噪音情况下语音端点检测方法的研究

朴春俊,马静霞,徐 鹏
PIAO Chun-jun,MA Jing-xia,XU Peng

燕山大学 电气工程学院,河北 秦皇岛 066004
Yanshan University,Qinhuangdao,Hebei 066004,China
E-mail:majing230@yahoo.com.cn

PIAO Chun-jun,MA Jing-xia,XU Peng.Research on voice activity detection method in noisy environment.Computer Engineering and Applications,2007,43(8):49-50.

Abstract: Spectral subtraction for speech signal processing is a classic method for decreasing noise which easily leads to error by regard fixed silence fragment as noise sample.Spectral entropy is an effective voice activity detection method.But detect effect will be much lower with these methods in low signal noise ratio environment,and threshold estimation make use of fixed silence fragment.Therefor,this paper presents a synchronous method of decreasing noise and voice activity detection.The experiment result proves that the method has better detection result.

Key words: spectral subtraction;spectral entropy;decreasing noise;voice activity detection

摘 要:语音信号处理中减谱法是一种传统的降噪方法,但减谱法利用固定的无音片段作为噪声样本容易产生误差。谱熵法是一种有效的端点检测方法,但在低信噪比环境下,检测效果将大大降低,并且门限估计也采用初始的固定无音片段。为此,提出了一种降噪和端点检测同步的方法。实验结果表明,该方法可以得到较高正确率的端点检测结果。

关键词:减谱法;谱熵;降噪;语音端点检测
文章编号:1002-8331(2007)08-0049-02 **文献标识码**:A **中图分类号**:TP391

语音端点检测就是检测语音信号的起点和终点,因此也叫起止点识别[1]。其目标是要在一段输入信号中将语音信号同其它信号(如背景噪声)分离开来。随着语音识别技术的发展和逐步走向应用,语音识别的稳健性问题已经逐步成为语音识别研究的热点。实用性的语音识别系统必须能够应付千差万别的噪声环境,但是现有语音识别系统的性能并不稳健,它们在噪声环境下的性能会极大的下降,其中一个主要的原因就是错误的语音检测。因此,稳健、精确、可靠的语音检测算法是语音识别系统必需的。

噪声中的语音检测是一个比安静环境中的语音检测复杂的问题。传统的端点检测方法,如短时能量、过零率等算法、基于熵以及熵与能量结合的改进算法,在平稳噪声或高信噪比时性能较好,但复杂环境低信噪比下易发生漏检或虚检情况[2]。因此,本文提出了一种降噪和端点检测同步的方法,结果表明此方法的可行性。

1 语音信号的降噪

噪声主要分为加性噪声和乘性噪声。加性噪声叠加在语音信号波形上,用下式表示:

$$x(t)=s(t)+n(t) \tag{1}$$

其中,$x(t)$表示含噪语音信号,$s(t)$表示语信号,$n(t)$表示噪声信号。

乘积性噪声又称为卷积噪声,乘积性噪声可以通过同态变换成为加性噪声,因此对加性噪声的讨论具有代表性。

减谱法是处理宽带噪声较为传统和有效的方法之一[3],其基本思想是在假定加性噪声与短时平稳的语音信号是不相关的,并且它们在频域上是加性的,因此,从带噪语音的功率谱中减去噪声功率谱,从而得到较为纯净的语音频谱。其推导如下:

对式(1)作傅里叶变换,并用 $X(\omega)$、$S(\omega)$、$N(\omega)$表示 $x(t)$、$s(t)$、$n(t)$对应的傅里叶变换,则有

$$X(\omega)=S(\omega)+N(\omega) \tag{2}$$

$$|X(\omega)|^2=|S(\omega)|^2+|N(\omega)|^2+S(\omega)N(\omega)^*+S(\omega)^*+N(\omega) \tag{3}$$

其中,$|X(\omega)|^2$、$|S(\omega)|^2$、$|N(\omega)|^2$ 表示对应的功率谱,$S(\omega)^*$、$N(\omega)^*$分别表示 $S(\omega)$、$N(\omega)$的复共轭。按照上面的假设,语音信号 $s(t)$与噪声 $n(t)$是不相关的,所以 $S(\omega)$、$N(\omega)$二者的乘积项为零,于是有:

$$|X(\omega)|^2=|S(\omega)|^2+|N(\omega)|^2 \tag{4}$$

$$|S(\omega)|^2=|X(\omega)|^2-|N(\omega)|^2 \tag{5}$$

然而,在具体运算时,为了防止出现负功率谱的情况,减谱时当$|X(\omega)|^2<|N(\omega)|^2$ 时,令$|S(\omega)|^2=0$,即完整的减谱运算公式如下:

基金项目:河北省研究性发展计划(No.04213537)。
作者简介:朴春俊(1946-),男(朝鲜族),副教授,主要研究方向:智能控制、语音识别。马静霞(1980-),女(汉族),硕士研究生,主要研究方向:语音识别。徐鹏(1981-),男(汉族),硕士研究生,主要研究方向:视觉伺服。

$$|S(\omega)|^2=\begin{cases}|X(\omega)|^2-|N(\omega)|^2 & |X(\omega)|^2\geqslant|N(\omega)|^2\\ 0 & |X(\omega)|^2<|N(\omega)|^2\end{cases} \tag{6}$$

由于在大多数情况下只能获得含噪语音,所以式(6)中$|N(\omega)|^2$无法直接计算出来。通常利用最开始的10帧信号的平均功率谱$E[|N(\omega)|^2]$来近似代替$|X(\omega)|^2$。式(6)可改写为

$$|S(\omega)|=\begin{cases}(|X(\omega)|^2-|N(\omega)|^2)^{\frac{1}{2}} & |X(\omega)|^2\geqslant|N(\omega)|^2\\ 0 & |X(\omega)|^2<|N(\omega)|^2\end{cases} \tag{7}$$

然后对$|S(\omega)|$作傅里叶反变换,就可以得到降噪后的语音信号$s(t)$。

减谱法有一个重要的缺陷,即始终用最开始的10帧信号中的噪声来估计整段语音的噪声,因为噪声是随机变化的,所以肯定会产生很大的误差。

2 谱熵的基本原理

对带噪语音信号$s(n)$经分帧、加窗求解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概率密度函数定义为

$$p_i=\frac{Y_m(f_i)}{\sum_{k=0}^{N-1}Y_m(f_i)}\quad i=1,\cdots,N \tag{8}$$

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250 Hz至6 000 Hz,为了增强概率密度函数区分语音和非语音段的能力,对式(8)引入约束条件:

$$Y_m(f_i)=0,f_i<250\text{ Hz}\ \text{或}\ f_i<6\ 000\text{ Hz} \tag{9}$$

考虑上述约束条件后,每个分析语音帧的短时谱熵定义为

$$H_m=-\sum_{k=1}^{N}p_k\log p_k \tag{10}$$

为了进一步提高端点检测正确率,本文采用短时能量加权的改进算法,将式(10)改为

$$H'_m=-E_m\times\sum_{k=1}^{N}p_k\log p_k \tag{11}$$

其中$E_m=\sum_{k=1}^{N}x^2(n)$表示第m帧语音信号的时域能量。$H_i(X)'$可看成是谱熵H_m和时域能量E_m的加权。

3 端点检测方法

用谱熵法进行语音端点检测的步骤:

(1)确定噪声的门限值

无音片段主要包含的是背景噪声。由于录音开始阶段往往有一段无音区,所以在实验审环境下通常取最开始的10帧信号作为对背景噪声的分析。对这10帧信号的按照式(11)计算每帧的谱熵值。通过多帧平均H,就得到其平均值,并按照下式确定噪声的门限值TH。

$$TH=k\times H \tag{12}$$

式中k为经验值,通常取1.4。

(2)利用谱熵值进行语音端点检测

计算每帧语音信号的谱熵值,与噪声的门限阈值做比较。大于TH,就以该帧的帧号作为有音片段的起点N_1,表明进入了有音片段。如果由过去帧已经得到了N_1,那么当小于TH时,就以该帧的帧号作为有音片段的终点N_2。相反,如果N_1还未得到,那么当小于TH时,表明当前帧仍处于无音片段。

