文章编号:1002-8684(2005)07-0052-03

# 基于短时能量的语音端点检测算法研究

·论文·

张仁志，崔慧娟
(清华大学 电子工程系 微波与数字通信国家重点实验室，北京 100084)

【摘　要】研究了噪声环境下,利用短时能量为特征进行语音端点检测的问题。在采用短时全带能量为特征的基础上,提出的算法将短时高频能量作为辅助特征,同时使用了最优边沿检测滤波以及双门限-三态转换判决机制,从而保证了算法在噪声环境下的端点检测准确性和对信号绝对幅度变化的稳健性。实验结果表明,与传统的能量阈值法以及 G.729 中使用的 VAD 算法相比,提出的算法在噪声环境下具有更好的性能,是一个简单、高效和稳健的语音端点检测算法。

【关键词】端点检测；短时能量；边沿检测滤波；三态转换判决机制
【中图分类号】TN912　　【文献标识码】A

**Speech Endpoint Detection Algorithm Analyses Based on Short-term Energy**

ZHANG Ren-zhi, CUI Hui-juan
(State Key Laboratory on Microwave & Digital Communications,
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

【Abstract】 This paper analyzes speech endpoint detection based on short-term energy feature in the presence of noise. Besides short-term full band energy feature, short-term high band energy is employed as an accessorial feature in the proposed algorithm. It also uses an optimal edge detection filter plus a three-state transition and judgment mechanism based on double thresholds, which ensure the accuracy in noisy environment and the robustness to changes in absolute levels. Experiments show that the proposed algorithm outperforms traditional energy threshold and G.729 VAD for speech endpoint detection in noisy environments and proves its accuracy, simplicity and robustness.

【Key words】 endpoint detection; short-term energy; edge detection filter; three-state transition and judgment mechanism

## 1　引言

语音信号的端点检测在语音识别、语音增强以及语音编码等语音信号处理系统中有着重要的应用[1]。正确地检测语音信号的端点，不仅可减少语音处理的运算量,还可有效地提高系统的性能。

一般来讲，不同的应用需要根据其对计算准确度、算法复杂度、稳健性以及响应时间等的不同要求来选择合适的算法。通常使用的方法有能量阈值[2]、基音检测[3]、频谱分析、倒谱分析[4]以及 LPC(Linear Prediction Coefficients)预测残差[5]等。

传统的能量阈值法根据语音信号的短时能量,采用双门限判决的方法进行端点检测，有时会辅以短时过零率信息进行判决。这些方法在高信噪比(SNR)时具有良好的性能,而在低信噪比时性能急剧恶化,而且缺乏对信号绝对幅度变化的稳健性。然而,语音信号处理系统通常工作在不同的噪声环境下，在语音处理系统中采用的端点检测应当适应可能出现的各种情况,以便在实际应用中达到好的性能。

笔者提出了一种基于短时能量特征的简单、有效和稳健的语音端点检测新算法。新算法在采用短时全带能量为特征的基础上，将短时高频能量作为辅助特征,同时使用了最优边沿检测滤波[6]以及合理的双门限-三态转换判决机制[7],从而保证了算法在噪声环境下的端点检测准确性和对信号绝对幅度变化的稳健性。实验表明,笔者提出的算法具有优越的性能。

## 2　算法描述

### 2.1　整体算法描述

算法的整体结构框图见图 1。根据经过预处理的输入语音信号，可得到短时全带对数能量和短时高频

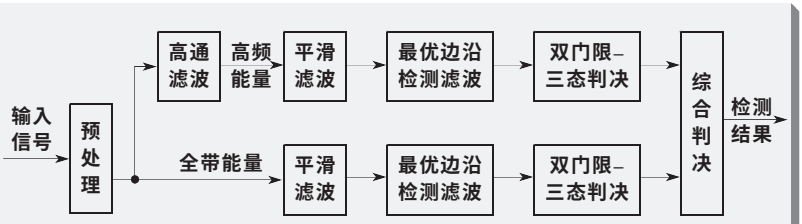


图 1　语音信号端点检测的整体结构

对数能量。短时全带及高频对数能量序列经过平滑滤波、最优边沿检测滤波和双门限-三态转换判决，分别得到全带能量端点判决和高频能量端点判决，然后经过以全带能量判决为基础，并辅以高频能量判决的综合判决就可得到最终的端点检测结果。

采用短时能量为特征进行语音信号的端点检测的传统方法在信噪比较低时，无法取得好的检测结果，主要原因是：(1)低信噪比时，语音与噪声的短时全带能量特征区分不明显；(2) 缺少一个有效的端点检测、判决方法。因此，笔者(1)根据语音信号的谱特征，引入短时高频能量，作为辅助特征来确保语音信号过渡信息的有效检测，提高正确检测率；(2)利用最优边沿检测滤波可靠地检测短时能量序列的边沿，并制定了合理的三态转换判决机制来保证端点检测结果的正确性与可靠性。

