文章编号:1001-5132（2010）03-0027-05

心理声学模型及其在 MP3 编码中的应用

张力光，王让定*

（宁波大学 纵横智能软件研究所，浙江 宁波 315211）

摘要：心理声学模型是音频感知编码的核心部分，其直接影响着音频编码的质量及压缩比. 基于心理声学的基本原理、绝对听觉阈值、掩蔽效应及临界频带等相关内容，并结合心理声学的数学表达，按照 MP3 标准编码程序中的各个模块来详细分析算法流程. 最后通过相应算法来描述预回声产生机理及其在 MP3 编码中的抑制方法.

关键词：心理声学模型；掩蔽效应; MP3 编码；预回声抑制

中图分类号：TP391.42 文献标识码：A

随着计算机网络、无线通信、电子消费产品、高清晰度电视、数字音频广播等新兴技术的迅猛发展，数字音频技术发展将面临着一些挑战，比如如何在较小的信道带宽、有限的存储空间以及高性价比等要求下寻求更一种有效的压缩方法，以获得低码率高品质数字音频. 目前传统的音频压缩技术很多，它们主要力求输入和输出信号波形一致，这种编码技术以数学意义上的最接近来进行量化和编码[1]. 因此这种编码器的码率很高，压缩比低. 为了获得更高的压缩比，一种期望在主观感知意义上更接近的高质量、低码率的音频编码技术越来越成为数字音频压缩技术的主导. 这种编码器对失真的考虑是基于人类对输出信号的有效感知，因而此种编码也被称为感知音频编码(Perceptual Audio Coder, PAC)[2].

现在一些比较成熟的音频编码技术都使用了感知编码，如 MPEG 系列标准等. 感知音频编码通过模拟人的发音器官的特性，利用人的听觉系统，运用分析技术和频率相关比特分配技术，使量化噪声和听觉特性相匹配[3]. 因而心理声学模型的好坏直接影响了音频压缩效率和音频感知质量.

1 心理声学原理

人的听觉系统能否感知到音频信号主要取决于音频信号的频率和强度，人们能感知的频率范围一般在 20~20 000 Hz. 音频信号强度一般用对数形式表示，单位为分贝(dB)，即：$SPL = 10\lg(I / I_0)$，其中，I_0 [4]为$10^{-12} W / m^2$.

1.1 绝对听觉阈值

绝对听觉阈值描述在无噪声环境下，人耳对不同的声音频率分量能够感知的最小声压级. 绝对听觉阈值的经验公式为[4]:

$$T_Q(f) = 3.64 f^{-0.8} - 6.5e^{-0.6}(f - 3.3)^{-2} + 10^{-3} f^4,$$

收稿日期：2009-07-17. 宁波大学学报（理工版）网址: http://3xb.nbu.edu.cn
基金项目：国家自然科学基金（60672070, 60873220）；浙江省自然科学基金（Y108022）.
第一作者：张力光（1983－），男，浙江嘉兴人，在读硕士研究生，主要研究方向：信息隐藏. E-mail: zlg4585192@sina.com
*通讯作者：王让定（1962－），男，甘肃天水人，博士/教授，主要研究方向：音频信息隐藏及语音识别. E-mail: wangrangding@nbu.edu.cn

其中，f 为音频信号频率；T_Q 为绝对听觉阈值. 一般人耳最敏感的频率段在 500~5 000 Hz 范围内.

1.2 音频信号的掩蔽效应

所谓掩蔽效应就是一个音频信号可使人的听觉系统感觉不到另一个声音的存在. 掩蔽效应主要可划分为时域掩蔽和频域掩蔽: (1)时域掩蔽是指能量较强的音频信号，可掩蔽同时或其前后出现能量较弱的音频信号的现象，所以又称异时掩蔽. 异时掩蔽又分为超前掩蔽(Pre-masking)和滞后掩蔽(Post-masking)，如图 1(a)所示，前掩蔽持续时间约为 20 ms，后掩蔽持续时间为 150 ms. (2)频域掩蔽是指掩蔽声与被掩蔽声同时作用时发生的掩蔽效应，也称同时掩蔽(Simultaneous Masking)，如图 1(b)所示. 掩蔽作用的大小可用信掩比(Signal-to-Mask Ratio, SMR)来衡量，其定义为掩蔽信号的能量 SPL 与该信号所产生的掩蔽阈值的能量之差. 其值越小，掩蔽效果越好.

