文章编号:1005－0523(2007)05－0135－04

噪声背景下语音识别中的端点检测

张跃进,刘邦桂,谢 昕

(华东交通大学 信息工程学院,江西 南昌 330013)

摘要:整个实验系统主要研究的是语音识别中在噪声背景下的端点检测.在实际环境中,加性背景噪声对语音识别的影响非常大,在信噪比小的情况下,无法进行端点检测.在此研究过程中,首先进行加性噪声模型的建立,而后通过滤波对带噪语音信号进行预处理,这里采用维纳滤波和自适应 LMS 滤波进行滤波.在进行语音信号的预处理后,使用 VAD 算法进行端点检测.建立整个实验系统后,通过试验比较维纳滤波器和自适应 LMS 滤波器的优劣,找出最佳方案.

关 键 词:语音识别;噪声;滤波;端点检测

中图分类号:TP391.9 **文献标识码**:A

1 引言

目前的语音识别系统对纯净语音可以达到非常高的识别精度,但是无处不在的噪声带来了训练模型和测试语音之间的失配,识别器的性能在噪声环境中将会急剧下降.语音识别已取得重大的进展,正在步入实用阶段.但目前的识别系统大都是在安静环境 下工作的.在噪声环境中尤其是在强噪声环境下,语音系统的识别率将受到严重影响.因此滤除噪声问题是语音识别达到真正实用所必须解决的关键[1].

在包含语音信号处理的信息和信号处理学科中,维纳滤波、卡尔曼滤波和自适应滤波是几种基础和较为重要的方法.维纳滤波和卡尔曼滤波是最优线性最小二乘滤波,他们广泛应用于预测、估计、内插、信号和噪声的过滤等领域.自适应滤波理论和技术是统计信号和非平稳随机信号处理的主要内容,它具有维纳滤波和卡尔曼滤波的最佳滤波性能,但不需要先验指示的初始条件,它通过自学习来适应外部自然随机环境,因而自适应滤波器可以用来检测确定性信号,也可以检测平稳的或非平稳的随机过程.

2 语音识别概念

语音识别是机器将语音信号转变为相应的文本文字或命令的技术,即将语音信号逐字逐句的翻译为相应的书面语言,或对语音所包含的要求和命令做出正确的响应.其根本目的是研究出一种具有听觉功能的机器,这种机器能直接接受人的语音,理解人的意图,并做出相应的反应[2].

语音识别的系统根据不同的要求,可以有不同的分类方法:

1) 按词汇量的大小分:通常分为小词汇量、中词汇量和大词汇量.

2) 按发音的方式分:语音识别可以分为孤立词识别、连接词识别、连续语音识别以及关键词检出等.

3) 按识别对象的类型来看:语音识别可以分为特定说话人识别和非特定说话人.

4) 按语音识别的方法分:有模板匹配法、随机模型法和概率语法分析这些方法都属于统计模式识别方法.

语音识别技术主要包括端点检测技术、特征提取技术、模式匹配准则及模型训练技术三个方面.端点检测的目的就是在复杂的应用环境下的信号流中分辨出语音信号和非语音信号,并确定语音信号的开始及结束. 一般的信号流都存在一定的背景噪声,而语音识别的模型都是基于语音信号训练的,语音信号和语音模型进行模式匹配才有意义.

收稿日期:2007－05－14

基金项目:华东交通大学校立科研基金资助(06ZKXX08)

作者简介:张跃进(1978－),男,湖北钟祥人,华东交通大学信息工程学院讲师,硕士,研究方向:语音识别、通信系统安全.

3 端点检测

3.1 端点检测介绍

语音信号起止点的判别是任何一个语音识别系统都必不可少的组成部分.因为只有准确的找出语音段的起始点和终止点,才有可能使采集到的数据是真正要分析的语音信号,这样不但减少了数据量、运算量和处理时间,同时也有利于系统识别率的改善.人的声音分为清音和浊音两种,浊音为声带振动发出,对应的语音信号有幅度高、周期性明显的特点,而清音则不会有声带的振动,只是靠空气在口腔中的摩擦、冲击或爆破而发声[3].

实际上,一般是用短时能量的概念来描述语音信号的幅度的.对于输入的语音信号 $x(n)$,其中 n 为采样点,首先进行分帧的操作,将语音信号分成每20～30毫秒一段,相邻两帧起始点之间的间隔为10毫秒,也就是说两帧之间有10～20毫秒的交叠.由于采样频率的差异,帧长和帧移所对应的实际采样点数是不同的.对于8KHz采样频率,30毫秒的帧长对应240点,记为N,而10毫秒的帧移对应为80点,记为M.

