基于倒谱距离的语音端点检测改进算法

王　博，郭　英，李宏伟，韩立峰
(空军工程大学 电讯工程学院，陕西 西安　710077)

摘　要:在讨论传统倒谱距离语音端点检测算法不足的基础上,提出了一种改进方案,该方法首先估计短时信噪比,然后由统计方法确定短时信噪比与门限的关系,进而完成正确的语音端点判决。通过对 3种典型噪声环境下信噪比从 - 5 dB到 20 dB的带噪语音信号进行的仿真实验结果表明,所提方法能更为准确地检测到语音端点。

关键词:端点检测;倒谱距离;判决准则;语音增强
中图分类号: TN912.34　　**文献标识码**: A　　**文章编号**: 1009 - 3516(2006)01 - 0059 - 05

准确的语音信号端点检测 (VAD - Voice Activity Detection)可以实现对噪声谱的实时更新,从而提高谱减法语音增强系统的性能。传统的检测方法采用短时能量、过零率和自相关参数,在高信噪比环境下可以获得较好的检测效果,但是在低信噪比环境下其检测性能却急剧下降。本文提出了一种基于倒谱距离的改进方法,通过分析信号的倒谱参数来进行带噪语音的端点检测。仿真结果表明,在低信噪比环境下较之传统的方法能更准确地检测出语音的端点。

1　基于倒谱距离的端点检测算法

1.1　倒谱距离定义

设信号 $s(n)$,其倒谱变换为 $c(n)$。信号倒谱的一种定义是信号的能量谱密度函数 $s(\omega)$ 的对数的傅里叶反变换,或者可以将信号 $s(n)$ 的倒谱 $c(n)$ 看成是 $\log S(\omega)$ 的傅里叶级数展开[1],即

$$\log S(\omega) = \sum_{n=-\infty}^{\infty} c(n) e^{-jn\omega}, \quad c(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log S(\omega)\, d\omega \tag{1}$$

式中, $c(n)$ 为倒谱系数,且 $c(n) = c(-n)$ 是实数。

假设信号 $s(n)$ 的 Z变换具有有理函数的形式

$$S(z) = A z^{r} \frac{\prod_{k=1}^{m_1}(1 - a_k z^{-1}) \prod_{k=1}^{m_0}(1 - b_k z)}{\prod_{k=1}^{p_1}(1 - c_k z^{-1}) \prod_{k=1}^{p_0}(1 - d_k z)} \tag{2}$$

式中 a_k、b_k、c_k和 d_k的模都小于 1, m_1 和 m_0分别表示单位圆内和外的零点数目, p_1 和 p_0分别表示单位圆内和外的极点数目。对 $\log S(z)$ 取逆 Z变换得到倒谱系数的另一种表达式为

$$c(n) = \begin{cases} \log|A| & n = 0 \\ -\sum_{k=1}^{m_1} \frac{a_k^n}{n} + \sum_{k=1}^{p_1} \frac{c_k^n}{n} & n > 0 \\ -\sum_{k=1}^{m_0} \frac{b_k^{-n}}{n} + \sum_{k=1}^{p_0} \frac{d_k^{-n}}{n} & n < 0 \end{cases} \tag{3}$$

收稿日期: 2005 - 06 - 03
基金项目:军队科研基金资助项目
作者简介:王　博 (1981 -),男,陕西商州人,硕士生,主要从事语音信号处理研究;
郭　英 (1961 -),女,山西临汾人,教授,博士生导师,主要从事自适应信号处理和信息对抗技术研究.

很明显当 n趋向无穷大时倒谱的幅度值是收敛的。

$$|c(n)| < \xi\frac{\lambda^{|n|}}{|n|}, \text{当} \ |n| \to \infty \text{时} \tag{4}$$

式中 λ 是 a_k、b_k、c_k和 d_k模的最大值，ξ为一实常数。由式(4)可以看出，倒谱是一个快速衰减序列，其衰减速率至少为 $1/|n|$[1]，所以在误差允许的范围内可以用有限阶(比如 p阶)的倒谱系数近似无限阶的倒谱系数。