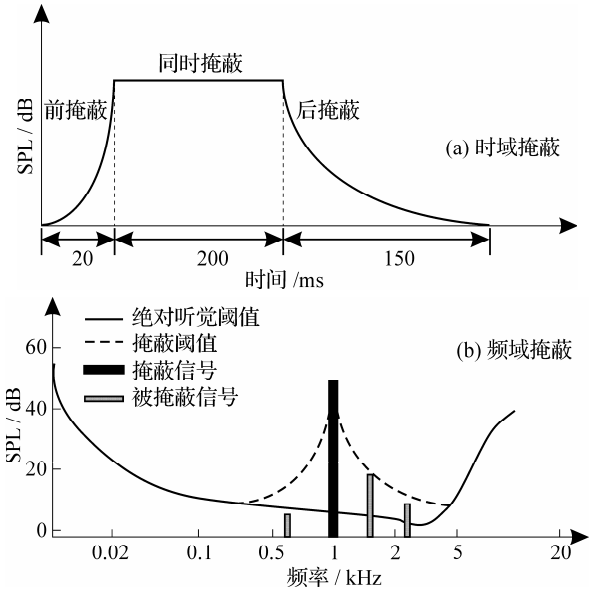

图 1 音频信号的掩蔽效应

1.3 听觉系统的临界频带

人的听觉系统实际上可以看成一组多通道滤波器，其幅度响应为不对称和非线性的. 掩蔽阈值在以掩蔽信号频率为中心的狭小频带内是个常数，而这个狭小频带的宽度即称为临界频带. 在实际应用中，将可感知频率范围内划分为 24 个临界频带，临界频带的单位为巴克(bark)，bark 和 Hz 的转换公式为[5]:

$$z(f)=13\arctan\frac{0.76f}{1000}+3.5\arctan\left(\frac{f}{7500}\right)^2,$$

而每个临界带的带宽可以由下式求得:

$$BW(f)=25+75\times\left(1+1.4\times\frac{f_c}{1000}\right)^2,$$

其中，f_c 为该临界带的中心频率；$BW(f)$ 为该临界带的带宽. 临界带在频率 500 Hz 以下几乎是等带宽，大约为 100 Hz；但当频率超过 500 Hz，临界频带的带宽随着频率的增加而递增.

2 心理声学模型的应用

心理声学模型主要有 2 种. 模型 I 比较简单，计算复杂度低，但是精度不高，特别在高频段失真较大，主要应用于 MEPG-1 layer1 和 layer2. 而模型 II 比较复杂，但计算精度高，适合应用于高保真的压缩编码[6]，如 MP3 和 AAC.

2.1 MP3 概述

MP3 是第 1 个高保真音频数据压缩国际标准，支持 32 KHz、44.1 KHz 或 48 KHz 的采样率，有 4 种声道模式. MP3 的编码框架如图 2 所示.

输入声音信号经 32 个子带滤波器组和 MDCT 变换进行时频转换处理，同时通过“心理声学模型 II”计算出每个子带的信号能量和 SMR. “量化和编码”利用 SMR 来决定分配给子带信号的量化位数，最后通过“数据流帧包装”将子带的样本及其

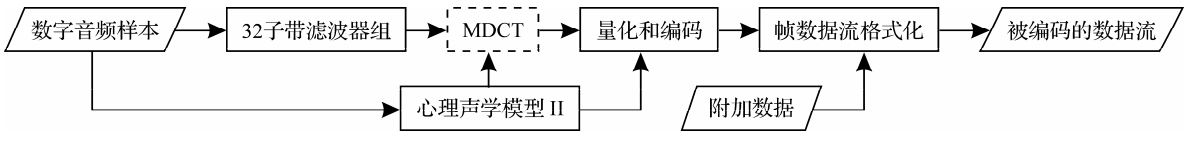

图 2 MP3 编码框架

他附加数据按帧的格式组装成位比特流.