按照谱熵法对语音端点检测进行仿真研究。下面是语音数字信号"5"的仿真结果,如图1所示,语音信号是在实验室环境下录制的。

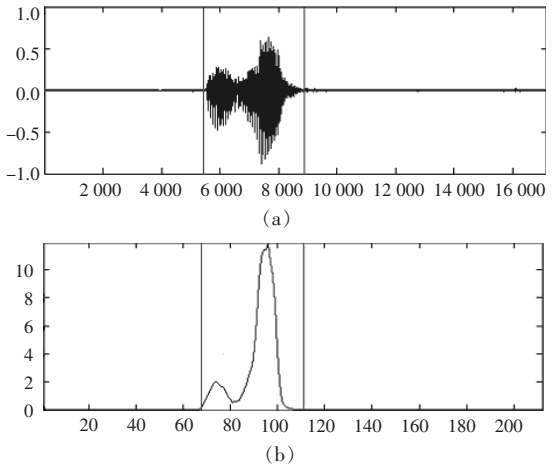

图1 数字"5"的检测结果

在实验室环境下,利用上述方法进行语音端点检测可以达到比较好的检测效果。但是当背景噪声较强时,有音片段的开始部分和结束部分容易被噪声淹没,从而会大大影响语音端点检测的准确性。

下面以谱熵法为基础给出应用减谱法进行端点检测的详细步骤:

(1)初始噪声估计:

在开始阶段,取前10帧作为无音片段,用来估计噪声。对无音片段的每一帧数据计算256点FFT,然后计算多帧的平均值,计算出该无音片段的平均功率谱$E[|N(\omega)|^2]$,即为噪声的估计值。

(2)进行降噪处理:

按照式(7)计算出当前帧语音信号的频谱值,再做256点IFFT,就得到了当前帧降噪后的语音信号$s(t)$。但由于本文采用谱熵法,因此可以直接用式(7)得到的语音信号频谱值来计算谱熵值。

(3)按照上面介绍的谱熵法进行语音端点检测。

(4)噪声值的更新。

当数据已经取完时,就结束语音端点检测,否则继续进行操作。如果步骤(3)表明当前帧仍处于有音片段,就取出下一帧数据,然后转步骤(2)执行。如果步骤(3)表明当前帧仍处于无音片段,就取当前帧的数据,并将它们与上一次用到的无音片段的数据按下式作加权平均,实现了对噪声的更新:

$$N=\alpha N+(1-\alpha)A \tag{13}$$

式中N为原噪声平均功率谱,A为当前帧信号平均功率谱,α为调节参数,本文α=0.3。

4 实验与结论

实验条件:在实验室环境下录制采样频率为8 KHz的数据和16 bits量化,加Hamming窗,窗长为256,帧移为80,FFT变换长度为256。端点检测过程基于Matlab仿真实现[5]。白噪声是一种非常典型的加性噪声,因此语音端点检测方法在白噪声条件下进行测试。

(下转53页)

从图 3 中可以发现,对于相同的发音,其曲线变化规律十分相似,不同的发音则有所差异,本文正是利用这个特性来进行识别。

然后利用 PNN 对语音样本进行训练,采用 MFCC+△MFCC 作为特征参数,输入神经网络进行训练。

最后将训练好的两个语音库对各个测试样本进行检测,检测算法如下:

当得到一个测试样本的 IPCNN 参数以及 PNN 概率参数后,设值 2 个域值 *theta*、*beta*。首先求出其对于 PNN 语音库的概率值,检测其中最大的一个,当其大于 *beta* 时,则其所在位置对应于待检测的语音序号。否则求取测试样本与 PCNN 语音库内不同发音数据的平均欧式距离,然后检测其中最小的一个,当其小于 *theta* 时,则其所在位置对应于待检测的语音序号。实验表明通过这种预处理能够提高检测准确度。从大量实验中得到的经验参数为 *theta*=0.03,*beta*=0.8。

当测试样本无法满足上述条件,则需要进行二次判断:

求取其与平均欧式距离差的绝对值,选择其中最小的 3 个结合 PNN 的参数进行判断,采用混合参数 $Mixc=\frac{pnn\ 参数}{ipcnn\ 参数}$,当对应的 PNN 参数越大、IPCNN 参数越小,则混合参数 *Mixc* 越大,最后可以通过峰值检测得出最终识别结果。

现采用单独的 IPCNN 和 PNN 方法对同样的一组数据进行识别,然后将 IPCNN 和 PNN 结合起来采取上述的算法对数据进行识别,识别结果如表 1 所示。

