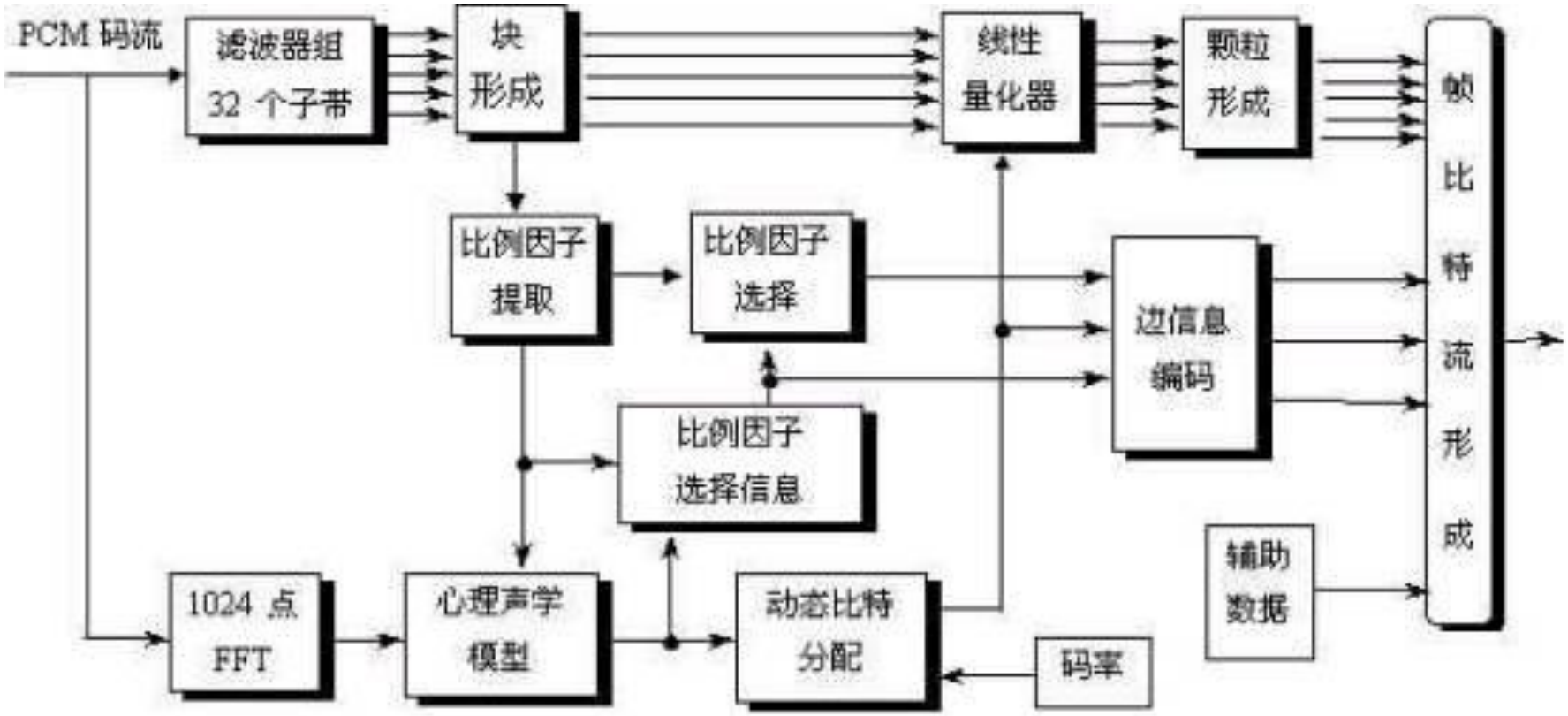
MPEG音频编码实验

MPEG-1 Audio LayerII编码器原理



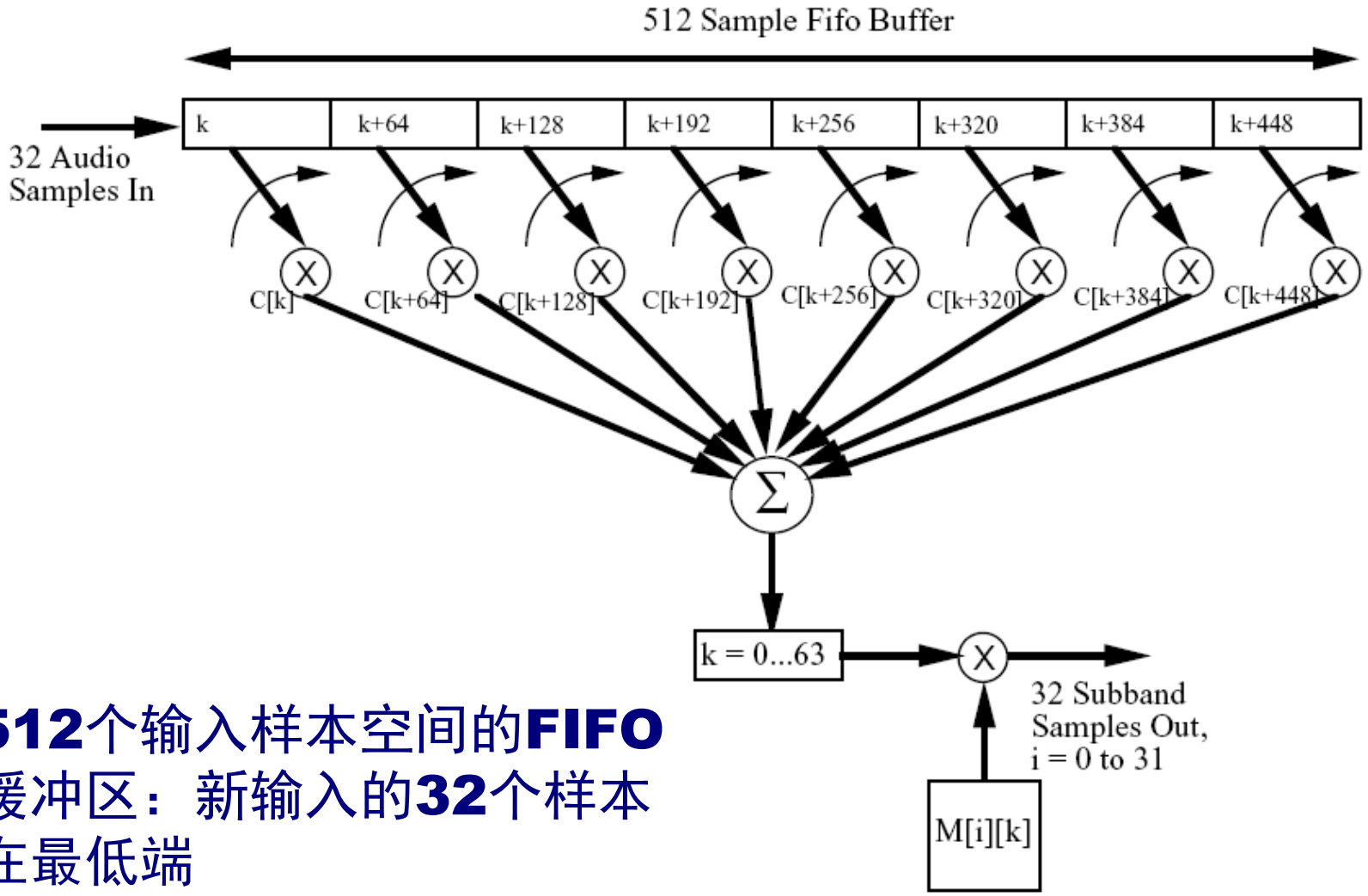
MPEG-I 心理声学模型

- 通过子带分析滤波器组使信号具有高的时间分辨率，确保在短暂冲击信号情况下，编码的声音信号具有足够高的质量
- 又可以使信号通过FFT运算具有高的频率分辨率，因为掩蔽阈值是从功率谱密度推出来的。
- 在低频子带中，为了保护音调和共振峰的结构，就要求用较小的量化阶、较多的量化级数，即分配较多的位数来表示样本值。而话音中的摩擦音和类似噪声的声音，通常出现在高频子带中，对它分配较少的位数

MPEG-1音频编码器框架图

- 多相滤波器组(Polyphase Filter Bank)：将PCM样本变换到32个子带的频域信号
  - **如果输入的采样频率为48kHz，那么子带的频率宽度为48/（2*32）=0.75Hz**
- 心理声学模型(Psychoacoustic Model)：计算信号中不可听觉感知的部分
  - 计算噪声遮蔽效应
- 比特分配器(Bit Allocator)：根据心理声学模型的计算结果，为每个子带信号分配比特数
- 装帧(Frame Creation)：产生MPEG-I兼容的比特流

多相滤波器组



512个输入样本空间的FIFO缓冲区：新输入的32个样本在最低端

多相滤波器组(2)

- for i=0…511，计算窗口内的样本： $Z[i] = C[i] \cdot X[i]$
  - 标准中规定了分析窗口的512个系数 C[i]
- 样本点分组，计算64个Y_k的值：$Y[k] = \sum_{j=0}^{7} Z[k+64j]$
- 计算32个子带样本：$S[i] = \sum_{k=0}^{63} Y[k] \cdot M[i][k]$
- 总计算公式： $S[i] = \sum_{k=0}^{63} \sum_{j=0}^{7} M[i][k]\left(C[k+64j]X[k+64j]\right)$
  - 其中分析矩阵： $M[i][k] = \cos\left(\frac{(2i+1)(k-16)\pi}{64}\right)$
  - 共需： 512 + 32x64 = 2560 次乘法
  - 每个子带的带宽为π/32T ， 中心为π/64T的奇数倍