2.2 心理声学模型 II 的计算流程

心理声学模型 II 在 MP3 中的计算流程如图 3 所示.

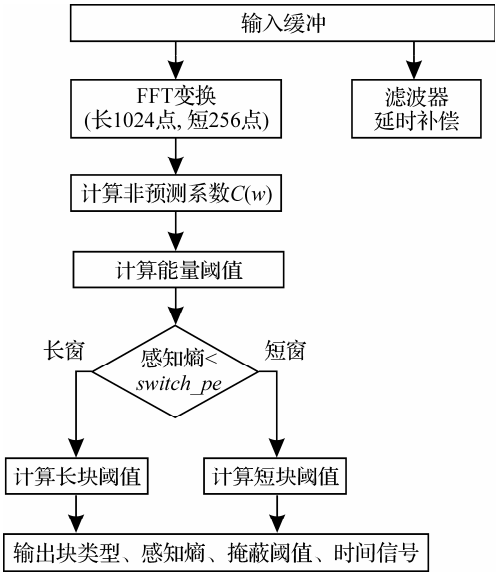

图 3　心理声学模型 II 计算流程

2.2.1 快速傅立叶变换(FFT)

在对输入信号做 FFT 变换之前, 先使用哈宁窗进行滤波, 然后做 2 次 FFT 变换. 1 次为 1024 点的长 FFT 变换, 每次变换使用 576 个新音频样本. 另 1 次为 256 点的短 FFT 变换, 每次变换使用 192 个新样本, 并重复做 3 次变换(3×192 = 576). 2 次 FFT 的运算结果将用来计算信号的声音含量指数.

2.2.2 计算不可预测指数 $C(w)$

$$C(w)=\begin{cases} c\_l(w), & 0\leqslant w<6,\\ c\_s((w+2)/4), & 6\leqslant w<206,\\ 0.4, & 206\leqslant w<1\,023, \end{cases}$$

其中, w 为频率值; $c\_l(w)$ 由长 FFT 系数计算得到; 而 $c\_s(w)$ 由短 FFT 系数得到; 不可预测指数具体定义可参照文献[4].

2.2.3 计算声音能量阈值

心理声学模型 II 是以频段(Partition)为单位计算阈值, 每个频段宽带接近临界频带宽度的 1/3. 按音频信号的采用率不同一般分为 59~63 个频段. 阈值的计算步骤按如下:

(1) 计算每个频段的能量及不可预测指数.

$$eb(b)=\sum r(w)^2,$$

$$cb(b)=\sum cw(w)\times r(w)^2,$$

其中, $eb(b)$ 和 $cb(b)$ 分别表示频段 b 的能量和不可预测指数.

由于掩蔽函数的作用, 各频段之间的掩蔽效应也会互相影响, 因此需要对每个频段的 eb 和 cb 分别和掩蔽函数[7]进行卷积运算来修正:

$$ecb(b)=eb\times sprdngf(z_i,z_b),$$

$$ctb(b)=cb\times sprdngf(z_i,z_b),$$

其中, $sprdngf(z_i,z_b)$ 是掩蔽曲线函数[4], 其意义表示临界频带 z_i 在临界频带 z_b 处的掩蔽值.

(2) 计算每个频段的 SNR 值.

将不可预测指数 ctb 转换成可预测指数 tbb. tbb 所反映的信号特征恰好与 ctb 相反, tbb 越大, 则其对应的频段中声音信号越强.

$$tbb(b)=-0.299-0.43\lg e^{\frac{ctb(b)}{ecb(b)}}.$$

每个频段的 SNR 值可以利用可预测指数进行插值计算得到:

$$SNR(b)=\max\{snr_b,29tbb(b)+6(1-tbb(b))\},$$

其中, snr_b 是心理声学模型 II 标准规定的最小信噪比值的补偿; 常数 29 和 6 分别为分别表示噪声信号掩蔽声音的信掩比和声音掩蔽噪音的信掩比.

