基于 G.729 的自适应实时语音活动检测方法研究

刘思伟,吕海波,慕德俊
LIU Si-wei,LV Hai-bo,MU De-jun

西北工业大学 自动化学院,西安 710072
Institute of Automatization,Northwestern Polytechnical University,Xi'an 710072,China
E-mail:lvhaibolaoda@mail.nwpu.edu.cn

LIU Si-wei,LV Hai-bo,MU De-jun.Adaptive and real-time voice activity detection method on G.729.Computer Engineering and Applications,2007,43(34):57-60.

Abstract: On the basis of short time energy and short time cross zero rate,a simple voice activity detection model on G.729 is presented.Some adjust methods are discussed for the model to adapt to different environments automatically.These rules can be realized simply and be used to detect activated voice without any delay.The experimental results show that these rules are effective in voice activity detection and in different voice environments they show better adaptability.

Key words: G.729;voice activity detection;short time energy;short time cross zero rate

摘 要: 提出了一种运用短时能量和短时过零率两个参数进行静音检测的模型,针对不同情况下的语音数据动态调整模型,实现了不同噪声环境下对语音片段的有效检测。该算法实现简单,可实时的对活动语音进行检测不引入延迟。实验结果表明所采用的方法能够比较准确地检测出语音片断,对于噪声环境和音量低的语音都有很好的自适应性。

关键词: 语音编码器;语音活动检测;短时能量;过零率
文章编号: 1002-8331(2007)34-0057-04 **文献标识码:** A **中图分类号:** TN912.3

1 引言

G.729 语音编码算法[1]是国际电信联盟(ITU-T)1996 年公布的语音压缩方法,它以压缩比高,速率低,恢复的音质清晰,编码延迟小等特点,在利用分组交换网传输语音数据的应用领域比如 IP 电话、IPTV、远程数字监控等方面是应用的首选,在其它一些需要保存语音数据的场合比如录音笔等方面的应用该算法也有很高的实用价值。

由于语音数据具有间歇性,在应用中可以只对有效语音数据进行编码压缩、数据传输、保存等处理工作,对静音和噪声不编码或简化处理,这就是语音活动检测技术(Voice Activity Detection,VAD)也称为静音压缩技术,它检测语音数据流中有效语音数据的同时,去除了不必要的静音和噪声或只提供少量的噪声指示信息到编码流中。结合 G.729 应用该技术可以进一步降低编码的数据率,提高通信介质的利用率,降低编码器的计算复杂度,也有利于减少终端设备电力的消耗。在语音识别上应用该技术也有助于增强识别的准确性。

评价语音活动检测技术的主要技术指标在于它应能够准确的检测出有效语音片断,尽可能多的滤除静音片断而对音质不产生影响。而在 G.729 的应用中另一项重要的技术指标就是算法的复杂性。G.729 算法尽管在低速率、低延迟、高音质等方面取得了很好的成绩,但这些是以算法的复杂性为代价的,复杂性已经使得它难于利用微处理器在嵌入式应用领域直接实现,从而影响了该算法的实际应用,因此,基于 G.729 的活动语音检测算法在有效的前提下应力求简单。在对活动语音检测的研究中,ITU-T 公布的 G.729 的附件 B2 是比较早的算法,该算法音质好,但复杂度高于 G.729,难于实用。且该算法过于谨慎,对于短时间静音片断滤除效果一般,不能去除背景噪声。基于小波变换的算法[3]和噪声谱自适应估计的算法[4]同样存在计算复杂的问题,实用性受到影响。单纯利用短时能量的算法[5]虽然简单,但是由于语音信号的复杂性,算法的有效性难以保证,且固定的能量阀值难以适应不同环境下进行活动语音检测的需求。文献[6]提出的双门限法不具有实时性,应用受到局限,且使用的能量幅值在应用中也是固定值,有效性、适应性难以保证。