通过表 1 的实验结果表明,采用 PCNN 和 PNN 结合使用方法能够有效提高对数字语音的识别率。

表 1　实验结果

识别	IPCNN(各个数字识别结果,共 100 个测试数据)										准确率
方案	0	1	2	3	4	5	6	7	8	9	75%
结果	8	10	9	7	6	9	1	10	5	10	
识别	PNN(各个数字识别结果,共 100 个测试数据)										准确率
方案	0	1	2	3	4	5	6	7	8	9	66%
结果	10	9	10	10	1	10	3	0	3	10	
识别	IPCNN+PNN(各个数字识别结果,共 100 个测试数据)										准确率
方案	0	1	2	3	4	5	6	7	8	9	92%
结果	10	9	10	9	8	10	10	10	6	10	

5　结论

通过以上实验表明利用改进的 PCNN 和 PNN 神经网络相结合能够实现说话人数字语音识别,识别率达 92%,并且其识别精度会随着样本库的增大而不断提高。但如果外部出现强烈噪声的情况下,利用 PCNN 识别语谱图的准确度会下降,对 PNN 的模式识别能力也会产生很大干扰,如何在强干扰下还能够保证较高的识别率,对于这方面的工作还有待于进一步研究。(收稿日期:2006 年 8 月)

参考文献:

[1] Eckhorn R,Reitboeck H J,Arndt M,et al.Feature linking via synchronization among distributed assemblies:simulation of results from cat visual cortex[J].Neural Computing,1990,2(3):293-307.

[2] Johnson J L,Padgett H J,Arddt M,et a1.PCNN model and applications[J].IEEE Transaction on Neural Networks,1999,10(3):480-498.

[3] Kuntimad G,Ranganath H S.Perfect image segmentation using pulse coupled neural networks[J].IEEE Transactionon Neural Networks,1999,10(3):591-598.

[4] 马义德,袁敏,齐春亮,等.基于 PCNN 的语谱图特征提取在说话人识别中的应用[J].计算机工程与应用,2005,41(20):81-84.

[5] Specht D F.Probabilistic neural networks[J].Neural Networks,1990,3(2):109-118.

(上接 50 页)

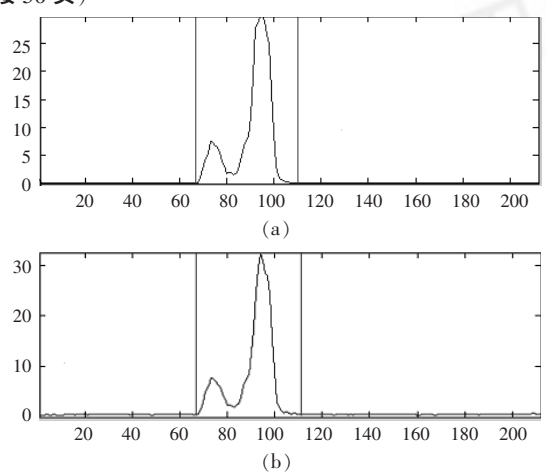

图 2　低信噪比下的检测结果

对上面的数字信号"5"与不同电平的白噪声混合作为测试样本,图 2 为 SNR 分别为 5 dB、0 dB 情况下端点检测结果。由图可知在信噪比比较低的情况下端点的检测结果与图 1 中的检测结果基本吻合,进一步说明使用了本文的降噪方法后,语音信号可以得到了有效的恢复。(收稿日期:2006 年 7 月)

参考文献:

[1] 陈尚勤,罗承烈,杨雪.近代语音识别[M].成都:电子科技大学出版社,1991:7-44.

[2] Wilpon J G,Rabiner L R,Martin T B.An improved word detection algorithm for telephone quality speech incorporating both syntactic and semantic constraint[J].AT&T Tech Journal,1984,63(3):479-498.

[3] 胡航.语音信号处理[M].哈尔滨:哈尔滨工业大学出版社,2000:191-194.

[4] Huang L S,Yang C H.A novel approach to robust speech endpoint detection in car envi-ronments [C]//IEEE International Conference on Acoustics,Speech,and Signal Processing,2000,3:1751-1754.

[5] 何强,何英.MATLAB 扩展编程[M].北京:清华大学出版社,2002:293-300.