DOI：10.13700/j.bh.1001-5965.2005.12.011
2005年 12月 北京航空航天大学学报 December 2005
第31卷 第12期 Journal of Beijing University of Aeronautics and Astronautics Vol.31 No.12

一种阈值自适应调整的实时音频分割方法

李 超 熊 璋 薛 玲 刘 云

（北京航空航天大学 计算机学院，北京 100083）

摘 要：基于内容的音频分析近年来引起了较多学者的关注.对自动分割方法进行了分析，分割是对音频信号进行深入分析的首要步骤，通常根据能量、幅值、基音等时域或频域的声音特征，设定若干固定阈值来实现.由于实时音频信号来源复杂，环境和采集设备的变化都会导致阈值的波动，从而直接影响到分割的质量.提出了一种基于声音背景学习的阈值初值计算方法，主要针对实时音频应用，设计了环境因子作为对外界环境进行检测的度量，并利用其自适应调节分割阈值，采用查表法，通过状态转移进行分片类型判断以在效率和精度之间取得平衡，并设计了多组分割实验对上述方法分别进行了验证.

关 键 词：实时；自适应性；音频；分割；背景

中图分类号：TP 391

文献标识码：A **文 章 编 号**：1001-5965(2005)12-1317-05

Adaptive threshold method for real-time audio segmentation

Li Chao Xiong Zhang Xue Ling Liu Yun

(School of Computer Science and Technology, Beijing University of Aeronautics and Astronautics, Beijing 100083, China)

Abstract: Content-based audio analysis has become an interesting direction for many researchers. Deep analysis on audio signal segmentation was reviewed. Conventionally, automatic segmentation can be implemented by calculating some audio features like short-term energy, amplitude, fundamental frequency or others, in time-domain or frequency-domain, via referencing to several constant thresholds established in advance. But these methods were found lack of reliability in such applications, because of the complexity of real-time audio signals, together with the fluky changing of environment and various models of acquiring devices. An adaptive threshold adjusting method based on background learning was introduced. On condition of real-time environment, a so-called environment factor was computed iteratively through background learning, and then it was used as a measure to control the fluctuating of real thresholds. To make a balance between efficiency and precision, a state table was introduced to help judging on the types of audio clips. Validity of the methods was proved by a group of experiments.

Key words: real-time; adaptivity; audio; segmentation; background

在音频检索、声音识别、语音传输等应用中，首先需要在分析前端对声音信号进行分割处理，分割的效果会直接影响到所提取特征的质量和对压缩编码算法的选择.以语音编码为例，将语音信号中的噪声、静音和信号段分别提取出来实施不同的参数化编码策略可大大提高编码的效率；与此同时，分割的长度对改善传输效果也有很大的影响，分片(clip)越长则平均的帧头开销越小，相对地传输延迟就越大，反之则帧头开销大而延迟小，因此合适的分割算法需要在负载效率和传输延迟之间求得平衡.

1 相关研究

由于具体应用侧重有所不同，声音自动分割可在多种策略中遴选，如基于距离计算或基于模

收稿日期：2004-09-22

作者简介：李 超(1974—)，男，四川乐山人，博士生，licc@buaa.edu.cn.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

型计算,进行实时分割或非实时分割,选择固定阈值或自适应阈值,在时域或频域来完成,甚至可以直接针对时域幅值包络,在图像领域实现,如Subramanya利用二值图像方法对音频信号进行分割[1].

