文章编号:1002-0411(2004)01-0077-05

一个基于谱熵的语音端点检测改进方法

王让定[1,2], 柴佩琪[2]

(1. 宁波大学纵横智能软件研究所, 浙江 宁波 315211; 2. 同济大学人工智能研究室, 上海 200092)

摘 要:本文提出了基于谱熵和谱减法相结合的带噪语音端点检测改进算法以及端点检测的判决准则. 仿真实验表明,在语音信号受到强噪声的干扰后(5dB ≤SNR ≤15dB),所提方法可检测到准确的语音端点. *

关键词:带噪语音;端点检测;谱熵;判决准则

中图分类号: TN912.3 **文献标识码**:A

An Improved Speech Endpoint Detection Method Based on Spectral Entropy

WANG Rang-ding[1,2], CHAI Pei-qi[2]

(1. *CKC Software Lab, Ningbo University, Ningbo* 315211, *China*; 2. *AI Lab, Tongji University, Shanghai* 200092, *China*)

Abstract: In this paper, we propose a new approach based on spectral entropy and spectral subtraction for noisy speech endpoint detection, and discriminative rules with robustness. Simulation experiments show that the accurate speech endpoint can be detected by the proposed method when the detected noisy speech SNR is between 5dB and 15 dB.

Keywords: noisy speech; endpoint detection; spectral entropy; discriminant rules

1 引言(Introduction)

在自动语音识别(ASR)、语音编码的研究中,如何在背景噪声中准确地检测语音的端点(endpoint)是提高识别精度和编码效率的关键. 在早期的基于实验室背景的孤立字识别系统中[1],采用基于能量和过零率的方法可检测到准确的语音端点,此背景环境可认为是无噪的"干净"的语音背景.

随着语音识别技术逐步从实验室走向实际应用,语音识别系统的工作环境带有一定噪声干扰. 一般来讲,这些噪声是加性(additive)的,即背景噪声叠加在语音信号之上,由于噪声的时域波形与辅音波形非常相似,当噪声的幅值等于或大于语音信号的幅值时,利用传统的能量和过零率的检测方法很难比较准确地检测语音端点. 另外,当语音识别系统工作于非平稳噪声背景环境时,传统端点检测方法更是无能为力.

在传统的基于短时能量和短时过零率检测语音端点的基础上,针对不同的应用需求,研究者提出了许许多多语音端点检测改进算法,有基于能量、过零率与状态决策(state decision)相结合[2]的改进算法、基于多特征结合的端点检测方法[3]、基于变换域(如 MFCC、Wavelet 等)[4]的改进算法、基于 HMM 模型[5]的检测算法,基于熵(entropy)[6]以及熵与能量结合[7]的改进算法等等. 最近发表的论文[8,9]提出了基于熵和倒谱特征的端点检测改进算法.

如何准确地检测带噪语音的端点至今仍是一个难题,目前已有的算法都适于特定的应用环境,而且在强背景噪声下,算法无法检测到准确的端点,研究自适应于各种应用背景的语音端点检测算法是解决噪声背景下语音识别的关键. 本文尝试性提出了在背景噪声(平稳或非平稳)环境下的熵与谱减增强相结合的语音端点检测方法,算法估计背景噪声能量,利用谱减法进行语音增强,然后通过谱熵判决语音端点. 实验表明,我们所提方法在 SNR ≥5dB 以上时,可准确地检测到语音端点.

