基于临界带功率谱方差的端点检测

张春雷，曾向阳，王曙光
(西北工业大学航海学院，西安 710072)

摘要：端点检测作为语音信号处理的关键技术，其准确性直接影响到语音识别系统的计算复杂度和识别能力。在人耳听觉特性理论研究的基础上，利用语音段和背景噪声段临界带功率谱上的差异，提出了一种基于临界带功率谱方差的端点检测方法。通过自适应门限值的选取，该方法对背景噪声具有良好的跟踪性能。在不同的信噪比条件下，进行了端点检测实验。结果表明：该方法与传统的短时能量和短时平均过零率方法、谱熵方法相比，可以有效降低背景噪声的影响，具有更好的鲁棒性和正确率。

关键字：端点检测；临界带；功率谱方差；自适应门限
中图分类号：TN912.34 **文献标识码**：A **文章编号**：1000-3630(2012)-02-0204-05
DOI 编码：10.3969/j.issn1000-3630.2012.02.017

A voice activity detection algorithm based on the variance of critical band power spectrum

ZHANG Chun-lei, ZENG Xiang-yang, WANG Shu-guang
(*College of Marine Engineering*, *Northwestern Polytechnical University*, *Xi'an* 710072, *China*)

Abstract: As the key technology of speech signal processing, the accuracy of voice activity detection directly affects the computational complexity and identification capabilities of speech recognition system. Based on theoretical research of human auditory characteristics, this paper presents a voice activity detection algorithm which involves the variance of critical band power spectrum. By using adaptive threshold, the presented method has a good performance on the tracking of background noise. Experiments are taken at different signal to noise ratios. The results show that comparing with short time features-based algorithm and entropy based algorithm, this method can effectively reduce the influences of background noise. Furthermore, it also has better robustness and higher accuracy.

Key words: voice activity detection; critical band; variance of power spectrum; adaptive threshold

0 引 言

端点检测就是从包含语音的一段声信号中找出语音的起点及终点，剔除背景噪音段，保留语音段。准确的端点检测不仅可以减少系统的数据存储量和处理时间，降低硬件消耗，提高处理的实时性，而且能排除背景噪音段的干扰，从而使后续的语音分析、合成、增强和识别的性能大大的提高。

端点检测方法在近年来取得较快的发展，目前端点检测的常规方法分为两类[1]：一类基于语音信号的时域处理，如短时能量、短时幅度谱以及短时平均过零率等，这一类方法原理比较简单和直观，运算量小，在高信噪比状态下端点检测的效果比较好，但在噪声环境下抗干扰能力和稳定性下降较快；另一类方法是基于语音信号的频域处理，如倒谱系数、谱熵、近似熵[2]以及子带能量特征等，此类方法在噪声环境下有一定的能力，但硬件消耗较大，处理时间较长，且在强噪声环境效果仍然较差。此外，近年来研究人员提出许多新的、综合性的端点检测方法，如基于多特征语音端点检测、新型分形理论、时频分解以及基于经验模态分解和 Teager 峭度的语音端点检测等[3-6]。

仅仅利用声音信号时域和频域特性的端点检测方法已日趋成熟，但是在选择合适的声学特征来提高端点检测正确率这一方面的研究还有待进一步加强[7]。结合前人的研究成果[8]，本文提出了一种基于临界带功率谱方差的端点检测算法。实验结果表明，在背景噪声频谱较为平坦的情况下，该方法具有良好的端点检测效果。

本文介绍了两种常用的端点检测方法：短时能量与短时平均过零率结合的方法和功率谱熵法[9]，提出基于临界带功率谱方差的端点检测原理和算

收稿日期：2011-04-07；**修回日期**：2011-07-08
基金项目：教育部新世纪优秀人才支持计划基金(NCET-08-0459)；西北工业大学本科毕业设计重点扶持项目
作者简介：张春雷(1989－)，男，江苏海门人，研究方向为声信号处理。
通讯作者：张春雷，E-mail: heimanba89@yahoo.cn