多相滤波器组(3)

- 多相滤波器组的公式 $S[i]=\sum_{k=0}^{63}\sum_{j=0}^{7}M[i][k]\left(C[k+64j]X[k+64j]\right)$
- 可用卷积公式代替：

$$S_t[i]=\sum_{n=0}^{511}X[n-t]h_i[n]$$

- 其中 $h_i[n]=h[n]\times\cos\left(\frac{(2i+1)(n-16)\pi}{64}\right)$

低通　　调制

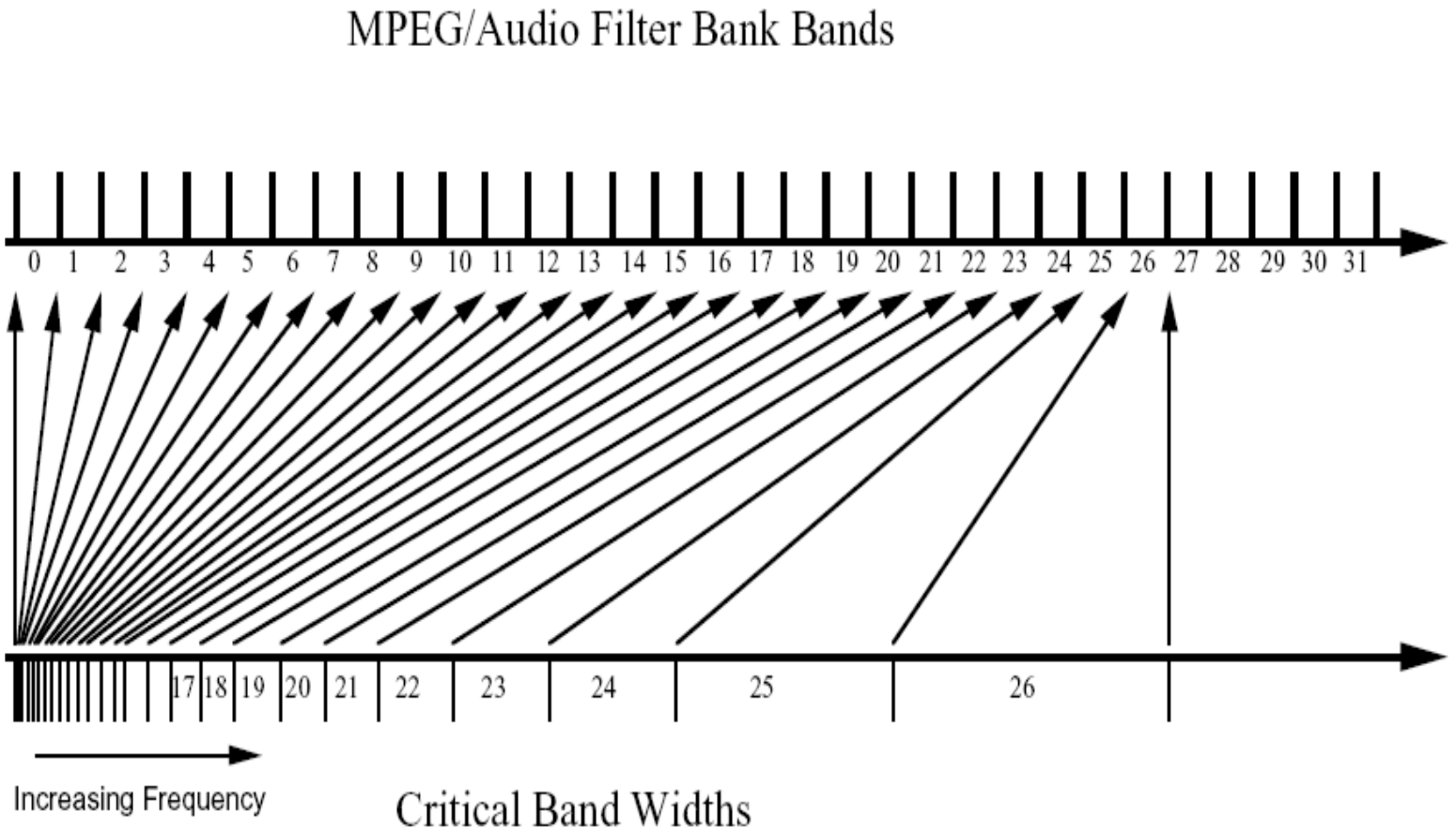
  - 当n/64 的整数部分为奇数时，令h[n]=-C[n]
  - 反之，h[n]=C[n]
  - 每个子带有自己的带通滤波器脉冲响应
    - 直观但效率不高：16384次乘法和16352次加法

多相滤波器组(4)

- 缺点：
  - 等带宽的滤波器组与人类听觉系统的临界频带不对应
    - 在低频区域，单个子带会覆盖多个临界频带。在这种情况下，量化比特数不能兼每个临界频带
  - 滤波器组与其逆过程不是无失真的
    - 但滤波器组引入的误差差很小，且听不到
  - 子带间频率有混叠
    - 滤波后的相邻子带有频率混叠现象，一个子带中的信号可以影响相邻子带的输出

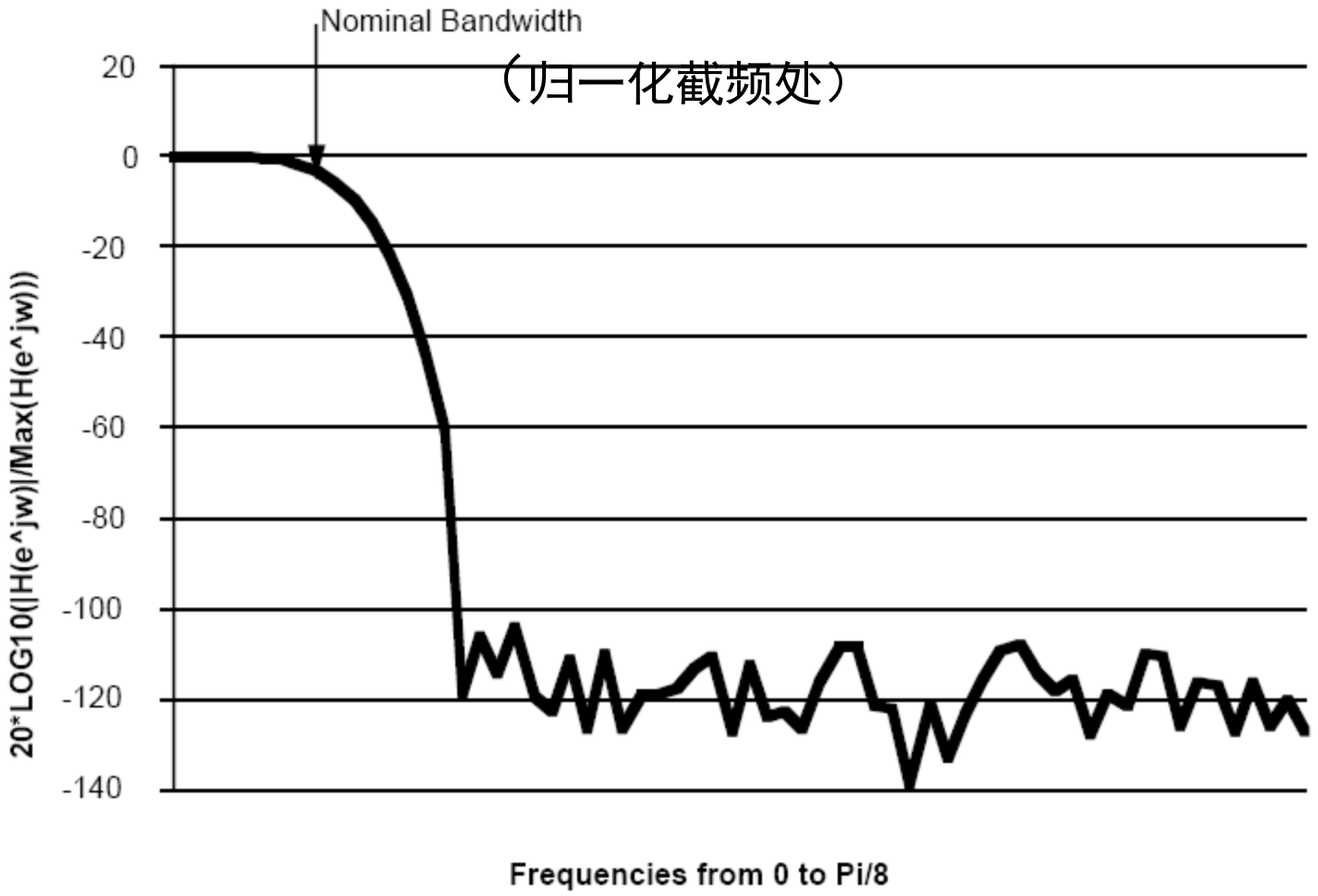
多相滤波器组(5)

- 滤波器组与临界频带比较：



多相滤波器组(6)

- 一个子带的频率响应：



多相滤波器组(6)