### 2.2　能量特征

短时全带能量在高信噪比时是区分语音与噪声的有效特征，但是其在低信噪比时，不能有效地区分语音信号低能量的过渡信息和噪声，因而无法实现有效的端点检测。语音信号的能量主要集中于 60~3 400 Hz 之间，而过渡信息的清音段能量则多集中在高频区域，同时考虑到有色噪声的能量集中于低频区域，笔者在采用短时全带能量的同时，引入了短时高频能量作为辅助特征

$$g(t)=10\lg\sum_{j=n_t}^{n_t+I-1}o^2(j) \tag{1}$$

$$g_{HP}(t)=10\lg\sum_{j=n_t}^{n_t+I-1}o_{HP}^2(j) \tag{2}$$

其中，$o(j)$是经过截止频率为 140 Hz 的高通滤波的语音采样数据，$O_{HP}(j)$是 $o(j)$经过高通滤波后的语音采样数据，高通滤波器采用的是 9 阶 Chebyshev Ⅱ型滤波器，截止频率为 2 000 Hz；$g(t)$是当前帧的全带对数能量，$g_{HP}(t)$是当前帧的高频对数能量；$t$ 表示当前帧为第 $t$ 帧，$I$ 是窗长，$n_t$ 是第 $t$ 帧的第一个样点在所有样点中的位置。

### 2.3　平滑滤波

为减少噪声抖动引起的可能误判，提高系统的稳健性，笔者在进行边沿检测滤波之前引入了平滑滤波

$$g'(t)=\frac{g(t-1)+g(t)+g(t+1)}{3} \tag{3}$$

3 帧联合的平滑滤波可以较好地平滑对数能量特性曲线，从而在一定程度上消除因噪声抖动引起的误判。

### 2.4　最优边沿检测滤波

低信噪比时，语音与噪声的短时能量特征区分不明显，通过简单的双门限判决将无法有效地检测端点，而且门限值需要根据背景噪声的绝对幅度进行设定。为在低信噪比时能够有效地区分语音和噪声，且解决门限值对背景噪声绝对幅度的依赖性，需要对短时能量特征进行一定的变换。

笔者采用了如下形式的最优边沿检测滤波器[6]以满足上述要求。

$$f(x)=e^{Ax}[K_1\sin(Ax)+K_2\cos(Ax)]+e^{-Ax}[K_3\sin(Ax)+K_4\cos(Ax)]+K_5+K_6e^{Sx} \tag{4}$$

$$h(x)=\begin{cases}f(x) & -W\leqslant x\leqslant 0\\ -f(-x) & 1\leqslant x\leqslant W\end{cases} \tag{5}$$

其中，$A$ 和 $K_i$ 是滤波系数，$W$ 为滤波器阶数。当 $W=7$ 时，$S=1$，$A=0.41$，$[K_1,K_2,K_3,K_4,K_5,K_6]=[1.583, 1.468,-0.078,-0.036,-0.872,-0.56]$[6]；笔者选用了 13 阶滤波器，$S=7/W=0.538\ 5$，$A=0.41S=0.220\ 8$，$K_i$ 取值不变，$h(i)$如图 2 所示。

设定了滤波器参数以后，就可使用

$$F(t)=\sum_{i=-W}^{W}h(i)g'(t+i) \tag{6}$$

的滑动窗的形式进行边沿检测滤波。

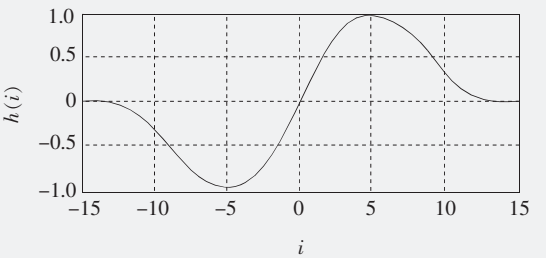


图 2　最优边沿检测滤波器

### 2.5 双门限-三态转换机制

根据语音信号的特点，笔者将语音信号划分为3种状态：静音、语音及过渡，并制定了合理状态转换判决机制。三态的划分使得判决更符合语音信号的特点，合理的状态转换判决机制将确保准确、有效地检测出语音端点。

双门限-三态转换判决图表见图3。起始状态是静音或语音，可以3种状态中的任意一种结束。状态的转换及端点的检测需要根据 $F(t)$ 与预先设置的判决门限 $T_L$ 和 $T_U$ 的比较结果，并辅以过渡常量来判定。