法，然后在不同信噪比条件下对三种方法的效果进行了对比分析。最后，对本文方法的适用环境和局限性进行了讨论。

1　常用端点检测方法原理

1.1　短时能量和短时平均过零率

时域语音信号通过加窗分帧处理后得到第 n 帧的语音信号为 $x_n(m)$，$x_n(m)$ 的短时能量用 E_n 表示，公式为

$$E_n=\sum_{m=0}^{N-1}x_n^2(m) \tag{1}$$

短时过零率表示一帧语音信号 $x_n(m)$ 波形穿过横轴(零电平)的次数。对于连续语音信号，过零即意味着时域波形通过时间轴，而对于离散信号，如果相邻的取样值改变符号则称为过零。过零率就是样本改变符号的次数。定义语音信号 $x_n(m)$ 的短时过零率为

$$Z_n=\frac{1}{2}\sum_{m=0}^{N-1}\left|\mathrm{sgn}[x_n(m)]-\mathrm{sgn}[x_n(m-1)]\right| \tag{2}$$

式中，sgn[] 是符号函数，即：

$$\mathrm{sgn}[]=\begin{cases}1, & (x\geq 0)\\ -1, & (x<0)\end{cases} \tag{3}$$

过零率实际上是信号频率的度量，高频就意味着较高的过零率，低频意味着较低的过零率。通常发清音时多数能量落在较高的频率上的，而浊音时具有较低过零率。虽然这种关系没有精确的数值关系，仅仅是相对的，但由于人的发音很多是以清音开始的，所以利用过零率来判断语音的起始段是可行的。

短时能量和过零率的结合在一定条件下具有很好的实用价值，因而经常被使用。

1.2　谱 熵

基于谱熵的端点检测引入了信息论上熵函数的概念。由于语音谱的固有特性使得谱熵能够有效地区分语音信号和噪声信号。求熵的步骤如下：

(1) 求信号的短时功率谱；

(2) 利用求得的功率谱计算概率密度函数：

$$p_i=\frac{|S(f_i)|^2}{\sum_{k=1}^{(N/2+1)}|S(f_k)|^2} \tag{4}$$

$S(f_i)$ 为信号经 FFT 变换之后的频谱分量，考虑到 FFT 的对称性，计算 p_i 时只要一半的分量点。进而可以定义谱熵为：

$$H(x)=-\sum_{i=1}^{N/2+1}p_i\log p_i \tag{5}$$

这里 x 为每一帧的序号，且规定如果 $p_i=0$，则 $p_i\log p_i=0$。

在实际应用中，通常会对其进行一些改进。如：设计一个带通滤波器将人耳不那么灵敏的频率分量滤去；设计一个谱熵平滑函数以防止突然出现噪声等造成的谱熵值的跳变或不连续。在这些改进的基础上，将每一帧语音信号对应的谱熵与事先设定的合适的门限值进行比较，以进行语音帧和非语音帧的区分。

2　基于临界带功率谱方差的端点检测

2.1　临界带功率谱方差的原理

利用短时傅里叶变换求得语音信号的短时谱，是按实际频率分布的的线性谱。然而人耳所听到的声音的高低与声音的频率并不成线性正比关系。为了模拟人耳的听觉特性，研究人员尝试将实际的线性频谱转化为更符合人耳听觉特性的临界带频率分布，取得良好的效果，增强了语音信号处理系统的性能。

本文提出的方法基于语音信号是非平稳信号而背景噪声相对平稳这一特征，利用两者临界带功率谱方差上的显著差异来进行区分。通过自适应门限的设计，以期获得合适的阈值，从而提高检测方法的实用性。

将每一帧语音的功率谱按一定的频带划分关系分为若干个子带，然后对每一个子带的功率谱求和，得到临界带特征矢量。此时，每一帧语音信号对应若干维临界带特征矢量。最后，对每一帧的特征矢量求方差，得到的结果即为功率谱方差。方差的大小反映的是语音频带内功率谱的变化，方差值越大表明功率谱变化越剧烈，这符合实际语音的特点；反之，对于平稳背景噪声，功率谱变化平缓，方差值就越小。

