第3章

数字音频压缩技术

本章首先介绍数字音频压缩的可行性和音频压缩方法，再详细介绍数字音频的各种压缩编码方法和音频压缩编码的标准，最后介绍常用的数字音频文件的格式。

【本章学习目标】

- 了解数字音频压缩的可行性、压缩方法的主要分类和数字音频文件的常见格式。
- 掌握数字音频的压缩方法和主要的音频压缩编码标准。

3.1 数字音频压缩技术概述

3.1.1 数字音频压缩的可行性

为利用有限的资源，压缩技术从一出现便受到广泛的关注。信号(数据)之所以能进行压缩，是因为信号本身存在很大冗余度。根据统计分析结果，音频信号中存在着多种冗余，可分别从时域和频域来考虑。另外，由于音频主要是给人听的，所以考虑人的听觉机理，也能对音频信号进行压缩。

1. 时域冗余

音频信号在时域上的冗余主要表现为以下几个方面。

1) 幅度分布的非均匀性

统计表明，在大多数类型的音频信号中，小幅度样值出现的概率比大幅度样值出现的概率要高。人的语音中，间歇、停顿等出现了大量的低电平样值；实际讲话的功率电平也趋向于出现在编码范围的较低电平端。

2) 样值间的相关性

对语音波形的分析表明，相邻样值之间存在很强的相关性。当采样频率为 8kHz 时，相邻样值之间的相关系数大于 0.85。如果进一步提高采样频率，则相邻样值之间的相关性将更强。因此，根据较强的一维相关性，可以利用差分编码技术进行有效的数据压缩。

3) 周期之间的相关性

虽然音频信号分布于 20Hz～20kHz 的频带范围，但在特定的瞬间，某一声音却往往只

是该频带内的少数频率成分在起作用。当声音中只存在少数几个频率时，就会像某些振荡波形一样，在周期与周期之间存在一定的相关性。利用音频信号周期之间的相关性进行压缩的编码器，比仅利用邻近样值间的相关性的编码器效果好，但要复杂得多。

4）基音之间的相关性

语音可以分为清音和浊音两种基本类型。浊音由声带振动产生，每一次振动使一股空气从肺部流进声道。激励声道的各股空气之间的间隔称为基音周期。浊音的波形对应于基音周期的长期重复波形。对浊音编码是对一个基音周期波形进行编码，并以它作为其他基音段的模板。

5）静止系数

两个人之间打电话，平均每人讲话时间为通话时间的一半，并且在这一半的通话过程中也会出现间歇停顿。分析表明，话音间隙使全双工话路的典型效率约为 40%（或称静止系数为 0.6）。显然，话音间隔本身就是一种冗余，若能正确检测出这些静止段，可插空传输更多信息。

6）长时自相关函数

统计样值、周期间的一些相关性时，在 20ms 时间间隔内进行统计，称为短时自相关函数。如果在较长的时间间隔（如几十秒）内进行统计，则称为长时自相关函数。长时统计表明，当采样频率为 8kHz 时，相邻样值之间的平均相关系数可高达 0.9。

2. 频域冗余

1）长时功率谱密度的非均匀性

在相当长的时间间隔内进行统计平均，可以得到长时功率谱密度函数，其功率谱呈现明显的非平坦性。从统计的观点看，这意味着没有充分利用给定的频段，或者说存在固有的冗余度。功率谱的高频成分能量较低。

2）语音特有的短时功率谱密度

语音信号的短时功率谱，在某些频率上出现峰值，而在另一些频率上出现谷值。这些峰值频率，也就是能量较大的频率，通常称为共振峰频率。共振峰频率不止一个，最主要的是前三个，由它们决定不同的语音特征。另外，整个功率谱也是随频率的增加而递减的。更重要的是整个功率谱的细节以基音频率为基础，形成了高次谐波结构。

3. 听觉冗余

人是音频信号的最终用户，因此，要充分利用人类听觉的生理和心理特性对音频信号感知的影响。利用人耳的频率特性、灵敏度以及掩蔽效应，可以压缩数字音频的数据量。可以将会被掩蔽的信号分量在传输之前就去除，因为这部分信号即使传输了也不会被听见；可以不理会可能被掩蔽的量化噪声；可以将人耳不敏感的频率信号在数字化之前滤除，如语音信号只保留 300～3400Hz 的信号。

3.1.2 数字音频压缩方法的分类

语音编码技术是一种很重要的技术。各种编码技术对语音信号进行处理，目的都是降低传输码率和提高音质。

1. 根据压缩品质的不同分类

一般来讲，可以将音频压缩技术分为无损（Lossless）压缩和有损（Lossy）压缩两大类。

1）无损压缩编码

无损压缩编码是一种可逆编码，其特点是利用数据的统计特性进行压缩编码，出现概率大的数据采用短编码，概率小的数据采用长编码，以去掉数据中的冗余，而且编码后的数据在解码后可以完全恢复，压缩比较小。无损压缩编码主要用于文本数据、程序代码和某些要求严格不丢失信息的环境中，常用的有霍夫曼编码、行程（游程）编码、算术编码和词典编码等。

2）有损压缩编码

有损压缩编码损失的信息是不能恢复的，所以是一种不可逆编码。编码后的数据在解码以后所复原的数据与原数据相比有一定的可以容忍的误差。它考虑到编码信息的语义，可获得的压缩程度取决于媒介本身。其压缩比比无损编码大得多。常用的有损压缩编码技术有预测编码、变换编码、矢量量化编码、分层编码、频带分割编码和混合编码等。

2. 根据编码方式的不同分类

根据编码方式的不同，音频编码技术分为 4 种：波形编码、参数编码、混合编码和感知编码。一般来说，波形编码的话音质量高，编码速率也很高；参数编码的编码速率很低，产生的合成语音的音质不高；混合编码使用参数编码技术和波形编码技术，编码速率和音质介于它们之间。

1）波形编码

波形编码是指不利用生成音频信号的任何参数，直接将时域信号转换为数字代码，使重构的语音波形尽可能地与原始语音信号的波形形状保持一致。波形编码的基本原理是在时间轴上对模拟语音信号按一定的速率采样，然后将幅度样本分层量化，并用代码表示。

波形编码方法简单，易于实现，适应能力强且语音质量好。不过因为压缩方法简单，也带来了一些问题：压缩比相对较低，需要较高的编码速率。一般来说，波形编码的复杂程度比较低，编码速率较高，通常在 16kb/s 以上，质量相当高。但编码速率低于 16kb/s 时，音质会急剧下降。

最简单的波形编码方法是脉冲编码调制（Pulse Code Modulation，PCM），它只对语音信号进行采样和量化处理。其优点是编码方法简单，延迟时间短，音质高，重构的语音信号与原始语音信号几乎没有差别，不足之处是编码速率比较高（64kb/s），对传输通道的错误比较敏感。

2）参数编码

参数编码是从语音波形信号中提取生成语音的参数，使用这些参数通过语音生成模型重构出语音，使重构的语音信号尽可能地保持原始语音信号的语义。也就是说，参数编码是把语音信号产生的数字模型作为基础，然后求出数字模型的模型参数，再按照这些参数还原数字模型，进而合成语音。

参数编码的编码速率较低，可以达到 2.4kb/s，产生的语音信号是通过建立的数字模型还原出来的，因此重构的语音信号波形与原始语音信号波形可能会存在较大的区别，失真会比较大。而且因为受到语音生成模型的限制，增加数据速率也无法提高合成语音的质量。不过，虽然参数编码的音质比较低，但是保密性很好，一直被应用在军事上。典型的参数编

码方法为线性预测编码(Linear Predictive Coding,LPC)。

3) 混合编码

混合编码是指同时使用两种或两种以上的编码方法进行编码。这种编码方法克服了波形编码和参数编码的弱点,并结合了波形编码的高质量和参数编码的低编码速率,能够取得比较好的效果。

4) 感知编码

感知编码是利用人耳听觉的心理声学特性(频谱掩蔽特性和时间掩蔽特性)以及人耳对信号幅度、频率、时间的有限分辨能力,凡是人耳感觉不到的成分不编码,不传送,即凡是对人耳辨别声音信号的强度、音调、方位没有贡献的部分(称为不相关部分或无关部分)都不编码和传送。对感觉到的部分进行编码时,允许有较大的量化失真,并使其处于听阈以下,人耳仍然感觉不到。简单地说,感知编码是以建立在人类听觉系统的心理声学原理为基础,只记录那些能被人的听觉所感知的声音信号,从而达到减少数据量而又不降低音质的目的。

3.2 数字音频压缩方法

3.2.1 波形编码

波形编码是将时域信号直接转换为数字代码,由于这种系统保留了信号原始样值的细节变化,从而保留了信号的各种过渡特征,所以波形编码系统的解码音频信号质量一般较高。波形编码系统的不足之处是传输码率比较高,压缩比不高。

1. 脉冲编码调制

脉冲编码调制(PCM)是各种数字编码系统中最规范的方法,也是应用最广泛的系统。PCM 是"数字化"的最基本的技术,模拟信号正是通过 PCM 转换成数字信号的,其具体过程是:通过采样、量化和编码 3 个步骤,用若干代码表示模拟形式的信息信号(如图像、声音信号),再用脉冲信号表示这些代码进行传输/存储。

2. 差分脉冲编码调制

PCM 方式是把采样值变成二进制数后进行传输或存储的。如果不记录采样值本身,而是记录采样值之间的差值,这种方式就称为差分脉冲编码调制(Differential Pulse Code Modulation,DPCM)。通常自然界的声音是频率越高,声压级越低,人耳特性也是随着频率的增高,灵敏度急剧下降。频率越高,声音的动态范围越小。PCM 与 DPCM 的电平分布如图 3-1 所示,可见 DPCM 方式是非常符合自然界规律的。3 位 DPCM 的系数法如表 3-1 所示。

3. 自适应脉冲编码调制

音频信号的振幅和频率分布是随时间比较缓慢但大幅度变化的。因此出现了根据邻近信号的性质使量化步长改变的编码,这就是自适应脉冲编码调制(Adaptive Pulse Code Modulation,APCM)。准瞬时压扩和动态加重就可以看作是一种 APCM。APCM 组成框图如图 3-2 所示。

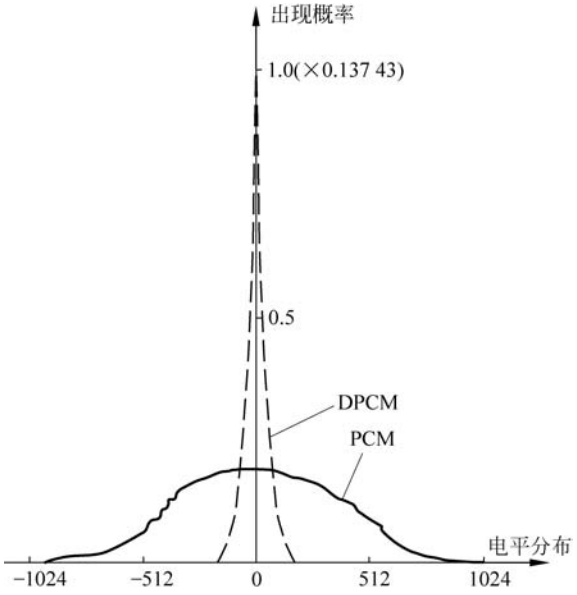

图 3-1　PCM 与 DPCM 的电平分布

表 3-1　3 位 DPCM 的系数法

DPCM 码(正值)	系　数	DPCM 码(负值)	系　数
011	1.75	111	0.90
010	1.25	110	0.90
001	0.90	101	1.25
000	0.90	100	1.75