2 语音端点检测及改进算法(Speech endpoint detection and the improved algorithm)

2.1 谱熵(Spectral entropy)的基本原理

定义 对带噪语音信号 $s(n)$ 经分帧、加窗,按帧间 50 %的重叠求解 FFT 变换,得其某频率分量 f_i 的能量谱为 $Y_m(f_i)$,则每个频率分量的归一化谱

* 收稿日期:2003 - 05 - 07

概率密度(pdf)函数定义为:

$$p_i = \frac{Y_m(f_i)}{\sum_{k=0}^{N-1} Y_m(f_k)} \quad i = 1, \cdots, N \qquad (1)$$

其中 p_i 为某频率分量 i 对应的概率密度, N 为 FFT 变换长度, m 为分析的某一帧语音. 由于语音的能量主要集中在 250Hz 至 4500Hz, 为了增强 pdf 区分语音和非语音段的能力,对(1)式引入约束条件:

$$Y(f_i) = 0, \quad \text{if } f_i < 250\text{Hz or } f_i > 4500\text{Hz} \qquad (2)$$

考虑上述约束条件后,每个分析语音帧的短时谱熵定义为:

$$H_m = -\sum_{K=1}^{N} p_k \log p_k \qquad (3)$$

2.2 谱熵特征分析

按照(3)式所求的谱熵具有如下特征:

第一,语音信号的谱熵不同于噪声信号的谱熵;

第二,理论上,如果谱的分布保持不变,语音信号幅值的大小不会影响(1)式的归一化 pdf. 但实际上,语音谱熵随语音随机性而变化,与能量特征相比,谱熵的变化是很小的,如图 1 所示. 从图 1 可知,虽然前段"关机"语音的幅值比后段语音的幅值大很多,但谱熵的变化不大;

第三,在某种程度上讲,谱熵对噪声具有一定的鲁棒性,图 2 为同一语音段在不同噪声干扰下的谱熵曲线,从图中可知,当 SNR 下降时谱熵的形状保持不变,但谱熵降低,利用谱熵进行端点检测变得较为困难.

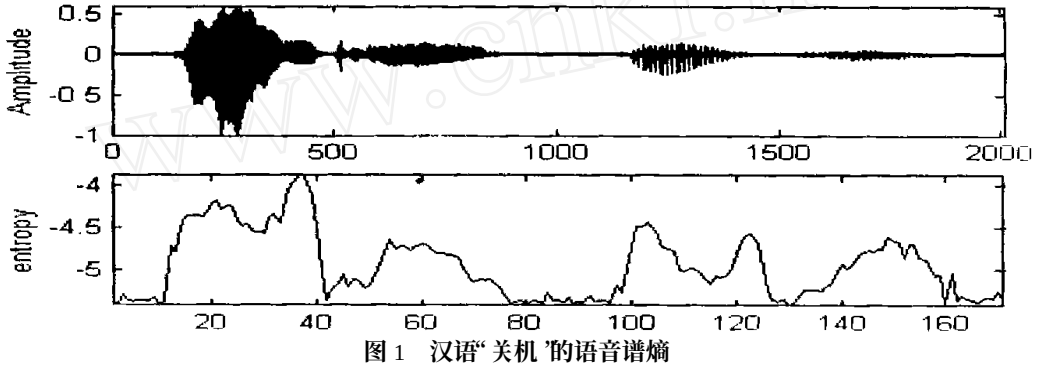

图 1 汉语"关机"的语音谱熵

Fig 1 Spectral entropy of mandarain "Guanji"

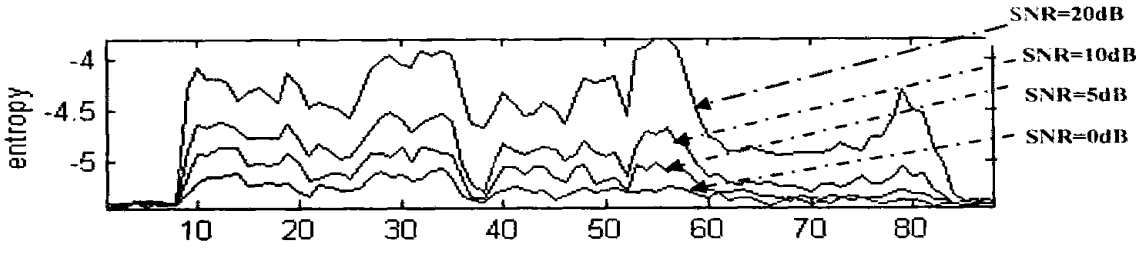

图 2 不同噪声环境下(SNR = 20dB - 5dB)的谱熵曲线

Fig 2 The spectral entropy curve diagrams under different noisy environments (SNR = 20dB - 5dB)

