文章编号:1007-9432(2009)05-0487-03

一种利用多参数进行实时语音边界检测与音节分割算法

段淑斐
(太原理工大学 信息工程学院,山西 太原 030024)

摘 要:在对语音信号 5 种特征参数:短时能量、平均过零率、相对能频比、相对能频积、短时自相关函数语音分段效果详细对比的基础上,提出了利用多参数结合进行语音边界检测与音节分割。同时较之当前主流一帧 20 ms 的处理方式,提出以 2.5 ms 为一帧处理,确保在 20 ms 内检测到信号边界,缩小了搜索时间,提高了实时性。

关键词:实时;多参数;边界检测;音节分割
中图分类号:TN912.34 **文献标识码**:A

语音段的边界检测是语音处理中一个重要环节,只有准确判定语音信号端点,才能为后续语音处理提供扎实基础。当前国内外通用的边界检测是对语音的基本特征参数进行处理,不同参数对于语音的分段效果各有不同。短时能量、平均过零率、短时自相关等特征参数对于纯净语音的分段效果相对比较好,但对带噪语音信号,鲁棒性就相对较差[1]。为此笔者提出了多个参数联合判定边界检测和音节分割解决方案,实验结果表明在低信噪比下分段准确率达到了 90 %以上。

对于短时能量、平均过零率、短时自相关等特征参数来说,帧长度对于它们能否反映语音信号的幅度变化起着决定性影响[2]。当前主流方法都采用长度为 20 ms 的帧,但是这需要多帧回溯搜索才能确定信号边界,实时性不强。为了提高算法的实时性,笔者将帧长修改为 2.5 ms 进行处理,确保在 20 ms 内检测到信号边界,缩小了搜索时间。

1 几种语音特征参数分段效果

对各种特征参数分别进行分段,效果对比如表 1 所示。可以看出单一参数的分段方法简单但各有局限性,基本上也都无法达到实时性。各种参数对于环境要求也高,在噪声背景下单参数分段的适应性很差。因此,本文将多种参数结合分段。

表 1 各类方法优缺点比较

方法	优点	缺点
短时平均过零率法	较简单	难以识别弱爆破音等,难以达到实时处理要求
短时能量法	较简单	弱摩擦音与结尾鼻音易和噪声混淆,难以达到实时处理要求
自相关方法	浊音检测精度较高	开端清音检测精度不够
相对能频积法	无声和清音检测精度较高	对浊音检测精度不够,难以识别弱摩擦音,难以达到实时处理要求
相对能频比法	浊音检测精度较高	对无声和清音检测精度不够,难以达到实时处理要求

2 多参数实时语音分段算法

2.1 各种特征参数性能介绍

表 2 展示的是各种特征参数在不同语音段的性能比较[3]。

表 2 各参数在语音不同段值的比较

特征参数	无声	清音	浊音
短时能量	低	较低	高
平均过零率	低	高	较低
相对能频比	低	较高	高
相对能频积	低	较高	高
自相关函数	无周期性	无周期性有最大值	有周期性,有最大值,基因周期上有极大值

收稿日期:2008-12-26
基金项目:国家自然科学基金资助项目(60772101)
作者简介:段淑斐(1983 -),女,山西清徐人,硕士生,主要从事语音预处理方面的研究,(Tel)13734036520

从表中可以看出不同的参数在各个语音段的值有高低之分。利用相对能频积对于无声和清音的分割较好;相对能频比对于清浊音的分割较好;短时能量以及平均过零率对无声和清音的分割较好,将这几种参数结合起来作为清浊音的判决门限。

根据一帧信号内出现极值点数以及极值的峰值,同时采用回溯寻找的方法,找到与当前新样点帧所在窗相差整数倍窗长的帧的自相关值进行比较设定阈值,进而判定起止点的分界点。

2.2 算法描述

1) 预加重。将语音信号通过一个 μ 取 0.97 的一阶高通滤波器 $H(z) = 1 - \mu z^{-1}$。预加重目的在于消除低频干扰,对语音识别更有用的高频部分频谱进行提升。

2) 加窗。预加重后的语音信号要加窗,窗的长短对于能否由短时能量等参数表征语音信号的幅度变化起着决定性作用。如果窗选的过长,等于几个基因周期值,这就等效于很窄的低通滤波器,短时能量等参数不能反应语音信号幅度变化。反之,短时能量等参数随着信号的细微变化而快速变化起伏以至于参数值不够平滑。因此,通常选取 $N = 100 \sim 200$[3]。实验中选择 $N = 160$ 的矩形窗。

