文章编号: 1000-5013(2008)02-0225-04

清浊音分段判决的递推最小二乘自适应算法

谢维波[1], 王永初[2], 戴在平[1], 吴恬盈[3]

(1. 华侨大学 信息科学与工程学院; 2. 华侨大学 机电工程与自动化学院, 福建 泉州, 362021;
3. 泉州师范学院 理工学院, 福建 泉州, 362000)

摘要: 在"零-能"判决法的基础上, 结合递推最小二乘(RLS)对非平稳信号的自适应跟踪能力, 提出自适应的清浊音分段算法. 算法能够快速实现语音信号清浊音的精确分段, 不需要通过样本集训练进行参数调整. 其自适应能力是在单一话音样本上实现的, 由RLS算法在清音段、浊音段及清浊音段交界处不同的跟踪能力来判别清/浊音段. 与基于阈值的方法不同, 算法基于极值点的识别, 避免各种基于样本集训练的自适应学习算法在泛化能力上的缺陷, 对于不同采样率、说话人、音量、背景噪声等变化因素, 具有较强的自适应处理能力.

关键词: 清浊音分段; 递推最小二乘; 短时过零率; 短时能量
中图分类号: TN 912.3 文献标识码: A

在语音信号预处理中, 清浊音判决的复杂性和准确度对后续的语音处理有很大影响. 目前, 清浊音判决算法都是通过样本集训练来确定阈值, 以提高算法的自适应学习能力, 主要有组合参数法[1]、小波分析方法[2-3]和神经网络方法[4-8]. 在国际电信联盟电信组织(ITU-I)创立的G.729B标准中, 对语音的清浊音判决采用组合参数法(能量、过零率、线谱频率及低通能量), 但参量个数的增加会影响所需的实时性[1]. 语音信号具有非常明显的变化性, 不同话者对于同一音素的声学实现存在明显差异, 在实现清浊音分段时, 必须有效地削减语音信号变化的影响. 基音周期差异、话者音量及环境噪声的变化, 是影响算法非特定人性质和鲁棒性的关键因素. 本文在"零-能"判决法的基础上, 结合递推最小二乘(RLS), 提出了一种不需要通过样本集训练进行参数调整, 可实现语音信号清浊音分段判决的快速自适应算法.

1 "零-能"判决法及清/浊音段的时域特征

1.1 清浊音段的时域特征

清/浊音段的判决是语音信号基于特征提取的模式识别, 因而特征的选取至关重要. 研究表明, 清/浊音段的主要时域特征有短时能量(Short-Period Energy, SPE)和短时过零率(Short-Period Cross-Zero-Rate, SRCZR), 而频域特征则包括线谱频率及低通能量. 鉴于时域频域的对偶性, 文中选取时域特征进行清/浊音段的识别. 对于语音序列$\{x(n)\}$, n时刻的时域特征SPE定义为$E_{SP}(n)$, 即

$$E_{SP}(n)=\sum_{m=-\infty}^{\infty}[x(m)w(n-m)]^2. \tag{1}$$

式(1)中, $w(n-m)$是加权窗函数. 式(1)定义的$E_{SP}(n)$函数对高电平信号非常敏感. 为此, 可以采用定短时平均幅度$M_{SP}(n)$函数, 来度量语音信号的局部能量. 即

$$M_{SP}(n)=\sum_{m=-\infty}^{\infty}\left|x(m)w(n-m)\right|=\sum_{m=n}^{n+N-1}\left|x_w(m)\right|. \tag{2}$$

式(2)中, N为语音帧的长度. 此外, 另一个时域特征SPCZR定义为$R_{SPCZ}(n)$, 即

$$R_{SPCZ}(n)=\frac{1}{2}\sum_{m=n}^{n+N-1}\left|\mathrm{sgn}[x_w(m)]-\mathrm{sgn}[x_w(m-1)]\right|. \tag{3}$$