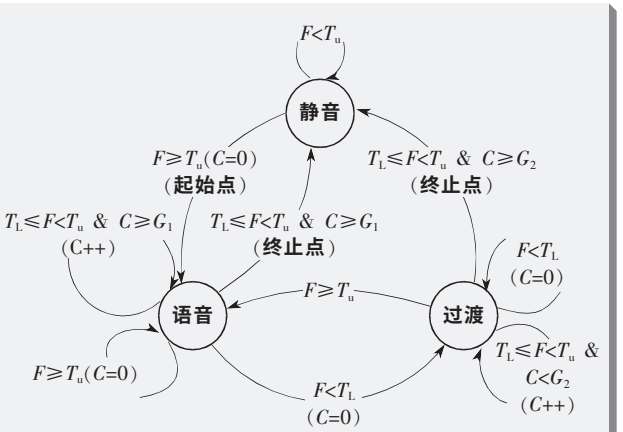


图3 双门限-三态转换判决图表

如图3所示，输入是 $F(t)$，输出是起始点或终止点标识；$C$ 是一个判决计数器，用于记录 $F(t)$ 连续处于2个阈值之间的帧数；$G_1$ 和 $G_2$ 是用于判定语音的终止点过渡常量。$G_1$ 的引入是为了防止因为噪声扰动而导致 $F(t)$ 暂时突破 $T_U$，并随后一直处于 $T_L$ 和 $T_U$ 之间而引起的长时间误判，将 $G_1$ 设置为一个合理的数值，可以及时、有效地跳出上述误判。$G_2$ 是检测终止点与语音实际终止点之间的帧数差，通过适当的设定值可以充分地保留语音的过渡信息。

如图4所示，起始时，状态转换判决图表起始处于静音态，到 $F(t)\geqslant T_U$[图4(b)中的A点]时，输出语音起点标识[图4(a)中的左垂直线]，并转换至语音态；到 $F(t)<T_L$[图4(b)中的B点]时，转换至过渡态，$C$ 开始计数；$C$ 经过若干次置零，直到达到图4(b)中的B′点；当 $C=G_2$[图4(b)中的C点]时，输出语音终点标识[图4(a)中的右垂直线]，并转换至静音态。如果在语音态，且 $F(t)$ 持续处于 $(T_L,T_U)$ 的计数大于 $G_1$，则立即转换至静音态，并输出终止点标识，以防止长时间误判；如果在过渡态，且出现 $F(t)\geqslant T_U$，则结束过渡态，跳转回语音态。

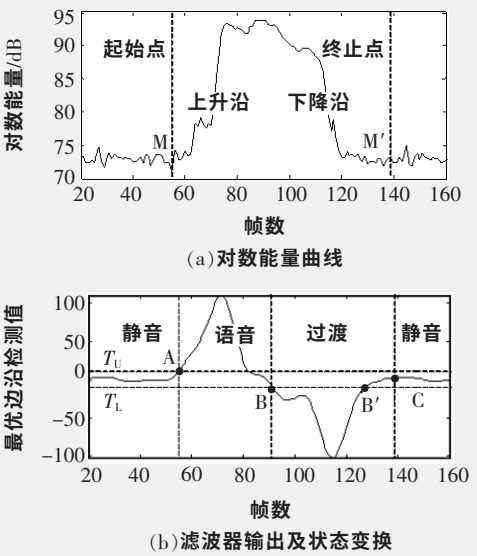


图4 端点检测示例

$T_L,T_U$ 以及 $G_1$ 和 $G_2$ 通过实验分析设定。由于最优边沿检测的滤波结果 $F(t)$ 与噪声绝对幅度无关，因此，$T_L$ 和 $T_U$ 不需要根据背景噪声的绝对幅度以及信噪比的不同进行调整。适当地设置短时全带能量判决和短时高频能量判决的 $T_U,T_L$ 以及 $G_1,G_2$，可以满足不同的应用要求。一般来讲，短时高频能量判决的 $T_L$ 和 $G_1$ 设低一些，$T_U$ 和 $G_2$ 设高一些，可有效地检测过渡信息，并避免将噪声扰动误判为语音。另外，在实际应用中，可以使用不同阶数的边沿检测滤波器分别进行上升沿和下降沿的检测，也可以将检测起始点前推若干帧作为语音信号的实际起始点，以适应更恶劣的情况。

## 3 实验

笔者将标准语音库的语音材料与NOISEX-92噪声库中白噪声进行合成作为语音信号端点检测的测试材料，并在不同信噪比的情况下，进行了端点检测对比测试的实验。

实验结果表明，笔者提出的算法在低信噪比环境下仍可有效地检测出语音端点。相比较而言，在低信噪比环境下，G.729的VAD算法[8]以及传统短时能量阈值法都有明显的切音现象，该算法的短时全带能量判决也有一定程度的切音。图5是不同信噪比的白噪声环境下，笔者提出的算法(NEFH)、笔者提出的算法的短时全带能量判决（NEOF)、G.729VAD的修正算法(G.729）以及传统的能量阈值法(EOF)（下转第59页）