- 混叠：一个单频正弦信号输入可能在两个子带中产生非零信号

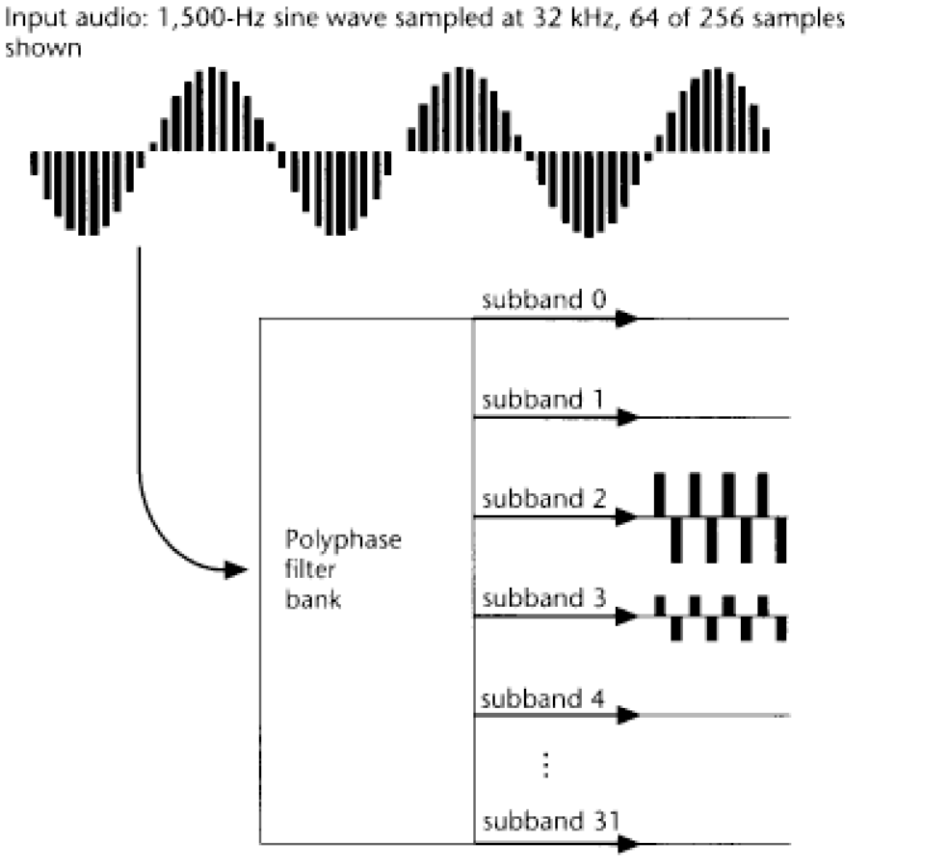
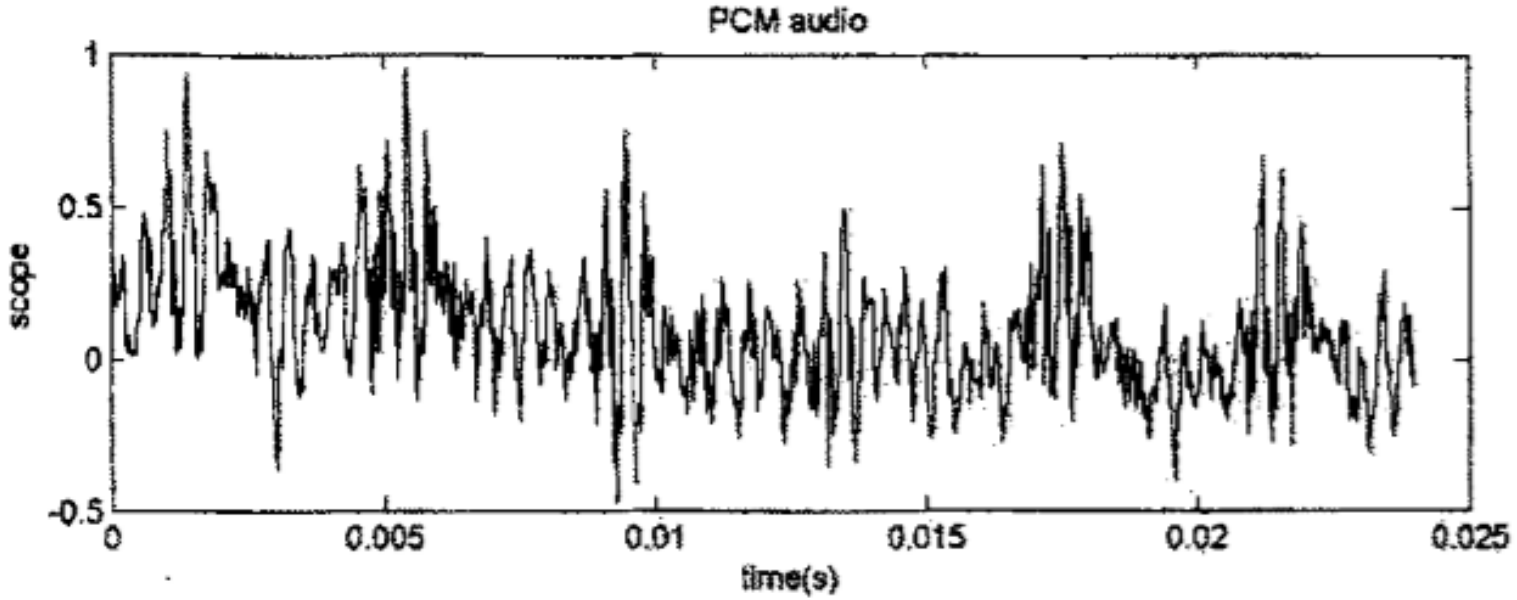
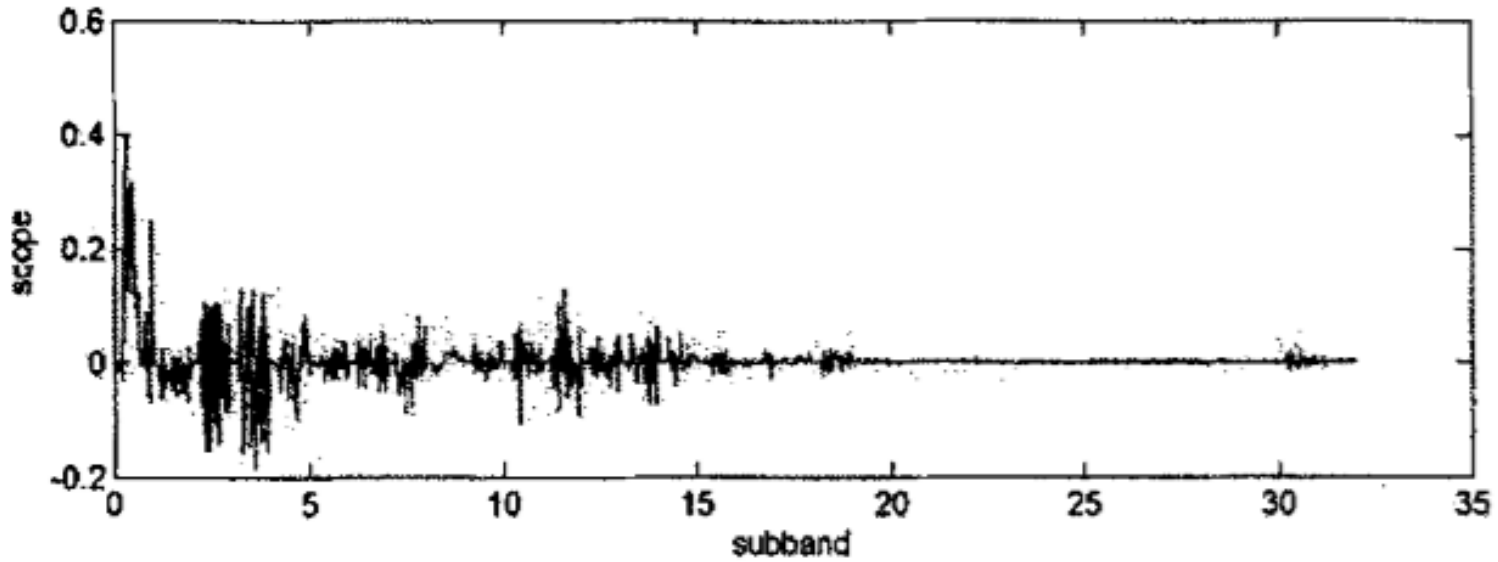


Figure 5. Aliasing: Pure sinusoid input can

多相滤波器组(6)