3) 分帧。分帧的长度一般取 20～30 ms,帧与帧之间的偏移通常取为帧长的 1/2 或 1/3,是为了避免相邻两帧的变化幅度过大,所以帧与帧之间要重叠一部分[2]。但是选取 20～30 ms 为一帧实时性效果不好,因此提出改进选取一帧长为 2.5 ms,20 个样点作为帧移。

4) 求取各特征参数。分别求取特征参数,对于自相关函数,还要进行归一化处理。从自相关的输出结果可以很明显地看到滤除了清音和噪音的成分,强调了浊音成分。

5) 阈值的设定。一般情况下语音信号前 50 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,再将这一段的参数值乘以相应的系数作为阈值。系数值经过大量的实验得到。

将前 50 ms 的 20 帧短时能量 E,平均过零率 Z,相对能频比 A,相对能频积 B 分别求和取平均记为 $E_{av,sum}$,$Z_{av,sum}$,$A_{av,sum}$,$B_{av,sum}$,然后乘上相应的系数。本实验中,经过大量计算得出经验值系数作为清浊音判定的阈值。记为 $E_{th} = 4E_{av,sum}$,$Z_{th} = 1.3Z_{av,sum}$,$A_{th} = 2.5A_{av,sum}$,$B_{th} = 1.5B_{av,sum}$。

流程图中指针 i 表示语音样点,s_{flag} 和 e_{flag} 用来动态标记语音的分段开始点和结束点,设定数组 $R[8]$ 用来记录当前窗内 8 帧数据每帧的自相关函数峰值点的和,8 帧 R 的和 R_{sum} 是每窗的特征参数,R_{av} 是从 e_{flag} 或 s_{flag} 开始与该窗相差 8 的整数倍的所有窗的自相关参数 R_{sum} 的平均值。

起止点的判定阈值用的是计算得出的 R_{av} 乘上相应的系数,经过大量实验得出经验值为:开始点判决门限为 $R_{av} \times 5$,结束点判决门限为 $R_{av} \times 0.08$。

6) 起止点的判定。判定过程详见流程图 1。如果 $s_{flag} \geq e_{flag}$,从 s_{flag} 开始求与该窗相差 8 的整数倍的所有窗的自相关参数 R_{Sum} 的平均值 R_{av},此时判断 $R[0]/R_{av}$ 是否小于 $R_{av} \times 0.08$(结束点判决门限),如果为真,则将该帧标记为结束点,然后新进一帧数据,继续查找下一个。反之新进一帧语音数据,返回 4) 继续处理,直到语音结束。