应幅值的变化，而相位变化很小。只有在传声器严重偏离原来位置，盒子拉伸到几乎脱离的情况下才会出现相位偏差超过90°的。当出现这种情况时，采用加上鲁棒约束的算法，对于避免系统出现发散的情况有所帮助。而对于通常的对象变化，无论是仿真还是试验，2种算法都有很好的控制效果。另外加强鲁棒性后，算法对于较高频率周期信号的控制有一定效果。

**参考文献**

[1] Adaptive Active Noise Control for Headphone Using the TMS320C30 DSP: Application Report[R], literature number SPRA160, Texas Instruments, 1997.

[2] 陈克安. 有源噪声控制[M]. 北京：国防工业出版社，2003.

[3] Design of Active Noise Control System With the TMS320 Family: Application Report [R], literature number SPRA042, Texas Instruments, 1996.

[4] Boaz Rafaely. Feedback Control of Sound[D]. Doctor's thesis of University of Southampton, 1997.

[5] Doyle, J.C., Francis B.A., Tannenbaum A.R. Feedback control theory. Maxwell MacMillan International. 1992.

[6] Elliott, S.J., Sutton, T.J.. Performance of feedforward and feedback systems for activecontrol. IEEE Transactions on Speech and Audio Processing, 1996, 4(3): 214—223.

[7] Rafaely B., Elliott S.J. An Adaptive and Robust Feedback Controller for Active Control of Sound and Vibration. Proceedings of CONTROL'96 conference, University of Exeter, UK, 1996.1 149—1 153.

**作者简介**

**王克杰**，硕士研究生，主要研究方向为基于智能结构的有源耳罩研究.

**张奇志**，教授，硕士生导师，主要研究方向为有源噪声、振动控制.

**周雅莉**，副教授，主要研究方向为智能结构.

**[收稿日期] 2005-06-24**

（上接第54页）

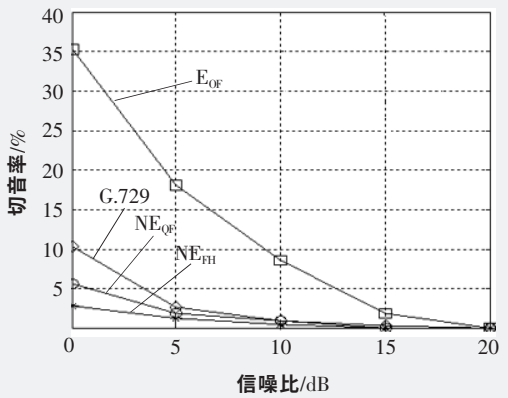


图5 不同信噪比(dB)白噪声环境下的各算法端点检测切音率(%)的比较

的语音端点检测性能对比。

## 4 结论

笔者提出的，以短时全带能量并辅以短时高频能量为特征参数，通过最优边沿滤波器进行边沿检测滤波，并引入基于双门限的三态转换判决机制的语音端点检测算法，与其它一些端点检测算法相比，更加简单、有效和稳健，并且能够在低信噪比环境下保持良好的性能。实验也证明了该算法的优越性能。

**参考文献**

[1] 杨行峻，迟惠生，等. 语音信号数字处理. 北京：电子工业出版社，1995.

[2] Wilpon J G, Rabiner L R, Martin T. An Improved Word-detection Algorithm for Telephone-quality Speech Incorporating Both Syntactic and Semantic Constraints. AT&T Bell Labs. Tech. J., 1984, 63: 479—498.

[3] Chengalvarayan R, Robust Energy Normalization using Speech/nonspeech Discriminator for German connected Digit Recognition. Proc. Eurospeech'99, Budapest, Hungary, 1999. 61—64.

[4] Haigh J A, Mason J S. Robust Voice Activity Detection Using Cepstral Features. in Proc. IEEE TENCON, 1993. 321—324.

[5] Deller J R, Proakis J G, Hansen J H L. Discrete-Time Processing of Speech Signals [M]. New York: Macmillan, 1993.

[6] Petrou M, Kittler J. Optimal Edge Detectors for ramp Edges, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(5):483—491.

[7] Qi Li, Jinsong Zheng, Tsai A, Zhou Qiru. Robust Endpoint Detection and Energy Normalization for Real-time Speech and Speaker Recognition, IEEE Transactions on Speech and Audio Processing, 2002, 10(3): 146—157.

[8] Annex B to ITU-T Recommendation G.729: A silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70, 1996.

**[收稿日期] 2005-04-27**