PCM audio
scope
time(s)
1
0.5
0
-0.5
0
0.005
0.01
0.015
0.02
0.025


scope
subband
0.6
0.4
0.2
0
-0.2
0
5
10
15
20
25
30
35

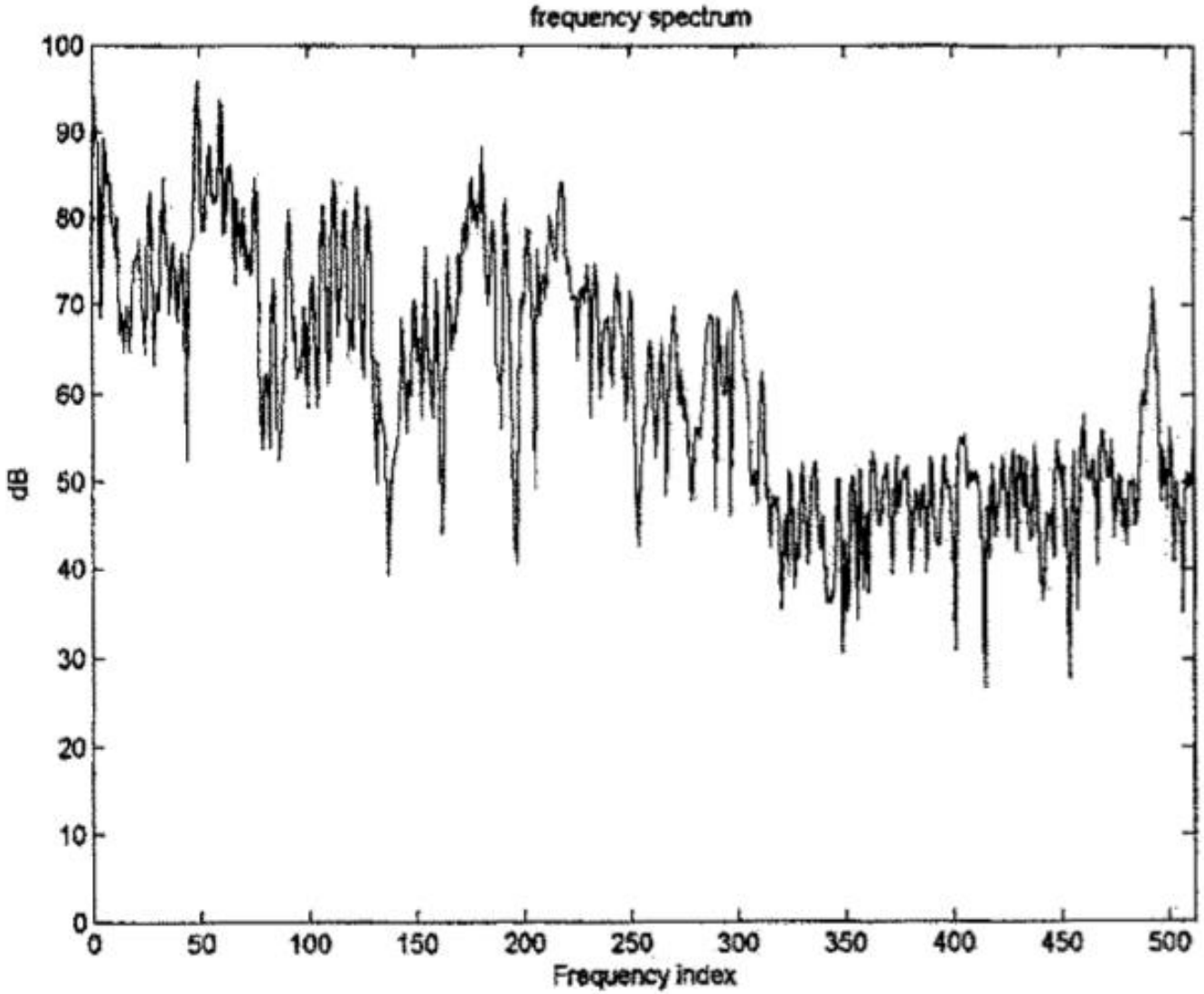
MPEG-I 心理声学模型

- MPEG-I 标准定义了两个模型
- 心理声学模型1:
  - 计算复杂度低
  - 但对假设用户听不到的部分压缩太严重
- 心理声学模型2 :
  - 提供了适合Layer III编码的更多特征
- 实际实现的模型复杂度取决所需要的压缩因子
  - 如大的压缩因子不重要，则可以完全不用心理声学模型。此时位分配算法不使用SMR（ Signal Mask Ratio ），而是使用SNR

心理声学模型I

- 1、将样本变换到频域
- 32个等分的子带信号并不能精确地反映人耳的听觉特性。引入FFT补偿频率分辨率不足的问题。
  - 采用Hann加权和DFT
    - Hann加权减少频域中的边界效应
    - 此变换不同于多相滤波器组，因为模型需要更精细的频率分辨率，而且计算掩蔽阈值也需要每个频率的幅值
  - 模型1：采用512 (Layer I) 或1024 (Layers II and III)样本窗口
    - Layer I：每帧384个样本点，512个样本点足够覆盖
    - Layer II 和Layer III：每帧1152个样本点，每帧两次计算，模型1选择两个信号掩蔽比（SMR）中较小的一个

心理声学模型I


frequency spectrum
dB
Frequency index
0
10
20
30
40
50
60
70
80
90
100
0
50
100
150
200
250
300
350
400
450
500

心理声学模型

- 2、确定声压级别

子带 n 中的声压级别 L_{sb} 计算如下：

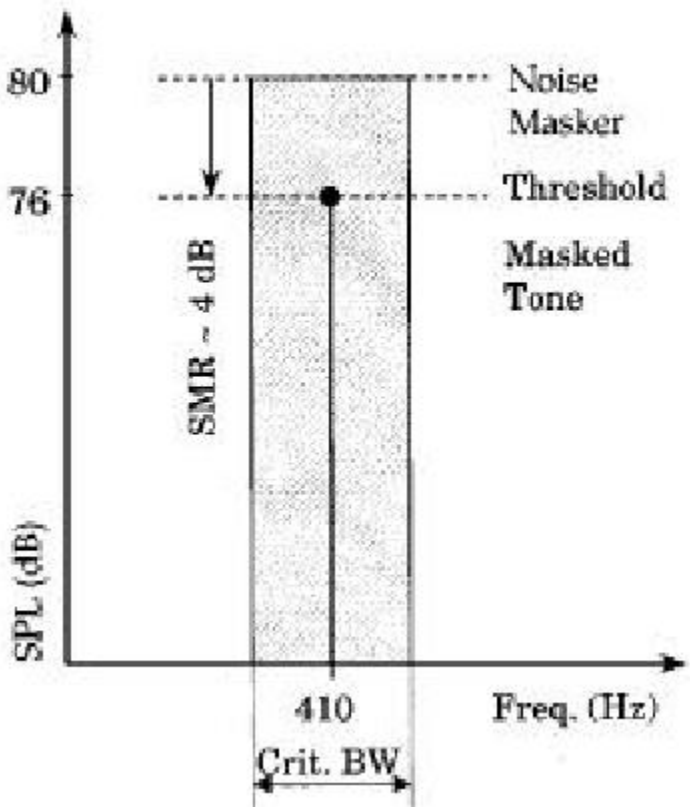
$$L_{sb}(n) = MAX[X(k), 20 \times \log_{10}(scf_{max}(n) \times 32768) - 10]dB$$

其中 $X(k)$ 是在子带 n 中的频谱线的声压级别，$scf_{max}(n)$ 在是一帧中子带 n 的三个缩放因子中最大的一个。

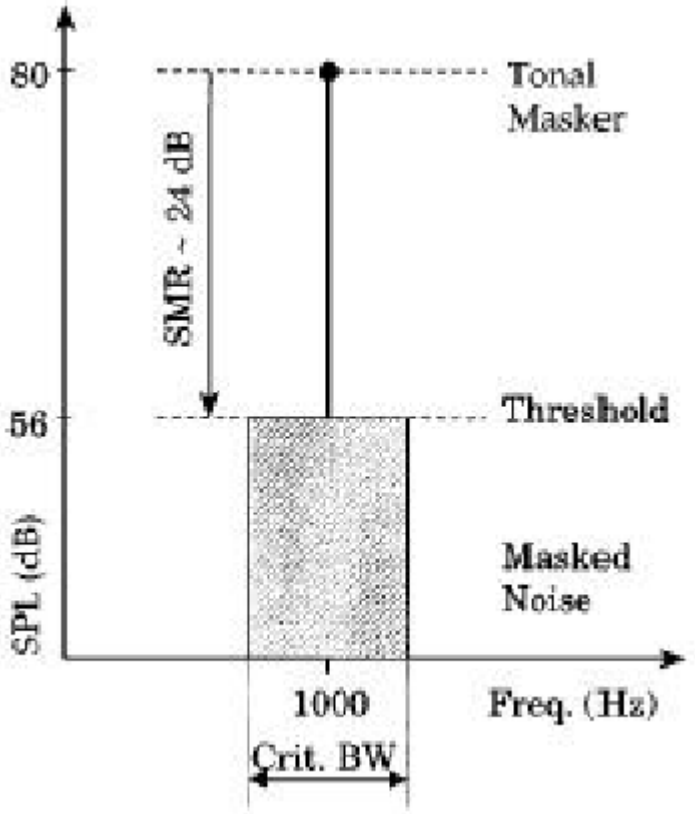
- 3、考虑安静时阈值
  - 也即绝对阈值。在标准中有根据输入PCM信号的采样率编制的“频率、临界频带率和绝对阈值”表。此表为多位科学家经多次心理声学实验所得。

心理声学模型

- 4、将音频信号分解成“乐音(tones)”和“非乐音/噪声”部分：因为两种信号的掩蔽能力不同



同一临界频带内噪声掩蔽乐音



同一临界频带内乐音掩蔽噪声

心理声学模型

- 4、将音频信号分解成“乐音(tones)”和“非乐音/噪声”部分：因为两种信号的掩蔽能力不同
- 模型1：根据音频频谱的局部功率最大值确定乐音成分
  - 局部峰值为乐音，然后将本临界频带内的剩余频谱合在一起，组成一个代表噪声频率（无调成份）

要列出谱线x(k)的有调和无调，需执行下面三个步骤：
（1）标明局部最大。如果x（k）比相邻的两个谱线都大，则x（k)为局部最大值；
（2）列出有调成份，计算声压级。如果x(k)-x(k+j)>=7dB,则x(k)列为有调成份。j随谱线的位置不同。
（3）列出无调成分，计算功率。在每个临界频带内将所有余留谱线的功率加起来形成临界频带内无调成分的声压级。并列出以下参数：最接近临界频带几何平均值的谱线标记k，声压级以及无调指示。

心理声学模型

- 5、音调和非音调掩蔽成分的消除
  - 利用标准中给出的绝对阈值消除被掩蔽成分；考虑在每个临界频带内，小于0.5Bark的距离中只保留最高功率的成分
- 6、单个掩蔽阈值的计算
  - 音调成分和非音调成分单个掩蔽阈值根据标准中给出的算法求得。

