音频编码中瞬态信号的时域检测方法

阎建新 窦维蓓 董在望

(清华大学电子工程系 北京 100084)

摘 要 在低比特率音频感觉编码中，预回声失真更为突出，而对其处理的前提是瞬态信号的有效检测。在时域，基于峰值功率与平均功率之比(PMR)定义了瞬态强度，并以此为判决函数提出一种新的瞬态信号时域检测算法。由于考虑了时域掩蔽效应来设置检测门限和有效瞬态点间隔，非常适用于感觉音频编码。与当前典型的基于感觉熵的频域瞬态检测方法相比，具有时间分辨率高、准确和算法简单等优点。

关键词 语音处理，瞬态检测，预回声失真，音频感觉编码

中图分类号：TN912.3 **文献标识码**： A **文章编号**：1009-5896(2006)02-0307-05

Time-Domain Detection Method of Transient Signals in Audio Coding

Yan Jian-xin Dou Wei-bei Dong Zai-wang

(*Dept. of Electronic Eng., Tsinghua University, Beijing* 100084, *China*)

Abstract In perceptual audio coding with low bit rate application, the pre-echo artifact is more distinct, and the effective detection of transient signals is the precondition. Transient intensity is defined in term of the Peak-Mean power Ratio(PMR), accordingly, a novel detecting algorithm of transient signals on time domain is presented while PMR is used as a criterion function. Because of considering the temporal masking effects to set up a threshold and the valid hop size of transient points, this method is very suitable for perceptual audio coding. When compared with the typical perceptual entropy detecting on frequency domain，it has some advantages, such as higher time resolution, more accuracy and simple algorithm.

Key words Speech processing, Transient detection, Pre-echo artifacts, Perceptual audio coding

1 引言

感觉音频编码的一般过程是：(1)使用滤波器组或变换将输入信号从时域转换到频域；(2)使用感觉模型计算出信号的掩蔽门限，给出了允许的最大量化误差；(3)根据允许的最大量化误差，量化并编码频谱系数，这样可以控制量化噪声低于掩蔽门限，从而能够维持感觉上无损伤的主观声音质量；(4)所有相关信息打包传输。

近年来，尽管感觉音频编码可以在压缩比高于 12 时，仍可得到高质量音频信号用于传输或存储，并形成一系列的MPEG标准和一些工业标准，但是在音频编码技术中的预回声失真一直是一个相当棘手的问题[1]，特别当比特率较低时，亦即压缩比较高时，预回声将变得更加明显和严重。预回声失真产生的关键原因是：时间分辨率的不足造成量化噪声的时域扩散。特别当一个瞬态信号被分块变换(或滤波)到频域进行量化编码时，由于量化噪声被扩散到整个变换块(或滤波器组)范围上，而且如果不能被信号掩蔽，就会出现预回声。预回声造成信号波形失真的例子如图 1 所示，显然在突发信号前出现了明显的量化噪声，且人耳对此类失真很敏感。

除了量化噪声的强度决定预回声失真影响声音质量的程度，人耳的时域掩蔽也起着重要的作用。时域掩蔽现象有两种情况[2]：预掩蔽和后掩蔽，预掩蔽的作用时间为可达

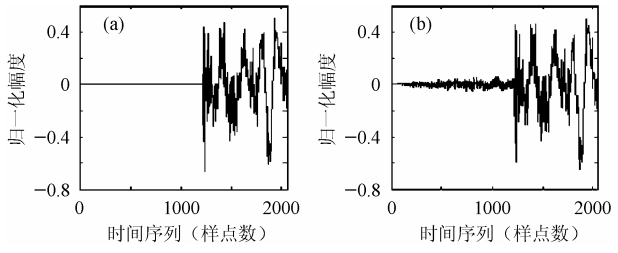

图 1 在长块 AAC 编码时原始音频信号(a)和预回声失真的情况(b)

2004-07-22 收到, 2004-12-08 改回

20ms，实际作用时间一般考虑在 0.5～2ms 以内有效；后掩蔽有更长的持续时间，约达到 200ms，而实际考虑 10～50ms 之内有效。由于后掩蔽作用时间较长，量化噪声一般能够被很好地掩蔽掉而不影响主观声音质量，因此在感觉编码器中较少考虑这种情况。相对于后掩蔽，预掩蔽能力较弱，需要仔细地设计一个合适的量化噪声的时域特性，使其不超过人耳的预掩蔽电平，这样预回声失真不能被人耳检测到，从而保证透明编码声音质量。