2.3 谱熵和谱减法相结合的端点检测改进算法

由上分析可知,对带噪语音而言,谱熵特征的语音端点检测优于能量特征,但当带噪语音的 SNR 很低时,其带噪语音的谱熵与噪声谱熵非常相似,严重影响了算法的鲁棒性. 为了能有效地检测较低 SNR 的带噪语音的端点,我们尝试性地提出利用谱减法和谱熵相结合的端点检测算法. 算法的基本思想是:在利用谱熵检测语音端点前,先用语音增强(谱减法)去除噪声信号,然后对其谱减后的语音利用谱熵进行端点检测.

2.3.1 基于背景噪声的谱减算法

谱减法是语音增强中最常用的一种算法,由于该算法的计算复杂度低、实时性强,一直受到了语音增强研究者的广泛重视,在基本谱减算法的基础上,产生了很多改进算法. 本文的改进谱减算法利用高频区估计背景噪声[10](平稳或非平稳). 算法原理为:

$$|\hat{S}_m(f)|^2 = \begin{cases} |Y_m(f)|^2 - \gamma |D_m(f)|^2, \\ \quad |Y_m(f)|^2 \geqslant |D_m(f)|^2 \\ 0.015 \times |Y_m(f)|^2, \\ \quad |Y_m(f)|^2 \leqslant |D_m(f)|^2 \end{cases} \qquad (4)$$

式中 γ 因子的取值为 1 或 2,$|\hat{S}_m(f)|$ 为增强后的能量谱,$|D_m(f)|$ 为估测的噪声能量谱,估测算法为对带噪语音进行上采样,如上采样至 20 KHz 以上并求解:

$$\mu = \text{average}[|X(f)|] \quad 10\,\text{KHz} < f < F_S/2 \tag{5}$$

然后得其估计的背景噪声:

$$|D_m(f)| = \mu \tag{6}$$

将估测的$|D_m(f)|$带入(4)式可求得增强后的语音信号谱$\hat{S}_m(f)$. \ =2. **3. 2 改进的谱熵端点检测算法**

为了能进一步调整谱熵的频率分量,利用谱减后的$\hat{S}_m(f)$求谱熵时,我们提出了加权的谱熵改进算法:

$$H'_m = -\sum_{K=1}^{N} w_k p'_k \log p'_k \tag{7}$$

式中 w_k为加权因子,$0 < w_k \leq 1$, p'_k是利用$\hat{S}_m(f)$按(1)、(2)、(3)式求得的 pdf.

3　算法的实现(Implementation of the algorithm)

设 ξ为用式(8)所求得的谱熵集,即:

$$\xi = \{H_1, H_2, \cdots, H_M\} \tag{8}$$

其中 M 为分析语音的总帧数. 利用 ξ求解判别语音断点的门限 γ,本文的门限设定算法为:

$$\gamma = \frac{\max(\xi) - \min(\xi)}{2} + \delta\min(\xi),\ \delta > 0 \tag{9}$$

当 γ确定后,可用下面的判决准则判断本帧是否为语音帧:

$$\xi = \begin{cases} \xi_j & \text{如果 } \xi_j \geq \gamma \\ 0 & \text{其它} \end{cases} \quad j = 1, \cdots, M \tag{10}$$

即当 $\xi_j = H_j$ 时,第 j 帧为语音帧. 从式(10)可知,高于门限 γ的帧都判断为语音帧,但由于声道干扰以及在谱减增强后的语音信号中可能存在残渣噪声,经(10)式判断的语音段可能有些不是真正的语音段. 众所周知,语音段至少应有一定的最小长度,语音段间存在一定的距离,由此,我们提出了基于门限的语音端点判决准则.

