第三部分 音频压缩

第三部分 音频压缩

有损的音频压缩完全是基于人类的听觉特点，所以在阐述压缩之前必须先考虑人类听觉的问题。令人惊奇的是，人类的听觉要比视觉灵敏得多，尤其在立体声环境中更是如此，所以音频压缩更要小心进行。与视频压缩相似，音频压缩按照所要求的压缩系数不同，需要的复杂程度也不同。

3.1 听觉机制

听觉由耳部的物理处理和构成声音印象的神经/大脑处理组成。我们所接收到的印象并非与耳道中的声波完全一致，这是因为丢失了一些熵。如果音频压缩系统仅丢失听觉机制中丢失的那部分熵，那么该音频压缩系统会产生很好的效果。物理听觉机制由外耳、中耳和内耳组成。外耳包括耳道和耳鼓、耳鼓将瞬间的声音转换为象麦克风膜片那样的振动。内耳则感知通过液体传来的振动。由于液体的阻抗比空气大得多，所以中耳就好比阻抗匹配转换器，提高了声音传播的能力。

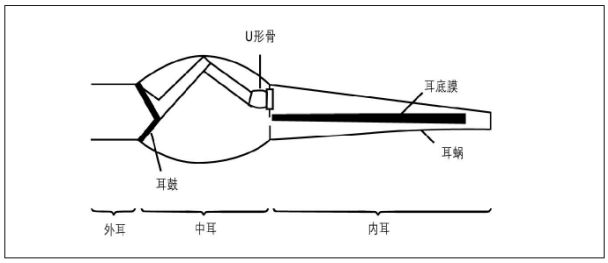

图 3.1

图3.1 显示了有镫骨传到内耳的振动，镫骨在椭圆形窗上起作用。液体中的振动在耳蜗中传播，耳蜗是骨骼中的螺旋腔。(图3.1为清楚起鉴未显示螺旋状)。耳蜗内撑有基本膜。该膜的质量和硬度根据长度会发生变化。在靠近椭圆形窗的一端，该膜硬而轻，所以共振频率较高。该膜在远端则重而软，共振频率较低。所能得到的共振频率范围决定了人类听觉的频率范围，通常从20Hz 到大约15kHz。

输入声音的不同频率会引起膜的不同区域发生振动。每个区域有不同的神经末梢来辨别声音高低。基本膜上还有由神经控制的小肌肉来共同组成一种主动反馈系统，以便改善共振的Q 系数。

基本膜的共振作用与变换分析仪的作用是完全一致的。根据变换的不确定理论，所知的信号频域越精确，则所知的时域就越不精确。因此，变换越能辨别两种频率，则越难以辨别两个事件发生的时间区别。人类的听觉在时间不稳定辨别和频率辨别之间采取了一种折衷，但其能力不是完美无缺的。

不完美的频率辨别造成无法区分邻近的频率，这就叫作听觉遮蔽，定义为在另一声音存在时对一声音的灵敏度降低。

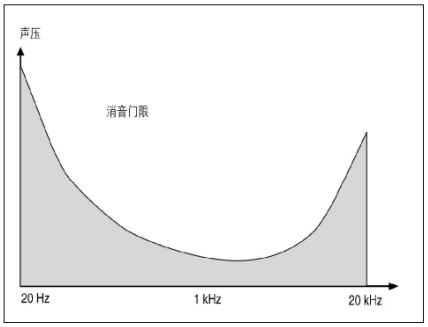

图 3.2a

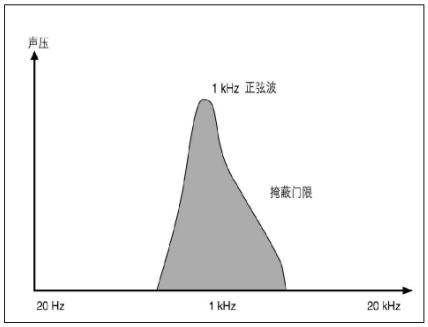

图 3.2b

图3.2a显示听觉的阀值是频率的函数。毫无疑问，最佳的灵敏度存在于语音范围内。只存在一个声调时阀值变为如图3.2b所示。请注意，阀值在较高频率和某些较低频率情况下会上升。在如音乐那样较复杂的输入频谱中，阀值在几乎所有频率上均上升。该情况的一种结果就是模拟音带中的哗哗声只有在音乐中安静的段落中才能听到。压缩就是运用该原理，在录制或传送前先将低电平音频信号放大，而后再将它们返回到正确的电平。