目前已经研究了许多方法来解决预回声问题，如比特池方法、时域噪声整型、长短块切换、混合滤波器组、增益控制以及以上几种的组合方法等。这些技术都是基于正确地检测到瞬态信号的后续处理方法，因此检测引起预回声失真的瞬态信号是有效抑制预回声失真的前提。至今已研究的几种音频信号瞬态检测方法大致可分为两类：时域能量检测法[3]和频域能量检测法[4]。这些方法存在各自的不足：如复杂度高、瞬态定位时间分辨率过大和不准确以及没有更好地考虑预掩蔽特性等。其中基于感觉熵(PE)的瞬态检测方法[5]可归类于频域能量检测法，已被广泛地应用到MPEG标准定义的系列感觉音频编码算法中。PE的瞬态检测法的基本原理是：当在分析音频段出现瞬态信号时，由于频谱中产生大量高频分量，使得计算的感觉熵明显增大，通过判断PE是否大于某一门限(标准中此值设置为 1800)，来决定当前块是否含有瞬态音频分量。一般对固定长块计算感觉熵，并检测瞬态性，具体步骤如下：

(1) 对长块(如 2048 样点)进行 FFT，得到信号功率谱；

(2) 计算各个谱系数的不可预测度，其值越大越类似噪声信号；(3) 计算掩蔽门限：在临界频带上，进行功率谱与扩散函数卷积，并计算纯音度；

(4) 计算感觉熵 PE：高于掩蔽门限的信号分量编码需要的平均信息量；

(5) 瞬态性指示：如果感觉熵大于某个门限，则认为本块属于瞬态块，进行瞬态帧的相关处理；否则以稳态帧处理。

基于感觉熵的瞬态检测法有两个根本性的缺陷：如果确实有大量高频信息，会造成错误启动预回声控制，降低编码效率；另外不能较准确地指明瞬态的起始位置。本文下面提出了一种新的时域能量检测方法，能够更准确、精确和有效地检测瞬态信号。

2 峰值均值功率比(PMR)时域能量检测法

2.1 音频瞬态信号检测有关的几个指标定义

首先给出感觉音频编码中，瞬态信号检测中要考虑的几个指标如下：

(1) 瞬态检测的精度 指检测到瞬态信号的时间分辨率时间位置的精度；对于感觉音频编码，时间分辨率达到 0.5ms(对 44.1kHz 采样频率，相当于 22 个样点长度)是足够的；

(2) 瞬态信号的强度 瞬态强度(TI)指瞬态特征的强度级别，值域被归一化到[0,1] 范围，具体公式在后面给出；用此值可以控制在编码时的某些条件下是否采用短块编码，以及采取相应策略降低预回声问题；

(3) 瞬态检测的复杂度；

(4) 瞬态检测的准确度 可分为漏检和虚检的概率，两者是一个矛盾的关系，一般设置高的严格门限，容易产生漏检；而低门限会出现大量的虚检；

(5) 对编码声音质量的影响。

2.2 基本原理

PMR 瞬态检测方法和其他时域检测方法出发点基本一致，这也是一种非常自然的想法：瞬态信号一般在时域信号发生幅度的突变，因此可以在时域很准确地检测到信号的瞬态特性。本检测方法的基本原理是首先滤除信号中的低频分量，降低平稳信号对瞬态检测的不利影响，并突出瞬态信号的特征。然后合理地对滤波后的信号进行分段处理，分段的原则是考虑瞬态检测的时间分辨率要求和时域预掩蔽特性。一般计算每段信号的平均功率及峰值功率，使用当前段峰值功率与前一段的平均功率比值有关的瞬态强度作为判决函数，能有效地反映瞬态的变化情况。为了加快检测速度(或降低复杂度)，同时保证检测时间分辨率的精度，分粗细两级搜索(这实际上是以一定的时间分辨率为代价换取检测速度)。在第一级当判决函数比值低于第一门限直接定义为稳态段；高于此门限则进入精细搜索以便获取最大比值，对瞬态变化给出更准确的时间定位。接着要考虑后掩蔽特性，确定是否两个瞬态段相距过近，从而可以精简掉第二个瞬态段标记，将其改为稳态段。计算瞬态段瞬态强度的好处是：在实际音频编码中，即使瞬态检测过程标记某些段为瞬态段，但当某些编码参数变化：如编码比特率提高时，就有可能使用更精细分辨率量化系数而降低块内量化噪声，这样对某些瞬态强度小的段，人耳不再会感觉到预回声失真，从而可以