根据 Parseval定理，对于两个不同信号 $s_0(n)$和 $s_1(n)$，其倒谱差异的均方值可用倒谱距离表示：

$$d_{cep}^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\log S_1(\omega) - \log S_0(\omega)|^2 d\omega = \sum_{n=-\infty}^{\infty} (c_1(n) - c_0(n))^2 \tag{5}$$

式中 d_{cep}为倒谱距离，$c_0(n)$和 $c_1(n)$分别是对应于谱密度函数 $S_0(\omega)$和 $S_1(\omega)$的倒谱系数。用 p阶倒谱系数近似无限阶倒谱系数，式(5)可以近似为[1]

$$d_{cep} = 4.3429\sqrt{(c_1(n) - c_0(n))^2 + 2\sum_{n=1}^{p}(c_1(n) - c_0(n))^2} \tag{6}$$

信号与其倒谱是一一对应的变换，因此倒谱的均方距离可以反映两个信号(比如语音与背景噪声)谱的区别，倒谱距离可以作为端点检测的一个判决参数，属于相似距离范畴。

1.2 传统的倒谱距离检测算法流程[5,7]

1) 预处理：对 8 kHz采样信号进行预加重处理，然后分帧加窗，帧长取 30 ms(240个采样点)，帧移 10 ms，对每一帧信号加 240点的 Hamming窗。

2) 估计噪声倒谱系数和倒谱距离 D_{cepsil}：阶数 p取 12，首先假定抽样信号起始 10帧是背景噪声，利用这 10帧的前 5帧倒谱系数的统计平均值作为背景噪声倒谱系数的估计值，用向量 c_0表示。同时采用式(6)计算这 10帧的后 5帧倒谱距离平均值作为背景噪声倒谱距离的估计值，其中 $c_1(n)$表示当前帧的倒谱系数，$c_0(n)$为对应于 C_0的倒谱系数。

3) 逐帧计算 d_{cep}值：逐帧计算倒谱系数，然后由每帧信号的倒谱系数和噪声倒谱系数的估计值通过式(6)计算倒谱距离。

4) 确定判决门限：采用类似于短时能量检测法所使用的动态门限判决准则，设定两个门限 G_1和 G_2

$$G_i = d_{cepsil}k_i, \quad i = 1, 2 \tag{7}$$

式中 d_{cepsil}为噪声倒谱距离估值，k_1，k_2分别为两个门限的乘系数，且 $k_2 > k_1$，以保证 $G_2 > G_1$，这里取 $k_1 = 1.0$、$k_2 = 1.3$。

5) 根据各帧的 d_{cep}值进行端点检测：如果当前帧的 d_{cep}值大于 G_1，则记录该帧位置为 start，然后继续计算后面各帧的 d_{cep}值，若在该帧之后若干帧以内，有连续 3帧的 d_{cep}值都大于 G_2，则认为 start为语音信号的起点，否则继续搜索。终点的检测可类比起点的检测得到。

6) 背景噪声倒谱系数和倒谱距离的更新：检测过程中为使背景噪声倒谱系数和倒谱距离的估计值能适应噪声的变化，当某帧已被确认为噪声帧时，按照

$$C_0 = \gamma C_0 + (1 - \gamma)C_{0l} \tag{8}$$

$$d_{cepsil} = \eta d_{cepsil} + (1 - \eta)d_{cep}(i) \tag{9}$$

对噪声倒谱系数和倒谱距离进行更新。以上二式中 C_{0l}为当前噪声帧倒谱向量，$d_{cep}(i)$为当前噪声帧倒谱距离，γ、η为更新因子，这里取 $\gamma = \eta = 0.99$。

在检测过程中，为避免突发噪声的影响，需在检测前对 d_{cep}参数值进行 5点中值平滑后处理。同时引入时滞机制[2]，即设定语音和噪声的最小持续期。当检测出的语音段长度小于语音最小持续期时，则认为该段为偶发的脉冲干扰，而非真正的语音段；同样，当检测出的噪声段长度小于噪声最小持续期时，则认为该段为说话中间的微小停顿或换气，而非静音段。语音和噪声的最小持续期可以通过分析截短错误(即将语音判为噪声)和扩展错误(即将噪声判为语音)长度来确定。本文分别取 100 ms和 150 ms。