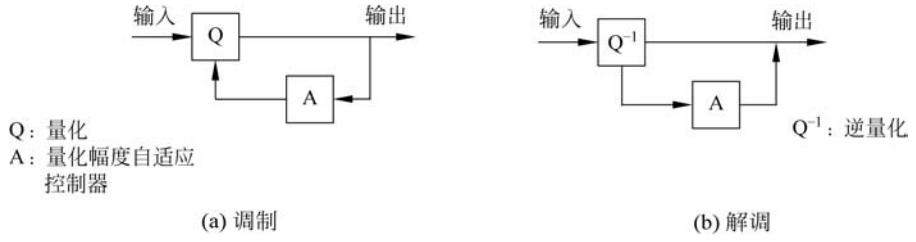

图 3-2　APCM 组成框图

4. 自适应差分脉冲编码调制

自适应差分脉冲编码调制(Adaptive Differential Pulse Code Modulation，ADPCM)就是把自适应量化步长引入 DPCM，即不是把信号 $x(n)$ 直接量化，而是把它和预测值 $x(n)$ 的差 $d(n)$ 进行量化，比前述的 APCM 效率高。ADPCM 组成框图如图 3-3 所示。多功能电话机的留言录音等短时间录音、不同磁带的固体录音机和向导广播、自动售货机以及多媒体技术应用领域的 CD-I 中，都是采用 4～8b 的 ADPCM。

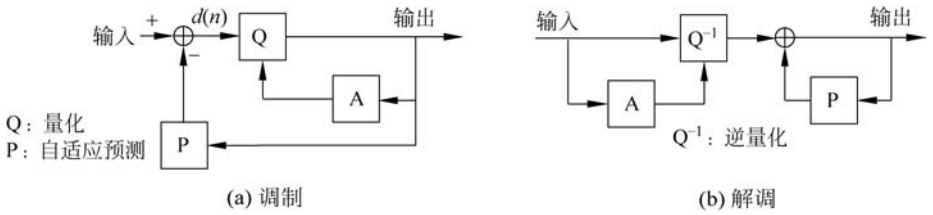

图 3-3　ADPCM 组成框图

5. 增量调制和自适应增量调制

1）增量调制

增量调制(Delta Modulation)是用一位二进制码表示相邻模拟采样值相对大小的 A/D 转换方式，ΔM 就是增量调制方式的代号。它将信号瞬时值与前一个采样值之差进行量化，并且对这个差值的符号进行编码，而不对差值的大小进行编码。因此，量化只限于正负两个电平，只用 1b 传输一个样值。如果差值为正，就发送“1”码；如果差值为负，就发送“0”码。数码“1”和“0”只是信号相对于前一时刻的增减，不代表信号的绝对值。简单增量调制原理和编码原理如图 3-4 和图 3-5 所示。

图 3-4 中，$x(t)$为一个模拟信号，$x'(t)$为本地解码器输出的前一时刻的量化信号。

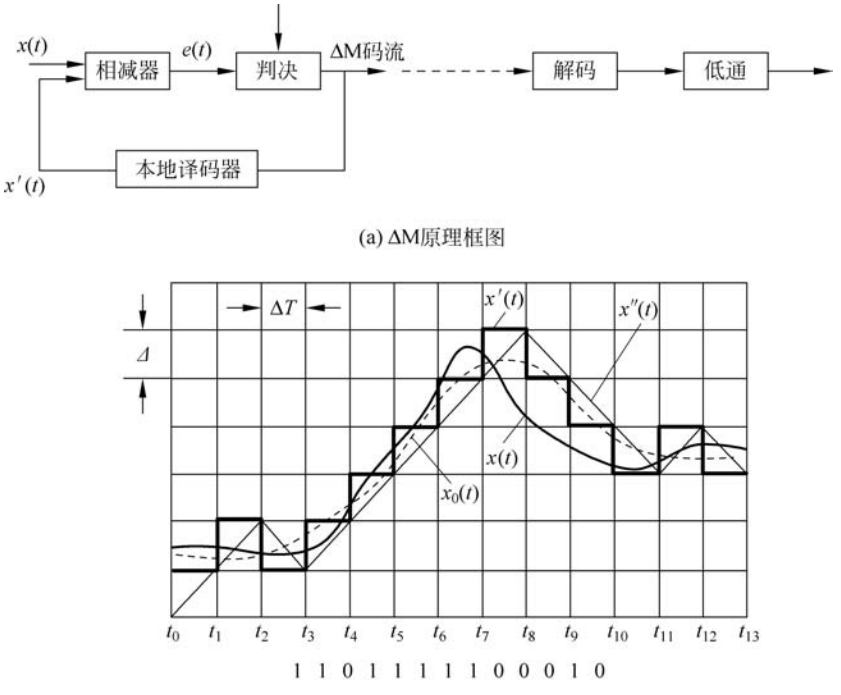

图 3-4　简单增量调制原理图

增量调制存在的主要问题有以下两个方面。

一个问题是斜率过载。当语音信号大幅度发生变化时，阶梯波形的上升或下降有可能跟不上信号的变化，因而产生滞后，这种失真称为“过载失真”。在斜率过载期间的码字将是

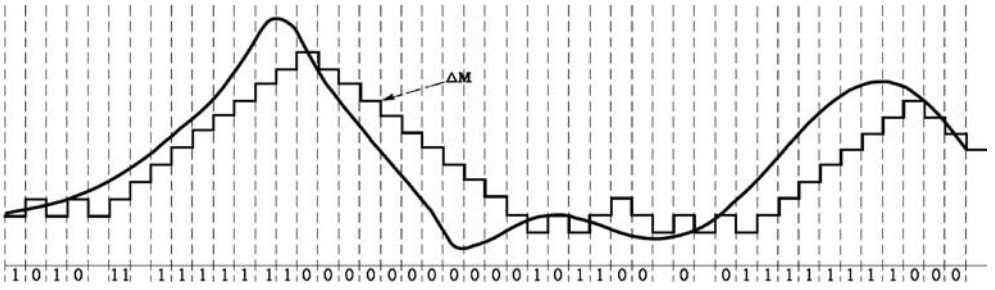

图 3-5　ΔM 编码原理

一连串的 0 或一连串的 1，如图 3-6 所示。为避免斜率过载，要求阶梯波的上升或下降的斜率必须大于或等于语音信号的最大变化斜率。

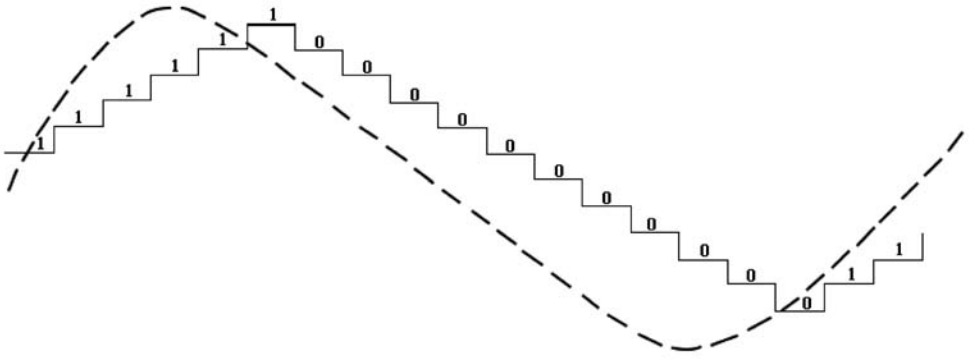

图 3-6　增量调制系统过载失真示意图

另一个问题是当话音信号不发生变化或变化很缓慢时，预测误差信号将等于零或具有很小的绝对值，在这种情况下，编码为 0 和 1 交替出现的序列。在解码器中得到的是等幅脉冲序列，这样形成的噪声称为颗粒噪声，如图 3-7 所示。

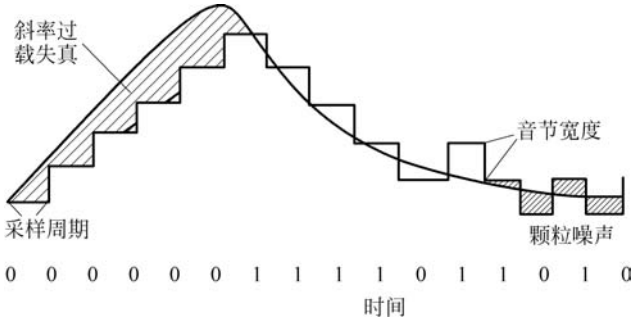

图 3-7　两种噪声的形式

2）自适应增量调制

自适应增量调制是一种改进型的增量调制方式，它的量化级随着音节时间间隔（5～20ms）中信号平均斜率而变化。由于这种方法中信号的平均斜率是根据检测码流中连“1”或连“0”的个数确定的，所以又称为数字检测、连续可变斜率增量调制（Continuously

Variable Slope Delta，CVSD)，简称为数字压扩增量调制，其原理如图 3-8 所示。

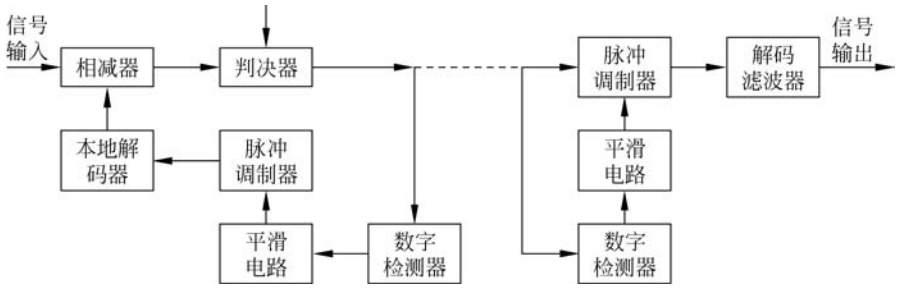

图 3-8　数字压扩增量调制原理

自适应增量调制与简单增量调制相比，编码器能正常工作的动态范围有很大提高，信噪比比简单增量调制优越。

6. 子带编码

子带编码(Subband Coding，SBC)是将一个短周期内的连续时间采样信号送入滤波器中，滤波器组将信号分成多个(最多 32 个)限带信号，以近似人耳的临界频段响应。由滤波器组的锐截止频率来仿效临界频段响应，并在带宽内限制量化噪声。子带编码要求处理延迟必须足够小，以使量化噪声不超出人耳的瞬时限制。子带编码通过分析每个子带的采样值并与心理声学模型进行比较，编码器基于每个子带的掩蔽阈值能自适应地量化采样值。子带编码中，每个子带都要根据所分配的不同比特数独立进行编码。在任何情况下，每个子带的量化噪声都会增加。当重建信号时，每个子带的量化噪声被限制在该子带内。由于每个子带的信号会对噪声进行掩蔽，所以子带内的量化噪声是可以容忍的。子带编码的原理如图 3-9 所示。