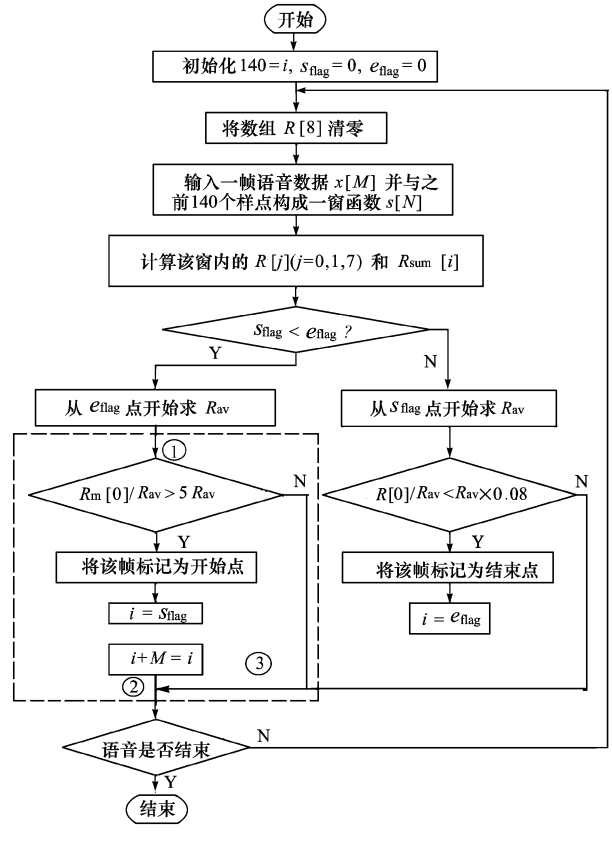

图 1 边界检测分段算法流程图

如果 $s_{flag} < e_{flag}$,从 s_{flag} 开始求与该窗相差 8 的整数倍的所有窗的自相关参数 R_{sum} 的平均值 R_{av},判断 $R[0]/R_{av}$ 是否大于 $R_{av} \times 5$(开始点判决门限),如果为真,则将该帧标记为开始点,然后新进一帧数据继续查找下一个。反之新进一帧语音数据,返回 4) 继续处理,直到语音结束。

7) 音节分割点的判定。上一步过程找到每个音的开始点后,进一步做音节分割,详见流程图 2。

此时 $i = s_{flag}$,判断从 i 开始连续 6 帧的 E, Z, A, B 是否大于清浊音判决门限 E_{th}, Z_{th}, A_{th}, B_{th},如果大于,则将 6 帧中的第一帧作为浊音的开始点 z_{flag},然后继续从下一个音的开始点查找下一个。反之新进一帧数据,重新做清浊音判定,直至语音结束。

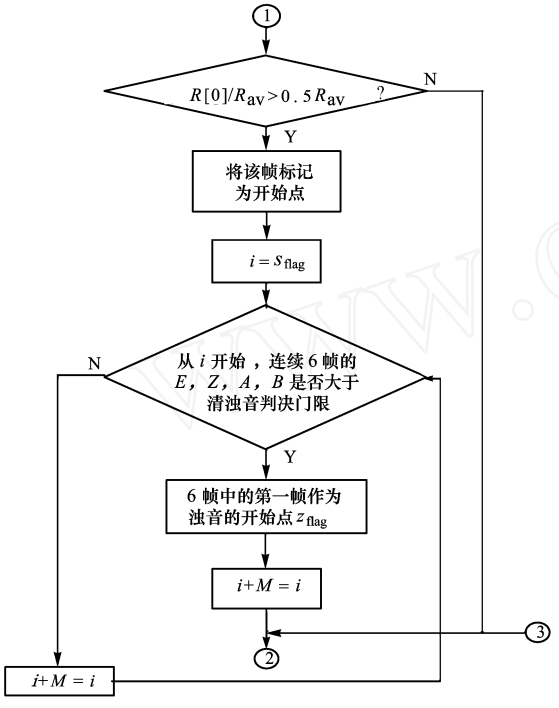

图 2　音节分割算法流程图

3　实验结果与分析

实验的语音信号采用从中科院购买的语音数据库中的一部分。采样率 8 kHz,16 bit 量化。语句 a 为纯净语音"大家都说普通话";语句 b 为带噪语音"向市场经济体制过渡的过程中",信噪比 12.006 9;语句 c 为"采用了国际上最先进的技术",信噪比 10.852。图 3～图 5 是边界检测和音节分割的输出结果图。

在手工标记左右 3 帧范围内出现算法标记则判为正确。误判定义为算法标记左右 3 帧的范围内没有出现手工标记。漏判定义为在手工标记左右 3 帧范围内没有出现算法标记。表 3 是本实验数据的正确率、误判率、漏判率统计。

从实验结果可以看到本文所提的实时分段算法,能在 20 ms 时间内很快捕捉到语音信号的突变点,这样提高语音处理实时性。与手工标记相比,误判率和漏判率都很小,正确率可达到 90 %以上。同时,对于带噪语音,在低信噪比的条件下,仍然能达到 90 %以上的正确率。