1.3 改进型算法

由于传统的倒谱距离检测方法采用恒定的门限乘系数，很难调整出同时适合低信噪比和高信噪比环境下的门限乘系数。所以有必要建立随信噪比(SNR)变化的检测门限，从而更加准确地检测语音的端点。改进型算法的基本思想如下：

1) 逐帧估计短时 SNR：根据 Francesco Beritelli提出的方法[3]，在每一帧采用对数功率估计 SNR，由于所

估计的 SNR主要由当前帧的功率值决定,所以不妨称之为短时 SNR。具体的估计方法如下所述:

在每一帧用下式计算该帧信号的对数功率 p_n

$$p_n = 10\lg\left[\frac{1}{M}\sum_{i=1}^{M-1}(x(i)w(i))^2\right] \tag{10}$$

式中 M 为 30 ms帧的采样点数, $x(i)$ 为滤波后的信号采样点, $w(i)$ 为平滑窗函数[3,4],

$$w(i) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi i}{399}\right) & 0 \le i < 200 \\ \cos\left(\dfrac{2\pi(i-200)}{159}\right) & 200 \le i \le 239 \end{cases} \tag{11}$$

在输入信号初始段,用前 10帧的对数功率均值作为噪声的对数功率估计值 P_N

$$P_N = \frac{1}{10}\sum_{n=1}^{10} p_n \tag{12}$$

从第 10帧开始检测语音是否激活。首先用噪声对数功率平均值 P_N和直到当前帧的输入信号语音帧对数功率估计值 P_{S+N}更新 SNR估计值[3]

$$\mathrm{SNR} = 10\lg\left[10^{\frac{P_{S+N}}{10}} - 10^{\frac{P_N}{10}}\right] - P_N \tag{13}$$

需要注意的是输入信号语音帧对数功率 P_{S+N}只能在语音激活后进行计算,这样在信号初始段认为 SNR的估计值为 0 dB。此外,如果估计值受诸如 $P_{S+N} \le P_N$这样的错误影响,将不计算 SNR(使用最近一个有效值)。SNR的估计值用来更新门限,然后以更新后的门限为标准进入下一帧的检测。

此外,在确认的语音段和噪声段对 P_{S+N}和 P_N进行更新。具体更新过程为:当检测到语音终止点时(当前帧可能是背景噪声段),计数量 c_{N1}(初始值为 0)开始计数,每读取一帧信号 c_{N1}加 1,直到饱和值 10。若在 c_{N1}计到饱和值前检测到语音起始点(当前帧可能是语音段),则 c_{N1}置 0,由于该终止点和起始点之间持续期小于 10帧,所以不认为是背景噪声段,不对 P_N进行更新;若 c_{N1}计到饱和值,则计数量 C_{N1}(初始值为 10)开始计数,每读取一帧信号 c_{N2}加 1,直到饱和值 110,然后利用计数量 c_{N2}对 P_N进行更新。

$$P_N = \frac{1}{c_{N1}}P_N + \frac{c_{N2}-1}{c_{N2}}p_n \tag{14}$$

同样的,当检测到语音起始点时(当前帧可能是语音段),计数量 c_{S1}(初始值为 0)开始计数,每读取一帧信号 c_{S1}加 1,直到饱和值 10。若在 c_{S1}计到饱和值前检测到语音终止点(当前帧可能是背景噪声段),则 c_{S1}置 0,由于该起始点和终止点之间持续期小于 10帧,所以不认为是语音段,不对 P_{S+N}进行更新;若 c_{S1}计到饱和值,则计数量 c_{S2}(初始值为 10)开始计数,每读取一帧信号 c_{S2}加 1,直到饱和值 110(与变量 c_{N2}不同的是 c_{S2}不需要重新初始化),然后利用计数量 c_{S2}对 P_{S+N}进行更新。

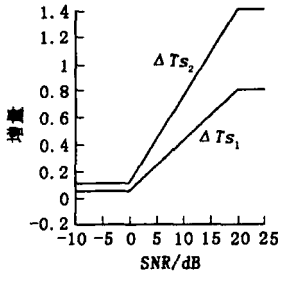
$$P_{S+N} = \frac{1}{c_{S1}}P_{S+N} + \frac{c_{S2}-1}{c_{S2}}p_n \tag{15}$$