收稿日期: 2007-10-25
作者简介: 谢维波(1964-), 男, 副教授, 主要从事数字信息处理与理论的研究. E-mail: xwblxf@hqu.edu.cn.
基金项目: 福建省自然科学基金资助项目(A0540005)

在式(3)中, $\mathrm{sgn}[x_w(m)]=\begin{cases}1, & x_w(m)\geqslant 0,\\ -1, & x_w(m)<0.\end{cases}$

1.2 “零-能”判决法

“零-能”判决法利用SPE和SPCZR作为特征来进行检测. 对于清音段, SPCZR比较高, SPE比较低; 对于浊音段, SPCZR比较低, SPE则比较高. 其第 n 帧语音定义为

$$X_n=\{x(m)\mid m=(n-1)\tau+1,(n-1)\tau+2,\cdots,n-1)\tau+N\}. \tag{4}$$

式(4)中, τ 为帧移. 对第 n 帧语音 X_n, 分别设定 $E_{SP}(n)$ 的阈值为 $\delta_E(n)$, $R_{SPCZ}(n)$ 的阈值为 $\delta_Z(n)$. “零-能”判决法的规则为: 若 $R_{SPCZ}(n)<\delta_Z(n)$ 且 $E_{SP}(n)\geqslant\delta_E(n)$, 则第 n 帧语音 X_n 为浊音帧; 反之, X_n 为清音帧. $\delta_E(n)$ 和 $\delta_Z(n)$ 阈值的设定对于“零-能”法起到决定性作用. 如果 $\delta_E(n)$ 选得太小或者 $\delta_Z(n)$ 选得过大, 就会把清音归到浊音里去; 反之, 如果 $\delta_E(n)$ 选得过大或者 $\delta_Z(n)$ 选得太小, 就会把浊音划分到清音里去. 由于语音信号的复杂性, 即使所选阈值 $\delta_E(n)$ 和 $\delta_Z(n)$ 刚好能够给出正确的清浊音划分, 在清浊音段的非分界处常常伴随出现多个“虚假”的清浊音划分. 图1给出了语音信号“k”的SPE判决. 从图1可以看出, 该阈值能较好地划分清浊音(在横坐标为2 000处是正确的清浊音划分, 左面是清音帧, 右面是浊音帧). 然而, 从图1的右上角局部放大图可以看到, SPE的曲线在阈值线上下波动, 无论如何调整阈值, 总是无法避免多个分界线的产生, 导致图2的判决结果. 即在横坐标[7 000, 9 000]的区间多了3条“虚假”的清浊音划分. 图2纵坐标为语音信号的幅度值 A. 结合“短时能量”和“过零率”的判决结果, 本文的判决法依据“判决误差”的最大值, 能够唯一确定清浊音的分界线.

