊音的判定阈值,则 8 帧中的第一帧作为浊音的开始。

⑤判断音节结尾。在前面判定中记下当前音节的能量最大值,当连续 5 帧的短时能量降为最大值的 $\frac{1}{16}$ 时认为语音已结束,5 帧中的第 1 帧作为音节结尾点。

⑥判断下一音节的开始。判断语音的结尾点减开始点是否大于 28,如果大于继续③,否则丢弃已判定的清浊音分界点和音节结尾点,继续④。

4 实验结果与结论

应用笔者提出的算法分别对噪声较小的室内环境和含有明显噪声环境下的样本进行实验,实验数据为 8 kHz 采样,16 bits 量化,普通传声器采集的 2 男 2 女共 200 个样本(每个样本 2~6 个字)。实验结果表明,端点检测和音节分割的效果与应用单一参数的算法相比,正确率提高了 20%以上,尤其在噪声较明显的环境下,以往算法显得无能为力,而该算法正确率可达到 90%以上。

图 1 是应用笔者提出的算法对单字的端点检测和声韵母分割效果,图 2 是应用笔者提出的算法对 4 个连续字的端点检测和音节分割效果,同时对每一个音节的声韵母也进行了有效的分割。从图中可以看出,检测和分割的效果较好,基本接近于手工分割效果,可以提高语音识别等后续工作的正确率。

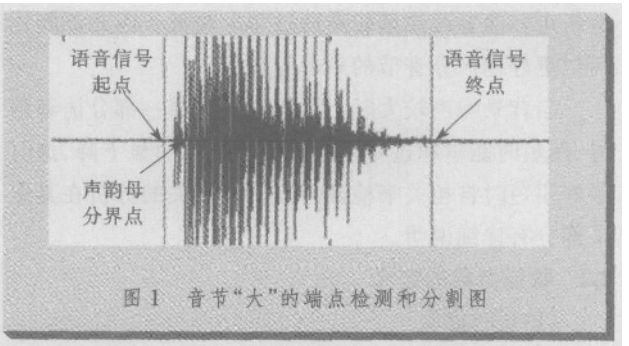
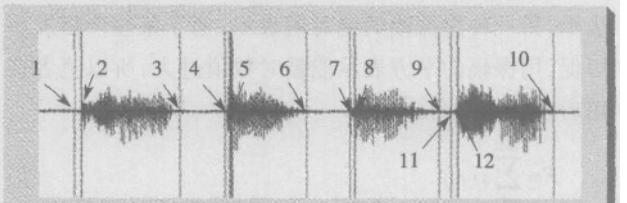


图 1 音节"大"的端点检测和分割图



注:1.语音信号起点;2.声韵母分界点;3.第一个音节终点;4.第二个音节起点;5.声韵母分界点;6.第二个音节终点;7.第三个音节起点;8.声韵母分界点;9.第三个音节终点;10.语音信号终点;11.第四个音节起点;12.声韵母分界点

图 2 四音节"西北大学"的端点检测和分割图

笔者提出的端点检测和音节分割算法克服了以往算法利用单一特征参数的缺陷,提高了算法的鲁棒性。由于算法第一步为判断背景噪声强弱,然后根据噪声的强弱选择应用能量还是自相关作为参数,因此该算法不仅适用于安静环境,而且在噪声环境下的效果也比较好,基本可以满足语音识别的要求。

参考文献

[1] Rabiner L R, Sambur M R. An Algorithm for Determining the Endpoints of Isolated Utterance. Bell System Tech J, 1975, 54(2):297—315.

[2] Rabiner L R, Juang B H. Fundamentals of Speech Recognition(影印版)[M]. 北京:清华大学出版社,1999. 82—85.

[3] 郭巧,张立伟,陆际联. 用于汉语语音信号端点检测与切分的有效方法. 计算机工程与应用,2000(5):92—95.

[4] 易克初,田斌,付强. 语音信号处理. 北京:国防工业出版社,2000.76—78.

[5] 王炳锡. 语音编码[M]. 西安:西安电子科技大学出版社,2002.48—50.

作者简介

卢艳玲,硕士研究生,现从事语音信号数字处理和 DSP 应用系统开发方面的研究.

[收稿日期] 2005-04-12

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net