设第 j 个语音段的长度 $\pi_j = e_j - s_j$,其中 e_j、s_j为第 j 个语音段终点和起点,第 j 个语音段和紧接其后的第 1 个语音段之间的距离为 d_{jl}. 并设语音段的最小长度为 ρ,两语音段间的最短长度为 η(ρ、η依实验设定,即若是语音段,其长度不得低于 ρ,并且两语音段间的最小长度不得低于 η). 则语音端点的判决准则为:

准则 1:如果 $\pi_j < \rho$,则第 j 帧为非语音帧,将其合并至 d_{jl}中,即 $d_{jl} = d_{jl} + \pi_j$,修改 ξ和并调整语音帧序号;

准则 2:如果 $\pi_j > \rho$或$\pi_l > \rho$并且 $d_{jl} < \eta$,则第 j、l 帧为同一语音帧,并将其合并,修改 ξ和并调整语音帧序号;

准则 3:如果 $\pi_j > \rho$或$\pi_l > \rho$并且 $d_{jl} > \eta$,则第 j、l 帧分别为独立的语音段.

按上述准则对所分析的语音进行合并、归类,最后可检测到正确的带噪语音的端点.

本文所述的语音端点检测器(endpoint detector)的实现框图如图 3 所示.

4　实验结果及分析(Experimental results and analysis)

语音端点的检测方法是在不同的背景噪声环境下进行的,为了评价上述语音端点检测算法的效果,我们分别进行了在平稳和非平稳噪声环境下的实验.

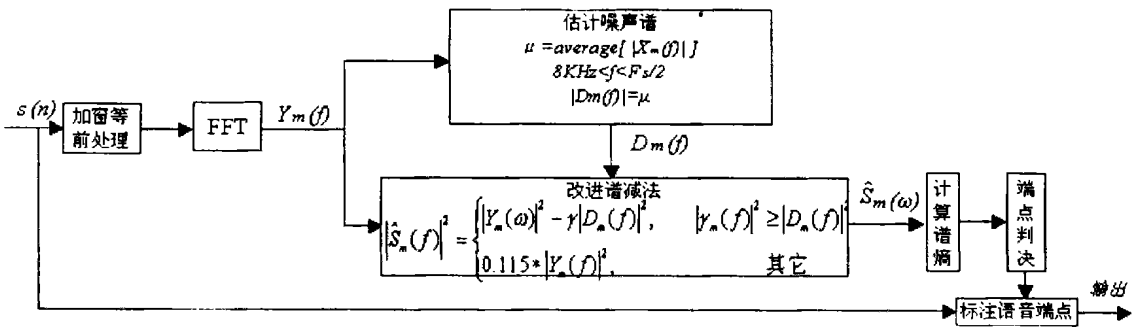

图 3　语音端点检测器的实现框图

Fig. 3　Block diagram of speech endpoint detector

本文实验条件是:采样频率为 11.025 KHz,16bit 量化,加 Hamming 窗,窗长为 256,帧间重叠 50 %,FFT 变换长度为 512,最短语音的长度 ρ 设为 10 帧,两语音间的最短长度 η 设为 8 帧. 采集说话人语音时,要求为正常发音,即音间有正常停顿. 平稳噪声为高斯白噪声,非平稳噪声为工厂噪声. 图 4 给出了某段汉语语音信号在受不同幅度的平稳噪声干扰下所测得的语音端点,图 4(a)为干净语音信号及手工标注的语音端点,图 4(b)、(c)、(d)、(e)、(f)分别为语音信号受到不同幅值的平稳噪声干扰后所测得的语音端点,其中图 4(b)到(f)对应的语音信号的 SNR 分别为 15dB、10dB、5dB、0dB、- 5dB.

