ISSN 1000-0054 清华大学学报（自然科学版）2005 年 第 45 卷 第 10 期 26/36
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2005, Vol. 45, No. 10 1397-1400

低信噪比下基于谱熵的语音端点检测算法

李 晔, 张仁智, 崔慧娟, 唐 昆

(清华大学 电子工程系, 微波与数字通信技术国家重点实验室, 北京 100084)

摘 要: 为提高语音端点检测系统在低信噪(0 dB 以下)下检测的准确率, 提出了一种基于谱熵的端点检测算法。将每帧信号分为 16 个子带, 选取频谱分布在 250～3.5 kHz 并且能量不超过该帧总能量 90% 的子带, 计算经过语音增强后的子带能量以及各子带信噪比, 根据各子带信噪比的不同调整其在整个谱熵计算过程中的权重, 然后平滑谱熵, 以最终的谱熵作为端点检测的依据。实验结果表明, 此方法在较低的信噪比下能够显著地提高端点检测的准确率。对坦克噪声, 检测效果明显优于 G.729 中的端点检测算法, 即使在 -5 dB 的信噪比下, 仍然可以达到 95% 以上的检测率。

关键词: 语音信号处理; 端点检测; 谱熵; 语音增强; 信噪比

中图分类号: TN 912.3 **文献标识码**: A

文章编号: 1000-0054(2005)10-1397-04

Voice activity detection algorithm with low signal-to-noise ratios based on the spectrum entropy

LI Ye, ZHANG Renzhi, CUI Huijuan, TANG Kun

(State Key Laboratory of Microwave and Digital Communications, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Voice activity detection (VAD) in low signal-to-noise ratio (SNR) environments is improved with an algorithm based on the spectrum entropy. Each frame is first divided into 16 bands with selection of bands with frequencies between 250 Hz and 3.5 kHz and energies below 90% of the total energy. The energy and the SNR of each band after speech enhancement are then calculated with the entropy band weight adjusted according to it's SNR. The smoothed entropy is then used for the voice activity detection. Test results show that the method significantly increases the voice activity detection ratio. For example, it works the detection accuracy is above 95% even with -5 dB noise whish is better than the G.729 algorithm for tank noise.

Key words: speech signal processing; voice activity detection; spectrum entropy; speech enhancement; signal to noise ratio

端点检测算法的研究在语音信号处理领域中一直有着重要的意义。作为语音识别的前端, 准确的端点检测可以提高识别的准确率; 用于语音增强系统中, 可以进行准确的噪声模型估计; 在语音编码领域中可以降低编码的平均比特率并降低功耗。目前的端点检测算法在较高的信噪比下均能给出较高的检测率, 但是在低信噪比下却不够理想。传统的能量和过零率特征在低信噪比下已不再稳健。许多新的特征被提出, 比如: 基于多特征联合的方法[1], 基于频域能量的特征[2], 基于差分能量和差分过零率的特征[3], 基于排序幅度谱的特征[4], 基于高阶统计量的特征[5], 基于高频能量和低频能量的特征[6]等等。以上的多种特征在低信噪比噪下检测准确度仍然不够理想。考虑到, 即使在很低的信噪比下, 语音帧中仍然存在信噪比较高的子带, 而噪声帧则不具备这个特点, 本文提出了一种新的基于子带选取, 带有加权因子的子带能量计算, 模糊子带加权和谱熵自适应平滑的算法, 大量的试验结果表明, 在不同的噪声环境和信噪比下, 算法具有很好的的性能。

1 算法描述

1.1 子带选取准则

语音抽样频率为 8 kHz, 帧长为 20 ms, 进行 256 点的 FFT 变换。首先将 0～4 kHz 的全频段化成 16 个频带, 每 8 点(250 Hz)构成一个频带。计算每一个频带的能量

$$S_i = \sum R_k^2 \qquad (1)$$

其中 R_k 是对应子带的相应的 Fourier 变换第 k 个幅度值。

收稿日期: 2004-10-11

基金项目: 国家自然科学基金资助项目 (60272020)

作者简介: 李晔(1981-), 男(汉), 山东, 博士研究生。

通讯联系人: 崔慧娟, 教授,

E-mail: cuihuijuan@mail.tsinghua.edu.cn

相应的概率密度函数为

$$P_i = S_i \Big/ \sum_{j=1}^{n} S_j. \tag{2}$$

然后对P_i和S_i做进一步的改进,语音信号的频谱一般分布在250～3 500 Hz,不属于这个频率范围的S_i,将其设为0。另外,如果某一个频带的能量超过总能量的90%,为了消除一些集中在特殊频率的噪声,可以限定$P_i < 0.9$,即采用下述约束关系:

$$S_i = 0, \quad \begin{cases} \text{若 } F_i > 3\,500 \text{ 或 } F_i < 250; \\ \text{若 } P_i > 0.9 \end{cases}$$