心理声学模型

- 7、全局掩蔽阈值的计算

某一频率点 i 的总掩蔽阈值可通过该点的绝对掩蔽阈值与单独掩蔽阈值相

$$LTg(i) = 10\lg(10^{LTq(i)/10} + \sum_{j=1}^{m} 10^{LTtm((z(j),z(i))/10} + \sum_{j=1}^{n} 10^{LTnm(z(j),z(i))/10})$$

加来获得。即：

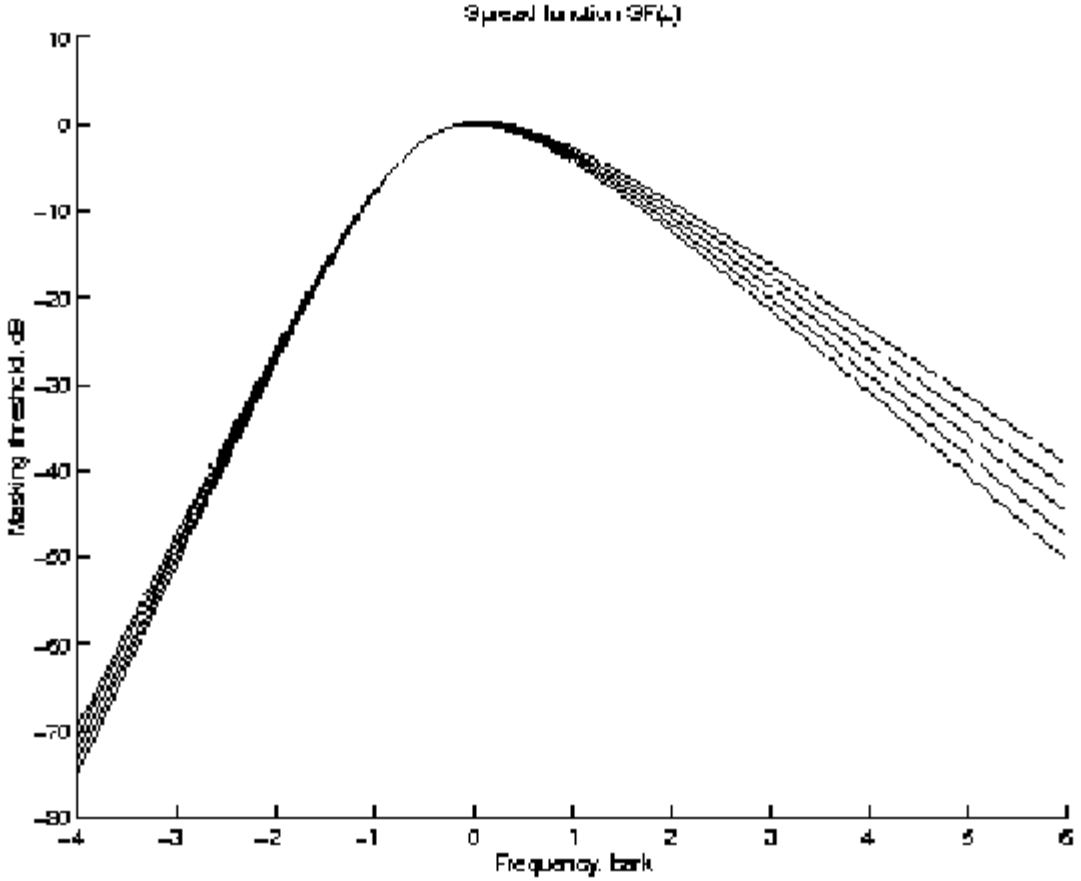
其中，LTq(i)是频率点 i 的绝对掩蔽阈值；LTtm(z(j),z(i))是第 j 个音调掩蔽成分对频率点 i 的掩蔽阈值，对频率点 i 有掩蔽效应的音调掩蔽成分共 m 个；LTnm(z(j),z(i))是第 j 个非音调掩蔽成分对频率点 i 的掩蔽阈值，对频率点 i 有掩蔽效应的非音调掩蔽成分共 n 个。