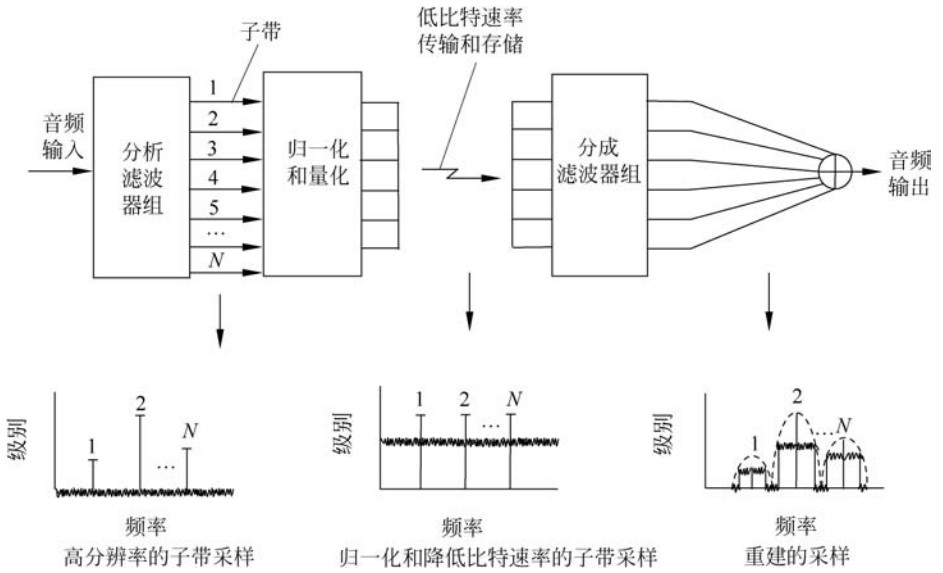

图 3-9　生成窄带高分辨率的子带编码

子带编码的主要特点如下。

(1) 每个子带对每一块新的数据都要重新计算，并根据信号和噪声的可听度对采样值

进行动态量化。

(2) 子带感知编码器利用数字滤波器组将短时的音频信号分成多个子带(对于时间采样值可以采用多种优化编码方法)。

(3) 每个子带的峰值功率与掩蔽级的比率由所做的运算决定,即根据信号振幅高于可听曲线的程度分配量化所需的比特数。

(4) 给每一个子带分配足够的位数,保证量化噪声处于掩蔽级以下。

可将子带编码与 ADPCM 编码相结合(SB-ADPCM),其编码和解码框图如 3-10 所示。

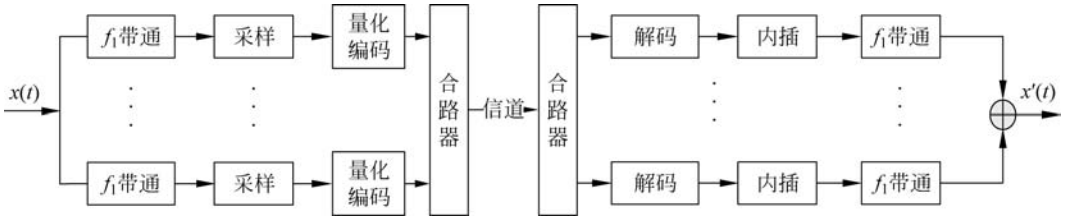

图 3-10　SB-ADPCM 编、解码框图

7. 矢量编码

标量量化(Scalable Quantization,SQ)是指独立地对一个样值量化编码的方式,由于对每一个样值单独编码处理,因此系统码率不可能低于采样频率。而矢量量化(Vector Quantization,VQ)是对若干个音频样值一起量化编码。在矢量量化编码中,把输入数据几个一组地分成许多组,成组地量化编码,即将这些数看成一个 k 维矢量,然后以矢量为单位逐个进行量化。VQ 的基本原理如图 3-11 所示。矢量量化是一种限失真编码,其原理仍可用信息论中的率失真函数理论来分析。率失真理论指出,即使对无记忆信源,矢量量化编码也总是优于标量量化。

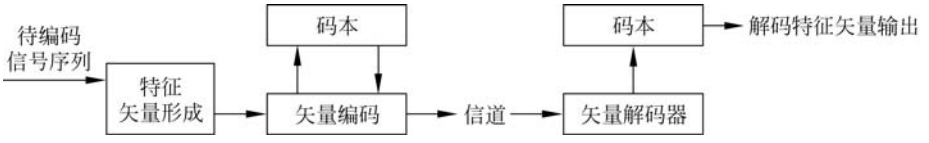

图 3-11　VQ 的基本原理

3.2.2　参数编码

参数编码技术以语音信号产生的数学模型为基础,根据输入语音信号分析出表征声门振动的激励参数和表征声道特性的声道参数,然后在解码端根据这些模型参数恢复语音。这种编码算法并不忠实地反映输入语音的原始波形,而是着眼于人耳的听觉特性,确保解码语音的可懂度和清晰度。基于这种编码技术的编码系统一般称为声码器,主要用在窄带信道上提供 4.8kb/s 以下的低速率语音通信和一些对时延要求较宽的场合。当前参数编码技术主要的研究方向是线性预测声码器和余弦声码器。

1. 语音生成模型

参数编码的基础是人类语音的生成模型。语音学和医学的研究结果表明,人类发音器官产生声音的过程可以用一个数学模型来逼近。人的语音发声过程是:气流从肺呼出后经

过声门时受声带作用，形成激励气流，再经过由口腔、鼻腔和嘴组成的声道的作用而发出语音。从声门出来的气流相当于激励信号，而声道可以等效成一个全极点滤波器，称为声道滤波器或合成滤波器。在讲话过程中，激励信号和滤波器系数不断地变化，从而发出不同的声音。通常认为激励信号和滤波器系数 5～40ms 更新一次。人们在发声母时，声带不振动，激励信号类似白噪声，这类声音称为清音；发韵母时，声带振动，激励信号呈周期性，这类声音称为浊音。因此，如图 3-12 所示，用白噪声或周期性脉冲信号激励声道滤波器就能合成语音，这就是 LPC 声码器的工作原理。

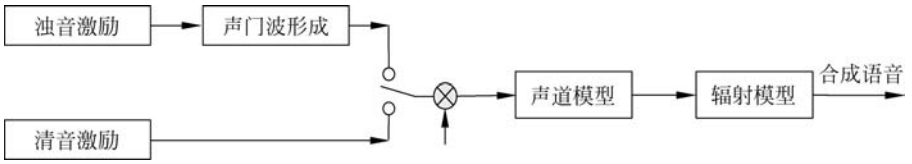

图 3-12　人类发音模型

这个模型的物理含义是：人类通过嘴讲出来的话，也可以用它来再生，条件是要合理地选择模型中的参数。人的讲话随着时间而变化，那么，模型的参数也是变化的。这个用模型参数代替原语音波形进行传输/存储的系统就是声码器。对该发声模型的参数进行编码传输，称为参数编码。人的发声是很复杂的，上面的模型只是一种近似，忽略了不少因素，这个模型也叫简化发声模型，它合成出的语音质量不高，后来又有许多改进。

2. 线性预测编码

线性预测编码(LPC)是一种非常重要的编码方法，线性预测方法分析和模拟人的发音器官，不是利用人发出声音的波形合成，而是从人的语音信号中提取与语音模型有关的特征参数。在语音合成过程中，通过相应的数学模型计算去控制相应的参数合成语音，这种方法对语音信息的压缩是很有效的，用此方法压缩的语音数据所占用的存储空间只有波形编码的十分之一甚至几十分之一。LPC 声码器是一种低比特率和传输有限个语音参数的语音编码器，它较好地解决了传输数码率与所得到的语音质量之间的矛盾，广泛地应用在电话通信、语音通信自动装置、语音学及医学研究、机械操作、自动翻译、身份鉴别、盲人阅读等方面，其原理如图 3-13 所示。

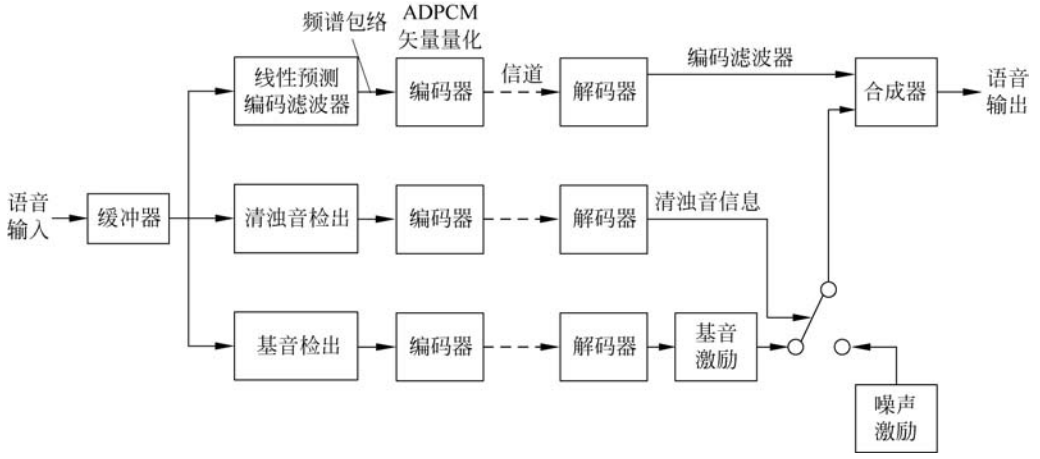

图 3-13　LPC 声码器的原理图

LPC 声码器在众多的声码器中是最为成功的，也是应用最为广泛的。它属于时域声码器，这类声码器从时间波形中提取重要的语音特征。

3.2.3　混合编码

混合编码是波形编码和参数编码的综合，既利用了语音生成模型，通过模型中的参数（主要是声道参数）进行编码，减少波形编码中被编码对象的动态范围或数目，又使编码的过程产生接近原始语音波形的合成语音，保留说话人的各种自然特征，提高了合成语音质量。目前得到广泛研究和应用的码激励线性预测编码法，以及它的各种改进算法，是混合编码的典型代表。

简单声码器由于激励形式过于简单，与实际差别较大，导致系统合成出的语音质量不好。是否可以对经过语音合成系统的逆系统——预测滤波器产生的预测误差信号，直接逼近产生新的激励形成，这样，问题的解决就容易得多了。但理论和实验表明，这样做并不能产生高质量的合成语音。因为人耳听见的只是合成语音，不是激励，即使新的激励与原来的预测误差信号很像，经过合成系统后，合成的语音与原来的语音仍有相当大的距离，因为激励部分的误差可能被合成滤波放大。解决这个问题的唯一办法，只能是改变激励信号的选择原则，使得最优激励信号的产生，不是去追求与预测误差信号接近，而是使它激励合成系统的输出，即合成语音尽可能接近原始语音。这样的编码方式叫作分析/合成（A/S）编码，即编码系统大都是先“分析”输入语音提取发声模型中的声道模型参数，然后选择激励信号去激励声道模型产生“合成”语音，通过比较合成语音与原始语音的差别选择最佳激励，追求最逼近原始语音的效果。所以，编码的过程是一个分析加合成的过程，原理如图 3-14 所示。