2.2　基于临界带功率谱方差的端点检测算法

基于功率谱方差的语音信号端点检测算法步骤如下：

(1) 分帧加窗得到语音帧 $x_n(m), m=0\sim(N-1)$，再用 FFT 变换求出语音帧对应的功率谱 $|X_n(k)|^2$。

(2) 划分临界带，按照公式：

$$f(i)=1960\times\frac{i+0.53}{26.28-i} \tag{6}$$

在 $f=0\sim f_s/2$ 之间确定临界带频率分割点 $f(i)$。

(3) 将每个临界带中的 $|X_n(k)|^2$ 取和即可得到相应的临界带特征矢量。如果用 $\boldsymbol{G}=[g_1,g_2,\cdots,g_l,\cdots,g_L]$

表示每一帧的临界带特征矢量，那么 g_l 可表示为：

$$g_l=\sum_{f_l<f_k\le f_{l+1}}|X_n(k)|^2 \tag{7}$$

其中，$f_k=\frac{f_s}{512}k$，512 为本文所使用的 FFT 点数，f_s 为采样频率。

(4) 求临界带功率谱方差。对临界带特征矢量 $\boldsymbol{G}$ 求均值，用 $\bar{E}$ 表示：

$$\bar{E}=\frac{1}{L}\sum_{l=1}^{L}g_l \tag{8}$$

再求临界带功率谱方差，方差 D 的定义为：

$$D=\frac{1}{L}\sum_{l=1}^{L}(g_l-\bar{E}) \tag{9}$$

(5) 自适应门限值的设定，端点检测。通过大量计算机仿真实验，D 的值随着信噪比的下降而增大，固定的门限值在这种情况下就不适用了。为此，需要设计一个与信噪比 SNR 有关的门限值，使其对噪声具有良好的跟踪性能。门限值 Th 由初始项和调整项两部分构成，如式(10)所示：

$$Th=f(SNR)\times Th_0 \tag{10}$$

其中：$f(SNR)$ 为调整系数；Th_0 为初始项。

通过大量仿真实验，得出 $f(SNR)$ 和 Th_0 的经验值为

$$f(SNR)=\begin{cases}2, & SNR<-5\\ SNR+7.5, & -5\le SNR\le 30\\ 50, & SNR>30\end{cases} \tag{11}$$

$$Th_0=\text{mean}\{\min[D(1{:}20)]\} \tag{12}$$

其中，mean{min[D(1:20)]} 为最小的 20 个 D 值的平均。

3 实验结果及对比

为了与文中提及的其他方法进行对比，实验采用相同条件：采样频率为 22.05 kHz，16 bit 量化，每帧 500 个采样点(23 ms)，帧移为 250 个点，PCM 编码。图 1 为连续四声“救命”的原始波形及加噪声后的波形，噪声信号为计算机模拟的零均值高斯白噪声，信噪比为−5 dB。

3.1 与短时能量和短时平均过零率、谱熵方法的实验对比

不同信噪比下，谱熵波形、短时能量和短时过零率波形，如图 2 和图 3 所示。由图可见，在高信噪比且语音信号平稳的情况下，谱熵方法与短时能量和短时平均过零率结合的方法能够很好地确定阈值，区分语音段、背景噪声段。但随着信噪比的降低，谱熵方法下的语音段的起始和结尾帧与