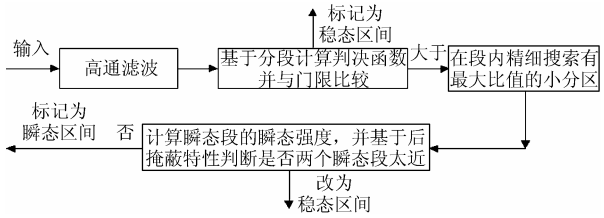

图 2 PMR 瞬态检测算法原理

设置为稳态段，从而提高编码效率。

2.3　检测算法的具体步骤

首先给出两个常量：判决门限和最小瞬态段间隔。这两个参数都是源于时域掩蔽特性，下面基于负指数衰减函数，给出时域掩蔽特性的近似解析函数表达式：

$$F_m(t)=\begin{cases}C\cdot e^{t/10}, & t<0\\ C, & 0\le t<T\\ C\cdot e^{-t/40}, & t\ge T\end{cases}$$

其中 C 为一与掩蔽信号电平有关的常数。

根据预掩蔽曲线图 3 以及下面提及的判决函数和瞬态强度计算公式，可以估算出预掩蔽决定的最小检测门限为 $\mathrm{Th}_{\mathrm{pm}}\approx 0.38$，仅当高于此门限才进行后续的瞬态检测处理。同样使用后掩蔽决定相邻两个瞬态的最小间隔设置，这里保守取 $T_{\min}=5\mathrm{ms}$。

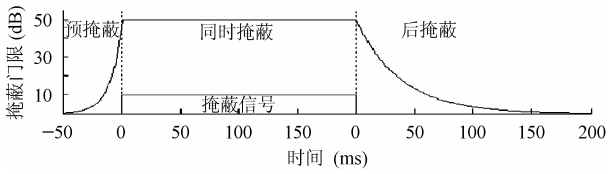

图 3　人耳的 3 种掩蔽特性

瞬态信号段的具体检测步骤如下：

(1)滤波　由于对高通滤波指标要求不高，选择一级 Haar 小波分解，通过简单的小波变换，得到信号细节(即高频)分量部分，也是瞬态信号的主要成分。

$$x_h=\mathrm{Haar}(x) \tag{1}$$

(2)对信号分段处理　段长 L 应该与最小时间分辨率有关，由于本算法采用第一级粗搜索和第二级精细搜索策略，这样段长也可以相对放宽要求，实验中取 $L=32$；

(3)计算第 k 段的平均功率　令 $\boldsymbol{x}_k=[x_k(0),x_k(x),\cdots,x_k(L-1)]$，$k$ 表示段号，

$$P_m(k)=\frac{1}{L}H_2(\boldsymbol{x}_k)=\frac{1}{L}\left(\sum_{n=0}^{L-1}x_k^2(n)\right)^{1/2} \tag{2}$$

其中 $H_2(\cdot)$ 表示分段信号矢量的 L_2 范数。

(4)计算第 k 段的峰值功率

$$P_p(k)=H_\infty(\boldsymbol{x}_k)=\max_{0\le n<L-1}\{|x_k(n)|\} \tag{3}$$

其中 $H_\infty(\cdot)$ 表示分段信号矢量的 L_∞ 范数。

(5)计算瞬态强度　首先计算峰值功率与平均功率之比

$$C(k)=\frac{P_p(k)}{P_m(k-1)} \tag{4}$$

然后计算瞬态强度 $\mathrm{TI}(k)=\frac{1}{50}\cdot\left(10\cdot\log_{10}\min(10^5,\max\ (C(k),1))\right)$。如果 $\mathrm{TI}(k)>\mathrm{Th}_{pm}$，则继续　否则标记当前分析段为稳态段。