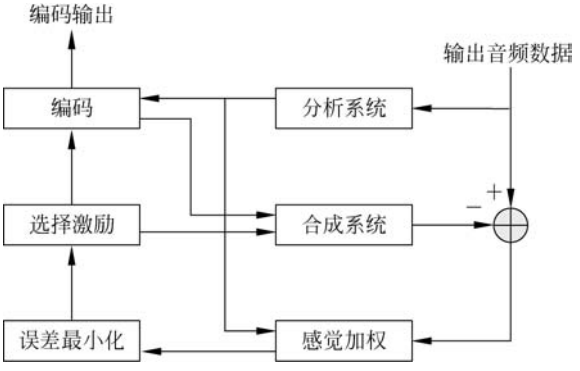

图 3-14　A/S 编码原理框图

1. 多脉冲线性预测编码

语音模型中的激励信号，可以通过分析 A/S 编码系统产生的预测误差获得。这个预测误差序列可由大约只占其个数十分之一的另一组脉冲序列来替代，由新脉冲序列激励 $H(z)$ 产生的合成语音仍具有较好的听觉质量。这个预测误差序列，尽管在大多数位置上都不等于零，但它激励合成滤波器所得的合成语音，与另一组绝大多数位置上都是零的脉冲序列激励同样的合成滤波器所得的合成语音具有类似的听觉。由于后者形成的激励信号序列中不为零的脉冲个数占序列总长的极小部分，所以编码时，仅处理和传输不为零的激励脉冲的位置与幅度参数，就可以大大压缩码率了。这种编码方法称为多脉冲线性预测编码

(Multi-Pulse Linear Predictive Coding,MPLPC)。

MPLPC 编码原理如图 3-15 所示,其主要任务就是寻找该脉冲序列中每个脉冲的位置和幅度大小,并对其编码。一般采用序贯方法,一个一个脉冲求解,寻求次优解。

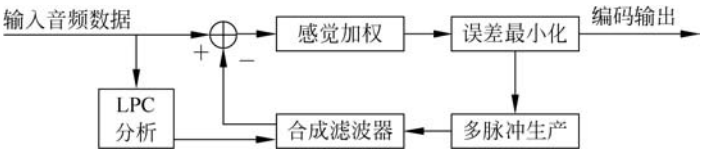

图 3-15 MPLPC 编码原理

2. 规则脉冲激励/长项预测编码

规则脉冲激励/长项预测(Regular Pulse Excitation/Long Term Prediction,RPE/LTP)编码是全球移动通信系统(Global System of Mobile Communications,GSM)标准中采用的语音压缩编码算法,其标准码率为 13kb/s,也叫作移动通信的全速率编码标准。GSM 语音压缩编解码器中的语音生成模型如图 3-16 所示。人们为进一步提高信道利用率,正在制定码率为 6～7kb/s,与 RPE/LTP 方案相当的语音压缩编码标准。新方案称为移动通信中的半速率语音编码算法。

图 3-16 GSM 语音压缩编解码器中的语音生成模型

RPE/LTP 语音压缩编码属于 A/S 编码方式,系统先分析,得到合成滤波器参数,再通过选择不同激励,判别它们的合成语音与原始语音的差别,得到最优的激励信号。RPE/LTP 采用了感觉加权滤波器。PRE/LTP 的各个非零激励脉冲,呈现等间隔的规则排列,只要使收方知道第一个脉冲的位置在何处(n 取什么值),其他激励脉冲的位置也就可以得知了,而且第一个脉冲的位置也是有限的几个可能性。所以,这种方案脉冲位置的编码所需码率非常低,非零激励脉冲个数可以增加许多。在一个编码帧内,GSM 方案的非零激励脉冲比 MPLPC 方案多了 3 倍,有利于提高合成语音质量。RPE/LTP 编码算法设置了基音预测系统以及相应的基音合成系统,线性预测处理语音信号可以去除语音信号样值间的相关性,大大降低信号的动态范围。

3. 码激励线性预测编码

码激励线性预测(Code Excited Linear Prediction,CELP)编码系统是中低速率编码领域最成功的方案。基本 CELP 算法不对预测误差序列个数及位置做任何强制假设,认为必须用全部误差序列编码传送以获得高质量的合成语音。为了达到降低传码率的目的,对误差序列的编码采用了大压缩比的矢量量化技术 VQ,也就是对误差序列不是一个一个样值分别量化,而是将一段误差序列当成一个矢量进行整体量化。由于误差序列对应语音生成模型的激励部分,现在经 VQ 量化后,用码字代替,故称为码激励。典型的 CELP 系统如图 3-17 所示。

4. 矢量和激励线性预测编码

矢量和激励线性预测(Vector Sum Excited Linear Prediction,VSELP)编码作为北美第

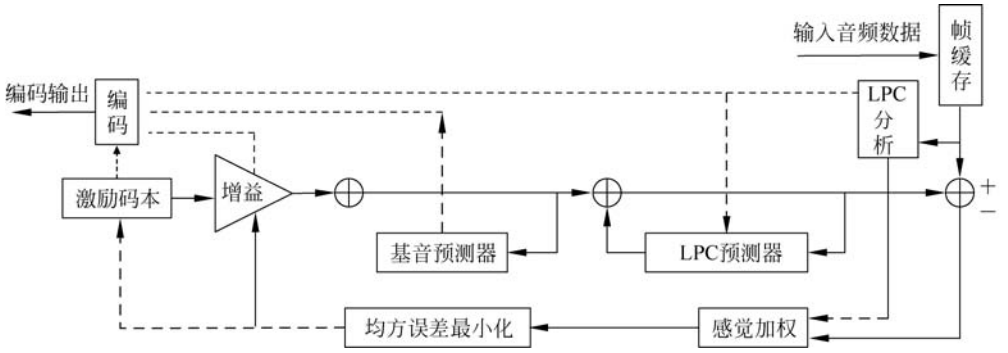

图 3-17　典型的 CELP 系统

一代数字蜂窝移动通信网语音编码标准，由摩托罗拉公司首先提出，其码率为 8kb/s。VSELP 编码系统结构如图 3-18 所示。

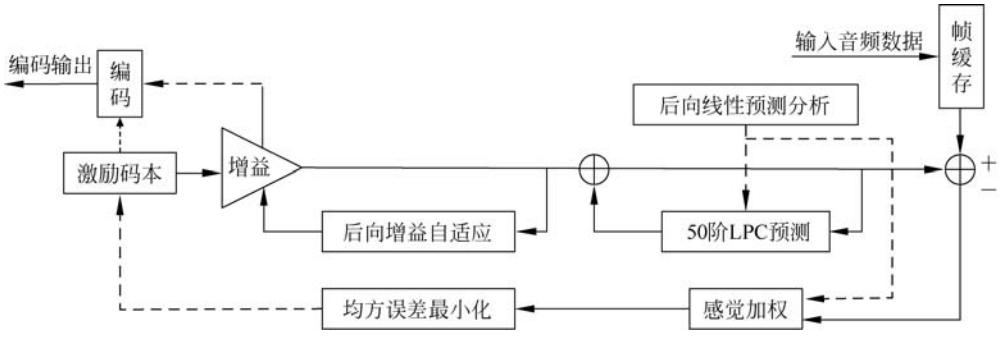

图 3-18　VSELP 编码系统结构

5. 多带激励语音编码

语音短时谱分析表明，大多数语音段都含有周期和非周期两种成分，因此很难说某段语音是清音还是浊音。传统声码器，如线性预测声码器，采用二元模型，认为语音段不是浊音就是清音。浊音段采用周期信号，清音采用白噪声激励声道滤波器合成语音，这种语音生成模型不符合实际语音特点。人耳听觉过程是对语音信号进行短时谱分析的过程，可以认为人耳能够分辨短时谱中的噪声区和周期区。因此，传统声码器合成的语音听起来合成声重，自然度差。这类声码器还有其他一些弱点，如基音周期参数提取不准确、语音发声模型同有些音不符合、容忍环境噪声能力差等，这些都是影响合成语音质量的因素。

多带激励（Multi-Band Excitation，MBE）语音编码方案突破了传统线性预测声码器整带二元激励模型，它将语音谱按基音谐波频率分成若干个带，对各带信号分别判断是属于浊音还是清音，然后根据各带清、浊音的情况，分别采用白噪声或正弦产生合成信号，最后将各带信号相加，形成全带合成语音。多带激励编解码器原理如图 3-19 所示。

6. 混合激励线性预测编码

混合激励线性预测（Mixed Excited Linear Prediction，MELP）编码算法对语音的模式进行两级分类。首先，将语音分为“清”和“浊”两大类，这里的清音是指不具有周期成分的强清音，其余的均划为浊音，用总的清/浊音判决表示。其次，把浊音再分为浊音和抖动浊音，用非周期位表示。在对浊音和抖动浊音的处理上，MELP 算法利用了 MBE 算法的分带思

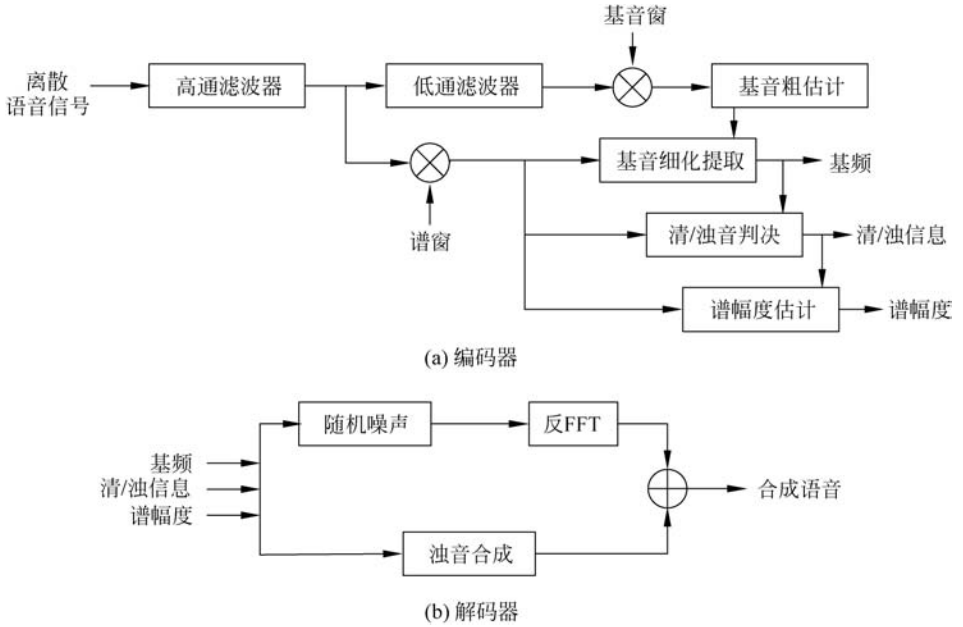

图 3-19　多带激励编解码器原理

想，在各子带上对混合比例进行控制。这种方法简单有效，使用的比特数也不多。如果使用 1b 对每个子带的混合比例参数进行编码，该参数也就简化为每个子带的清/浊音判决信息。另外，在周期脉冲信号源的合成上，MELP 算法要对 LPC 分析的残差信号进行傅里叶变换，提取谐波分量，量化后传到接收端，用于合成周期脉冲激励。这种方法提高了激励信号与原始残差的匹配程度。

MELP 的参数包括 LPC 参数、基音周期、模式分类参数、分带混合比例、残差谐波参数和增益。如图 3-20 所示，在 MELP 的参数分析部分，语音信号输入后要分别进行基音提取、子带分析、LPC 分析和残差谐波谱计算。MELP 算法的语音合成部分仍然采取 LPC 合成的形式，不同的是激励信号的合成方式和后处理。这里的混合激励信号为合成分带滤波