耳朵对时间辨别的不完全是由于耳朵的共振反应造成的。 Q系数就是指声音出现后至少一秒种后才能被听见。由于反应速度慢，所以即便两个信号非同时出现时也会产生遮蔽。当遮蔽声音在其实际长度前后继续遮蔽较低电平声音时就会发生前向或后向遮蔽。图3.3显示了这个概念。

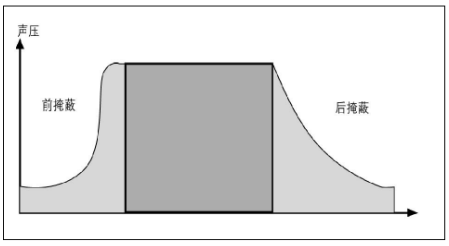

图 3.3

遮蔽提高了听觉的阀值。压缩器通过升高噪声底来利用这种效果，从而使音频波形可以用较少的数位来表达。噪声底只能够在有效遮蔽存在的频率处上升。为了使有效遮蔽最大化，必需将音频频谱分到不同的频带，在每个频带上注入不同量的压缩和噪声。

3.2 亚能带编码

图3.4 显示了频带分解压缩扩展器方框图。频带分解压缩扩展器是一套窄带线性相位滤波器，它们相互交叠，并具有相同的带宽。每个带的输出由表示波形的取样物组成。在每个频率带中，音频输入在传送之前先放大至最大电平。然后，每个电平再返回到其正确的值。传送中的噪声在每个带中均降低。如果将降低的噪声与听觉的阀值相比较，我们可以发现由于遮蔽作用使一些频带能容忍较大的噪声。因此，在经过压缩扩展的每个频带中可以降低取样的字长。由于因分辨力损失引入的噪声被遮蔽，所以该项技术能够实现一种压缩。

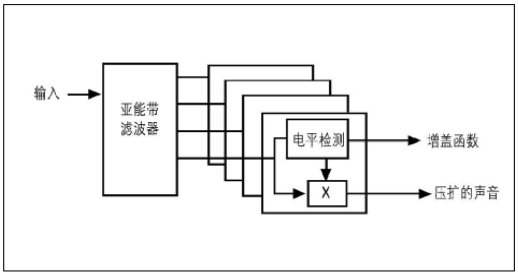

图 3.4

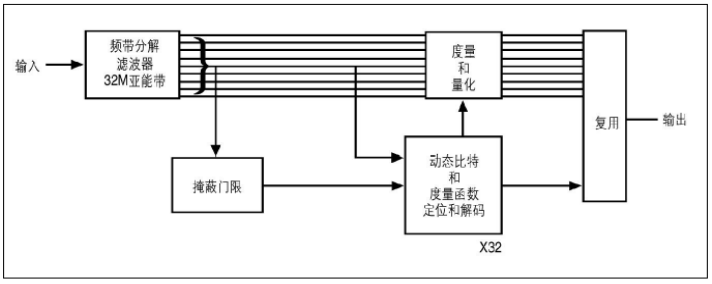

图 3.5

图3.5 显示了在MPEG 第一层面中使用的简单频带分解编码器。数字音频输入被送到频带分解滤波器中，滤波器将信号频谱分成若干个频带。在MPEG中这个数字是32。时间轴被分割成相同长度的块。在MPEG第一层面中共有384 个输入取样物，所以在滤波器输出的32个频带中每个有12个取样物。在每个频带中，电平通过乘法放大到最大值。所需的增益在一个块的长度中保持恒定，且单个度量系数与每个频带的每一块一起发送，以便在解码器中实现反向处理。