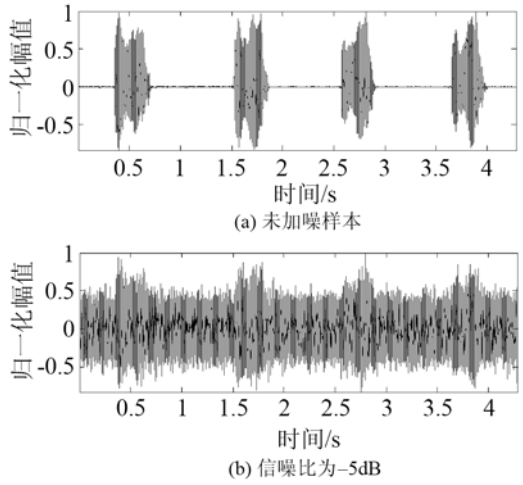

图 1 原始及加噪波形图

Fig.1 The signal of the raw speech and the mixing noise

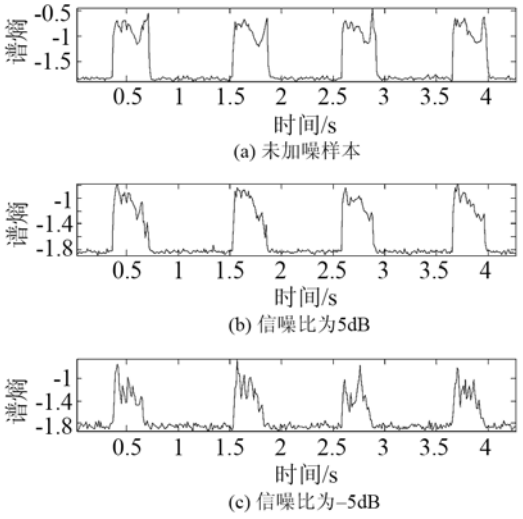

图 2 不同信噪比下的谱熵波形

Fig.2 Spectral entropy under different SNRs

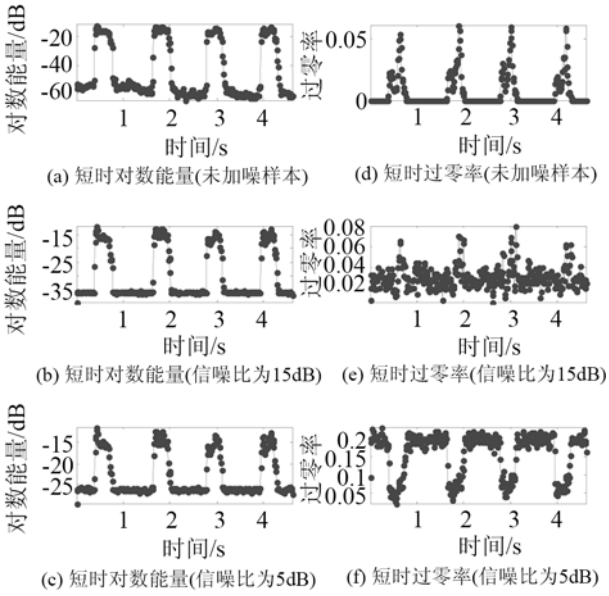

图 3 短时能量及过零率波形

Fig.3 Short-time energy and zero crossing rate under different SNRs

背景噪声段的差异变小，使剪切得到语音帧的长度减小。同时，非语音段谱熵值与语音段谱熵值之间的差距也在逐渐缩小，不利于阈值的确定。信噪比越低，语音段与背景噪声段的短时能量间的差异越小；短时过零率在一定信噪比的条件下，其值较凌乱，不能作为反应语音的声学特征。此时，短时能量与短时过零率结合的方法也将不可避免地受到其影响。

图 4 为不同信噪比下的归一化临界带功率谱方差。由图 4 可见，基于临界带功率谱方差的端点检测方法不仅在高信噪比情况下有很好的区分能力，在低信噪比情况下依然有强大的性能。即使在信噪比为–5 dB 时，语音段和背景噪声段的波形依然差异明显。同谱熵方法相比，背景噪声段波动很小，且本文使用的方法在低信噪比时的语音段方差波形同原始序列相比几乎不变，这样在设定阈值的时候主要关注背景噪声的方差值的变化，为阈值的确定提供了很好的条件。