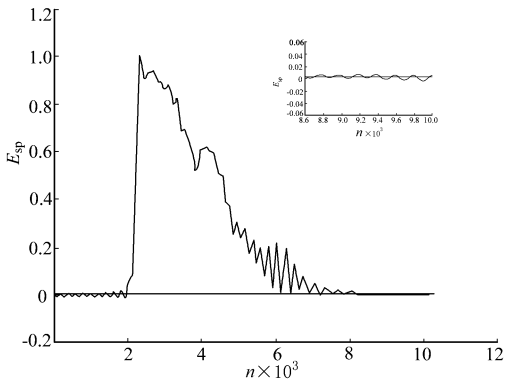


图1 “k”的短时能量判决

Fig. 1 ‘k’ of decision based on SPE

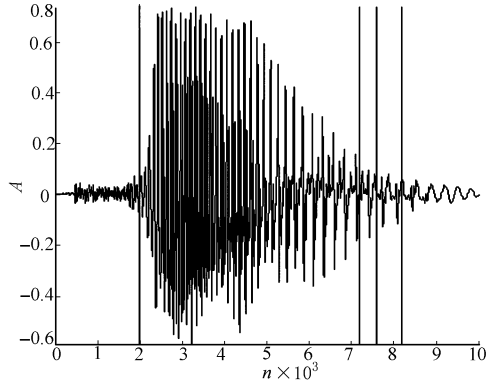


图2 “k”的“零-能”法判决

Fig. 2 ‘k’ of decision based on zero-energy

2 清浊音判决的RLS算法

2.1 RLS算法

在 $n-1$ 时刻滤波器的基础上, n 时刻的滤波器参数可以根据 n 时刻到来的数据 $x(n)$ 进行更新. 根据最小二乘(LS)准则, 滤波器的目标函数为

$$Q(n)=\sum_{i=1}^{n}e^2(i\mid n), \tag{6}$$

式(6)中 $e(i\mid n)=x(i)-\hat{x}(i)=x(i)-X_N^{\mathrm{T}}(i)W_N(n)$. $e^2(i\mid n)$ 是 n 时刻的权 $W_N(n)$ 对 i 时刻的数据 $x(i)$ 进行跟踪预测所得的误差, $X_N^{\mathrm{T}}=[x(i-1),\cdots,x(i-N)]$, $W_N(n)=[w_1(n),\cdots,w_N(n)]^{\mathrm{T}}$. 其中, $W_N(n)$ 为 n 时刻滤波器的系数, N 为滤波器的阶数. 一般地, 可在式(6)中添加 λ^{n-i} 因子(λ 一般取值在0.950 0到0.999 5之间), 使整体预测误差能量 $Q(n)=\sum_{i=1}^{n}\lambda^{n-i}e^2(i\mid n)$ 最小. 其中, λ 为遗忘因子, 对时间较近的数据加以较大的权, 时间较远的数据其权按指数减小, 从而加强对非平稳信号的适应性.

2.2 基于RLS的清浊音判决算法

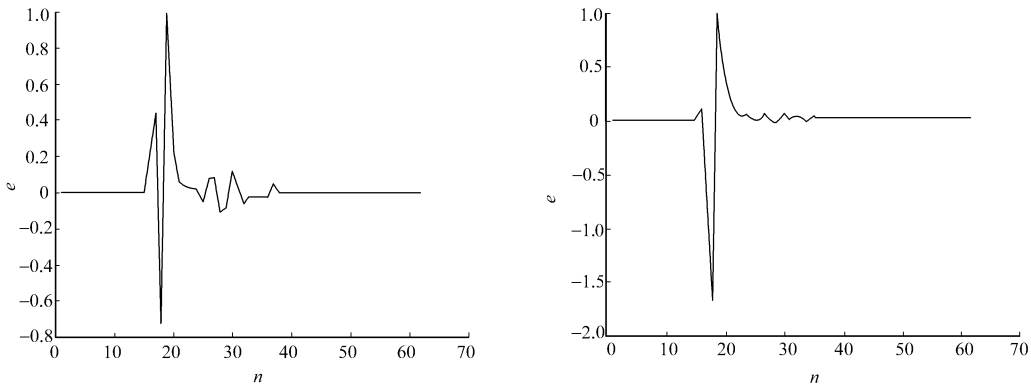
研究表明, 在清音段或浊音段内相邻帧间对应分量是强正相关的. 随着帧间距离的增加, 相关性也随之有所减小, 但还有很强的相关性, 并且还可看到相邻帧间的线性趋势. 语音帧间的依赖关系可以用线性模型来体现[3], 由于语音信号在适当的分段内相邻的帧之间具有线性关系, 可以考虑以帧为单位,

用线性预测的方法对语音信号进行预测. 但是语音信号只有在短时间段(10~ 30 ms)内具有平稳性, 相邻帧之间的线性关系持续时间很短. 显然在较长的时间段内, 用线性预测的方法来进行预测是不合适的, 而用递推最小二乘 RLS 对整段非平稳语音信号进行自适应预测是可行的. 语音在清音段和浊音段内是各自相对平稳的, 清/浊音段的特征数据 $E_{SP}(n)$ 和 $R_{SPCZ}(n)$ 也相对平稳, RLS 将给出良好的预测跟踪. 语音信号的非平稳性, 主要体现在清音段和浊音段的交界处, 在清/浊音段的交界处, RLS 算法的预测跟踪能力必然相对地减弱. 这种"相对性"正是算法不必设定阈值的根源所在.