从图中可知,随着 SNR 的降低,所检测的语音端点的准确度也有所降低,当 SNR 为 5dB 时,可检测到较为准确的语音端点(如图 4(d)所示);但当 SNR 低于 0dB 时,部分语音的端点检测不到,如图 4(f)所示,这是由于噪声的幅值已超出了语音所致.

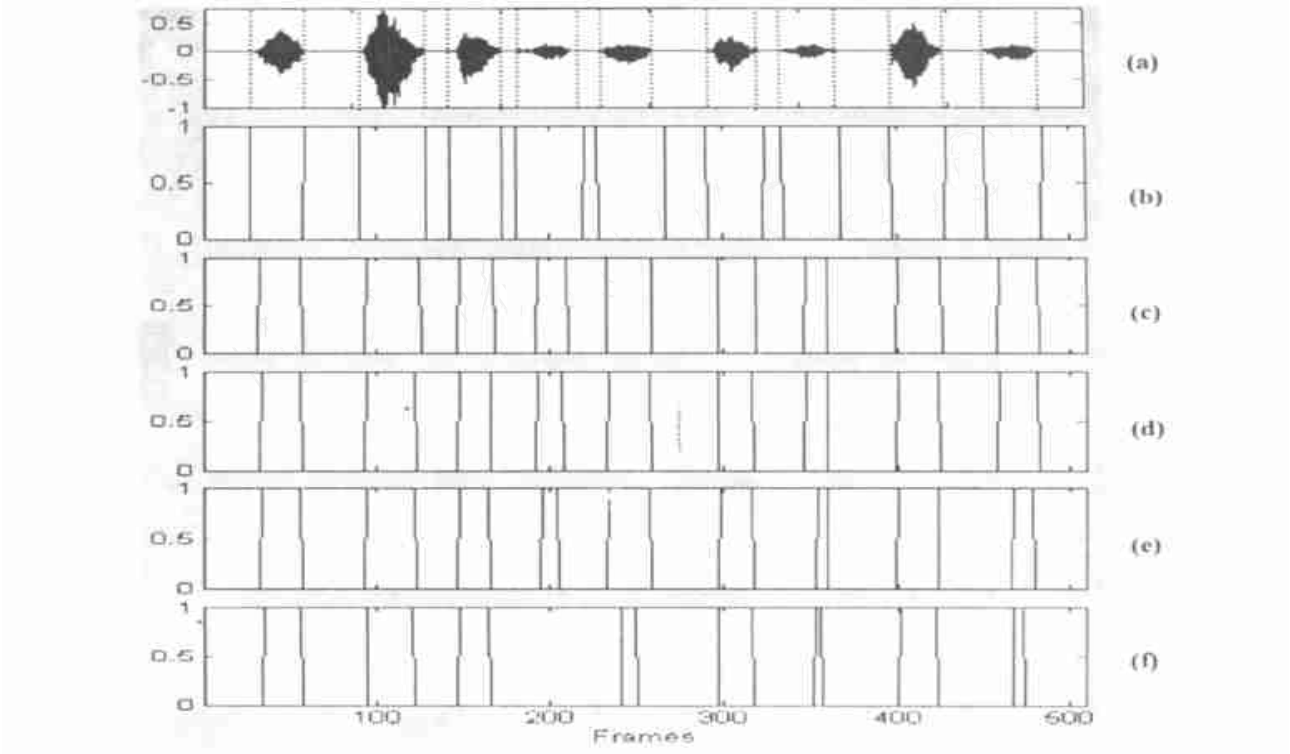

图 4 端点检测实验结果

Fig. 4 Experimental results of endpoint detection

为了说明本文所提算法的有效性,我们进行了大量的端点检测实验. 表 1 给出了在不同背景噪声干扰下,采用不同的端点检测方法所测得的准确率,准确率按下式测量:

P_{RATE} = 在不同背景噪声下测得的正确的语音端点数/ 总实验语音样本的端点数

其中,所测正确的语音端点数与手工标记的相比,允许有 ±10 %的误差.