对于第 i 帧,第 n 个样本,它与原始语音信号的关系为:

$$x_i(n)=x[(i-1)M+n] \qquad (3-1-1)$$

第 i 帧语音信号的短时能量可以用下面几种算法得到:

$$e(i)=\sum_{n=1}^{N}|x_i(n)|$$

$$e(i)=\sum_{n=1}^{N}x_i^2(n)$$

$$e(i)=\sum_{n=1}^{N}\log x_i^2(n) \qquad (3-1-2)$$

它们分别为绝对值的累加,平方累加和平方的对数的累加.选择其中的一种就可以.

将语音信号分帧后计算每帧的短时能量,再设定一个门限,就可以实现一个简单的端点检测算法.但是这样的算法是很不可靠的,因为人的语音分清音和浊音两种,浊音为声带振动发出,对应的语音信号有幅度高、周期性明显的特点,而清音则不会有声带的振动,只是靠空气在口腔中摩擦、冲击或爆破而发声,其短时能量一般比较小.如声母"s"、"c"等幅度很低,往往会被基于能量的算法漏过去.

3.2 端点检测的流程

语音信号序列先经过一系列预处理.首先将语音序列去直流(即减去平均值),再作归一化处理将幅值限制在1之内,然后通过一个预加重滤波器,滤去50Hz的电源干扰和超出一半采样率的频率分量.经过预处理后的语音序列即可进行短时能量计算amp和过零率计算zcr.

我们设置了两个短时能量的门限(一个高门限amp1和一个低门限amp2)和两个过零率门限(一个高门限和zcr1一个低门限zcr2).

首先为短时能量和过零率分别确定两个门限.两个短时能量的门限(一个高门限amp1和一个低门限amp2)和两个过零率门限(一个高门限和zcr1一个低门限zcr2).amp2、zcr2比较低的门限,其数值比较小,对信号变化比较敏感,很容易就会被超过.amp1、zcr1是比较高的门限,数值比较大,信号必须达到一定的强度,该门限才可能被超过.低门限被超过未必就是语音的开始,有可能是时间很短的噪音引起的.高门限被超过则可以基本确信是由于语音信号引起的[4].

整个检测阶段:在静音段,如果能量或过零率超过了低门限(amp>amp2或zcr>zcr2),就应该开始标记起始点,进入过渡段.在过渡段中,由于参数的数值比较小,不能确信是否处于真正的语音段,因此只要两个参数的数值都回落到低门限一下(amp<amp2且zcr<zcr2),就将当前状态回复到静音状态.而如果在过渡段中两个参数中的任一个超过了高门限(amp>amp1或zcr>zcr1),就可以确信进入语音段了.

4 噪声情况下的端点检测

当SNR=10dB时语音信号与噪音信号几乎无法辨别,当SNR为10dB～30dB之间,此时的噪声为强噪声环境;当SNR为40dB～50dB之间,此时的噪声为通常实际中一般环境;当SNR为50dB～60dB之间,此时的噪声为低噪环境.故在通常情况下,SNR在40dB～50dB之间,则该实验中将加入的高斯白噪声的能量以SNR=40dB大小表示,就此进行维纳滤波器和自适应LMS滤波器的方法研究和比较.

4.1 维纳滤波器

在MATLAB中实现维纳滤波的函数为wiener2:

实验结果:测试语音为汉字"开",该wiener大小为[80,80],实际程序见附录.

测试结果:

x1=79

x2=104(此次起止帧数)

由此可以看出,通过维纳滤波后,语音信号得到了主要特征地提取,对噪音有很大程度的消除.但是对于清音部分的语音能量特征不强,经过维纳滤波后,其语音特征被平滑,VAD对其不敏感,所以不能提取其清音部分,清音部分几乎没有被检测出;语音尾部有用信号的能量也较小,与噪音信号大小相近,经过滤波后,有用语音与噪音一起被平滑掉.所以经

过维纳滤波后，语音能量小的部分被滤除，失去了部分的有用语音信号[5].

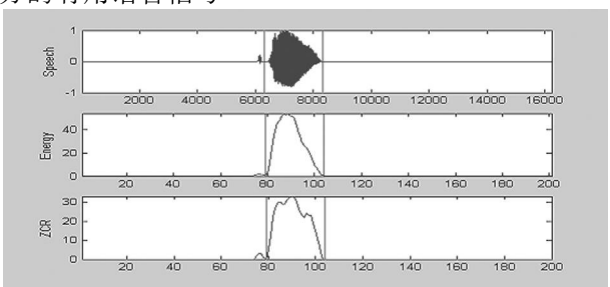

图 1　SNR＝40dB 实验通过维纳滤波的结果

4.2　自适应 LMS 滤波器

在 MATLAB 中实现维纳滤波的函数为 adaptlms：实验结果：测试语音为汉字“开”，该自适应 LMS 滤波器的设定：FIR 滤波器的阶数为 31，mu＝0.413.