RLS 算法的目标是对"非平稳性"的良好跟踪, 当然其对"平稳性"的跟踪也相对更好. 本文算法将这种"相对性"应用于清/浊音段的划分, 并将跟踪预测建立在 SPE 和 SPCZR 反映"清/浊音段"本质的特征粒度之上. 后者是十分重要的, 否则 RLS 的跟踪将不知所为. 粒度计算思想表明, 信息的提取要有一个恰当的"粒度". 基于 RLS 的清浊音判决算法有如下 5 个方面. (1) 将语音信号进行分帧, 每 10 ms 为一帧, 每帧位移为 5 ms. (2) 对每帧信号, 统计 SPCZR 和 SPE. (3) 对所得的 SPCZR 和 SPE 分别进行 RLS 跟踪预测. 设 p 为滤波器的阶数, 用前 p 帧的 $E_{SP}(n)$, $n=1, \cdots, p$, 或者 $R_{SPCZ}(n)$, $n=1, \cdots, p$, 来预测 $p+1$ 帧的 $E_{SP}(p+1)$ 或 $R_{SPCZ}(p+1)$. (4) 由式(6)分别计算 $Q_{SPE}(n)$ 和 $Q_{SPCZR}(n)$. 其中, $Q_{SPE}(n)$ 是基于 SPE 的整体跟踪预测误差能量, $Q_{SPCZR}(n)$ 是基于 SPCZR 的整体跟踪预测误差能量. 显然在清/浊音段的交界处, $Q_{SPE}(n)$ 和 $Q_{SPCZR}(n)$ 会出现最大值, 即后验误差最大. (5) 对所得 $Q_{SPE}(n)$ 和 $Q_{SPCZR}(n)$ 的极大值点进行逻辑与运算的判断, 滤除"虚假"的清/浊音段分界点, 得到更为精确的清/浊音判决.

3　实验结果

基于 SPCZR 和 SPE 的 RLS 分析误差, 如图 3 所示. 图 3 中对于语音信号的前面部分(清音), RLS



(a) SPCZR　　(b) SPE

图 3　"b"的 RLS 预测误差

Fig. 3　Prediction of the speech 'b'

算法能够较好地进行跟踪. 到了清/浊音的分界处, 跟踪误差出现了剧烈的抖动, 该抖动延续直到浊音段内逐渐消失, RLS 继续良好地跟踪浊音信号. 图 3 很好地诠释了本算法的基本思想. 对比图 4 和图 2 可知, 基于 RLS 的清/浊音分段方法, 比传统的"零-能"判决法得到更好的辨识效果. 图 2 中的"零-能"判决法需要设定阈值. 当设定的阈值调整到恰能正确地划分清/浊音段时, 该语音段内有多个"非清浊音分界点"的"零-能"值超过了所设定的阈值, 出现了多个虚假的"清浊音分界点". 因而, "零-能"判决法不能很好地进行清/浊音段的判决. 基于 RLS 的清/浊音分段方法, 是通过求取判决误差的最大值来确定清浊音的分界点, 因而能够唯一确定清浊音的分段. 基于 RLS 算法和"零-能"法的语音信号判决比较, 如表 1 所示. 其信号帧长为 10 ms, 每帧位移为 5 ms. 由 RLS 算法确定的最优滤波器阶数为 2 或 3. 从表 1 中可看出, 基于 RLS 的清/浊音分段方法能满足语音

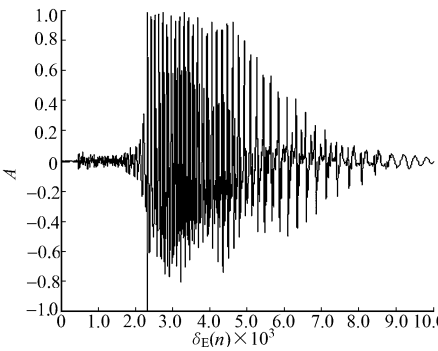


图 4　"k"的 RLS 判定

Fig. 4　'k' of decision based on RLS

处理的实时性要求,具有时间上的优势.