为了实现简单、有效、实时的进行活动语音检测的目的,本文提出了一种利用短时能量和短时过零率进行活动语音检测的模型,简单有效地实现了对活动语音检测的目的,通过调整模型的相关参数,该模型对高背景噪声、能量很高、能量很低的语音的检测都表现出了很好的自适应性。

2 参数选择依据及特点分析

2.1 参数选用依据

短时能量[7]直观地反映了一段时间内语音信号的强度,它指示了语音信号的存在,因此,短时能量可以作为活动语音检测的参数使用。在 G.729 算法中,需要对输入信号进行预处理,对加窗[7]后的语音信号做自相关运算求取自相关系数 $r(k)$,计算方法为:

$$r(k)=\sum_{n=k}^{239}s'(n)s'(n-k),k=0,\cdots,10 \quad (1)$$

可见当 k=0 时,自相关系数 $r(0)$正好是窗长为 240 的一

段语音信号的能量,因此,在 G.729 的应用中选用短时能量作为参数的好处在于它不增加计算量。

短时能量反映的是信号幅度上的变化,对于幅度小或背景噪声大的信号,单纯用短时能量难以确定语音的起点和终点,而短时过零率[7]在语音的起点终点判别上能起到很好的帮助作用,因此作为选用的一项参数。短时过零率的计算相对比较简单,只是引入自增运算和比较判断,增加的运算量和 G.729 的复杂性相比,可以忽略不计。

2.2 语音数据特点分析

人们在研究中把语音信号大体分为两类,一类是浊音,例如英语中的爆破音、汉语中的元音,浊辅音等大多数是浊音,另一类是清音,例如英语中的摩擦音、汉语中的清辅音大多数是清音。浊音的特点是能量相对比较大而过零率低,清音的特点是能量小过零率高。10 ms 内它们的过零率分布情况如图 1 所示。

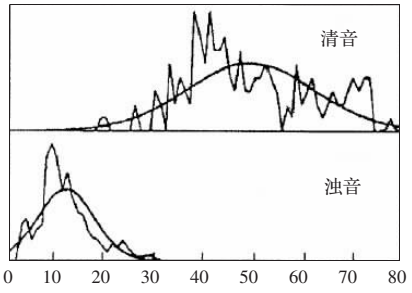


图 1 10 ms 内清音和浊音过零率分布情况

从图 1 中可见,浊音的过零率在30 以下,清音的过零率基本在 15 以上。

通过大量实验发现,语音数据还具有如下特征:

(1)静音片断的短时过零率和能量都很低;

(2)噪音片断在能量和过零率上它可以和语音信号的特点很相似难以区分,也可能存在一些不同。不过总体说来,在背景确定的情况下,背景噪音的短时能量基本上在一定的范围内,且不会发生突增或突减的情况;

(3)从静音片断到语音片断无论是浊音还是清音都会发生短时能量的突增,从清音或噪音到浊音的过渡短时能量的突增也比较明显;

(4)从浊音或静音过渡到清音过零率的增加会比较明显,而从清音或噪音过渡到浊音过零率不会有明显增加;

(5)另外在浊音的起始和结尾处往往会伴随着一小段清音;

这些特征是整套检测方法的依据。

3 检测方法分析

3.1 一般情况下的检测模型

一般情况指的就是背景噪音不复杂,信噪比高的情况。

记当前语音帧的能量为 *energy*,过零率为 *zero*,前一帧语音信号的能量为 *last*1_*eng*,过零率为 *last*1_*zero*,记 *last*0_*eng* 为前一个静音片断的能量,*last*0_*zero* 为前一个静音片断的过零率,*vad* 为当前语音帧的状态,*vad*=1 表示当前帧被判断为语音帧,*vad*=0 表示当前帧判断为静音,*last*_*vad* 为前一帧的状态。

在检测过程中首先以短时能量为依据,设定一个能量上限 *E*_max 和一个能量下限 *E*_min,对于能量大于上限 *E*_max 的语音帧判断为有效语音,对于能量小于下限 *E*_min 的语音帧不满足另外的条件时(后面介绍)被判断为静音,对于短时能量处于上下限之间的语音帧按照如下规则判断:

(1)判断当前帧的过零率相对于前一个静音片断的过零率 *last*0_*zero* 是否有明显的变化,大于 2 倍认为变化明显,且若同时满足该过零率大于 15,则认为当前帧为语音帧。否则判断下面的条件。

(2a)如果前一帧不是语音帧(*last*_*vad*=0),则判断当前帧的能量 *energy* 相对于前一帧的能量 *last*1_*eng* 变化是否明显,这里认为大于 2 倍为明显,如果为条件为真,则当前帧为语音帧;否则当前帧为静音片断,同时更新 *last*0_*eng* 和 *last*0_*zero* 的值。

(2b)如果前一帧是语音帧,判断当前帧的能量 *energy* 是否大于前一个语音帧的能量 *last*1_*eng* 或者是否大于前一个静音帧的能量 *last*0_*eng* 的 4 倍,如果任一条件为真,则当前帧为语音帧;否则当前帧为静音片断,同时更新 *last*0_*eng* 和 *last*0_*zero* 的值。

(3)不满足上述情况,判断为静音,更新 *last*0_*eng* 和 *last*0_*zero* 的值。

(4)记录 *last*1_*eng*,*last*1_*zero*,返回检测结果。

图 2 为一般情况下一段语音信号能量随时间分布的情况,下面以图 2 为例说明上述检测过程。

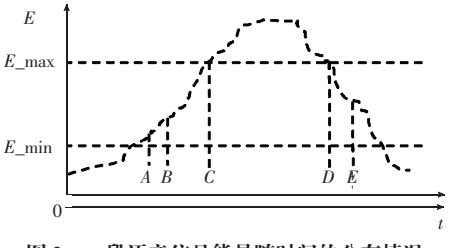


图 2 一段语音信号能量随时间的分布情况

假设 *A* 点以前的语音帧是静音,*B* 帧是紧接着 *A* 帧后面一帧,由于语音的开始多为清音过零率高,假设 *B* 点处的语音帧根据规则(1)被判断为语音帧,那么 *B*~*C* 之间的所有语音帧根据规则(2b)被判断为语音段,*C*~*D* 之间的语音帧因为能量大于能量上限 *E*_max,故这段也被判断为语音帧,假设到了 *E* 帧处,该帧的能量正好是 *last*0_*eng* 的 4 倍(这里 *last*0_*eng* 为 *A* 帧处的能量),则根据规则(2b)*D*~*E* 之间的语音帧也被判断为语音段,而 *E* 帧至能量下限 *E*_min 之前的语音帧被判断为静音片断,由此检测出有效的语音片断为 *B* 帧至 *E* 帧之间的部分。

3.2 信号能量偏大情况下的调整方法

在上面的检测方法中考虑的是一般情况,能量的上下限 *E*_max,*E*_min 也都是固定值,显然这样设定不能满足对所有语音信号的检测要求,如果存在连续的背景噪声且噪声的能量都大于 *E*_max,这种情况下应用上述方法必然是所有语音段包括噪声段都被检测为语音段,而噪声在检测过程中也是希望尽量去除的。

这时其语音能量和能量上下限 *E*_max,*E*_min 的关系大致如图 3 所示。

这时就应该通过一种机制检测出这种情况,然后调高 *E*_max 和 *E*_min 的值到一个合适的位置,使其大致符合检测模型图 2 中的关系,处理方法如下:

(1)检测机制:由于浊音的能量大,所以当信号帧的能量大于能量上限时,正常情况应该是浊音,而到浊音的过渡过零率

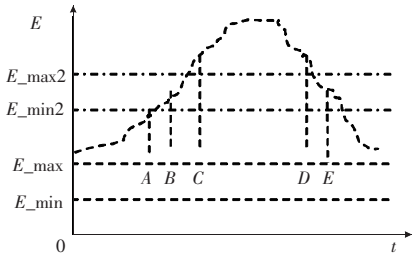


图 3　语音能量偏大时语音能量与设定能量上下限的关系

不会明显增加，因此若发生过零率明显增加的情况，可以认为该帧不是浊音，那么就只能是清音或噪音，而连清音和噪音的能量都高于设定的能量上限就说明上限设得太低了，应该调高能量的上限和下限。