2)由短时 SNR确定判决门限。用估计的短时 SNR确定语音起点和终点判决门限,仍然采用双门限判决,设定语音起点判决门限 Ts_1、Ts_2和终点判决门限 Tn_1、Tn_1为

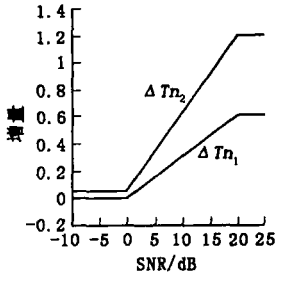
$$Ts_i = d_{\mathrm{cepsil}} + \Delta Ts_i \quad i = 1, 2 \tag{16}$$

$$Tn_i = d_{\mathrm{cepsil}} + \Delta Tn_i \quad i = 1, 2 \tag{17}$$

式中, d_{cepsil}为噪声倒谱距离估值, ΔTs_i和 ΔTn_i为门限增量,其值与 SNR有关,如图 1所示。



(a) 语音起点判决门限增量图



(b) 语音终点判决门限增量图

图 1　门限增量与 SNR关系曲线图

2 仿真试验结果分析

实验室条件下录制 200条相对纯净语音,男女各 100条,4男 3女朗读,采样频率 8 kHz、6 bit量化,长度 3～10 s。然后采用 NoiseX - 92专业噪声库中常见的 3种噪声——平稳高斯白噪声 (White)、M109坦克噪声 (M109)和 F16战斗机噪声 (F16)——将这 200条语音分别混成 SNR为 - 5 dB、0 dB、5 dB、10 dB和 20 dB的带噪语音信号作为测试语音库。

由于语音信号扩展错误比截短错误更易于被人们所接受,所以采用文献[1]中的加权错误测度 (WA),定量比较算法的准确性。定义加权错误测度为

$$WA = (k_c \cdot CLP + k_w \cdot WDN) / Frams \tag{18}$$

式中 CLP表示截短错误帧数,WDN表示扩展错误帧数,Frams表示采样语音数据的总帧数,k_c、k_w为加权系数。参照文献[6],取 $k_c = 1.4$, $k_w = 0.6$。

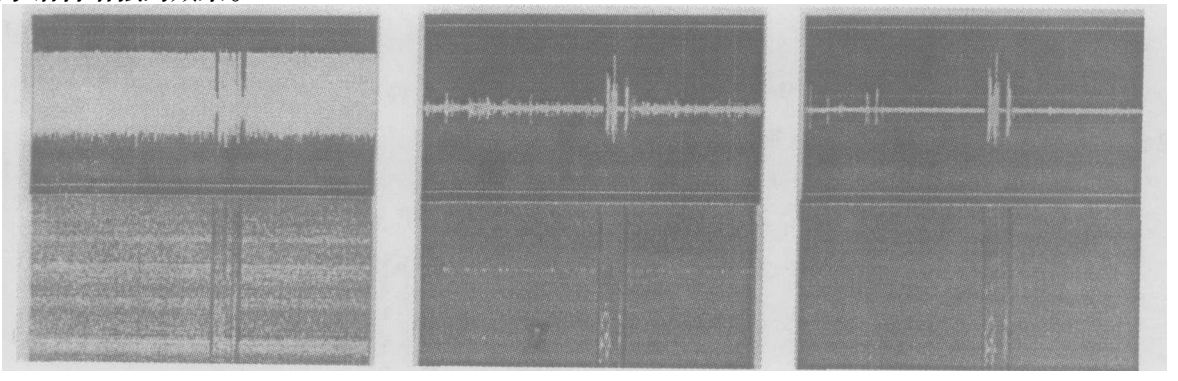
对所录制的所有语音采用传统倒谱距离算法、改进型算法以及短时能量算法进行仿真实验,统计结果如表 1所示。表 1中,Energy、Cepstral和 Cepstral_gai分别表示短时能量算法、传统倒谱距离算法和改进型算法。实验结果表明在不同 SNR环境下改进型算法比短时能量算法和传统倒谱距离算法的检测错误率都有明显下降。