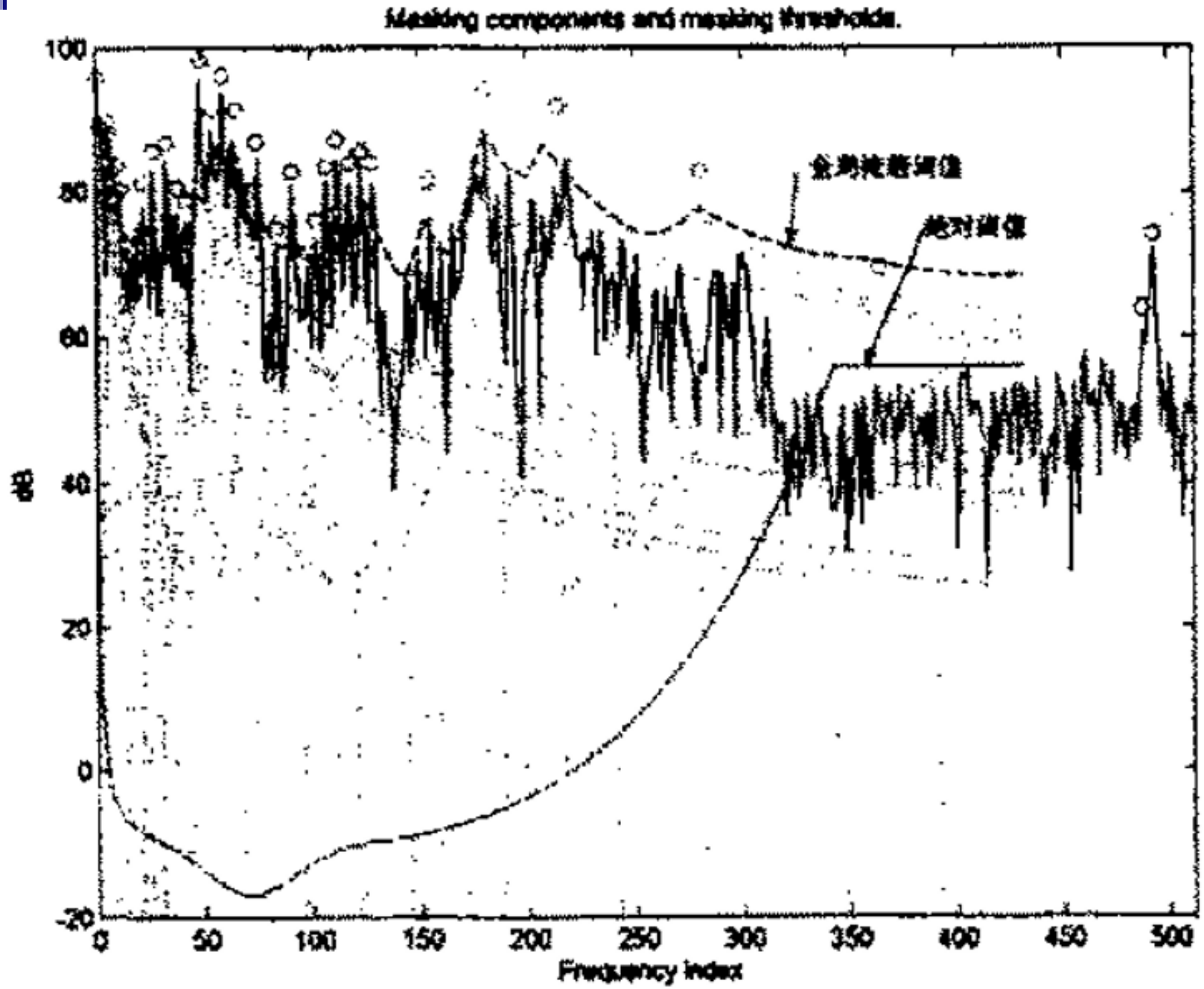
7、全局掩蔽阈值的计算

- 还要考虑别的临界频带的影响。一个掩蔽信号会对其它频带上的信号产生掩蔽效应。这种掩蔽效应称为掩蔽扩散。

$$SF_{dB}(x) = 15.81 + 7.5(x + 0.474) - 17.5\sqrt{1 + (x + 0.474)^2}\, dB$$

其中x的单位为巴克，SF(x)单位是dB。



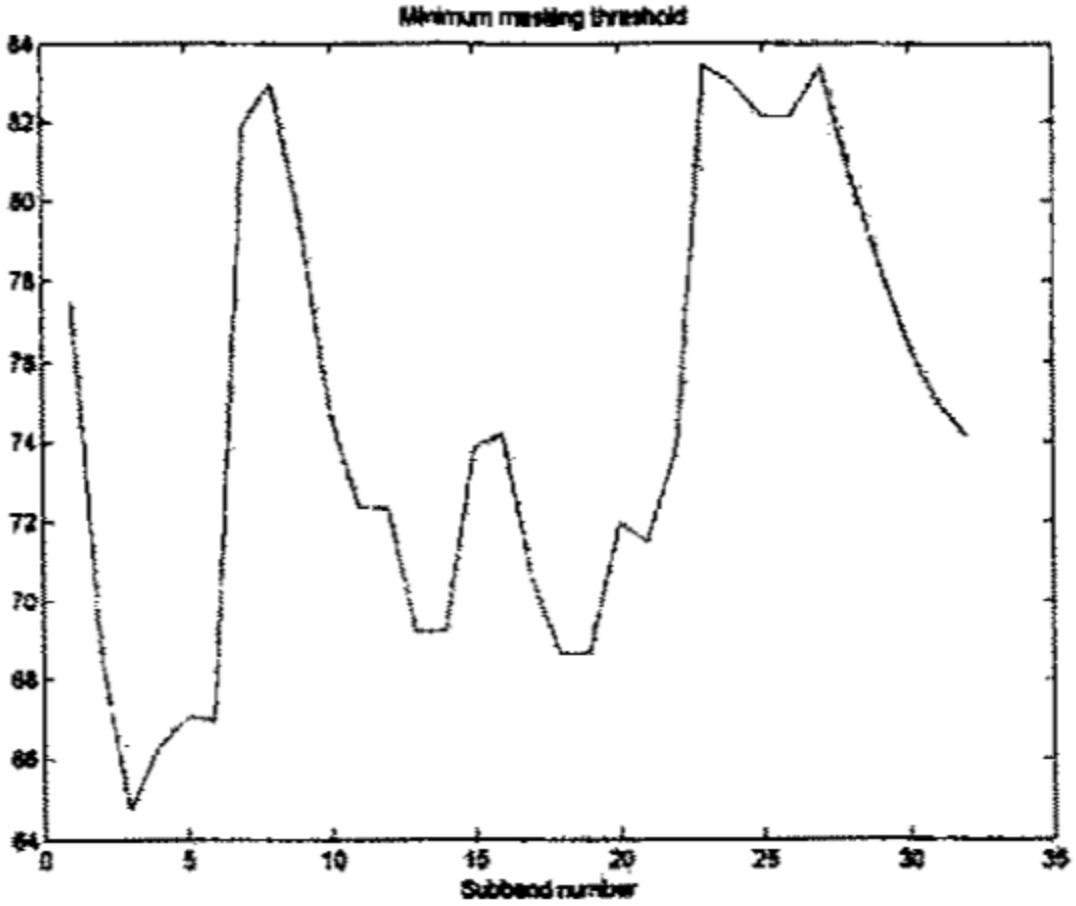

Masking components and masking thresholds.
100
80
60
40
20
0
-20
dB
0
50
100
150
200
250
300
350
400
450
500
Frequency index

心理声学模型

- 8、每个子带的掩蔽阈值
  - 选择出本子带中最小的阈值作为子带阈值
  - 对高频不正确——高频区的临界频带很宽，可能跨越多个子带，从而导致模型1将临界带宽内所有的非音调部分集中为一个代表频率，当一个子带在很宽的频带内却远离代表频率时，无法得到准确的非音调掩蔽值。但计算量低。

心理声学模型

- 8、每个子带的掩蔽阈值
  - 选择出本子带中最小的阈值作为子带阈值



心理声学模型

- 9、计算每个子带信号掩蔽比(signal-to-mask ratio, SMR)
  - SMR = 信号能量 / 掩蔽阈值
- 并将SMR传递给编码单元

Layer I 编码：码率分配

- **在调整到固定的码率之前**
  - **先确定可用于样值编码的有效比特数**
  - **这个数值取决于比例因子、比例因子选择信息、比特分配信息以及辅助数据所需比特数**
- **比特分配的过程**
  - 对每个子带计算掩蔽-噪声比MNR，是信噪比SNR–信掩比SMR，即：MNR = SNR–SMR
  - NMR=SMR-SNR

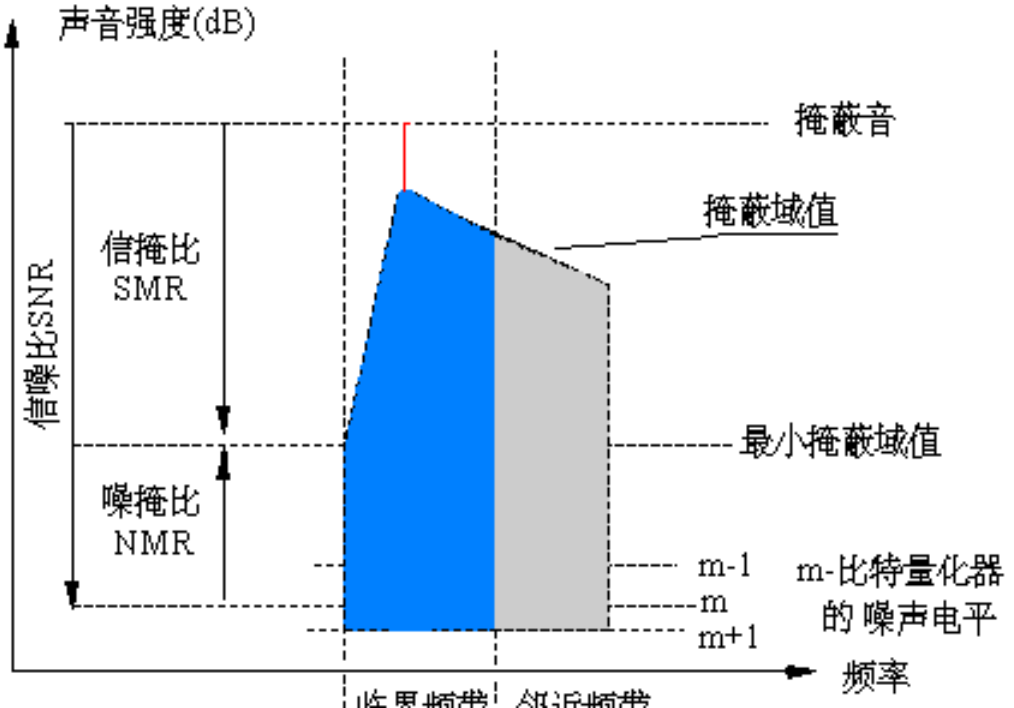
Layer I 编码：码率分配

- 算法：使整帧和每个子带的总噪声—掩蔽比最小
  - 计算噪声-掩蔽比(noise-to-mask ratio, NMR)：

$$NMR = SMR - SNR \qquad (dB)$$

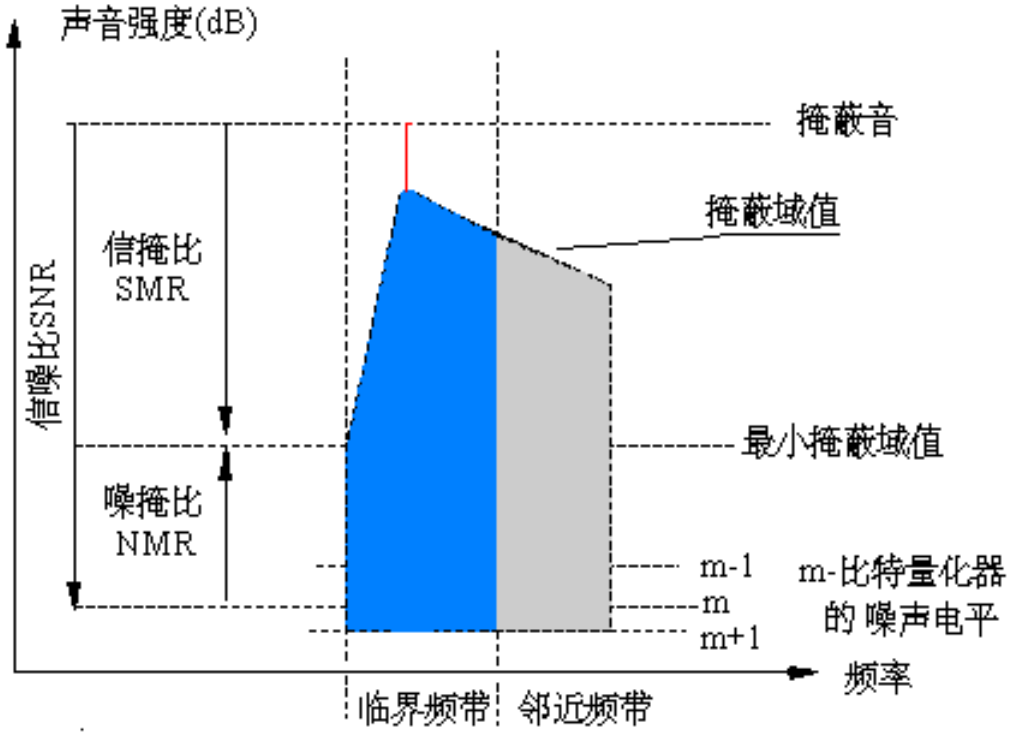
其中SNR 由MPEG-I标准给定 (为量化水平的函数)