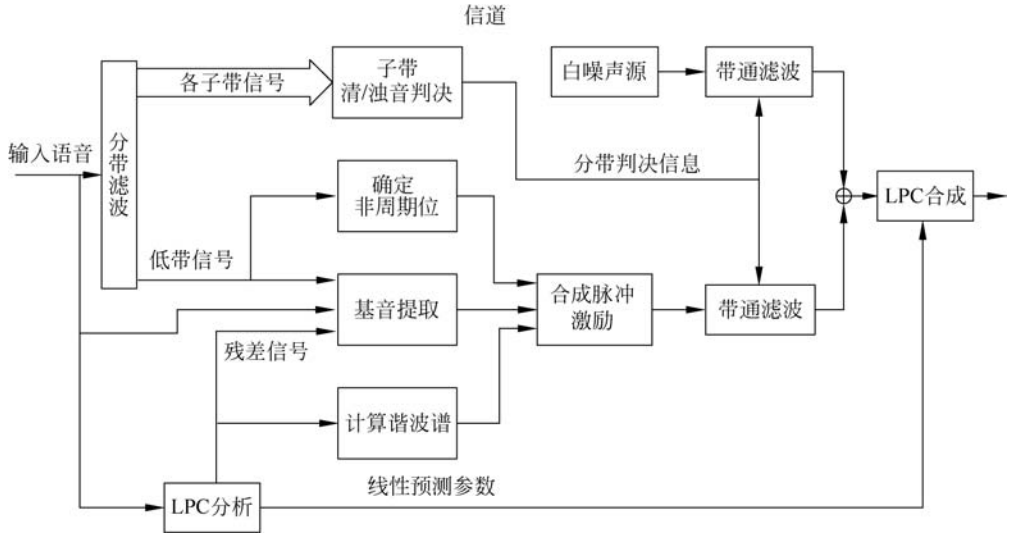

图 3-20　MELP 算法的分析/合成框图

后的脉冲与噪声激励之和。脉冲激励通过对残差谐波谱进行离散傅里叶反变换得出，噪声激励则在对一个白噪声源进行电平调整和限幅之后产生，两者各自滤波后叠加在一起形成混合激励。混合激励信号合成后经自适应谱增强滤波器处理，用于改善共振峰的形状。随后，激励信号进行 LPC 合成得到合成语音。

3.3 音频压缩编码标准

当前国际上数字音视频标准有两个系列，一个是声音信源编码中的运动图像专家组(Moving Picture Exports Group，MPEG)制定的音频编码，简称 MPEG 音频；另一个是先进电视系统委员会(Advanced Television System Committee，ATSC)的杜比 AC-3 音频编码。其中 MPEG 音频的应用涉及领域广泛，不仅用于数字电视、数字声广播，还有影音光盘、多媒体应用以及网络服务等，因此是主流。而杜比 AC-3 则仅用于多声道环绕立体声重放，包括 DVD 影音光盘及 ATSC 数字电视标准中的音频编码。在音频压缩标准化方面取得巨大成功的是 MPEG 数字音频压缩方案。

3.3.1 MPEG 数字音频压缩标准

MPEG 是一组由国际电工委员会(International Electrotechnical Commission，IEC)和国际标准化组织(International Organization for Standardization，ISO)制定发布的视频、音频、数据的压缩标准。MPEG 的声音数据压缩编码不是依据波形本身的相关性和模拟人的发音器官的特性，而是利用人的听觉系统的特性来达到压缩声音数据的目的，这种压缩编码属于感知编码，现已发展成为数字音视频的主流技术。

目前，MPEG 已经完成了 MPEG-1、MPEG-2、MPEG-4 第一版的音频编码等方面的技术标准，正在制定 MPEG-4 第二版、MPEG-7 及 MPEG-21 的音频编码技术标准。

1. MPEG-1 音频压缩编码标准

在音频压缩标准化方面取得巨大成功的是 MPEG-1 音频(ISO/IEC 11172-3)，它是 MPEG-1(ISO/IEC 11172)标准的第三部分。在 MPEG-1 音频中，对音频压缩规定了 3 个层次：层Ⅰ、层Ⅱ(即 MUSICAM，又称为 MP2)和层Ⅲ(又称为 MP3)。由于在制定标准时对许多压缩技术进行了认真的考查，并充分考虑了实际应用条件和算法的可实现性，因而这 3 种模式都得到了广泛的应用。

MPEG-1 音频压缩的基础是量化。虽然量化会带来失真，但是 MPEG-1 音频标准要求量化失真对于人耳来说是感觉不到的。在 MPEG-1 音频压缩标准的制定过程中，MPEG 音频委员会做了大量的主观测试实验。实验表明，采样频率为 48kHz，采样值精度为 16b 的立体声声音数据压缩到 256kb/s 时，即在 6∶1 的压缩比下，即使是专业测试员也很难分辨出是原始音频还是编码压缩后复原出的音频信号。

MPEG-1 音频使用感知音频编码来达到既压缩音频数据又尽可能保证音质的目的。感知音频编码的理论依据是听觉系统的掩蔽效应，其基本思想是在编码过程中保留有用的信息而丢掉被掩蔽的信号，其结果是经编解码之后重构的音频信号与编码之前的原始音频信号不完全相同，但人的听觉系统很难感觉到它们之间的差别。也就是说，对于听觉系统，这种压缩是无损压缩。

MPEG-1 音频编码标准提供 3 个独立的压缩层次，它们的基本模型相同。层Ⅰ是最基础的，层Ⅱ和层Ⅲ都是在层Ⅰ的基础上有所提高。每个后继的层次都有更高的压缩比，同时也需要更复杂的编码器。任何一个 MPEG-1 音频码流帧结构的同步头中都有一个 2b 的层代码字段(Layer Field)，用来指出所用的算法是哪一个层次。MPEG-1 音频码流按照规定构成“帧”(Frame)的格式，层Ⅰ的每帧包含 384 个采样值的码字，384 个采样值来自 32 个子带，每个子带包含 12 个采样值。层Ⅱ和层Ⅲ的每帧包含 1152 个采样值的码字，每个子带包含 36 个采样值。MPEG-1 层Ⅰ的压缩编码原理如图 3-21 所示。

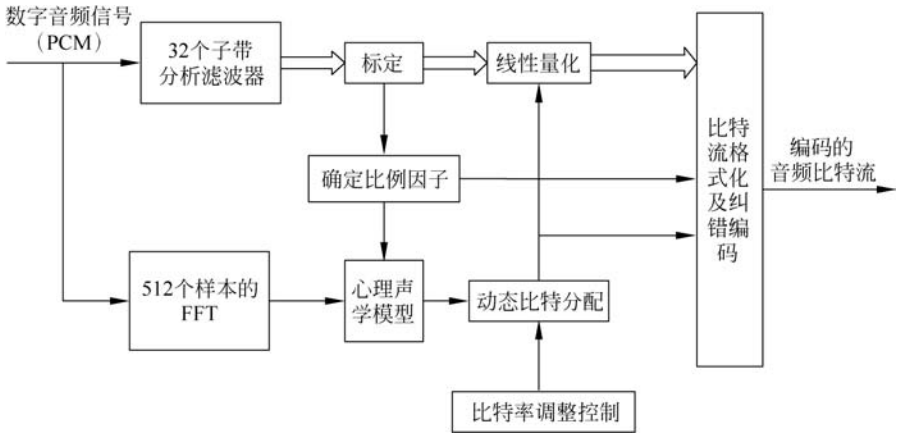

图 3-21　MPEG-1 层Ⅰ的压缩编码原理

1) MPEG-1 层Ⅰ

MPEG-1 层Ⅰ主要技术如下。

(1) 子带分析滤波器组(多相滤波器组)

编码器的输入信号是每声道 768kb/s 的数字化音频 PCM 信号，用多相滤波器组分割成 32 个子带信号。

多相滤波器组是正交镜像滤波器(Quandrature Mirror Filter，QMF)的一种，与一般树形构造的 QMF 相比，它可用较少的运算实现多个子带的分割。

层Ⅰ的子带是均匀划分的，它把信号分到 32 个等带宽的频率子带中。值得注意的是，子带的等带宽划分并没有精确地反映人耳的听觉特性。在低频区域，一个子带覆盖若干个临界频带，某个子带中量化器的比特分配以其中最低的掩蔽阈值为准。这样对低频的量化比较简单，容易引起低频端的量化误差。

(2) 标定

如果将子带信号直接原样量化，则量化噪声电平由量化步长决定，当输入信号电平低时，噪声就会显现出来。考虑到人耳听觉的时域掩蔽效应，将每个子带中连续的 12 个采样值归并成一个块，在采样频率为 48kHz 时，这个块相当于 $12\times32/48=8$ms。这样，在每一子带中，以 8ms 为一个时间段，对 12 个采样值并成的块一起计算，求出其中幅度最大的值，对该子带的采样值进行归一化，即标定，使各子带电平一致，然后进行适当的量化。

比例因子的作用是充分利用量化器的动态范围，通过比特分配和比例因子相配合，可以相对降低量化噪声电平。

(3) 快速傅里叶变换

数学角度上,信号由时域变换到频域表示的过程称为傅里叶变换。

快速傅里叶变换(Fast Fourier Transform,FFT)是计算离散傅里叶变换的一种快速算法,为了在频域精确地计算信号掩蔽比(Signal Mask Ratio,SMR),以及掩蔽音与被掩蔽音对应的频率范围和功率峰值,输入的 PCM 信号同时还要送入 FFT 运算器。这样,既可以通过多相滤波器组使信号具有高的时间分辨率,又可以使信号通过 FFT 运算具有高的频率分辨率。

足够高的频率分辨率可以实现尽可能低的码率,而足够高的时间分辨率可以确保在短暂冲击声音信号情况下,编码的声音信号也有足够高的质量。

(4) 心理声学模型

心理声学模型是模拟人类听觉掩蔽效应的一个数学模型,它根据 FFT 的输出值,按一定的步骤和算法计算出每个子带的信号掩蔽比。基于所有这些子带的信号掩蔽比进行 32 个子带的比特分配。

根据心理声学模型计算得到以频率为自变量的噪声掩蔽阈值。一个给定信号的掩蔽能力取决于它的频率和响度。MPEG-1 音频压缩标准提供了两个心理声学模型,其中模型 1 比模型 2 简单,以简化计算。两个模型对所有层次都适用,只有模型 2 在用于层Ⅲ时要加以修改。另外,心理声学模型的实现有很大的自由度。

(5) 动态比特分配

为了同时满足码率和掩蔽特性的要求,比特分配器应同时考虑来自分析滤波器组的输出样值以及来自心理声学模型的信号掩蔽比,以便决定分配给各个子带信号的量化比特数,使量化噪声低于掩蔽阈值。