表 1 3种噪声下加权错误测度 WA

	White(%)			F16(%)			M109(%)		
	Energy	Cepstral	Cepstral_gai	Energy	Cepstral	Cepstral_gai	Energy	Cepstral	Cepstral_gai
- 5 dB	54.96	39.96	29.21	58.07	38.94	27.15	49.51	36.95	19.94
0 dB	26.49	17.34	14.86	32.14	19.67	15.15	25.25	17.55	15.44
5 dB	9.16	7.59	7.32	13.06	7.61	7.55	9.79	7.99	7.54
10 dB	3.84	3.61	2.69	4.96	3.74	2.74	5.28	3.86	3.50
20 dB	1.72	1.77	1.94	2.13	2.08	1.94	4.54	3.47	2.56

3 在语音增强系统中的应用

将所提改进型算法应用于谱减法语音增强系统,所得实验结果如图 2所示。图 2为短波电台采集声音增强前和增强后的语音时域波形图和语谱图。由图可见,加入改进型端点检测算法并更新噪声谱,有效的提高了语音增强的效果。



(a)原始语音 (b)谱减法增强语音 (c)加入端点检测增强语音

图 2 语音增强效果图

4 结束语

仿真试验结果表明所提算法可以从背景噪声中有效地检测出语音的起点和终点,明显优于传统的短时能量检测算法,也优于传统倒谱距离检测算法。运算量较之短时能量算法和传统倒谱距离算法并没有明显

的增加,因此,对于平稳高斯白噪声和色噪声环境中的语音端点检测,改进型倒谱距离算法是一种很有效的方法。但是,由于改进型算法需要预先设定 SNR估值的初始值,因而 SNR估值的初始值设定是否合适对于检测结果的影响很大。在实际应用中,可根据背景噪声大小作适当的调整。当 SNR较高时,可以设定初始 SNR为 20 dB;当 SNR较低时,可以设定初始 SNR为 0dB。对于输入带噪语音信号 SNR变化的情况,由于式(15)对于噪声对数能量的更新能很好的跟踪噪声的变化,具有很强的自适应性,所以该算法仍然能较准确的检测到语音的起点和终点。

参考文献:

[1] Rabiner L R, Juang B H. Fundamentals of Speech Recognition[M]. New York: Prentice Hall PTR, 1999.

[2] Francesco Beritelli, Salvatore Casale, Alfredo Cavallaro. A Robust Voice Activity Detector for Wireless Communications Using Soft Computing[J]. IEEE Select Areas Commun, 1998, 16(9): 1818 - 1829.

[3] Francesco Beritelli, Salvatore Casale, Salvatore Serrano. A Low - complexity Speech - pause Detection Algorithm for Communication in Noisy Environments[J]. Euro Trans Telecomms, 2004, 15(1): 33 - 38.

[4] ITU - T Rec. G. 729 Annex B, 1996. A Silence Compression Scheme for G. 729 Optimized for Terminals Conforming to Recommendation V. 70[S].

[5] Haigh J A, Mason J S. Robust Voice Activity Detection Using Cepstral Features[J]. Computer, Communication, Control and Power Engineering. Proceedings of the IEEE Region 10 Conference TENCON, 1993, 3(3): 321 - 324.

[6] 徐　望,丁　琦,王炳锡. 一种基于特征空间能量熵的语音信号端点检测算法[J]. 通信学报, 2003, 24(11): 125 - 132.

[7] 胡光锐,韦晓东. 基于倒谱特征的带噪语音端点检测[J]. 电子学报, 2000, 28(10): 95 - 97.

(编辑:门向生)

An Improved Voice Activity Detection Method Based on Cepstrum Distance

WANG Bo, GUO Ying, LI Hong - wei, HAN Li - feng

(The Telecommunication Engineering Institute, Air Force Engineering University, Xi'an, Shaanxi 710077, China)

Abstract: On discussing the defects of the traditional voice activity detection method based on cepstrum distance, this paper proposes an improved project. In this method, an accurate SNR estimation is made at first, and the discriminative threshold is fixed according to the relationship between SNR and the threshold, then the accurate end - point can be confirmed, in the end the simulation experiments are made with a lot of noisy speech signals ranging from - 5dB to 20 dB in three representative noisy environments. The result shows that the comparatively accurate speech endpoints can be detected by the above method in different SNR environments.

Key words: voice activity detection; cepstrum distance; discriminative rules; speech enhancement