测试结果：

x1＝73

x2＝110(此次起止帧数)

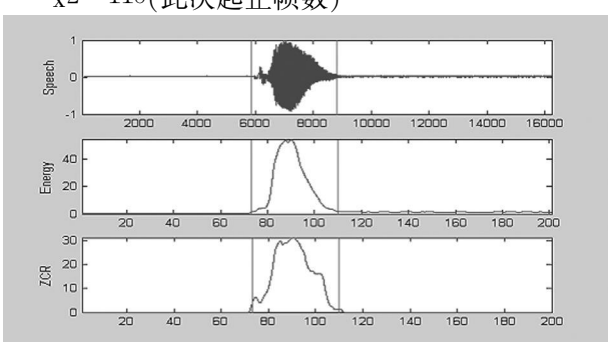

图 2　SNR＝40dB，mu＝0.413 实验通过自适应 LMS 滤波的结果

由图 2 可以看出，自适应 LMS 滤波的检测结果与无(低)噪声情况下的端点检测结果基本相近. 由于自适应 LMS 滤波器为一个性能反馈的自适应的系统，通过自身与外界环境的接触来改善自身对信号处理的性能，当 mu 值恰当时，其自适应的性能使得滤波效果非常好，对有用的语音信号得到了很好的提取.

在自适应 LMS 滤波器中可以通过对参数 mu 的改变而改变滤波器的性能. 而且其滤波的效果对于 mu 的改变很敏感，因此 mu 的确定是很重要的，选择 mu 值是非常重要和关键的一步. 如图 3 为 mu＝0.5 时的检测结果. 可以明显看出其滤波效果大大降低了.

检测结果：x1＝91，x2＝103

对于自适应系数是要通过大量实验所得，通过使用汲取经验，从而很好的确定自适应系数.

由实验结果对比可得，恰当的自适应 LMS 滤波器的检测效果比在维纳滤波器的检测效果要略好些，两者滤波都比较稳定.

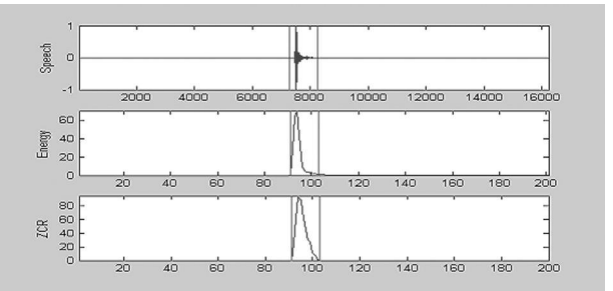

图 3　SNR＝40dB，mu＝0.5 时自适应 LMS 滤波器的检测结果

4.3　对比实验

为了能体现自适应 LMS 滤波器的自适应效果，对两者做了进一步的对比实验，将信噪比降到 10dB 较强的噪声环境下，进行实验.

维纳滤波器：(当 SNR＝10dB 时)

测试结果：

x1＝79

x2＝120(此次起止帧数)

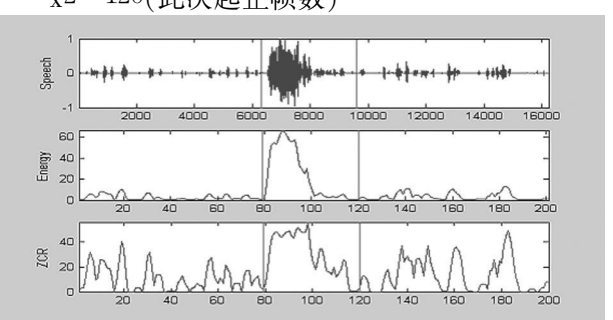

图 4　SNR＝10dB 实验通过维纳滤波的结果果

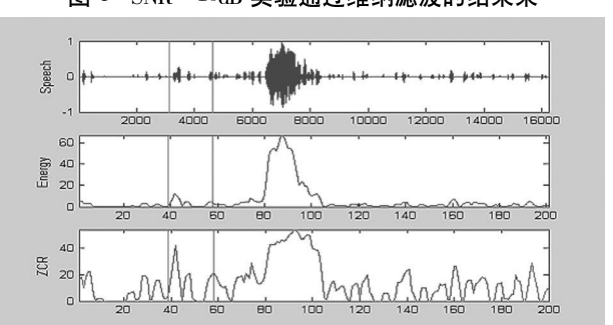

图 5　SNR＝10dB 实验通过维纳滤波的误判结果

由图 4 所示，当 SNR＝10dB 时，维纳滤波可以提取语音有用信号强度大的部分，但是清音部分还是被平滑掉了. 再者由于维纳滤波器为最优的线性最小二乘滤波器，对语音信号集体进行了线性最小二乘的计算，将此结果作为新的语音信号特征. 这种方式的弊端是滤波后的语音信号波形如图 4 有毛刺，噪音信号变化较大的地方通过维纳滤波后，部分特征被提取，即噪音没有很好的被滤除，这些毛刺则会导致在端点检测中被误判，如图 5.