Thomas等通过德语广播新闻的分割应用对基本音频分割方法进行了比较,将其划分为基于能量、基于度量和基于模型3个主要流派,其报告说明,其中前两者在实时分割中应用较多,后者多用于非实时处理[2].Foote等受到视频分析中关键帧提取的启发,在音频分割中也应用了类似方案,其工作主要在频域完成,通过比较频谱图的相似性来寻找声音变化的关键点[3].该方法需要进行频谱分析,同时要考虑长时特征,很难直接应用到实时系统中;孙文彦等提出了一种以语音时域特征为基础的动态实时分割方法,但初始阈值的选择与自适应调整问题仍悬而未决[4];卢坚等则在说话人识别研究中提出了一种基于一般似然比(generalized likelihood ratio)距离方差的自适应阈值调整算法,该方法实现较复杂[5].本文涉及的实时分割方法以文献[4]的部分工作为基础,分割以帧(frame)为基础,帧为窗口长度固定的若干连续样本的集合,连续的相同类型帧合并为一个分片(clip),故每个分片可包括一个或多个音频帧.

2 环境因子的提出

环境参数用于描述声音信号采集过程中所携带的与周围环境密切相关的信息.在实时音频分割中,影响阈值变化的环境因素众多,量化位数、背景噪声和采集设备是其中最重要的.一般而言,量化位数越多,背景噪声越强,则阈值越高,它们与声音的幅值和能量有关;而引入噪声与采集设备本身也有关,设备质量差则阈值高,如图1所示,在相同环境和量化位数的情况下,普通网络摄像头所附带拾音器(图1a)与专业麦克风(图1b)的采集质量就存在显著的差别,主要表现为信号的标准偏差迥异.环境参数应该能够及时反映出这些变化.



a 摄像头麦克风

b 语音麦克风

图1 相同环境下不同采集设备录入的声音信号

文献[4]、文献[5]都通过不断调整各项参数,利用函数拟合的方法来直接寻求阈值和这些因素之间可能存在的非线性关系.当采取的分割方法改变时就需要重新寻求所有因素同阈值之间的非线性关系,实验过程往往是非常复杂的.事实上,无论分割策略如何改变,环境本身的客观因素并未发生本质变化,因此,如果能够提取出一个较客观的环境参数,再通过将该参数考虑到新策略中则可降低应用新方法带来的未知性.文献[4]通过拟合方法用二次曲线对阈值问题进行处理,本文则通过类似实验并参考文献[5]的结果提出了环境因子 E 的一种对数曲线计算方法,其中对数关系的引入受到了Webber-Fechner法则的影响,该法则用于定性反映感觉与物理特性间的映射关系.

E 计算公式如式(1)所示:

$$E = \alpha \cdot \log_{10} P + \beta/\log_{10} P/V + \varepsilon \qquad (1)$$

式中,P 为时段$(0\sim T)$内的样本平均能量;V 为样本幅值绝对值的平均偏差,显然有 $0\leqslant V\leqslant P$;α, β, ε 为常数.

式(1)所反映的基本关系为:E 的提高预示着环境的恶化.其中,在能量一定的情况下,偏差越大则 E 越大;在偏差一定的情况下,能量提高一方面会导致 E 变大,一方面因偏差与能量之比变小则使 E 变小,这与实际情况相符,其平衡通过 α 和 β 来调节,根据多次实验调整,取 $\alpha=1-\beta=0.3$, $\varepsilon=0.1$.

3 分片类型判别

文献[4]将语音分割为信号、过渡音和静音3种类型,因此所设定阈值主要包括一个静音阈值和一个过渡音阈值.然而通过幅值分析可看到,过渡音存在前后差别,两者在时域上呈现不同的特性,一般而言,前过渡音短促,后过渡音相对缓长,语音信号表现更是如此.心理声学研究中也发现,声音掩蔽现象中前掩蔽的持续时间也远远短于后掩蔽的持续时间.因此,设置单一的阈值,对前后过渡音不加区别并不十分恰当.本文将分片类型区别为信号段、前过渡段、后过渡段和静音段4种类型,设置一个静音阈值和前后2个过渡音阈值.如图2所示,对分片类型的判断分2步,预判,调整,即并不单纯依靠阈值,还需参考前2个分片的结果,但由于并没有直接依赖后续分片,速度上未受明显影响,且进一步提高了判别的准确性.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