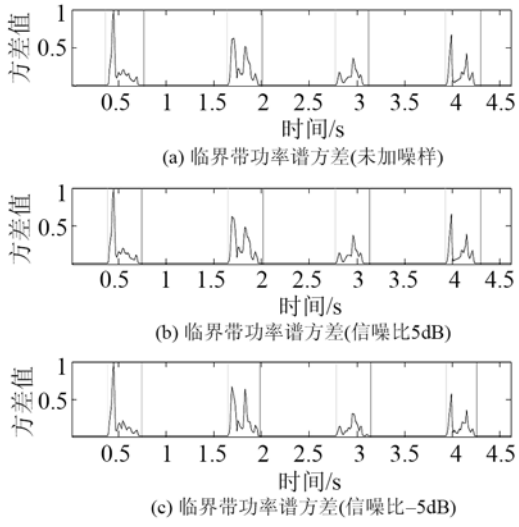

图 4　归一化临界带功率谱方差

Fig.4　Normalized variance of critical band power spectrum

3.2　不同信噪比不同检测方法的正确率比较

端点检测实验所用的语音通过普通个人计算机的录音机和麦克风进行录制，共录制了 86 个语音段。为便于对比，语音的录制条件与上文一致。端点检测正确的定义为自动端点检测与手工标定的误差在 4 帧之内，正确率即为判断正确的端点数和总的端点数的比值。检测结果见表 1。

从表 1 中可以看出，在无噪声以及高信噪比条件下，三种方法的结果相当。随着信噪比的降低，短时能量与短时过零率相结合的方法以及谱熵法的正确率都出现了较显著的下降。信噪比降到–5 dB 时，短时能量与短时过零率相结合的方法已经无法进行正确的检测。此时，谱熵法的正确率也降低至 37.2%，无法满足实际应用的需求。而本文提出的基于临界带功率谱方差的方法效果则较为稳定，仍然保持在 95%以上的正确率，这证明了本文方法的有效性。

表 1　不同信噪比下三种方法正确率统计

Table.1　Accuracy of all three methods under different SNRs

信噪比/dB	三种方法的检测正确率/%		
	能量&过零率	谱熵	临界带
未加噪声	98.8	97.7	98.8
15	93.0	87.2	98.8
5	74.4	79.1	97.7
–5	0.0	37.2	96.5

3.3　本文方法的适用环境和局限性

由于临界带功率谱方差的端点检测算法是基于语音谱的起伏特性和背景噪声频谱相对较平缓这一前提来实现语音段的检出，因此本文的端点检测适合持续稳定的噪声环境。

仿真实验表明，基于临界带功率谱方差的的端点检测对于语音起始音节声母为不送气音：如“b、d、g、z、zh、j”等有较好的检出效果；对于起始音节声母为送气音，如“q、t、x”等，端点检测性能会出现一定程度的下降，尤其是在低信噪比条件下，端点检测出错率增加。

本文利用 50 个以送气音为语音起始音节声母的普通话单词做测试，包括“西工大”、“抢劫”等。在不同信噪比条件下端点检测正确率见表 2。

表 2　以送气音为起始的单词的正确率统计

Table 2　Accuracy of word detection, the word starts at aspirated sound

信噪比/dB	不加噪	5	–5
正确率/%	96	94	86

从表 2 中数据可以看出，以送气音为语音起始音节声母的单词端点检测，在较高信噪比下依然有很好的检出效果，但是在较低信噪比条件下正确率下降较快。

图 5 为连续五声“抢劫”的端点检测结果，由图 5 可以解释造成这一问题的原因。可以看出，以送气音为起始音节声母的语音，其方差值在起始处较小，特别是在信噪比较低时，语音起始阶段的方差值相对于背景噪声的方差值区别更小，因此，导致端点检测的精度下降。