测试结果：

$x1=39, x2=58$(此次起止帧数)
自适应 LMS 滤波器:(SNR=10 mu=0.511)
测试结果:
$x1=76, x2=106$(此次起止帧数)

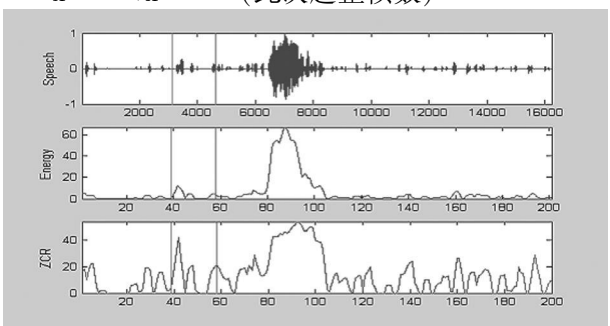

图 6 SNR=10dB, mu=0.511 实验通过自适应 LMS 滤波的结果

由图 6 所示,当 SNR=10dB 时,通过大量实验得到适当的 mu 后,通过自适应 LMS 滤波器噪声基本被滤除,没有毛刺现象,主要的有用语音信号也得到了恢复,虽信号特征与原语音的特征有些差异,但对端点检测的影响不是很大,总体来说自适应的滤波效果还是不错的.

当 SNR 改为 10dB 时,通过对比实验,维纳滤波器的效果明显降低,自适应 LMS 滤波器的效果相对来说较为稳定.由于自适应的系数对整个滤波器有很大的调整作用,使其在输出变化时,对滤波器做了相应的调节,使带噪语音能更好的恢复到纯净语音或是逼近于原信号[6].

5 结论

在通常情况下,信噪比为 40dB～50dB 之间,所以实验中取 SNR=40dB,来对两种滤波器的滤波效果进行研究.在加入 SNR=40dB 的噪声,通过维纳滤波后的端点检测的结果与通过自适应 LMS 滤波器的端点检测的结果进行对比,此时两者滤波效果相似,自适应 LMS 滤波器略胜一筹.通过对比实验,将 SNR 降为 10dB 时,通过维纳滤波器和自适应 LMS 滤波器,此时自适应 LMS 滤波器的效果要好得多.因为维纳滤波器通过最优的线性最小二乘算法来实现去除噪音的,但这种方式在低信噪比的情况下不能得到很好的体现.由于语音信号与噪音信号的幅度相近,很容易干扰其对实际语音信号的提取.而且在噪声变化比较大时,维纳滤波容易产生毛刺,此时端点检测容易产生误判.相较之下,自适应 LMS 滤波器为一个性能反馈的自适应的系统,通过自身与外界环境的接触来改善自身对信号处理的性能.自适应 LMS 滤波器的效果相对来说较为稳定,所以在噪声背景下的语音识别系统中可使用自适应 LMS 滤波器.

参考文献:

[1] 王炳锡,等.实用语音识别基础[M].北京:国防工业出版社,2005.
[2] 王炳锡.语音编码[M].西安:西安电子科技大学出版社,2002.
[3] 赵力.语音信号处理[M].北京:机械工业出版社,2003.
[4] 何方,朱杰,郁桦,等.一种语音信号端点检测方法及其在 DSP 上的实现[J].微型电脑应用,2002,18(5):48－50.
[5] 高瑞华,朱君波,王守觉.一种噪声环境下连续语音识别的快速端点检测算法[J].计算技术与自动化,2003,(23):95－97.
[6] Elliott S J, Nelson P A.Fundamentals of Speech Recognition [M].北京:清华大学出版社.1999.

Vertex Examination in Speech Recognition Under the Noise Background

ZHANG Yue-jin, LIU Bang-gui, XIE Xin

(School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: The entire experimental system discusses vertex examination in speech recognition under the noise background. In the actual environment, the additive background noise influences speech recognition greatly. It is impossible to carry on the vertex examination in less－noisy ratio situation. In this study, additive noise model is established, then pretreatment of the belt chirp pronunciation signal through the filter is conducted. The filter adopts the Wiener filter and the auto－adapted LMS filter. Finally, vertex examination is carried on through the VAD algorithm. After establishing the entire experimental system, the best pattern may be discovered by comparing the Wiener filter and the auto－adapted LMS filter.

Key words: speech recognition; Noise; Filter; Endpoint detection;