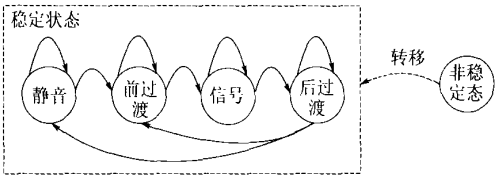


图 2　基本分片状态转移图

令 A, B, C, D 分别代表静音、前过渡音、后过渡音和信号音,考虑连续 3 片信号,组合可能性为 64 种,将每一种组合看作一个状态,合法的 19 种分片组合构成稳定状态,其余 45 种组合构成非稳定状态.规定非稳定状态必须向稳定状态转移,由此,所有的片组合构成一个 64×64 的稀疏矩阵,当出现非法组合时按照事先制定的原则根据状态转移图将其转化为合法组合,转移原则:①时序原则,尽量只改变最近一个分片类型而进入稳定状态;②步长原则,通过最少的步骤进入稳定状态.这种方法为分割过程注入了语义元素,保证了分割结果的合理性.与参考文献[2]、文献[3]中基于模型的方法如贝叶斯信息准则、隐马尔科夫模型或人工神经网络方法相比,本文采用的查表法速度快,无需长时间的训练,更适用于实时分割应用.

4　背景学习方法

如何确定初始阈值是文献[4]遗留的一个重要问题.声音背景学习的思想来自于视频运动对象检测[6]:在大多数视频监控场景中,背景图像变化往往很小,因此常通过求当前帧与背景帧之差提取运动对象,背景帧则根据环境的变化按一定的规则更新,这种方法称为背景减(background subtraction).文献[4]认为,语音间隙的静音段与环境音近似,可以用作阈值调整的依据.因此,本文对每一个静音段(按照滑动窗口大小,一般最少为 30 ms)分别计算其局部能量 P'、绝对幅值均值 M' 和绝对幅值的标准偏差 V',其中初值 P'_0, V'_0, M'_0 需要通过一个较长的背景学习过程(1～5 s)统计得到,环境越稳定学习速度越快,后续值则在分割过程中利用信号间隙的静音按照下面的步骤 2)和 3)反复进行自适应调整,每一步后将用计算出的局部参数更新全局参数:

1) 通过背景学习计算环境参数,并令 $P_0=P'_0$, $V_0=V'_0$, $M_0=M'_0$,计算环境因子 E_0 和初始阈值:

$$\left.\begin{aligned} T_0^{\mathrm{s}} &= M_0\cdot(1+\gamma_{\mathrm{s}}\cdot E_0)\\ T_0^{\mathrm{f}} &= M_0\cdot(1+\gamma_{\mathrm{f}}\cdot E_0)\\ T_0^{\mathrm{b}} &= M_0\cdot(1+\gamma_{\mathrm{b}}\cdot E_0)\end{aligned}\right\} \tag{2}$$

2) 利用静音帧更新各环境参数:

$$\left.\begin{aligned} P_n &= \eta\cdot P_{n-1}+(1-\eta)\cdot P'_n\\ V_n &= \eta\cdot V_{n-1}+(1-\eta)\cdot V'_n\\ M_n &= \eta\cdot M_{n-1}+(1-\eta)\cdot M'_n\end{aligned}\right\} \tag{3}$$

3) 按式(1)计算新的环境因子 E_n;

4) 重新计算各项局部阈值,并更新全局阈值.不同算法对环境因子的使用不同,本文仅以幅值法为例:

$$\left.\begin{aligned} T_n^{\mathrm{s}} &= \eta\cdot T_{n-1}^{\mathrm{s}}+T'^{\mathrm{s}}_n\cdot(1-\eta)\cdot(1+\gamma'_{\mathrm{s}}\cdot E_n)\\ T_n^{\mathrm{f}} &= \eta\cdot T_{n-1}^{\mathrm{f}}+T'^{\mathrm{f}}_n\cdot(1-\eta)\cdot(1+\gamma'_{\mathrm{f}}\cdot E_n)\\ T_n^{\mathrm{b}} &= \eta\cdot T_{n-1}^{\mathrm{b}}+T'^{\mathrm{b}}_n\cdot(1-\eta)\cdot(1+\gamma'_{\mathrm{b}}\cdot E_n)\end{aligned}\right\} \tag{4}$$