经过上述改进,第i帧的熵可以定义为

$$H_i = \sum_{j=0}^{m} P_i \log P_i. \tag{3}$$

选取熵作为判别的基本标准是因为实验中发现,经过平滑以后,它要比单纯利用较高信噪比子带的个数作为判别标准有效的多,尤其是对于坦克噪声。

1.2 带有加权因子的子带能量计算

首先,对每一个帧的各个子带的能量计算公式进行改进。进行能量计算时,根据各频率样点不同的信噪比给予不同的加权因子,从而提高子带能量计算的可信度。算法的噪声模型估计采用基于初始噪声段和自适应平滑的方法。加权因子的估计公式为[7]

$$G_k = \frac{\sqrt{\pi V_k}}{2r_k} M(-0.5; 1; -V_k). \tag{4}$$

式中$M()$为合流超几何函数,可以利用级数求和计算,如下:

$$M(a, c, x) = 1 + \frac{a}{c}\frac{x}{1} + \frac{a(a+1)}{c(c+1)}\frac{x^2}{2!} + \cdots, \tag{5}$$

而

$$V_k = r_k \xi_k / (1 + \xi_k), \tag{6}$$

$$\xi_k = \lambda_{sk} / \lambda_{nk}; \quad r_k = R_k^2 / \lambda_{nk}; \tag{7}$$

其中:ξ_k和r_k分别称为先验信噪比和后验信噪比,R_k为当前带噪语音帧Fourier变换的第k个幅度值,λ_{nk}为噪声帧Fourier变换的第k个幅度估值,λ_{sk}为当前语音帧所对应的干净语音的Fourier变换的第k个幅度值,文[7]中给出了它们的具体计算步骤。由此可以得到改进后的各子带能量的计算公式为

$$S_i = \sum_{k=\text{on}}^{\text{end}} (G_k R_k)^2. \tag{8}$$

其中on和end分别代表第i个子带所对应的Fourier变换幅度的起止点。

1.3 模糊子带加权准则

这一步是算法的关键所在,主要是基于这样一个事实:即使是在很低的信噪比下,语音帧仍然含有许多信噪比较高的频点,算法中语音分为16个子带,因此语音帧中会含有信噪比较高的子带,而噪声帧则不具备这个特点。首先计算每一帧的各个子带的信噪比,子带信噪比的估计要依赖于噪声子带的能量估算。噪声子带能量的估算依赖于初始噪声段以及后续的自适应平滑,若假设共取了l帧噪声段,则估算公式如下。

若取一帧(只考虑250～3 500 Hz范围内的子带),则各子带能量为

$$N_i, \quad i = 1, \cdots, 13$$

以后$l-1$帧无语音噪声段对噪声各子带能量进行重估。

$$\begin{aligned} &\text{for } j = 2, \cdots, l, \\ &\quad N_i = [N_i(j-1) + E_i^j]/j, \\ &\text{end}, \end{aligned} \tag{9}$$

其中E_i^j代表第j个重估帧的第i个子带的能量。

如上所述,每一个噪声子带的能量是通过多帧的平均得到的,在后面的端点检测过程中,每进行一次噪声与语音的判决,就对子带能量进行一次重估,从而可以跟踪噪声能量的变化。在谱熵的具体计算过程中,根据各子带的信噪比情况,决定其在谱熵法计算中的权重,信噪比越低,权重越小。假定信噪比的隶属度函数为:

$$U_{R_{\text{sn}}(i)} = \text{bell}[S_{\text{nr}(i)}, 3, 1, \max(R_{\text{sn}})], \tag{10}$$

$$\text{bell}(x, a, b, c) = \left[1 + \left(\frac{x-c}{a}\right)^{2b}\right]^{-1}. \tag{11}$$

其中:$U_{R_{\text{sn}}(i)}$代表对应子带在谱熵计算中所占的权重,$S_{\text{nr}(i)}$是对应子带的信噪比,$\max(R_{\text{sn}})$代表所有子带信噪比中的最大值。因此,总的谱熵的计算公式为:

$$\begin{aligned} &\text{for } i = 1, \cdots, 13 \\ &\quad R = \max[(S - N)/N]; \\ &\quad S_{\text{nr}(i)} = (S_i - N_i)/N_i; \\ &\quad C = \text{bell}(S_{\text{nr}(i)}, 3, 1, R); \\ &\quad H_j = H_j + C\left(\left(-S_i \Big/ \sum_m S_m\right) \lg\left(S_i \Big/ \sum_m S_m\right)\right); \\ &\text{end}; \end{aligned} \tag{12}$$

1.4 谱熵平滑准则

在计算谱熵的过程中,利用噪声谱熵的先验知识,对谱熵进行平滑:

如果

$$\text{abs}\left[H_j - \sum_{i=1}^{j-1} F_i/(j-1)\right] > T,$$

$$H_j = \mathrm{abs}\left[H_j - \sum_{i=1}^{j-1} F_i/(j-1)\right] + \sum_{i=1}^{j-1} F_i/(j-1),$$

$$H_j = (H_j + H_{j-1})/2,$$

$$F_j = \sum_{i=1}^{j-1} F_i/(j-1). \quad (13)$$

其中: H_j 是第 j 帧的谱熵, $\sum_{i=1}^{j-1} F_i/(j-1)$ 是对噪声谱熵的估计, 是对前 $(j-1)$ 个已经判决为噪声的帧的熵的平均。如果谱熵的变化超过门限值, 则判断为语音帧, 同时, 对前一帧的谱熵进行重估, 从而对谱熵进行自适应平滑, 减少语音前端的错误剪切, 噪声谱的估计值不变。否则进行如下操作:

$$F_j = H_j,$$

$$H_j = \sum_{i=1}^{j} F_i/j, \quad (14)$$

$$H_j = (H_j + H_{j-1})/2 \quad (15)$$

也就是说当前帧判断为噪声帧, 对噪声帧的谱熵进行重估, 同时, 对噪声帧的子带能量进行重新的估算:

$$\text{for } i = 1, \cdots, 13$$

$$N_i = N_i a + \sum_{n=8i}^{8(i+1)} D_n (1-a),$$

$$\text{end;} \quad (16)$$

加权因子 a 控制估计器遗忘的速率, 作用是"遗忘"掉久远的过去数据, 以便当滤波器工作在非平稳的环境中时, 能够跟踪噪声数据的统计变化。在语音变化比较剧烈的区域, a 值通常比较小, 而在语音较为平稳的区域, a 值通常比较大。在大多数文献中, a 通常在 0.95～0.99 之间。在本文的方法中, a 采用估计的方法得到, 在不同的区域, 取值不同。假定噪声变化平稳, 在帧与帧之间变化并不剧烈, 带噪语音的能量的急剧变化应该是由语音引起的。估计公式如下:

$$a = \sqrt{1 - \frac{|E_j - E_{j-1}|}{\max(E_{j-1}, E_j)}}, \quad (17)$$

E_j 为第 j 帧带噪语音的能量。

1.5 Hangover 过程设计原则

在语音检测的过程中, 不可避免的会出现语音剪切的情况, 包括前端剪切, 句中剪切和句末剪切。前端剪切引起的问题可以通过设置一个 buffer 来解决。具体过程是:

$$H_j > T_h \text{ 并且 } F_{l3} \neq 0,$$

$$V_{(i-1)(i-F_{l3})} = 1,$$

$$F_{l3} = 0. \quad (18)$$

i 为当前帧的标号, 若当前帧被判断为语音帧, 则 i 帧的前 F_{l3} 帧全被判断为语音帧, $\max(F_{l3}) = T_{h1}$, T_{h1} 需要根据试验的具体情况进行选取, F_{l3} 的更新是在当前帧被确切的判断为非语音帧的时刻, 即

如果 $V_i == 0$ 并且 $F_{l3} < T_{h1}$, 那么

$$F_{l3} = F_{l3} + 1. \quad (19)$$

句中语音剪切的结果是整个句子的中间可能会出现一小段被错误地判断成噪声的部分, 如何正确地区分这一小段被错误地判断为噪声的语音和真正的语音间歇间的噪声是改善语音质量的关键。根据统计结果, 正常的语音间歇往往要明显大于由于错判而造成的语音间歇, 所以采取了如下的步骤 (nF_{l1} 和 F_{l2} 的初始值均为 0):

如果 $H_j > T_h$,

$V_j = 1$; $F_{l1} = F_{l1} + 1$; $F_{l2} = n$,

另外, 若 $F_{l1} > 10$ 并且 $F_{l2} > 0$

$V_j = 1$, $F_{l2} = F_{l2} - 1$,

否则 $V_j = 0$; $F_{l1} = 0$。

如果当前帧被判断为语音, 则标志位 F_{l2} 赋值 n, n 为因为错判造成语音间歇的最大间隔, 通常这个值要明显小于正常的语音间歇。若当前帧被判断为噪声, 则检测 F_{l1} 和 F_{l2} 的状态, 若 F_{l1} 大于 10 且 F_{l2} 大于 n, 则仍判断当前帧为语音帧, 同时 F_{l2} 减 1, 否则, 当前帧判断为噪声帧。通过这样一个 hangover 过程, 可以明显减少正常语音剪切的情况。但是由于语音实时通信, 延时不能太长的苛刻要求, 会造成一定的噪声拖尾现象。为了能使算法在不同的信噪比的情况下具有"鲁棒性", F_{l1}、F_{l2} 和 F_{l3} 的值都是在算法中采取自适应的方法根据当前的信噪比确定的。