(6)精确检测瞬态点在当前段的起始位置　如果已确定当前段为瞬态段，需要将此段进一步分割成 $L/4$ 长的小分区，将时间分辨率精度提高 4 倍，在 4 个小区间上搜索判决函数取极大值的那个小分区，即为瞬态起始点(或转换点)的近似。

$$T_s=k\cdot L+l\cdot\Delta \tag{5}$$

使得在第 l 小区间 $C(\cdot)$ 取得最大值，其中 $0\le l<4$，步长 $\Delta=L/4$。

(7)如果相邻两个瞬态段的间距小于后掩蔽决定的最小值 $T_{\min}$，则去掉后一个瞬态检测标志，恢复成稳态指示。

3　仿真结果及分析

基于上节的PMR时域瞬态检测算法，在MATLAB上进行了仿真。实验设计如下：对 8 个含有瞬态特性的测试音乐片段分别进行了PE检测和PMR时域瞬态检测，其中PE检测直接采用MP3 编码算法标准中的方法，表 1 给出了两种检测方法的准确度比较；图 4 和表 2 给出两种检测方法应用MP3编码算法时，应用客观测试方法(符合ITU-R BS1387)[6]得到的结果比较。

下面就瞬态检测有关的 5 个主要指标逐项分析：

(1) 时间分辨率　PMR 方法明显具有更好的时间分辨率，MP3 中对 44.1kHz 采样信号的瞬态检测时间分辨率约为 13ms，MPEG AAC 的分辨率更低约为 46ms，因此这些编码算法处理瞬态信号时必须将长块分成多个短块(MP3均分成 3 段，AAC 均分为 8 段)，这样就可能造成某些稳态信号部分也强迫使用短块编码而降低了编码效率。

(2) 瞬态强度指示　PMR 方法给出了瞬态强度(TI)参数，由此可以结合其他编码参数调整某些原来标记为瞬态段的信号是否采用稳态编码方式，可以进一步提高编码效率；PE 方法没有此项指标，完全是固定的检测模式，一旦检测出瞬态信号，必定进行瞬态编码方式。

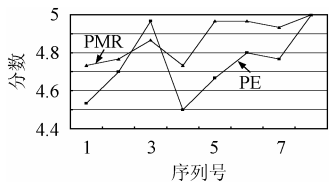

图 4　主观测试结果(表 1 中 8 个序列)

表 1 瞬态检测准确度比较

测试序列	瞬态特性描述	准确度			
		漏判		误判	
		PE	PMR	PE	PMR
响板 1	大量强(53)	18	3	8	4
响板 2	大量强(43)	11	0	5	0
英语女声	较多，中等强度(16)	9	4	10	15
蔡琴歌曲	较多，中等强度(17)	1	3	55	16
打击乐	较多，较强(87)	34	8	20	23
竖琴	较少，较强(9)	4	3	1	1
钟琴	较少，部分较强(6)	1	1	2	2
小号	瞬态很少(2)	0	0	3	0
总结	233	33.5%	9.4%	44.6%	26.2%

表 2 瞬态检测客观声音质量测试 ODG 和 NMR 的比较

序列	$\Delta_{ODG}=ODG_{PMR}-ODG_{PE}$			$\Delta_{NMR}=NMR_{PMR}-NMR_{PE}$		
	128	96	64	128	96	64
响板 1	− 0.01	− 0.04	0.03	− 0.0658	− 0.0255	0.0414
响板 2	− 0.01	0.04	0.04	− 0.238	− 0.1613	− 0.0754
英语女声	0.02	0.02	0.03	− 0.0621	− 0.0918	− 0.0362
蔡琴歌曲	0.04	0.03	0.02	− 0.0954	− 0.0781	− 0.0427
打击乐	0	0.01	0.05	− 0.0472	− 0.0677	− 0.1270
竖琴	0.02	− 0.01	− 0.01	0.0293	0.0413	− 0.0053
钟琴	0.04	0	− 0.09	0.2340	0.0722	0.0847
小号	0	0.07	0.01	− 0.0397	− 0.0177	− 0.0178
平均值	0.0125	0.0163	0.01	− 0.0356	− 0.0411	− 0.0223