NMR：表示波形误差与感知测量之间的误差

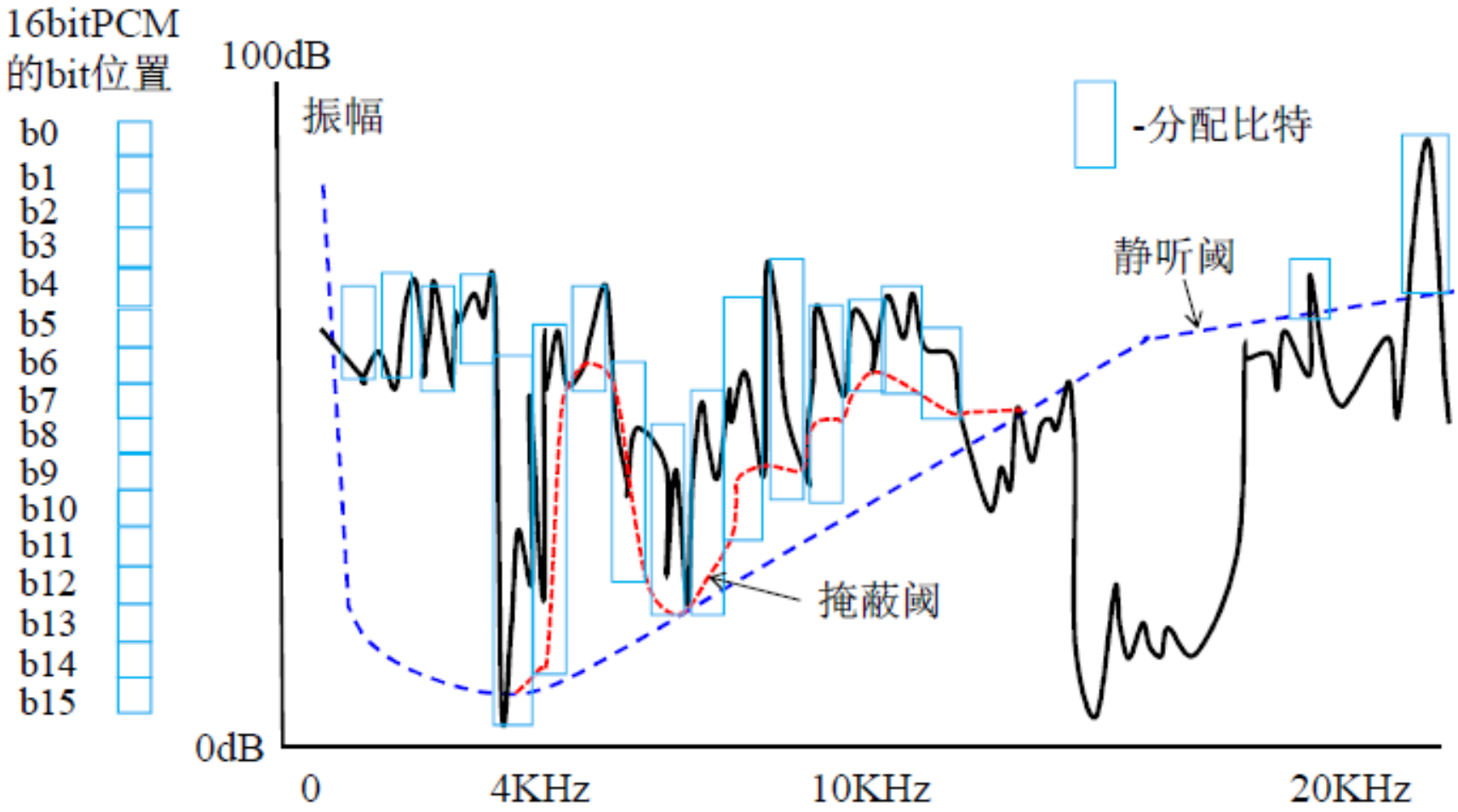


Layer I 编码：码率分配

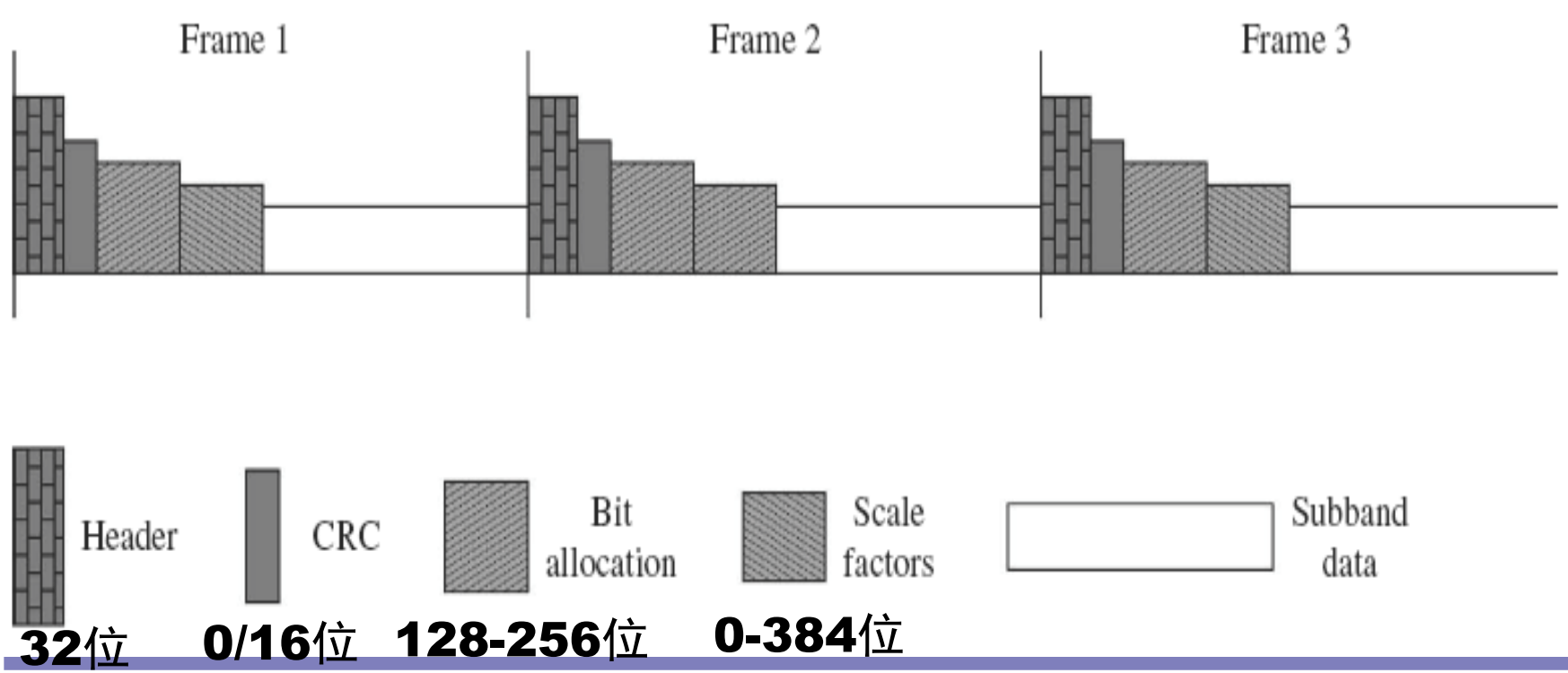
- 算法：循环，直到没有比特可用：
  - NMR = SMR– SNR (dB)
  - 对最高NMR的子带分配比特，使获益最大的子带的量化级别增加一级
  - 重新计算分配了更多比特子带的NMR



Layer I 编码：码率分配

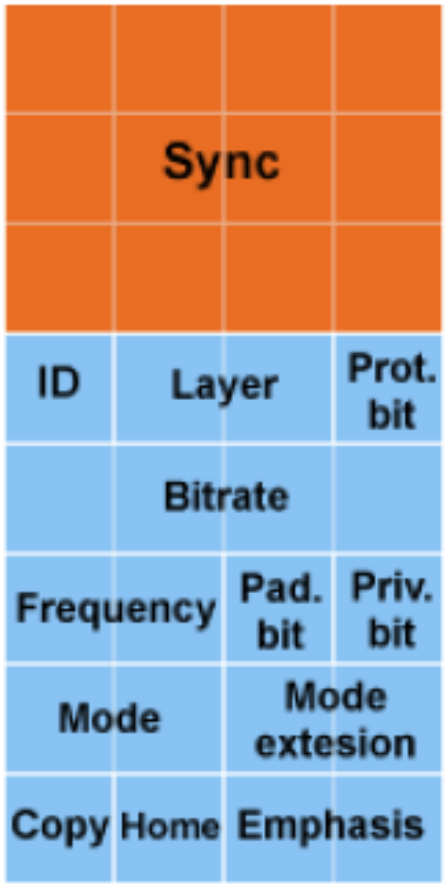


Layer I 编码：装帧


Frame 1
Frame 2
Frame 3
Header
CRC
Bit allocation
Scale factors
Subband data
32位
0/16位
128-256位
0-384位

Header：帧头格式

- 32比特：



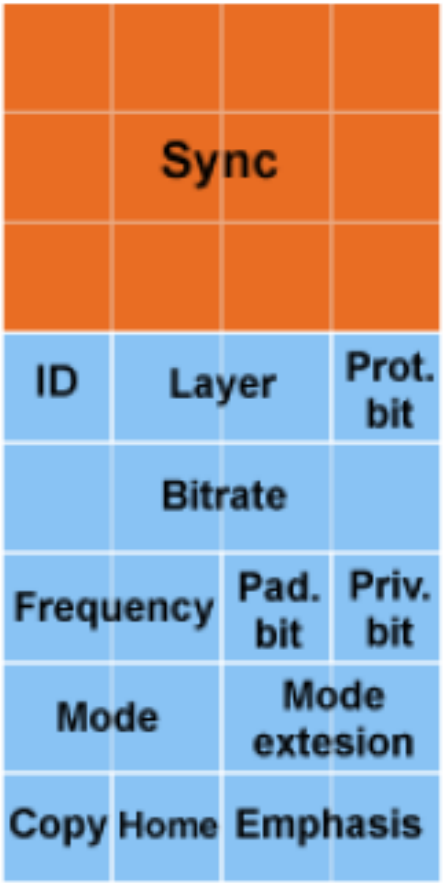
- Sync：同步码，取值为111111111111
  - 可以随机访问和回放
- ID=1：MPEG
- Layer：

00	reserved
01	Layer III
10	Layer II
11	Layer I

- Prot. bit：置位表示有CRC校验
- Bitrate：码率索引，表示15种固定的码率

Header：帧头格式

- 32比特：



- Frequency：取样频率

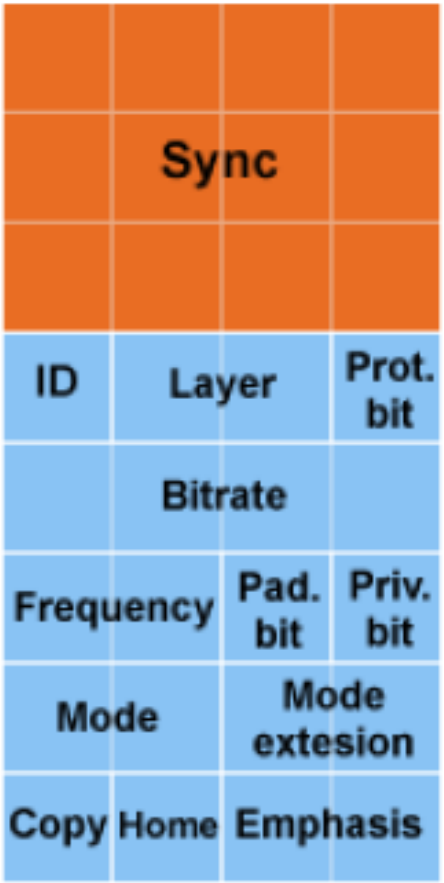
Bits	MPEG1	MPEG2	MPEG2.5
00	44100 Hz	22050 Hz	11025 Hz
01	48000 Hz	24000 Hz	12000 Hz
10	32000 Hz	16000 Hz	8000 Hz
11	reserv.	reserv.	reserv.