由于掩蔽效应的存在,降低了对量化比特数的要求,不同的子带信号可分配不同的量化比特数。对于各个子带信号而言,量化都是线性量化。

(6) 帧结构

将量化后的采样值和格式标记以及其他附加辅助数据按照规定的帧格式组装成比特数据流,如图 3-22 所示。

图 3-22　MPEG-1 层Ⅰ的音频码流的数据帧格式

2) MPEG-1 层Ⅱ

MPEG-1 层Ⅱ采用了 MUSICAM 编码方法,其压缩编码器原理如图 3-23 所示。

层Ⅱ和层Ⅰ的不同之处如下。

层Ⅱ使用 1024 点的 FFT 运算,提高了频率的分辨率,得到原信号的更准确瞬时频谱特性。

层Ⅱ的每帧包含 1152 个采样值的码字。与层Ⅰ对每个子带由 12 个采样值组成一块的

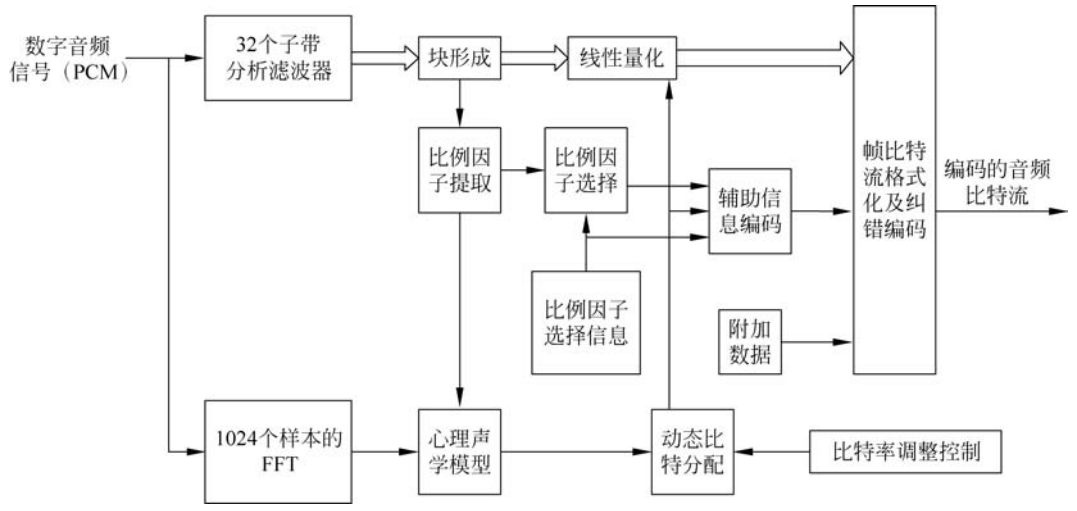

图 3-23　MPEG-1 层Ⅱ的压缩编码器原理

编码不同，层Ⅱ对一个子带的 3 个块进行编码，其中每块 12 个采样值。

描述比特分配的字段长度随子带的不同而不同。低频段子带用 4b 来描述，中频段子带用 3b 来描述，高频段子带用 2b 来描述。这种因频率不同而比特率不一样的做法，也是临界频带的应用，如图 3-24 所示。

图 3-24　MPEG-1 层Ⅱ的音频码流的数据帧格式

编码器可对一个子带提供 3 个不同的比例因子，所以，每个子带每帧应传送 3 个比例因子。比例因子是人们对音频信号统计分析和观察得出的特征规律的反映，在较高频率时频谱能量出现明显的衰减，因此，比例因子从低频子带到高频子带出现连续下降。比例因子的附加编码措施就是考虑到上述的统计联系和听觉的时域掩蔽效应，将一帧内的 3 个连续的比例因子按照不同的组合共同编码和传送。采用附加编码措施后，用于传送比例因子所需的码率平均可压缩约 1/3。

3）MPEG-1 层Ⅲ

层Ⅲ综合了 ASPEC 和 MUSICAM 算法的特点，比层Ⅰ和层Ⅱ都要复杂。层Ⅲ压缩编码器的原理如图 3-25 所示。MP3 采用 44.1kHz 的采样频率，压缩比能够达到 10∶1～12∶1，基本上拥有近似 CD 的音质。

层Ⅲ使用比较好的临界频带滤波器，把输入信号的频带划分成不等带宽的子带。按临界频带划分子带时，低频段取的带宽窄，即意味着对低频有较高的频率分辨率；在高频段时则有较低一点的分辨率。这更符合人耳的灵敏度特性，可以改善对低频段压缩编码的失真，但这样做需要较复杂的滤波器组。虽然层Ⅲ所用的滤波器组与层Ⅰ和层Ⅱ所用的滤波器组

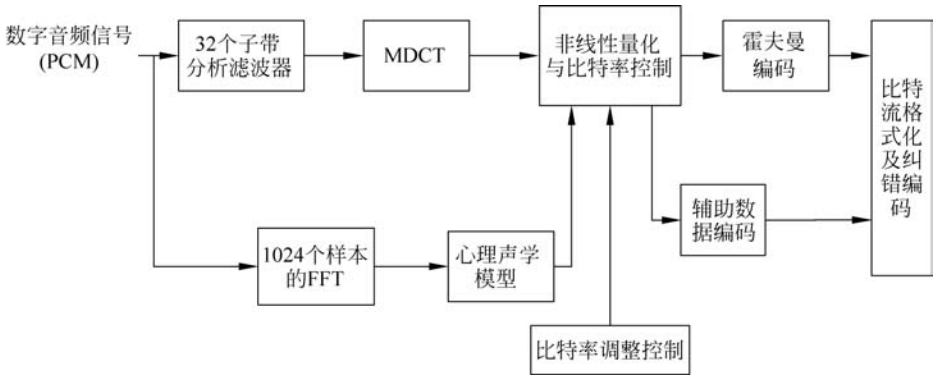

图 3-25　MPEG-1 层Ⅲ压缩编码器的原理

结构相同，但是层Ⅲ重点使用了改进离散余弦变换（Modified Discrete Consine Transform，MDCT），对层Ⅰ和层Ⅱ的滤波器组的不足做了一些补偿。MDCT 把子带的输出在频域里进一步细分，以达到更高的频域分辨率。通过对子带的进一步细分，层Ⅲ编码器已经部分消除了多相滤波器组引入的混叠效应。

层Ⅲ指定了两种 MDCT 的块长：长块块长为 18 个采样值，短块块长为 6 个采样值。相邻变换窗口之间有 50％的重叠，所以窗口大小分别为 36 和 12 个采样值。长块对于平稳的音频信号可以得到更高的频域分辨率，短块对瞬变的音频信号可以得到更高的时域分辨率。在短块模式下，3 个短块代替一个长块，而短块的块长恰好是一个长块的 1/3，所以 MDCT 的采样值不受块长的影响。对于给定的一帧音频信号，MDCT 可以全部使用长块或全部使用短块，也可以长、短块混合使用，这样既能保证低频段的频域分辨率，又不会牺牲高频段的时域分辨率。长块和短块之间的切换有一个过程，一般用一个带特殊长转短或短转长数据窗口的长块来完成长、短块之间的切换。

层Ⅲ采用霍夫曼编码进行无损压缩，进一步降低数码率，提高压缩比。据估计，霍夫曼编码以后，可以节省 20％的码率。

虽然层Ⅲ引入了许多复杂的概念，但是它的计算量并没有比层Ⅱ增加很多，增加的主要是编码器和解码器的复杂度，以及解码器所需要的存储容量。

4）MPEG-1 音频压缩编码的基本结构

MPEG-1 音频信号数据压缩过程可分为如下 4 步。编码和解码流程如图 3-26 和图 3-27 所示。

（1）时间/频率映射（滤波器组），用以将输入的信号转化为亚抽样的频谱分量分为子带。

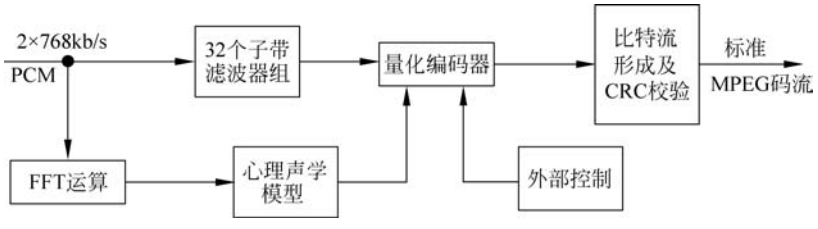

图 3-26　MPEG-1 的音频压缩编码流程

图 3-27　MPEG-1 的音频压缩解码流程

(2) 频域滤波器组或并行变换的输出，根据心理声学模型求出时变的掩蔽门限估值。

(3) 按量化噪声不超过掩蔽门限的原则将子带量化编码，从而听不到量化噪声。

(4) 按帧打包成码流(包括比特分配信息)。

2. MPEG-2 音频压缩编码标准

MPEG-2(ISO/IEC 13818)标准公布于 1995 年，是 MPEG-1 的一种兼容型扩展。MPEG-2 声音编码标准是 MPEG 为多声道声音开发的低码率编码方案，它是在 MPEG-1 标准的基础上发展而来的。

与 MPEG-1 相比，MPEG-2 声音主要增加了以下 3 个方面的内容。

(1) 支持 5.1 路环绕声。它能提供 5 个全带宽声道(左、右、中和两个环绕声道)，外加一个低频效果增强声道，统称为 5.1 声道。

(2) 支持多达 8 种语言或解说。

(3) 增加了低采样和低码率。在保持 MPEG-1 声音的单声道和立体声的原有采样率的情况下，MPEG-2 又增加了 3 种采样率，即把 MPEG-1 的采样率降低了一半(16kHz，22.05kHz，24kHz)，以便提高码率低于 64kb/s 时每个声道的声音质量。

MPEG-2 标准委员会定义了两种音频压缩编码算法：MPEG-2 后向兼容多声道音频编码(MPEG-2 Backward Compatible Multichannel Audio Coding)标准，简称为 MPEG-2 BC，与 MPEG-1 音频压缩编码算法是兼容的；MPEG-2 高级音频编码(MPEG-2 Advanced Audio Coding)标准，简称为 MPEG-2 AAC，与 MPEG-1 音频压缩编码算法是不兼容的，也称为 MPEG-2 非后向兼容(Non Backward Compatible，NBC)标准。

1) MPEG-2 BC

MPEG-2 BC(ISO/IEC 13818-3)主要是在 MPEG-1 音频和 CCIR775 建议的基础上发展起来的。标准规定的码流形式还可与 MPEG-1 音频的层Ⅰ和层Ⅱ做到前、后向兼容，并可依据 CCITT775 建议做到与双声道、单声道形式的向下兼容，还能够与杜比环绕声形式兼容。

正是与 MPEG-1 的前、后向兼容性成为 MPEG-2 BC 最大的弱点，使得 MPEG-2 BC 不得不以牺牲码率的代价换取较好的声音质量。

MPEG-2 BC 适用于数据比特率从 8kb/s 的单声道电话的音质到 160kb/s 的多声道高质量的全音域音频编码，也适用于 DVD，图像清晰度可达到 500 线，可提供两路立体声声道和高质量的 5.1 声道环绕立体声。

2) MPEG-2 AAC

MPEG-2 AAC(ISO/IEC 13818-7)是一种非常灵活的声音感知编码标准，主要使用听觉系统的掩蔽特性压缩声音的数据量，并且通过把量化噪声分散到各个子带中，用全局信号把噪声掩蔽掉。

MPEG-2 AAC 支持的采样频率为 8～96kHz，编码器的音源可以是单声道、立体声和多声道的声音，多声道扬声器的数目、位置及前方、侧面和后方的声道数都可以设定，因此能支

持更灵活的多声道构成。

MPEG-2 AAC 可支持 48 个主声道、16 个低频增强(Low Frequency Enhancement，LFE)声道、16 个配音声道(或称为多语言声道(Multilingual Channel)和 16 个数据流)。

MPEG-2 AAC 的压缩比为 11∶1，在每个声道的码率为 64kb/s，5 个声道的总码率为 320kb/s 的情况下，很难区分解码还原后的声音与原始声音之间的差别。