2 实 验

8 种不同的噪声环境 (白噪声, 粉色噪声, 坦克噪声, 工厂噪声, 群口噪声, 射频噪声, 机枪噪声, f16 噪声等) 下, 在 5 dB 到 -10 dB 的信噪比范围内, 对本文所提出的算法进行了验证 (见图 1), 并在坦克噪声下与 G.729 中的 VAD 算法进行了比较 (见图 2)。

假设:

错误帧数= 语音错判为噪声的帧数+ 噪声错判为语音的帧数,

准确率= (总帧数- 错误帧数)/总帧数,

算法的性能统计如图 1 所示 (仅取 4 类代表性噪

声)。

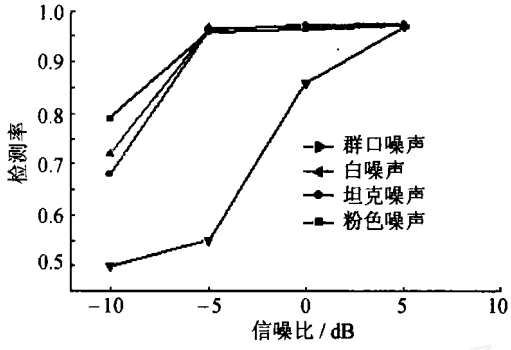

图1 算法在不同的噪声环境和信噪比下的检测结果

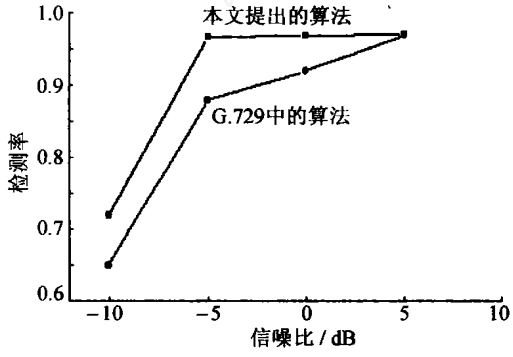

图2 算法与G. 729 中VAD 算法性能比较结果(坦克噪声)

群口噪声为众人说话噪声,随着信噪比降低,在一堆人声中检测特定人声的端点尤为困难,因此准确率较其他噪声环境更低。在- 10 dB 下算法性能下降较为厉害,在低于此信噪比的环境下,该算法检测率已不能满足实际应用,这也是以后研究改进的一个方向。文中的算法在坦克噪声下表现出了比G. 729 算法更为优越的性能。

3 结 论

即使在很低的信噪比下,语音帧仍然具有很多较高信噪比的频点(或子带),而噪声帧却没有。据此,论文提出了一种算法,根据子带信噪比不同而采用不同权重因子来计算谱熵,从而进行端点检测。能够在0 dB 以下的噪声环境中取得很好的检测效。以坦克噪声为例,在- 5 dB 的信噪比下,算法仍然可以取得95% 以上的检测率,显著优于G. 729 中的VAD 算法。

参考文献 (References)

[1] 徐大为, 吴边, 赵建伟, 等. 一种噪声环境下的实时语音端点检测算法[J]. 计算机工程与应用, 2003, **24**(1): 115-117.
XU Dawei, WU Bian, ZHAO Jianwei, et al. A real time algorithm for voice activity detection in noisy environment [J]. *Computer Engineering and Application*, 2003, **24**(1): 115-117. (in Chinese)

[2] Junqua J C, Mak B, Reaves B. A robust algorithm for word boundary detection in the presence of noise [J]. *IEEE Transactions on speech and Audio Processing*, 1994, **2**(3): 406-412.

[3] Beritelli F, Casale S, Ruggeri G, et al. Performances evaluation and comparision of G. 729/AMR/fuzzy voice activity detectors [J]. *IEEE Signal Processing Letters*, 2002, **9**(3): 85-88.

[4] Pencak J, Neloson D. The NP speech activity detection algorithm [J]. *Int Conf Acoustics, Speech and Signal Processing*, 1995. 381-384.

[5] Nemer E, Goubran R, Mahmoud S. Robust voice activity detection using higher-order statistics in the LPC residual domain [J]. *IEEE Trans Speech and Audio Processing*, 2001, **9**(3): 217-231.

[6] Woo K H, Yang T Y, Park K J, et al. Robust voice activity detection algorithm for estimating noise spectrum [J]. *Electronics Letters*, 2000, **36**(2): 180-181.

[7] 迟惠生, 杨行峻, 唐昆, 等. 语音信号数字处理[M]. 北京: 电子工业出版社, 1995.
CHI Huisheng, YANG Xingjun, TANG Kun, et al. Speech Signal Processing [M]. Beijing: Electronic Engineering Publishing House, 1995. (in Chinese)