(3) 计算每个频段的声音能量阈值.

$$nbb(b)=ecb(b)\times norm(b)\times 10^{\frac{-SNR(b)}{10}},$$

$$norm(b)=1/\sum_{b=0}^{p} sprdang(z_i,z_b).$$

由于预回声现象, 能量阈值应该取前 2 次计算值与本次计算值中最小的值, 再加上考虑到的静音阈值, 最终频段 b 的能量阈值应该为:

$$thr(b)=\max\{qthr(d),\min\{nbb(b),2\times nbb_{t-1}(b),16\times nbb_{t-2}(b)\}\},$$

其中, $qthr(d)$ 为该频段的静音阈值, 由心理声学模型 II 给出; $nbb_{t-1}(b)$, $nbb_{t-2}(b)$ 分别表示为前 2 次计算的阈值.

2.2.4 计算感知熵 PE(Perceptual Entropy)

$$PE=-\sum_{b=0}^{p} cbwidth(b)\times \lg\left(\frac{thr(b)}{eb(b)}\right).$$

PE 值反映数据块频谱的平坦性，*PE* 越大，则该数据块包含能量较强的高频分量，因而在时域内必有瞬时的剧烈变化. 心理声学模型 II 规定当 $PE > swith\_pe$ 时，数据块为短类型，反之为长类型. $swith\_pe$ 设为常数 1800.

2.2.5 计算长块掩蔽阈值

上述计算的阈值都是以频段为计算单位，但是 MP3 量化编码都是以比例因子带(Scale Factor Bands, SFB)为计算单位，因此最后需要将每个频段的阈值转化为比例因子带所对应的阈值:

$$en(sb)=w1\times eb(b_u)+\sum_{b=b_u+1}^{b=b_o+1} eb(b)+w2\times eb(b_o),$$

$$thrn(sb)=w1\times thr(b_u)+\sum_{b=b_u+1}^{b=b_o+1} thr(b)+w2\times thr(b_o),$$

其中，$w1, w2, b_u, b_o$ 值均由模型 II 标准给出；b_u, b_o 分别为比例因子子带样本的起始和终止值.

2.2.6 计算短块掩蔽阈值

若当前数据块类型为短块时，模型 II 会重新计算每个频段的声音能量阈值，计算方法与长块相同，见步骤 2.2.3 部分(按长块类型来计算). 但不同的是把整个数据块分成 3 个短块分别进行计算，短块中每个频段的 *SNR* 值不是计算得到，而是由模型 II 标准给出.

当计算完频段的能量阈值后，再将其转化为比例因子子带的阈值，转化方法与长块一样，见步骤 2.2.5 部分.

2.3 MP3 中的预回声控制技术

某块音频数据如图 4(a)所示，横坐标表示 1024 个采样点，纵坐标表示采样点的幅值. 该数据块前面采样点的幅值较小，而后面幅值突然变大. 如果对这整块数据进行 DCT 变换，即用 1024 点 DCT 变换，量化噪音就会扩展到整个数据块中去，如图 4(b)所示，这就是预回声效应. 控制预回声的有效方法就是把整个数据块分成 2 个小块，并分别做 512 点 DCT 变换，这样量化噪音就会被限制在 1 块数据中[6]，如图 4(c)所示. MP3 就是运用长短块切换来控制预回声效应，对于变化剧烈的信号使用短块，而变化缓慢的使用长块. 一般短块的长度为 8 ms 左右，而超前掩蔽效应时间为 20 ms，因而短块产生的预回音很容易被掩蔽.