滤波器组的输出也经过分析确定输入信号的频谱。该分析决定是遮蔽类型，即每个频带中能够预期的遮蔽程度。能够实现的遮蔽越大，在每个频带中取样物的精度可能越差。通过重新量化可降低取样物的精度，从而减少字长。字长的减少对频带中的每个字都是固定的，但不同频带则可能使用不同的字长。对每个频带，字长需比特串行定位编码后进行传送，从而使解码器正确地串并转换数据流。

3.3 MPEG 第一层面

图3.6 显示了一个MPEG 第一层面音频数据流。在同步图案和报头信息之后有每次4位共32位的定位编码。这些编码描述了每个亚能带中的取样物的字长。紧跟着的是在每个频带压缩扩展时使用的32 个度量系数。这些度量系数决定了解码器为了将音频返回到正确电平所需要的增益。度量系数后面依此跟着每个频带中的音频数据。

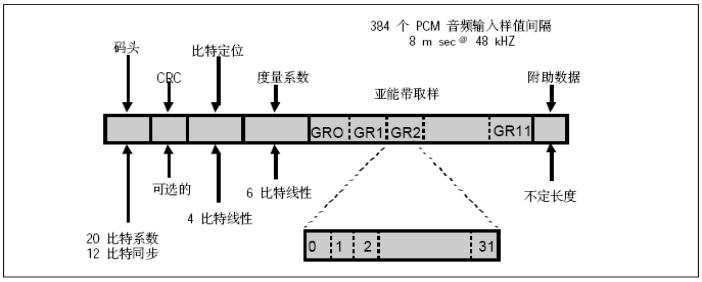

图 3.6

图3.7显示的是第一层面解码器。定时发生器检测同步图案，从而串并转换比特定位和度量系数。然后比特定位数据允许串并转换成可变长度的取样。通过度量系数数据将重新量化和压缩反变换，使每个频带都回到正确的电平。接着，32 个不同的频带在组合滤波器中被组合起来产生音频输出。

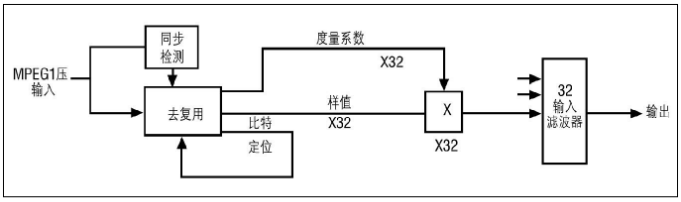

图 3.7

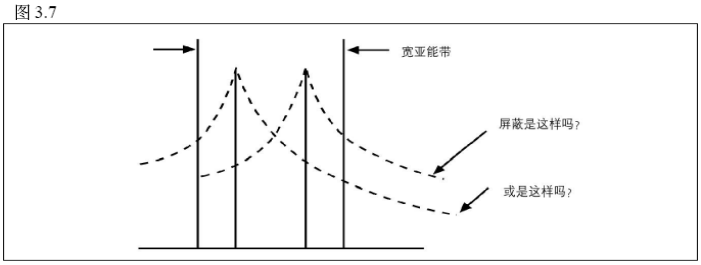

图 3.8

3.4 MPEG 第二层面

图3.8显示，当使用频带分解滤波器决定遮蔽类型时，频谱分析可能不是非常精确。这是因为只有32个频带，而能量可能在频带中的任何地方。因为在最坏的情况下遮蔽可能不起作用，所以噪声比无法上升很多。频谱分析越精确就越要求较高的压缩系数。在MPEG第二层面中，频谱分析由一些不同的处理过程组成。我们使用直接工作于输入上的512点FFT来替代决定遮蔽类型。为了更精确地分辨频率，变换的时间跨度不得不增加，即将块尺寸提高到1152 个样本。