表1 不同判决方法的语音信号时间列表(单位:s)

Tab.1 Comparing on the times of taking by two decisions(unit: s)

语音段	b	c	d	g	j	k	p
RLS方法	0.062 0	0.125 0	0.015 0	0.031 0	0.016 0	0.062 0	0.015 0
“零-能”法	6.235 0	7.203 0	2.063 0	3.469 0	3.437 0	3.391 0	1.907 0

4 结束语

本文算法是在传统的SPCZR和SPE判决法的基础上,针对语音信号“非平稳性”的特征,引入递推最小二乘RLS,实现清/浊音段的在线判决.方法的优点在于,无需对语音信号进行噪声统计得到SPCZR或者SPE的阈值,避免了阈值的选取对于判决准确性的影响.该算法具有很高的时间效率,适用于语音信号的在线处理.

参考文献:

[1] 郭 英,李雪娇.一种组合参数的语音信号清/浊音判决方法[J].空军工程大学学报:自然科学版,2002,3(4):18-20.

[2] 王毓芳.一种自适应的汉语普通话音节清/浊音分段方法[J].北京航空航天大学学报,2001,27(4):409-412.

[3] 马 霓,韦 岗,应益荣,等.基于线性预测和小波变换的语音基音周期检测新算法[J].西北建筑工程学院学报:自然科学版,1997(2):36-42.

[4] QI Ying-youg, HUNT B R. Voiced-unvoiced-silence classifications of speech using hybridfeatures and a network classifier[J]. IEEE Transactions on Speech and Audio Processing, 1993,1(2):250-255.

[5] ATAL B, RABINER L. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition[J]. IEEE Transactions on Signal Processing, 1976, 24(3):201-212.

[6] BASSON J A L, PREEZ J A. Adaptive estimation of speech parameters[C]// Communications and Signal Processing, Proceedings of the 1994 IEEE South African Symposium on, 1994:177-182.

[7] RABINER L R, SAMBUR M R. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem[J]. IEEE Transactions on Signal Processing, 1977,25(4):338-343.

[8] 周志杰,胡光锐.采用非线性网络实现清浊音判决[J].南京航空航天大学学报,1998,30(1):47-51.

Adaptive Voiced/Unvoiced Segmentation Based on RLS

XIE Wei-bo[1], WANG Yong-chu[2], DAI Zai-ping[1], WU Tian-ying[3]

(1. College of Information Science and Engineering, Huaqiao University;
2. College of Mechanical Engineering and Automation, Huaqiao University, Quanzhou 362021, China
3. College of Science and Engineering, Quanzhou Normal University, Quanzhou 362000, China)

Abstract: An adaptive voiced/unvoiced segmentation based on the traditional short-time analysis, with the adaptive tracking capacity of recursive least square (RLS) to the non-steady signal, has been presented. The algorithm can rapidly realize precise voiced/unvoiced segmentation, without parameter-adjustment by samples training. The adaptability comes from a single pronunciation sample, and deciding voiced/unvoiced segmentation based on the different tracking capacity of RLS in voiced/unvoiced section and the intersection point. It is different from the methods based on threshold, the algorithm based on recognizing the extreme value can avoid the drawback of various adaptive learning algorithms in generalization, which can better adapt for various varying factors of different sampling rate, speaker, volume, background noise.

Keywords: voiced/unvoiced segmentation; recursive least square; short-period cross-zero-rate; short-period energy

(责任编辑:黄仲一 英文审校:吴逢铁)