（2）调整方法：调整能量上限到比当前能量稍大一些的位置，下限的值也适当的调高一定的倍数，如图 3 中 *E*_max2，*E*_min2 所示。为防止程序对信号能量的增加过于敏感而经常进行上限提高的调整工作，可以设定当当前帧的能量大于另外一个比较大的值时才进行调整，本文在程序中取该值为 0x3fffffff。

调整完成后，根据一般情况下的检测模型，*B*~*E* 之间的语音段仍被检测为有效语音。

3.3　能量偏小情况的调整方法

由于语音帧比较复杂，当然也存在整体音量比较低，信号帧短时能量比较小的情况，比如打电话时的悄悄话。这时信号帧的能量和能量上下限 *E*_max，*E*_min 的关系大致如图 4 所示。

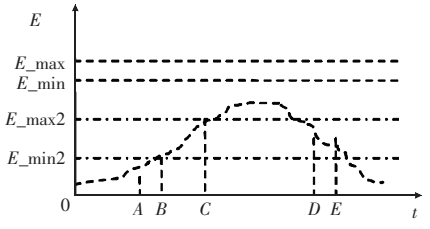


图 4　语音能量偏小时语音能量与能量上下限的关系

这种情况无法应用前面的检测方法，此时应该设定一个机制对这种情况进行检测并降低能量上限和下限的值到一个合适的位置，使其符合一般情况下检测模型中的关系，处理方法如下：

（1）判断当前帧的能量与前一帧的能量相比是否有突然增加，如果当前帧能量大于前一帧能量的 4 倍，则判断当前帧为语音帧，同时调整能量下限 *E*_min 为当前帧的能量 *energy*，能量上限减少一半。为了避免这种能量变化是能量很低的噪声的情况，仅当当前帧的能量大于一个最低值时才进行这种调整，最低值取 0xffff。同时检查上限 *E*_max 不低于一个上限最小值 0xfffff。

（2）如果当前帧的能量大于前一帧能量的 2 倍，认为当前帧为静音片断，但存在能量上下限偏高的情况，用同样的方法降低上限和下限。

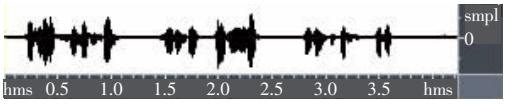
（3）不满足上述情况，判断当前帧为静音片断，更新 *last*0_*eng* 和 *last*0_*zero* 的值为当前帧的值，并只取 *last*0_*zero* 处于 8 至 15 之间的值，否则令 *last*0_*zero*=10。这样做是为了避免当前帧是高频噪声的情况。

以图 4 为例说明上述调整过程，假设设定的能量上限和下限都高于当前语音段的能量，如图中 *E*_max，*E*_min 所示。在 *B* 点处检测到该帧的能量大于 *A* 点处语音帧能量的 4 倍，则认为 *B* 帧为语音帧，调整 *E*_min 到当前帧能量的位置，如图 *E*_min2 所示，调整 *E*_max 到 *E*_max2 的位置。因为 *A* 帧为静音片断，所以 *last*0_*eng* 为 *A* 帧处的能量，这样在 *E* 帧之前的语音帧依据一般情况下的检测模型判别为语音帧。

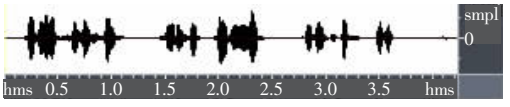
4　检测结果实例分析

4.1　一般情况下活动语音检测的结果

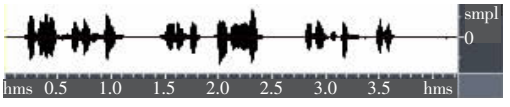
图 5 所示的是截取电视节目主持人的一段录音进行检测的比较结果，讲话内容为“记住了，是晚上，九点半，下周一晚，不见不散，再见”，该段录音信噪比比较高，G.729b 检测的活动语音为 88%，本算法检测的活动语音为 74%，静音压缩效果优于 G.729b。图 6 是局部放大后的波形对比情况，从图中可见，G.729b 没有去除语气停顿形成的静音段。