表 1 语音端点实验检测结果

Tab. 1 Experimental results of speech endpoint detection

检测率	正确率									
	白噪声(SNR)					工厂噪声(SNR)				
	15	10	5	0	- 5	15	10	5	0	- 5
ENG/ CRZ	0.96	0.76	0.64	0.43	0.12	0.94	0.71	0.57	0.38	0.07
SEP	0.97	0.78	0.70	0.56	0.30	0.95	0.73	0.64	0.47	0.22
SSP	0.99	0.86	0.79	0.70	0.64	0.96	0.83	0.73	0.62	0.43

表 1 中的 ENG/ CRZ 表示以能量和过零率测试语音端点的方法,SEP 表示采用谱熵方法[8]测量端点,SSP 为本文所提的方法.

5　结论(Conclusions)

本文提出的语音端点检测算法可有效地从背景噪声中检测语音信号的端点,但当语音信号基本上被强噪声干扰后,即人无法辨听时,本文所提算法也无能为力.目前我们在实用系统的研究开发中,因工作的背景并非是平稳的噪声环境,同时当语音命令控制器远离说话人时,或当串入干扰的背景语音噪声时,语音端点很难检测,直接影响了识别装置的推广应用.本文的端点检测算法仅是一种尝试性的解决方案,该算法已用于我们设计的有限人有限词汇的汉语语音命令(短语)的识别装置中.下一步的研究是从实时角度考虑,用 DSP 技术实现端点检测.

参　考　文　献(References)

[1] Rabiner L R, Sambur M R. An algorithm for determining the endpoints of isolated utterances [J]. Bell System Technical Journal, 1975, 54(2): 297～315.

[2] Li Q, Zheng J, Zhou Q, *et al*. A robust, real-time endpoint detector with energy normalization for ASR in adverse environments [A]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001) [C]. Salt Lake City: 2001, 1. 574～577.

[3] Shin W H, Lee B S, Lee Y K, *et al*. Speech/ non-speech classification using multiple features for robust endpoint detection [A]. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing [C]. Istambul, Turkey: 2000, 3. 1399～1402.

[4] Martin A, Charlet D, Mauuary L. Robust speech/ non-speech detection using LDA applied to MFCC [A]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001) [C]. Salt Lake City: 2001, 1. 685～688.

[5] Kosmides E, Dermatas E, Kokkinakis G. Stochastic endpoint detection in noisy speech [A]. International Workshop on Speech and Computer (SPECOM97) [C]. Cluj-Napoca, Romania: 1997. 109～114.

[6] Shen J L, Hung J W, Lee L S. Robust entropy-based endpoint detection for speech recognition in noisy environments [A]. International Conference on Spoken Language Processing (ICSLP-98) [C]. Sydney, Australia: 1998. 232～238.

[7] Huang L S, Yang C H. A novel approach to robust speech endpoint detection in car environments [A]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00) [C]. Istambul, Turkey: 2000, 3. 1751～1754.

[8] Jia C, Xu B. An improved entropy-based endpoint detection algorithm [A]. International Conference on Spoken Language Processing (ICSLP 2002) [C]. Taipei: 2002. 285～288.

[9] Bou-Ghazale S E, Assaleh K. A robust endpoint detection of speech for noisy environments with application to automatic speech recognition [A]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2002) [C]. Orlando, USA: 2002. 1753～1756.

[10] Yamauchi J, Shimamura T. Noise estimation using high frequency regions for spectral subtraction [J]. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2002, E85-A(3): 723～727.

作者简介

王让定(1962 -),男,副教授,博士研究生.研究领域为语音识别,语音编码,数字水印,计算机应用.

柴佩琪(1935 -),教授,博士生导师.研究领域为模式识别,人工智能等.