- Pad bit：填充指示位
- Priv. bit：应用特定
- Mode：通道模式

00	**Stereo**	分别编码，一起播放
01	**Joint Stereo**	联合编码：**L, R → M, S**
10	**Dual Channel**	分别编码，分别播放
11	**Single Channel**	

Header：帧头格式

- 32比特：



- Mode extesion：通道模式扩展，当通道模式为联合立体声时有效
  - 利用立体声双声道的相关性编码

Bits	Intensity stereo	MS stereo
00	Off	Off
01	On	Off
10	Off	On
11	On	On

  - M, S：middle-side

$$M = \frac{L+R}{2}, S = \frac{L-R}{2}$$

Layer II编码：概述

- 与Layer I类似，但对Layer I有增强
  - 装帧、缩放因子表示、量化
  - 缩放因子(比例因子)一般从低频子带到高频子带出现连续下降
- 帧：
  - 3 组/帧 x 12个样本/子带 x 32个子带/帧 = 1152个样本/帧

  → 每个样本的overhead更少
- 缩放因子：每个子带的3个组尽可能共用缩放因子
  - Layer I: 1个/12个样本
  - Layer II： 1个/（24/36）个样本
    - 1/2/3个缩放因子和缩放因子选择信息(scale factor selection information, SCFSI) （每子带2比特）一起传送
      - 如果缩放因子和下一个只有很小的差别，就只传送大的一个，这种情况对于稳态信号经常出现
      - 如果要给瞬态信号编码，则要在瞬态的前、后沿传送两个或所有三个比例因子

比例因子：① 掩蔽心理
学模型
② 帮助量化

Layer II编码：量化

- Layer I：每个子带从相同的量化集合中选择
  - 每个子带取共14个量化器中的一个

每子频带12个连续的样值都除以比例因子进行归一化，得到的值用X表示，进行量化计算：

$$A \times X + B$$ ，A和B：量化系数。

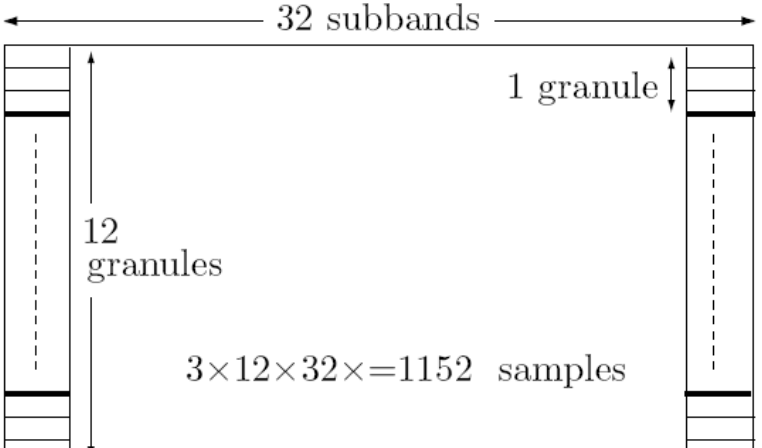
查量化表：根据Bit分配信息得量化级数，根据级数查量化表得A和B。

量化级	A	B	量化级	A	B
3	0.75	-0.25	1023	0.99902438	-0.000976563
7	0.625	-0.375	2047	0.999511719	-0.000488281
511	0.998046875	-0.001953125	65535	0.999984741	-0.000015259

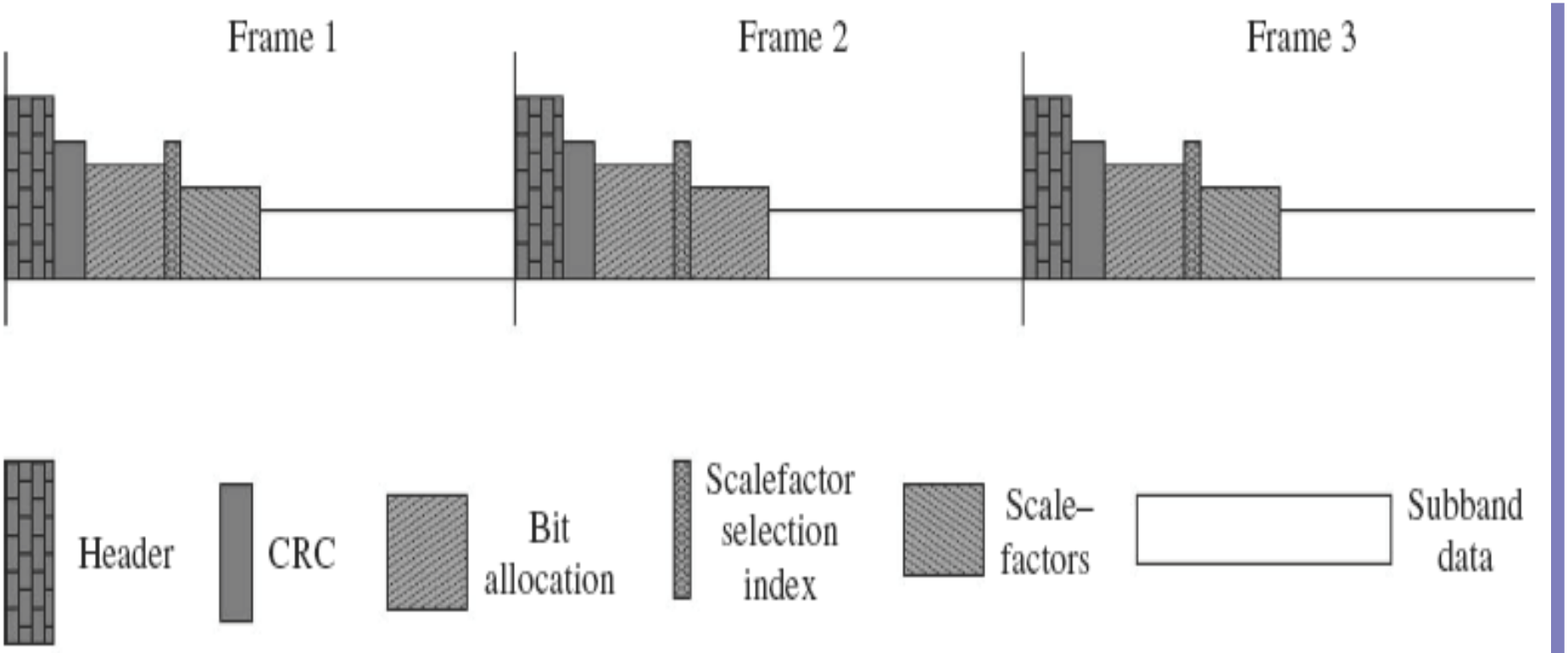
Layer II编码：量化

- Layer II：
  - 根据采样和码率量化，不同子带可以从不同的量化器集合中选择
    - 某些（高频）子带的比特数可能为0
  - 对量化级别在3、5、9级时，采用“颗粒” 优化
    - 颗粒= 3 个样本，根据颗粒选择量化水平
    - 例：3个样本 @ 3个量化水平 = 27种可能的值 → 5 比特
      - 不采用颗粒量化：1个样本 @ 3个量化水平 = 2比特 x 3 个样本 → 6 比特

可将压缩比从4：1增加到6：1至8：1



Layer II 编码：装帧



32位　0/16位　26-128位　0-60位　0-1080位

实验要求（基本）

- 理解程序设计的整体框架
- 理解感知音频编码的设计思想
  - 两条线
  - 时-频分析的矛盾！
- 理解心理声学模型的实现过程
  - 临界频带的概念
  - 掩蔽值计算的思路
- 理解码率分配的实现思路

主页

实验要求（基本）

- 输出音频的采样率和目标码率
- 选择三个不同特性的音频文件
  - 噪声（持续噪声、突发噪声）
  - 音乐
  - 混合
- 某个数据帧，输出
  - 该帧所分配的比特数
  - 该帧的比例因子
  - 该帧的比特分配结果

实验要求（进阶）

- 将Mp2音频编码器代码进行改造，生成动态链接库或静态链接库
- 编写调用上述动态/静态链接库的主程序
  - 可以选择输入的音频源文件
  - 可以调整各种编码参数
  - 可以选择编码后的文件，并保存
  - 具有界面——加分