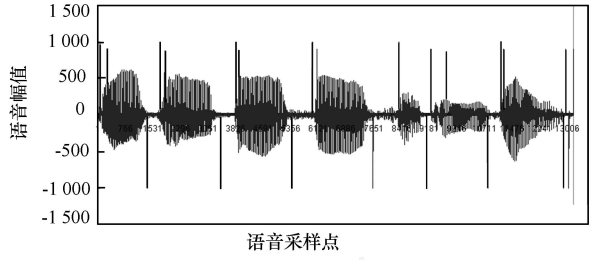

图 3　语句 a 分段结果

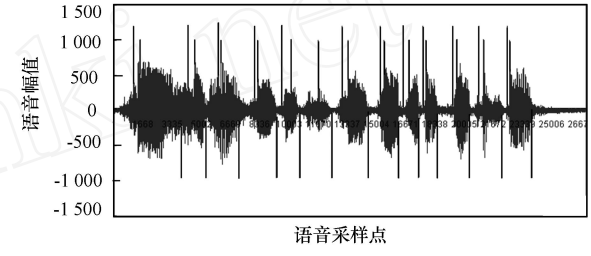

图 4　语句 b 分段结果

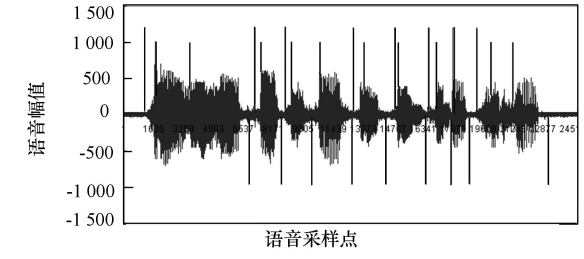

图 5　语句 c 分段结果

上三个图中,横轴上方标记长竖线为语音起点,短竖线为浊音起点,横轴下方标记竖线为语音终点。

表 3　实验数据结果统计

实验数据	正确率/ %	误判率/ %	漏判率/ %
语句 a	96.12	3.1	1.42
语句 b	93.3	3.92	2.81
语句 c	91.56	3.50	5.37

4　结束语

语音的边界检测和音节分割是语音处理中至关重要的阶段,本文所提算法极大地改善了传统单一参数在低信噪比下分段鲁棒性差的缺点,判决精度达到了 90 %以上,同时由于采用了 2.5 ms 为一帧处理语音,缩小了搜索时间,提高了实时性。

参考文献:

[1] 卢艳玲. 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[J]. 电声技术,2005(7):60-62.

[2] 易克初,田斌,付强. 语音信号处理[M]. 北京:国防工业出版社,2000.

[3] 张刚,张雪英,马建芬. 语音处理与编码[M]. 北京:兵器工业出版社,2000.

[4] 郝静. 基于粒计算的语音实时分段算法[D]. 太原:太原理工大学,2008.

(下转第 493 页)

[3] 杨福生. 小波变换的工程分析与应用[M]. 北京:科学出版社,2000.
[4] 陈武凡. 小波分析及其在图像处理中的应用[M]. 北京:科学出版社, 2002.
[5] Vincent L, Soille P. Watershed in digital spaces: an efficient algorithm based on immersion simulations[J]. IEEE Trans Pattern Analysis and Machine Intelligence, 1991, 13(6): 538-598.
[6] Donoho D L, Johnstone M I. Adapting to unknown smoothness via wavelet shrinkage[J]. J A SA, 1995, 90: 1200-1223.

A Method of Image Segment Based on Wavelet Transform and Mathematical Morphology

SHI Jian-fang, ZHANG Fu-jun, HAO Bao-feng
(College of Information Engineering of TUT, Taiyuan 030024, China)

Abstract: In order to restrain over-segmentation phenomenon caused by noise and close textures in conventional watershed transform, a method of image segment based on wavelet transform and mathematical morphology is presented. Wavelet transform is used to remove the mixed noises of infrared image. Morphological opening and closing by reconstruction operators are used to reduce local maximum value caused by gray irregular disturbance and noise on gradient image. Marker-Controlled Watershed algorithm is applied to segment the image. The simulation results show that the proposed method is better than traditional segmentation method and the segment result is effective even without subsequent region merging.

Key words: image segmentation; wavelet transform; mathematical morphology; watershed

(编辑:张红霞)

(上接第 489 页)

A Real-time Border Detection and Syllable Segmentation of Voice Based on Multi-Parameter

DUAN Shu-fei
(College of Information Engineering of TUT, Taiyuan 030024, China)

Abstract: On the basis of the detailed comparison of the five characteristics parameters of voice signal: short-term energy, average rate of zero, relative frequency, relative frequency to the plot, and short-time auto-correlation function, this paper made a real-time voice detection of the border with the syllable segmentation based on multi-parameter. Meanwhile, instead of a current mainstream frame 20 ms, this paper made 2.5 ms a frame to insure detecting signal border in 20 ms, thus improving the real-time processing.

Key words: real-time; multi-parameter; the border detection; syllable segmentation

(编辑:贾丽红)