式(2)～式(4)中,T_k^{s}, T_k^{f}, T_k^{b} ($k=0,\cdots,n$)分别为全局静音阈值、前过渡阈值和后过渡阈值;T'^{s}_k, T'^{f}_k, T'^{b}_k ($k=0,\cdots,n$)分别为局部静音阈值、前过渡阈值和后过渡阈值;η 为常数,控制阈值更新速度,此处取值为 0.99;γ 和 γ' 为常数,根据算法取值不同,但一般应满足 $\gamma_{\mathrm{s}}<\gamma_{\mathrm{b}}\leqslant\gamma_{\mathrm{f}}$;$n$ 为静音帧序号.

5　实验设计与结果

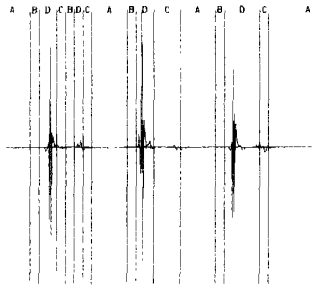
实验分为 4 组,共选择了 3 种常用的时域分割方法进行比较,分别是能量算子法,能量法和绝对幅值法.全部实验在 Pentium 4 台式机,Windows 2000 操作系统下实现,直接通过声卡连接专用语音麦克风实时采集环境声作为系统输入,其中样本采样率为 8 kHz,每样本量化为 16 bit,单声道;每个滑动窗口为 256 样本,即约 30 ms,为简单起见,实验 1～3 中均设窗口重叠率为 0%,具体参数见表 1,其中能量算子阈值分别为 T_0^{s}, T_0^{f}, T_0^{b},能量阈值分别为 $T_{\mathrm{p}}^{\mathrm{s}}$, $T_{\mathrm{p}}^{\mathrm{f}}$, $T_{\mathrm{p}}^{\mathrm{b}}$,绝对幅值阈值分别为 $T_{\mathrm{a}}^{\mathrm{s}}$, $T_{\mathrm{a}}^{\mathrm{f}}$, $T_{\mathrm{a}}^{\mathrm{b}}$,平均幅值为 M,平均能量为 P.传统分割方法可将音频流分割为静音和信号音,文献[4]则增加了过渡音,但由于没有考虑前后音段的相关性,无法区别前过渡或后过渡,本文引入状态转移矩阵则可解决这一问题.

实验 1　通过专用语音麦克风模拟在低噪声环境下的背景学习和分割效果,计算得环境因子为 2.29,采用能量算子法、能量法和幅值法进行比较,其分割结果如图 3 所示.

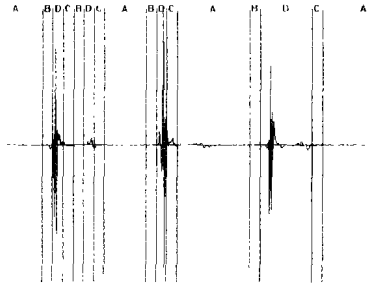
(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表1 实验1和实验2中的各项参数