与 MPEG-1 音频的层Ⅱ相比，MPEG-2 AAC 的压缩比可提高一倍，而且音质更好；在质量相同的条件下，MPEG-2 AAC 的码率大约是 MPEG-1 音频层Ⅲ的 70%。MPEG-2 AAC 编码器如图 3-28 所示。

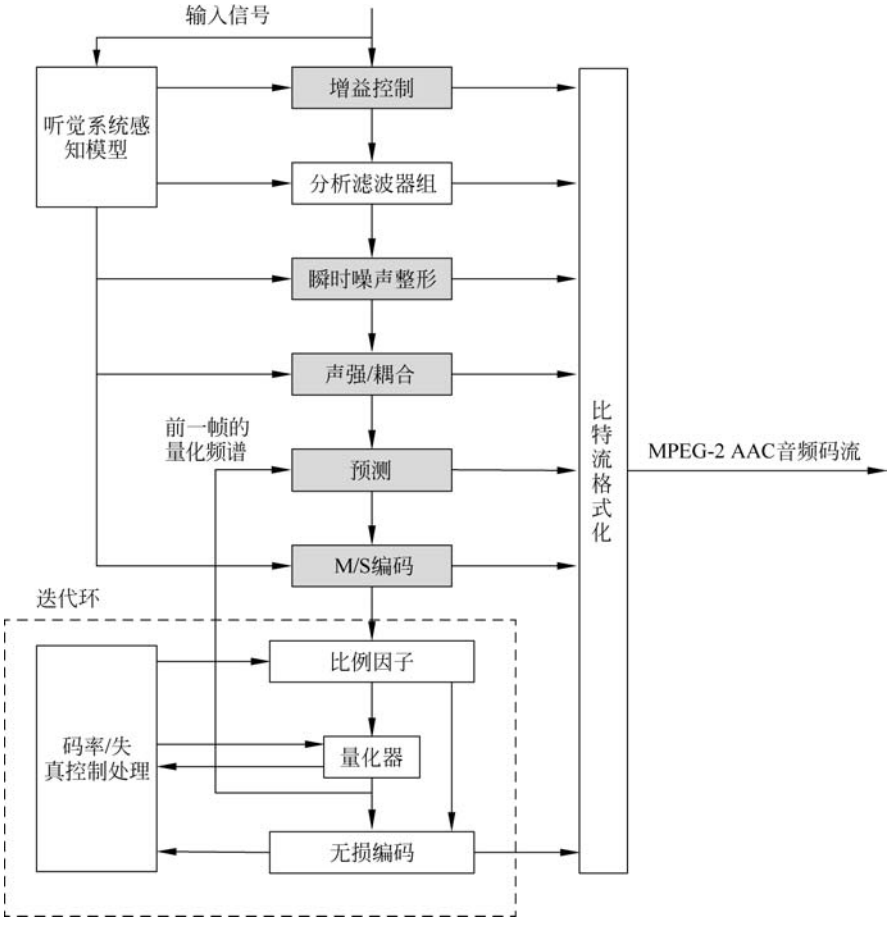

图 3-28　MPEG-2 AAC 编码器

(1) 增益控制

增益控制模块用在可分级采样率类中，它由多相正交滤波器(Polyphase Quadrature Filter，PQF)、增益检测器和增益调节器组成，把输入信号划分到 4 个等带宽的子带中。在解码器中也有增益控制模块，通过忽略多相正交滤波器的高子带信号获得低采样率输出信号

(2) 分析滤波器组

分析滤波器组把输入信号从时域变换到频域。该模块采用了 MDCT，是一种线性正交

交叠变换，使用时域混叠抵消（Time Domain Alias Cancellation，TDAC）技术，在理论上能完全消除混叠。

AAC 提供了两种窗函数：正弦窗和凯塞尔窗。正弦窗使滤波器组能较好地分离出相邻的频谱分量，适合具有密集谐波分量（频谱间隔小于 140Hz）的信号。频谱成分间隔较宽（大于 220Hz）时，凯塞尔窗 AAC 系统允许正弦窗和凯塞尔窗之间连续无缝切换。

（3）听觉系统感知模型

感知模型即心理声学模型，它是包括 AAC 在内的所有感知音频编码的核心。AAC 使用的心理声学模型原理上与 MP3 所使用的模型相同，但在具体计算和参数方面并不一样。AAC 用的模型不区分单音和非单音成分，而是把频谱数划分为“分区”，分区范围与临界频带带宽有线性关系。

（4）瞬时噪声整形

瞬时噪声整形（Temporal Noise Shaping，TNS）是用来控制量化噪声的瞬时形状的一种方法，解决掩蔽阈值和量化噪声的错误匹配问题，是增加预测增益的一种方法。其基本思想是：时域较平稳的信号，频域上变化较剧烈；反之，时域上变化剧烈的信号，频谱上则较平稳。TNS 是在信号的频谱变化较平稳时，对一帧信号的频谱进行线性预测，再将预测残差编码。在编码时判断是否要用 TNS 模块的依据是感知熵，当感知熵大于预定值时就用 TNS。

（5）声强/耦合编码

这是 AAC 编码器的可选项，又称为声强立体声（Intensity Stereo Coding）或声道耦合编码（Channel Coupling Coding），探索的基本问题是声道间的不相关性（Irrelevance）。人耳听觉系统在听 4kHz 以上的信号时，双耳的定位对左右声道的强度差比较敏感，而对相位差不敏感。声强/耦合就利用这一原理，在某个频带以上的各子带使用左声道代表两个声道的联合强度，右声道谱线置为 0，不再参与量化和编码。平均而言，大于 6kHz 的频段用声强/耦合编码较合适。

（6）预测

在信号较平稳的情况下，利用时域预测可进一步减小信号的冗余度。AAC 编码中的预测是利用前面帧的频谱预测当前帧的频谱，再求预测的残差，然后对残差进行编码。预测使用经过量化后重建的频谱信号。

（7）量化

量化模块按心理学模型输出的掩蔽阈值把限定的比特分配给输入谱线，要尽量使产生的量化噪声低于掩蔽阈值，达到不被听到的目的。量化时须计算实际编码所用的比特数，量化和编码是紧紧结合在一起的，AAC 在量化前将 1024 条谱线分成数十个比例因子频带，对每个子频带采用 3/4 次方非线性量化，起到幅度压扩作用，信号的信噪比和压缩信号的动态范围有利于霍夫曼编码。

（8）无损编码

无损编码实际上就是霍夫码编码，它对被量化的谱系数、比例因子和方向信息进行编码。

（9）码流打包组帧

AAC 的帧结构非常灵活，除支持单声道、双声道、5.1 声道外，可支持多达 48 个声道，

具有 16 种语言兼容能力。

3. MPEG-4

MPEG-4 以“各种音/视频媒体对象的编码”为标题，MPEG-4 第一版于 1998 年 12 月成为一项通用的国际标准（ISO/IEC 14496IV）；第二版于 1999 年 12 月完成；第三、四版于 2001 年开始制定。

MPEG-4 标准的目标是提供未来的交互式多媒体应用，它具有高度的灵活性和可扩展性。与以前的音频编码标准相比，MPEG-4 增加了许多新的关于合成内容及场景描述等领域的工作，增加了可分级性、音调变化、可编辑性和延迟等新功能。MPEG-4 将以前发展良好但相互独立的高质量音频编码、计算机音乐及合成语音等第一次合并在一起，在诸多领域内给予高度的灵活性。

为了实现基于内容的编码，MPEG-4 音频编码也引入了音频对象（Audio Object，AO）的概念。AO 可以是混合声音中的任一种基本音，如交响乐中某一种乐器的演奏音或电影声音中人物的对白。通过对不同 AO 的混合和去除，用户就能得到所需要的某种基本音或混合音。

MPEG-4 支持自然声音（如语音、音乐）、合成声音以及自然和合成声音混合在一起的合成/自然混合编码（Synthetic/Natural Hybrid Coding，SNHC），以算法和工具形式对音频对象进行压缩和控制，如以分级码率进行回放、通过文字和乐器的描述合成语音和音乐等。

如图 3-29 所示，为获取到所有比特率下的高音质，MPEG-4 音频定义了如下 3 类编码模式。

（1）低比特率的参数编/解码器。采样频率在 8kHz 时数据比特率为 2～4kb/s；采样频率为 8/16kHz 时为 4～16kb/s。

（2）中间比特率的码激励线性预测（CELP）编/解码器。采样频率为 8/16kHz 时，数据比特率为 6～24kb/s。

（3）高比特率的编/解码器，包含 MPEG-2 AAC 和矢量量化编码在内的时/频编/解码器。采样频率大于 8kHz，数据比特率为 16～64kb/s，采用 AAC。

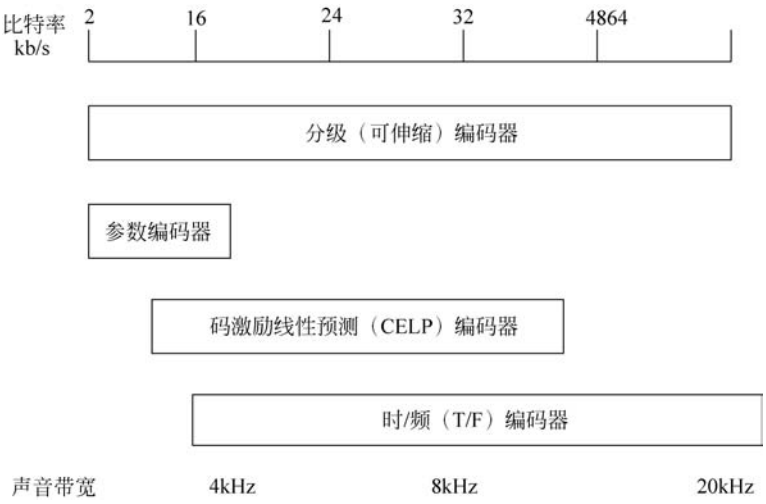

图 3-29　MPEG-4 声音编码及其码率带宽（2～64kb/s）关系图

4. MPEG-7

在信息社会中，可以利用的视听信息形式越来越多，如图像、视频、语音、3D 模型及图形等。而手段不仅是记录—存储—重放，尤其是随着网络的出现，特别是互联网多媒体服务、各项服务项目种类和大容量数据库等基于内容服务需求的快速增长，引发了对视听信息内容的检索、交换及传递的迫切要求。

MPEG-7 标准于 1996 年开始制定。MPEG-7 称为“多媒体内容描述接口”，主要描述多媒体素材内容的通用接口的标准化。MPEG-7 本质上与 MPEG-1、MPEG-2 及 MPEG-4 不同，后三者是论述音视频具体的编码，而 MPEG-7 是促进数据元的互操作性、通用性和数据管理灵活性。因此，MPEG-7 的目标是产生一个描述多媒体内容的标准，支持对多媒体信息在不同层面的解释和了解，从而将其依据用户需求进行传递和存取。

为了使人们在因特网上能够很快地搜索到所需要的内容，MPEG-7 多媒体接口应能支持如下功能。

(1) MPEG-7 可完成人耳听觉感知需要的内容，如频率轮廓线、音色、和声、频率特征(音调、音域)、振幅包络、时间结构等，即声音特性(音头持续时间及音尾)和文本内容。例如，通过唱一首歌曲的开始歌词或发出一篇文章开始一段的文字声音或声音近似值，即唱出歌曲的旋律或发出一种声音效果，就可以搜索到相应的全部原型声音或文本。