4　结 论

本文在前人研究的基础上，利用临界带模拟人

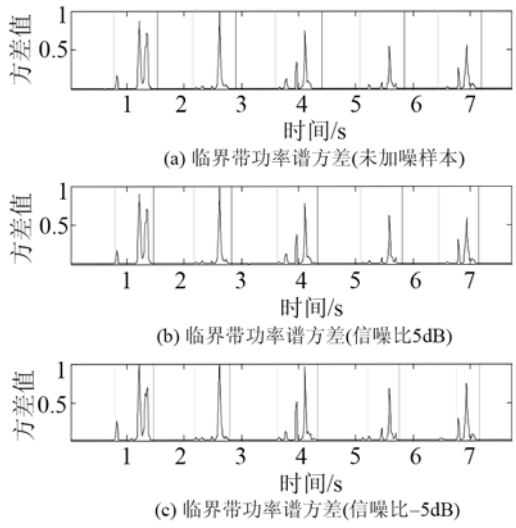

图 5 端点检测效果

Fig.5 Effects of endpoint detection

耳听觉特性，在背景噪声频谱平坦或稳定的环境下，利用语音段和背景噪声段功率谱方差的显著区别，对语音进行有效的区分，提出了基于临界带功率谱方差的端点检测方法。该方法在理论上易于理解，可操作性强，通过自适应门限的设定，能够更好地进行语音段的分拣，在平稳噪声环境下有很好的效果。实验证明，该方法在强噪声环境下较短时能量及过零率方法、谱熵方法有更好的正确率和鲁棒性。后续研究中，还将进一步改进算法，使其在语音起始音节声母为送气音的情况下仍然具有稳定的检测性能。

参 考 文 献

[1] 吴秀良, 范影乐. 基于排列组合熵的语音端点检测技术研究[J]. 计算机工程与应用, 2008, **44**(1): 240-242.
WU Xiuliang, FAN Yingle. Application of permutation entropy measure in detecting speech[J]. Computer Engineering and Applications, 2008, **44**(1): 240-242.

[2] 雷雄国, 曾以成. 基于近似熵的语音端点检测[J]. 声学技术, 2007, **26**(1): 121-124.
LEI Xiongguo, ZENG Yicheng. Noisy speech endpoint detection based on approximate entropy[J]. Technical Acoustics, 2007, **26**(1): 121-124.

[3] 王文延, 曾庆宁. 一种噪声环境下的语音端点检测方法[J]. 声学技术, 2007, **26**(3): 435-441.
WANG Wenyan, ZENG Qingning. New method for detecting endpoint of speech in noise environment[J]. Technical Acoustics, 2007, **26**(3): 435-441.

[4] 张君昌, 姜菲, 刘红. 多特征相结合的带噪语音端点检测算法的研究[J]. 计算机工程与应用, 2009, **45**(32): 114-116.
ZHANG Junchang, JIANG Fei, LIU Hong. Study on endpoint detection based on multi-characteristic jointed in noisy environment[J]. Computer Engineering and Applications, 2009, **45**(32): 114-116.

[5] 屈百达, 蒋纯纲. 基于一种新型分形理论的语音端点检测[J]. 山西大学学报(自然科学版), 2008, **31**(4): 508-511.
QU Baida, JIANG Chungang. Speech endpoints detection based on a new fractal dimension[J]. Journal of Shanxi University(Nat.Sci.Ed.), 2008, **31**(4): 508-511.

[6] 张德祥, 吴小培. 基于经验模态分解和 Teager 峭度的语音端点检测[J]. 仪器仪表学报, 2010, **31**(4): 493-499.
ZHANG Dexiang, WU Xiaopei. Endpoint detection of speech signal based on empirical mode decomposition and teager kurtosis[J]. Chinese Journal of Scientific Instrument, 2010, **31**(4): 493-499.

[7] 赵立. 语音信号处理[M](第二版), 北京: 机械工业出版社, 2009.
ZHAO Li. Speech signal processing[M](Second edition). Beijing: China Machine Press, 2009: 37-46.

[8] 武文娟, 顾宏斌, 潘秀林. 基于临界带特征矢量距离的端点检测算法[J]. 计算机科学, 2009, **36**(2): 220-222.
WU Wenjuan, GU Hongbin, PAN Xiulin. Voice activity detection method based on selected subbands vector distance[J]. Computer Science, 2009, **36**(2): 220-222.

[9] 刘荣, 刘珩. 低信噪比下基于功率谱熵的语音端点检测算法[J]. 计算机工程与应用, 2009, **45**(33): 122-124.
LIU Rong, LIU Heng. Power spectrum entropy based voice activity detection algoritlnn in low signal-to-noise ratio conditions[J]. Computer Engineering and Applications, 2009, **45**(33): 122-124.