当块压缩扩展方案与第一层面中的完全一样时，并非所有度量系数均被发送，这是因为度量系数中含有一些真实节目内容上的冗余。相同频带中连续块的度量系数超出了比时间的10%少2dB的限度，通过分析连续三个度量系数可以对这一特点加以利用。在内容固定的节目中，三个度量系数中只发送一个。随着亚能带中瞬变内容的增加，就要发送两个或三个度量系数。同时还要发送一个度量系数选择编码，使解码器能够确定每个亚能带中发送了什么。该项技术有效地使度量系数数据率降低了一半。

3.5 变换编码

第一层面和第二层面均以频带分隔滤波器为基础，信号仍然表现为波形。然而，第三层面采用了与视频编码相似的变换编码。如前面所述，耳朵实现了瞬间声音的频率变换，由于耳底膜的Q系数使反应不能快速增加或降低。因此，如果音频波形变换为频域，那么就不必经常传送系数。该理论是变换编码的基础。对于较高的压缩因素而言，系数会被重新量化，使精确度下降。该项处理会产生位于遮蔽最大的频率上的噪声。变换编码器的副产物是输入频谱能够精确获知，从而可以建立准确的遮蔽模式。

3.6 MPEG 第三层面

在需要最高的压缩系数时，我们就要求有这种复杂层的编码。它与第二层面有一定程度的相似性，使用离散余弦变换，每块含384个输出系数。该输出可以通过直接处理输入取样来获得，但在多层编码器中则可以使用与第一层面和第二层面32频带滤波相互配合的混合变换作为基础。如果这样做了，那么QMF(正交镜面滤波器)上的32个亚能带将每个由12 频带MDCT(修正离散余弦变换)进一步处理，从而获得384 个输出系数。

我们使用两个窗口尺寸以避免瞬变上的预先回声。窗口的切换由音质模式来操作。我们发现预先回声与上升到平均值以上的音频中的熵有关。为了获得最高的压缩系数，我们需要将系数的非统一量化与霍夫曼(Huffman)编码一起使用。该项技术分配最短的字长为最普通的编码值。

3.7 AC-3

AC-3音频编码技术与ATSC系统一起使用，用来替代MPEG音频编码方案之一。AC-3是以变换为基础的系统，其通过重新量化频率系数获得编码增益。

输入AC-3编码器的PCM被分为重叠的窗口块，如图3.9所示。这些块每个包含512 个取样物，但由于完全重叠，就有100% 冗余。在变换之后，每个块中有512 个系数，但由于冗余的原因，这些系数可以用所谓时域混叠消除(TDAC)的技术以十取一的方式变为256 个系数。

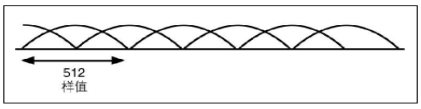

图 3.9

输入波形经过分析，如果块的后一半存在显著瞬变，则波形会一分为二以避免预先回声。在这种情况下，系数的数目保持不变，但频率分辨率会降低一半，而时间分辨率会增加一倍。在数据流中会设立一个旗标向解码器作出上述所做的提示。

系数如尾数和幂那样以浮点方式输出。再以科学计数的二进制等值数重现。幂是有效的度量系数。块中的幂的设置形成对输入的谱分析，具有对数尺度的有限精度，被称作频谱包络。这一频谱分析是遮蔽模式的输入，而该模式则确定了噪声在每个频率处上升的程度。

遮蔽模式驱动重新量化处理，通过将尾数四舍五入的方式降低了每个系数的精度。大部分传送数据都含有这些尾数。

幂也被传送，但并非象它们中间有可以进一步发掘的冗余那样进行直接传送。在一个块中，只有第一个(最低频率)幂被以绝对完整的形式传送，剩下的则以差异方式传送，解码器将差异添加到前面的幂上。在输入音频存在平滑频谱的地方，几个频率带中的幂可能是相同的。幂可以组合成带有旗标的两个或四个一套，用来描述做了些什么。

六个块组成一个AC-3同步帧。帧的第一块总是含有全幂数据，但在静止信号中，帧后面的块可以使用相同的幂。