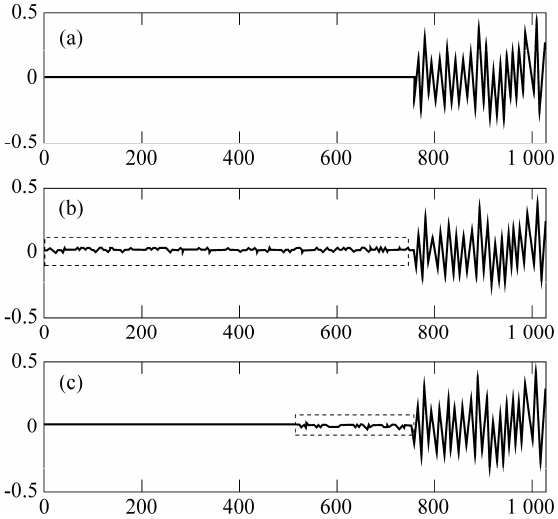

图 4 预回声产生及控制

MP3 在长块类型(long)与短块类型(short)的切换过程中，引入了过渡块数据类型，即起始类型(start)和结束类型(stop). 判断当前数据块最终类型的具体算法如下:

```
if (PE<swith_pe){
    switch (前一个数据类型) {
        case (长类型 || 结束类型):
        当前类型 = 长类型; break;
case (短类型):
        当前类型 = 结束类型; break;
  case (起始类型)
        error; break; }
else {
        当前类型 = 短类型;
    if (前一个类型 = = 长类型)
        前一个类型 = 起始类型;
    if (前一个类型 = = 结束类型)
```

前一个类型 = 短类型; }

这样通过过渡块类型的引入, MP3 在编码过程中就能既保证 MDCT 变换时数据的平滑过渡, 又有效地控制预回声产生.

3 结语

通过心理声学的基本原理详细地分析了心理声学模型在 MP3 编码中的参数计算流程, 以及结合算法来说明预回声的产生及控制. 通过对心理声学模型算法的研究, 发现算法还有进一步优化的可能, 比如可以用 MDCT 系数代替 FFT 系数来计算不可预测指数, 省去 2 次 FFT 变换; 不管是长块还是短块, 每次都要计算长块能量阈值, 其实对于短块, 计算是可以省略, 而这些所有优化的可能将是下一步工作的重点.

参考文献:

[1] 徐盛, 陈健. 数字音频编码技术的回顾与发展[J]. 电声技术, 1999(8):3-5.

[2] 何冬梅, 高文. 基于小波包和心理声学模型的音频编码算法[J]. 计算机研究与发展, 2000, 37(3):229-335.

[3] 高智衡, 韦岗. MP3 宽带音频压缩中的核心技术[J]. 电声技术, 2000(9):9-12.

[4] Edwards B, Sound I D, Alto P. Application of psychoacoustics to audio signal processing[J]. Signals, Systems and Computers, 2001(1):814-818.

[5] Painter T, Spanieas A. Perceptual coding of digital audio [J]. Proceeding of the IEEE, 2000, 88(4):542-462.

[6] 高成伟. 移动多媒体技术——标准、理论与实践[M]. 北京: 清华大学出版社, 2006.

[7] Schuller G. A low delay filter banks for audio coding with reduced pre-echo[C]//99th AES Convention, New York, 1995:6-9.

[8] Wang Ye, Yaroslavesky L. Some peculiar properties of the MDCT[J]. Proceedings of ICSP, 2000(1):61-64.

Psychoacoustic Model and its Application to MP3 Encoding

ZHANG Li-guang , WANG Rang-ding*

(CKC Software Lab, Ningbo University, Ningbo 315211, China)

Abstract: Psychoacoustic model involves core technology in perceptual audio encoding, and it directly affects the encoding quality and compress radio. This paper first introduces the basic principles of psychoacoustic, mainly including the absolute threshold of hearing, masking effect and critical bands. Then, combining with mathematical model of psychoacoustic, the algorithmic process in accordance with mp3 standard procedure modules is analyzed. In the end, the pre-echo producing mechanism and reducing method in mp3 encoding are described with a given number of experiments.

Key words: psychoacoustic model; masking effect; MP3 encoding; pre-echo

CLC number: TP391.42　　**Document code:** A

（责任编辑　章践立）