(3) 复杂度 一般来说 PE 方法的复杂度很高，涉及 FFT 频域变换和频域掩蔽等运算，但是在感觉编码中一般需要计算频域掩蔽门限，这样 PE 的计算就显得相对容易。而进一步分析会发现算法中不管当前段是稳态段还是瞬态段都要计算两种情况的掩蔽门限计算，以便得出基于 PE 的瞬态信号的检测结果，因此总体来说 PE 瞬态检测法具有很高的运算复杂度。PMR 时域检测方法由于是预处理，在心理声学模型分析前就给出稳态/瞬态指示，心理声学模型计算可以一次完成(一个长块或几个短块)。另外由于采用 Haar 变换和功率比等简单运算，此瞬态检测方法复杂度很低。

(4) 检测准确度 从表 1 中可以看到，PMR 方法比 PE 方法一般有更好的检测准确度，即漏判和误判概率要低。因为 PMR 方法的漏判率较小，因此可以有效抑止出现预回声失真的情况保证主观声音质量；而误判率比 PE 方法低，可以防止编码算法无谓进入瞬态段处理，降低编码效率。另外两种方法对含有语音的测试序列明显给出过多的误判情况，这是由于语音信号有很多瞬态强度中等的情况，即处于瞬态与非瞬态判决边界，一般这种情况给出的误判对声音质量影响较小，但会造成一定数量的比特率提高。

(5) 声音质量比较 从图 4 可以看到，除了对语音序列的主观声音质量稍差外，其他序列都有一定的改善。但改善非常有限，这是因为一般音频信号中稳态信号的出现更加频繁，基于 PE 的传统瞬态检测法已经能够鉴别出大部分瞬态信号，因此尽管 PMR 方法增加了检测精度和准确度，对最后主观声音质量提高也不会很明显。客观声音质量得到的结果通过表 2 给出，可以看到对 3 种不同码率下 PMR 方法的 NMR(噪声掩蔽比)值要稍好(即低于)于 PE 方法，因此能够有更好的量化噪声掩蔽能力；而 ODG(客观差别分)分数也要稍高。

4 结束语

本文给出了一种适用于感觉音频编码的时域瞬态信号

检测方法，即峰值功率与平均功率比(PMR)方法，这种方法利用时域预掩蔽曲线获得有效的检测门限，而后掩蔽用于精简不必要的瞬态指示。并通过定义与 PMR 有关的瞬态强度作为判决函数，能够有效检测音频信号的瞬态特性。实验结果表明此方法优于 MPEG 中使用的感觉熵(PE)方法，主要体现在：具有更精细的时间分辨率，对 44.1kHz 采样音频信号的时间分辨率可以精确到 0.5ms；通过对 8 个测试音频信号的仿真结果分析，其检测准确度(漏检和虚检)也好于 PE 感觉熵方法；主观声音质量稍有提高，客观测试结果的 ODG 和 NMR 也略好；最后由于时域检测方法位于感觉编码的最前端，避免了 PE 频域检测方法中进行感觉模型分析时必须计算长和短块感觉熵后再判断瞬态特性的的复杂过程，从而使得复杂度也有明显下降。进一步的研究重点是：对含有语音类音频信号，如何改善瞬态判决准确度；另外如何有效结合编码参数动态调整瞬态强度判决门限，提高编码效率。

参 考 文 献

[1] Painter T, Spanias A. Perceptual coding of digital audio [J]. *Proc. IEEE*, 2000, 88(4): 451 − 513.

[2] Zwicker E, Fastl H. Psychoacoustics: Facts and Models, Berlin: Springer-Verlag, 1999, Chap. 4.

[3] Verma T S. A perceptually based audio signal model with application to scalable audio compression, [PhD thesis], Stanford University, 1999.

[4] Masri P, Bateman A. Improved modeling of attack transients in music analysis-resynthesis. Proc. International Computer Music Conference (ICMC96), Hong Kong, China, 1996: 100 − 103.

[5] ISO/IEC JTC1/SC29. Information Technology-Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5Mbit/s-IS 11172-3(audio), 1992.

[6] ITU-R BS.1387. Method for Objective Measurements of Perceived Audio Quality, 1998.

阎建新：　男，1966 年生，博士生，研究方向为多率数字信号处理、音频信号处理等.

窦维蓓：　女，1963 年生，副教授，研究方向为数字信号处理和数字音频信号处理等.

董在望：　男，1937 年生，教授，博士生导师，研究方向为数字声音广播及音频信号处理、集成电路设计等.