（a）源文件的波形

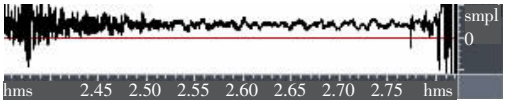


（b）G.729b 检测后输出的波形

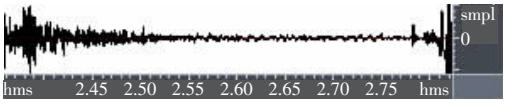


（c）本算法检测后输出的波形

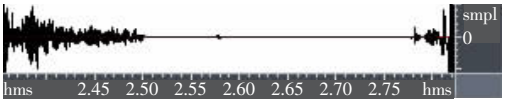
图 5　一般情况下活动语音检测的波形比较



（a）源文件局部放大的波形



（b）G.729b 检测后局部放大的波形



（c）本算法检测后局部放大的波形

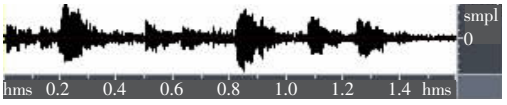
图 6　局部放大的对比

4.2　高背景噪声情况下活动语音检测的结果

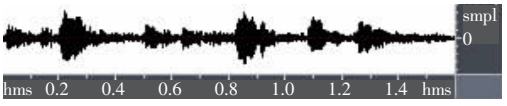
图 7 显示了在高背景噪声情况下进行活动语音检测的对比结果，截取了电影中列车行进中的一段语音，语气比较急促，内容为“boss，quick，come quick”，G.729b 检测到的活动语音为 85%，基本不能消除无用的背景噪声，本算法检测结果为 43%，保留有效语音信息的同时对噪声片断的去除效果很好。

4.3　音量很小情况下活动语音检测结果

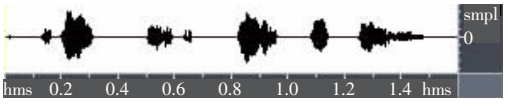
图 8 显示了在音量很小的情况下进行活动语音检测的比较结果，语音为模仿打电话时悄悄话语气的录音，内容为“喂，你好，干啥呢”，G.729b 检测的活动语音为 73%，本算法检测结



(a)源文件的波形

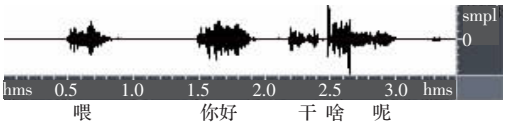


(b)G.729b 检测后输出的波形

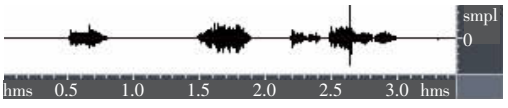


(c)本算法检测后输出的波形

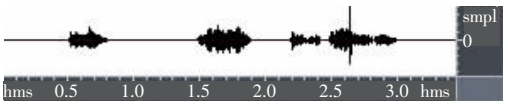
图 7 高背景噪声情况活动语音检测波形比较



(a)源文件的波形



(b)G.729b 检测后输出的波形



(c)本算法检测后输出的波形

图 8 音量很小情况静音检测波形比较

果为 45%,结果显示,尽管该段录音音量比较小,但本算法在检测中语音内容没有任何损失,静音压缩效果优于 G.729b。

5 结论

本文利用语音的短时能量和过零率两个参数设计了一种简单的语音活动检测模型,通过设定不同的检测机制,动态的调整检测模型使其适应不同环境下进行语音活动检测的需要,用很小的计算量实现了实时进行活动语音检测的目的,实验证明本文采用的检测方法能够准确的检测出活动语音数据,有效地去除静音和噪音片断,并能很好的适应不同环境下进行语音活动检测的需求。(收稿日期:2007 年 7 月)

参考文献:

[1] ITU-T Recommendation G.729.Coding of speech at 8kbit/s using conjugate-structure algebraic-code-excited linear-prediction(CS-ACELP)[S],1996-03.