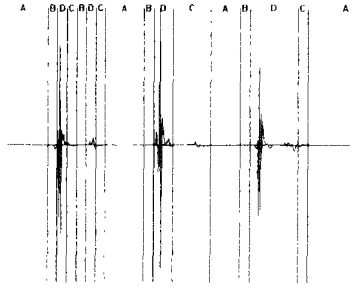
参数名称	实验1	实验2
采集设备	专用语音麦克风	专用语音麦克风
环境	封闭室内	嘈杂室内
E	2.29	2.63
M	1089	1084
V	120	60965
P	1185768	1236647
T_0^s	29045	319941
T_0^f	37516	413257
T_0^b	34491	379930
T_p^s	1764914	1849969
T_p^f	2082236	2202360
T_p^b	2027049	2141075
T_a^s	1296	1314
T_a^f	1548	1598
T_a^b	1510	1547



a 能量算子法



b 能量法

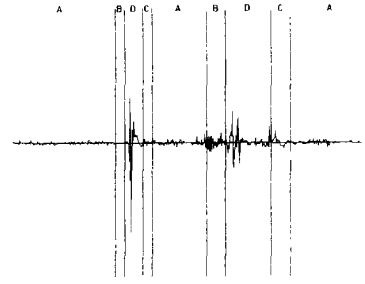


c 幅值法

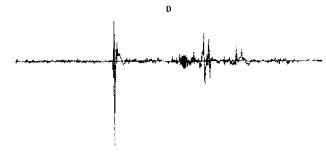
图3 实验1分割效果

实验中3种分割方法取得了基本相同的结果，说明在低噪声环境下，阈值的初始值计算是合理的，环境因子具有一定的普适性.

实验2 改变实验环境，引入机器噪音作为噪声音源模拟较嘈杂环境下的背景学习和分割效果，计算得到的环境因子从2.30增长到2.63，说明环境嘈杂度有了较大提高，采用自适应阈值变化的能量算子法和固定阈值法(具体阈值通过实验1确定)进行比较，分割结果见图4a和图4b.



a 自适应阈值法



b 固定阈值法

图4 实验2分割效果

实验2结果，自适应分割方法仍然能够达到合理的分割效果，说明在环境发生较大变化的情况下，通过本文提出的背景学习方法调整确定的各项阈值仍能够准确完成分割任务.更多重复实验表明，能量算子适应性更好，不同分割方法效果的差别还可通过仔细调整公式中各常数来进一步优化.而在环境变化中，传统固定阈值方法则已经失效，如图4b，所有帧都被认为是信号，故无法实现分割.

实验3 采用专用语音麦克风，以能量算子法为例，通过引入其它噪声源(如电机运转声)来模拟实际环境中的条件变化以测试阈值自适应调节的效果，初始环境因子为2.29，测试期间峰值为2.49，稳定在2.45左右，如图5～图7.

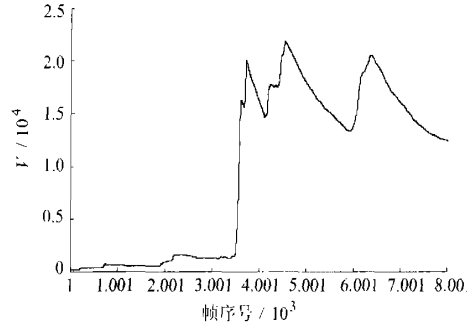


图5 V 变化曲线

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

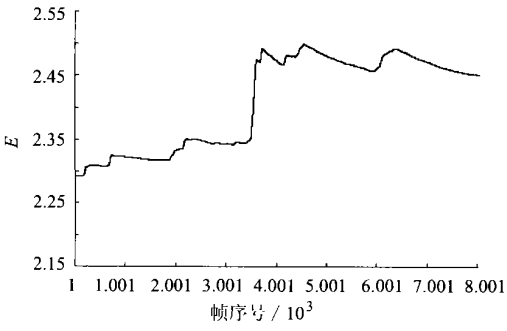


图 6　E 变化曲线

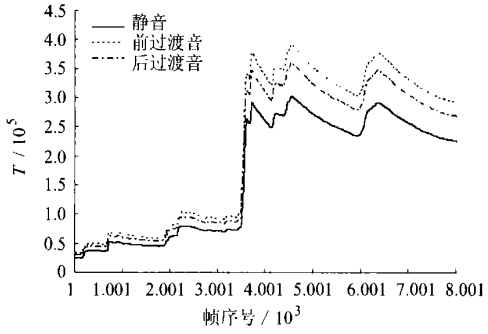


图 7　T 变化曲线

实验结果说明,环境因子可以迅速检测到环境的变化,主要体现为能量偏差显著增大,并能在环境改变后自适应更新阈值,速度通过动量因子 η 来调节,η 为经验值,根据对速度和收敛效果的要求设定,实验中取 η 为 0.99,分割效果达到了预期要求.