[2] ITU-T Recommendation G.729 Annex B.A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70[S],1996-11.

[3] 高戈,胡瑞敏,李德仁.改进的基于小波变换的语音活动检测算法[J].武汉大学学报:信息科学版,2002,1:85-88.

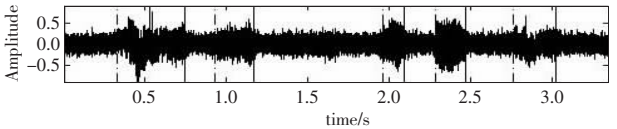
[4] 董恩清,万东辉.基于短时能量和噪声谱自适应估计的语音激活性检测方法[J].模式识别与人工智能,2004,2:227-231.

[5] 郑洪英,郭冬辉,黄国和,等.基于静音检测的 ITU-T G.729 算法[J].厦门大学学报:自然科学版,2002,4:431-434.

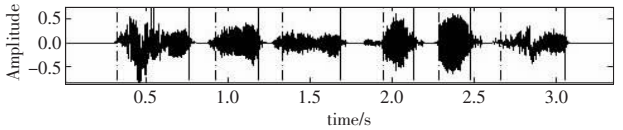
[6] 刘庆升,徐霄鹏,黄文浩.一种语音端点检测方法的研究[J].计算机工程,2003,3:120-121.

[7] 蔡莲红,黄德智,蔡锐.现代语音技术基础与应用[M].北京:清华大学出版社,2003-11.

(上接 56 页)



(a)文献[1]中的方法的端点检测



(b)本文方法的端点检测

图 2 汉语"湖南师范大学"加高斯白噪声 0 dB 时的端点检测比较

4 结论

在短时域端点检测的基础上,针对低信噪比的语音端点检测,提出了一种基于信号子空间法语音增强和信息复杂度结合的语音端点检测方法,给出了一种新的判断语音与非语音的信息复杂度的门限,并在实验中考虑到实际语音的情况,修正了语音端点检测的判断准则。实验仿真表明,本文提出的方法,能明显提高语音端点检测准确率,抗噪性能好,特别是在低信噪比条件下仍然具有较高的端点检测准确率,适用于实际工作环境,因此具有广泛的应用价值。(收稿日期:2007 年 8 月)

表 1 文献[1]方法和本文方法比较

SNR	White		*f*16		factory1	
	文献[1]方法	本文方法	文献[1]方法	本文方法	文献[1]方法	本文方法
15	0.96	0.98	0.94	0.97	0.93	0.96
10	0.86	0.97	0.85	0.95	0.83	0.91
5	0.78	0.95	0.73	0.90	0.68	0.87
0	0.65	0.91	0.57	0.83	0.52	0.79
-5	0.42	0.87	0.36	0.75	0.29	0.72

参考文献:

[1] 侯周国,钱盛友,姚畅.短时域语音端点检测中谱熵算法的改进[J].计算机工程与应用,2006,42(21):55-56.

[2] Ephraim Y,van Trees H L.A signal subspace approach for speech enhancement[J].IEEE Transaction Speech and Audio Processing,1995,3(4):251-266.

[3] Afshin Rezayee,Saeed Gazor.An adaptive KLT approach for speech enhancement[J].IEEE Transaction Speech and Audio Processing,2001,9(2):87-95.

[4] 张学文.组成论[M].合肥:中国科学技术大学出版社,2003:70-75.

[5] 王让定,柴佩琪.一个基于谱熵的语音端点检测改进方法[J].信息与控制,2004,33(1):77-81.