实验 4　采用实验 1 算法,改变滑动窗口大小和重叠率,考察不同算法对分割实时性能的影响.为了便于时间统计,一次处理样本数为 10 240 个(1 280 ms),窗口分别采用 256 和 512 样本,重叠率分别采用 0%,30%和 50%,对能量算子法和幅值法进行比较,在实验环境中得到算法造成的平均耗费为 14.6～59.2 ms,具体结果见表 2.

表 2　实验 4 计算性能考察

窗口类型		平均计算耗费/ms	
长度	重叠率/%	能量算子法	幅值法
512 样本	0	16.5	14.6
	30	25.1	21.7
	50	32.8	28.8
256 样本	0	30.6	28.8
	30	43.3	39.1
	50	57.2	59.2

对于实时语音传输类应用,合适的窗口大小应为 20～30 ms[4],可采用每窗口 256 样本,在无重叠情况下分割,若对每个窗口即时处理,则造成的延迟增加值约为 0.7 ms;对于音乐等信号一般采取更大的窗口,如每窗口 512 样本,则造成延迟增加值约为 0.8 ms.可见分割算法并不会对处理的实时性造成太大困扰.

6　结 束 语

本文提出以环境因子描述影响实时声音信号质量的主要环境因素,通过声音背景学习提取环境因子,从而实现了实时分割中初始阈值的确定与动态调整,解决了文献[4]中的遗留问题,降低了不同分割策略下的阈值调整代价.实验表明,若环境不发生剧烈恶化,则该方法能保持较好的灵活性和分割效果,可用作语音识别、实时音频分析、自动广播等系统的实时分割前端.存在的问题则是当环境突然恶化并稳定保持时,所有分片都判别为信号,则以静音间隙作为调整依据的阈值动态调整将失去动力.若引入长时特征会影响到系统的实时性,因此将来考虑结合低能分片比或如文献[7]引入部分频域特征来加以克服.

参考文献 (References)

[1] Subramanya S, Abdou Y. Segmentation of audio data based on the binary images of the audio samples[A]. Proc of Inter Conference on Intelligent Systems[C]. Denver: IEEE, 1999

[2] Thomas K, Michael S, Martin W, *et al*. Strategies for automatic segmentation of audio data [A]. Proc of ICASSP [C]. Istanbul: IEEE, 2000

[3] Foote J. Automatic audio segmentation using a measure of audio novelty[A]. Proc of ICME 2000[C]. NY: IEEE, 2000. 452～455

[4] 孙文彦,熊　璋,李　超,等. 语音信号实时传输中的动态变长分帧算法[J]. 通信学报,2001,22(7):80～86
Sun Wenyan, Xiong Zhang, Li Chao, *et al*. An dynamic variable length packetization algorithm in real-time speech transmission[J]. Journal of China Institute of Communications, 2001, 22(7): 80～86 (in Chinese)

[5] 卢　坚,毛　兵,孙正兴,等. 一种改进的基于说话者的语音分割算法[J]. 软件学报,2002,13(2):274～279
Lu Jian, Mao Bing, Sun Zhengxing, *et al*. An improved speaker based speech segmentation algorithm[J]. Journal of Software, 2002, 13(2): 274～279(in Chinese)

[6] Robert T, Alan J, Takeo K. A system for video surveillance and monitoring[R]. CMU-RI-TR-00-12, 2000

[7] George T, Perry C. Multi-feature audio segmentation for browsing and annotation[A]. Proc of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics[C]. Mohonk: IEEE, 1999

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net