(2) 支持数据音频(如 CD 唱片、MPEG-1 音频格式)、模型音频(如磁带介质、MPEG-4 的 SAOL)及 MIDI(包括一般 MIDI 及 Karaoke 格式)。

3.3.2 杜比 AC-3 音频压缩算法

美国杜比(Dolby)实验室开发的杜比 AC-3 数字音频压缩编码技术与高清电视的研究紧密相关。杜比 AC-3 环绕声系统共有 6 个完全独立的声音声道：3 个前方的左声道、右声道和中置声道以及两个后方的左、右环绕声道，这 5 个声道皆为全频带的(20Hz～20kHz)；另有一个超低音声道，其频率范围只有 20～120Hz，所以将此超低音声道称为“0.1 声道”，构成杜比数字(AC-3)的 5.1 声道。杜比 AC-3 环绕声播放系统如图 3-30 所示。

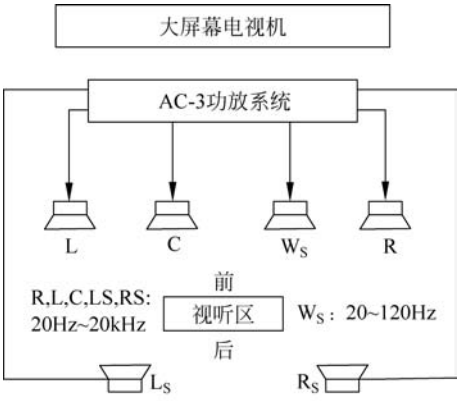

图 3-30 杜比 AC-3 环绕声播放系统

杜比 AC-3 可以把 5 个独立的全频带和一个超低音声道的信号实行统一编码，成为单一的复合数据流。各声道间的隔离度高达 90dB，两个环绕声道互相独立实现了立体化，超低音声道的音量可独立控制，可支持 32kHz、44.1kHz 和 48kHz 3 种采样频率。码率可低至单声道的 32kb/s，高到多声道的 640kb/s，以适应不同需要。

AC-3 是在 AC-1 和 AC-2 基础上发展起来的多声道编码技术，因此保留了 AC-2 的许多特点，如加窗处理、变换编码、自适应比特分配。AC-3 还利用了多声道立体声信号间的大量冗余性，对它们进行“联合编码”，从而获得了很高的编码效率。AC-3 编码器接收 PCM 声音数据，输出的是压缩后的码流。

杜比 AC-3 编码器原理如图 3-31 所示。

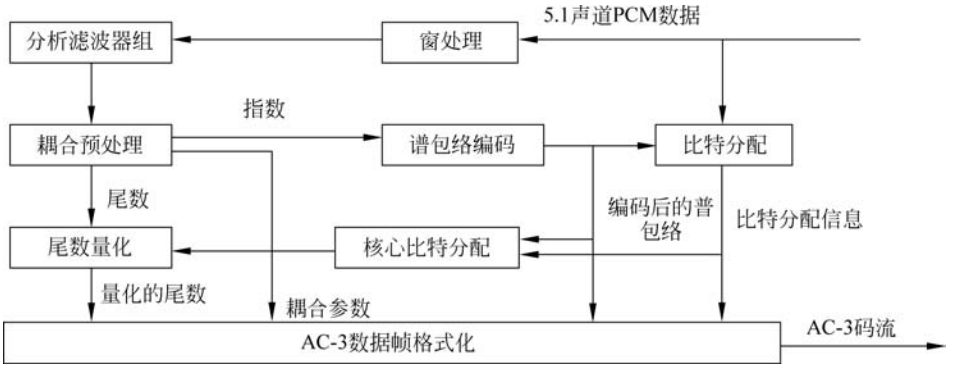

图 3-31　杜比 AC-3 编码器原理

1. 分析滤波器组

分析滤波器组的作用是把时域内的 PCM 采样值数据变换成频域内的一系列变换系数。在变换之前，要先将音频的采样值数据分成许多“块”。每块包含 512 个采样值，因为是重叠采样，所以每块有 256 个采样值是新的，另外 256 个采样值与上一块相同。每个音频的采样值会出现在两个块中，可以防止听得见的块效应。对音频采样值的分块是靠窗函数实现的。窗函数的形状决定了滤波器组中各滤波器的形状。

由时域变换到频域的块长度的选择是变换编码的基础，长的变换长度适合谱变化很慢的输入信号。长的变换长度提供比较高的频率分辨率，由此改进了这类信号的编码性能；短的变换长度提供比较高的时间分辨率，更适合时间上快速变化的信号。

人耳在时域和频域上存在掩蔽效应，因此，在进行变换编码时存在时间分辨率和频率分辨率之间的矛盾，不能同时兼顾，必须统筹考虑处理。对于稳态信号，其频率随时间变化缓慢，要求滤波器组有好的频率分辨率，即要求一个长的窗函数。反之，对于快速变化的信号，要求好的时间分辨率，即要求一个短的窗函数。

在编码器中，输入信号通过一个 8kHz 的高通滤波器取出高频成分，将它的能量和预先设定的阈值相比较，从而判断输入的信号是稳态还是瞬态。对于稳态信号，采样率为 48kHz 时的块长度选为 512，滤波器的频率分辨率为 187.5Hz，时间分辨率为 5.33ms；对于瞬态信号，块长度为 256，滤波器的频率分辨率为 375Hz，时间分辨率为 2.67ms。

通常的块长度为 512。对通常的加窗块做变换时，由于前后块重叠，变换后的变换系数可以去掉一半而变成每块包含 256 个变换系数。

比较短的块的构成是将 512 个采样值加窗的音频段分成两段，每段 256 个采样值。一个块的前半块和后半块分别进行变换，每半块产生 128 个单值的非“0”变换系数，对应从 0 到 $f_s/2$ 的频率分量，一块的变换系数总数是 256 个。这和单个 512 个采样值块所产生的系数数目相同，但分两次进行，从而改进了时间分辨率。两个半块得到的变换系数交织在一起，形成一个单一的有 256 个值的块。这个块的量化和传输与单一的长块相同。

AC-3 采用基于 MDCT 的自适应变换编码(Adaptive Transform Coding，ATC)算法，ATC 算法的一个重要考虑是基于人耳听觉掩蔽效应的临界频带理论，即在临界频带内一个声音对另一个声音信号的掩蔽效应最为明显。划分频带的滤波器组要有足够锐利的频率响

应，以保证临界频带外的衰减足够大，使时域内和频域内的噪声限定在掩蔽阈值以下。

2. 谱包络编码

频域变换系数都转换成浮点数表示，所有变换系数的值都定标为小于 1。

分析滤波器输出的是指数和被粗化的尾数，两者被编码后都进入码流。指数值的允许范围从 0（对应于系数的最大值，没有前导“0”）到 24，产生的动态范围接近 144dB。系数的指数凡大于 24 的，都固定为 24，这时对应的尾数允许有前导“0”。

减小指数的码率有两种方法。第一种方法是采用差分编码发送 AC-3 指数。每个音频块的第一个指数总是用 4b 的绝对值表示，范围为 0～15，这个值指明第一个（直流项）变换系数的前导“0”的个数，后续的（频率升高方向）指数的发送采用差分值。第二种方法是在一个帧内的 6 个音频块尽量共用一个指数集（在各个块的指数集相差不大时可以采用）。这样，只需在第一块传送该指数集，后面的块共享第一块的指数集，使指数集编码的码率降为原来的 1/6。

AC-3 将上述两种方法结合在一起，并且将差分指数在音频块中联合成组。联合方式有 3 种：4 个差分指数联合成一组，称为 D45 模式；两个差分指数联合成一组，称为 D25 模式；单个差分指数为一组，称为 D15 模式。3 种模式统称为 AC-3 的指数策略，在指数所需的码率和频率分辨率之间提供一种折中。D15 模式提供最精细的频率分辨率，D45 模式所需的数据量最少。在各个组内的指数数目仅取决于指数策略。

3. 比特分配

按照谱包络编码输出的信息确定尾数编码所需要的比特数，将可分配的比特按最佳的方式分配给各个尾数。分配给各个不同值的比特数由比特分配程序确定。在编码器和解码器中运行同样的核心比特分配程序，所以各自产生相同的比特分配。

4. 尾数量化

按照比特分配程序确定的比特数对尾数进行量化。分配给每个尾数的比特数可由一张对照表查到。对照表是按输入信号的功率谱密度（Power Spectral Density，PSD）和估计的噪声电平阈值建立的。PSD 可在细颗粒均匀频率尺度上计算，估计的噪声电平阈值在粗颗粒（按频段）频率尺度上计算。

对某个声道的比特分配的结果具有谱的粒度，它对应于所用的指数策略。具体地说，在 D15 模式的指数集中的每个尾数都要分别计算比特分配；在 D25 模式的指数集中的每两个尾数都要分别计算比特分配；在 D45 模式的指数集中的每 4 个尾数都要分别计算比特分配。

5. 声道组合

在对多声道音频节目编码时，利用声道组合技术可以进一步降低码率。组合利用了人耳对调频定位的特性。人的听觉系统可以跟随低频声音的各个波形，并基于相位差来定位。对于高频声音，由于生理带宽的限制，听觉系统只能跟随高频信号的包络，而不是具体的波形。组合技术只用于高频信号。通过组合将包括在组合声道中的几个声道的变换系数加以平均。

各个被组合的声道有一个特有的组合坐标集合，可用来保留原始声道的高频包络。组合过程仅发生在规定的组合频率之上。解码器将组合声道转换成各个单独的声道，只要用那个声道的组合坐标和频率子带乘以组合声道变换系数的值。

6. 矩阵重组

AC-3 中的重组矩阵是一种声道融合技术。在立体声编码中，左(L)、右(R)声道具有相关性，利用“和”和“差”的方法产生中间声道和边声道，即不是对双声道中的左右声道进行打包，而是首先进行如下变换，然后对中间声道(M)和边声道(S)进行量化编码和打包。

$$M=\frac{R+L}{2}$$

$$S=\frac{L-R}{2}$$

显然，如果原始的立体信号的两个声道是相同的，则这种方法会使 M 信号与原始的 L 信号和 R 信号相同，而 S 信号为 0。这样可以用很少的比特对 S 信号进行编码，同时用较多的比特对 M 信号进行编码。

在解码端，将 M、S 声道恢复至 L、R 声道。这种方法对保留杜比环绕声的兼容性尤其重要。

7. 动态范围控制

音频节目有很宽的动态范围，一般在广播前要先将动态范围缩小。当动态范围较宽时，响的部分会显得太响，而静的部分会变得听不见。AC-3 的语法允许每个音频块传送一个动态控制字，解码器用来改变音频块的电平。控制字的内容指明在信号响度高于对话电平时降低增益，在信号响度低于对话电平时提高增益，信号接近对话电平时就不需要调节增益。

8. AC-3 的帧格式

AC-3 的音频码流是由一个同步帧的序列组成的。每个同步帧包含 6 个编码的音频块(AB)，各个编码音频块由 256 个采样值的码字构成，在各帧的开始有同步头(SI)，包含获取和保持同步的信息。在 SI 之后是比特头(BSI)，包含描述编码的音频业务的参数。编码音频块可后跟一个辅助数据字段(AUX)。在每一帧的末尾是一个误码检测字段，如循环冗余校验码(CRC)。同步帧格式如图 3-32 所示。

图 3-32　AC-3 同步帧格式

各个编码音频块是一个可解码的实体。在对某个音频块进行解码时，并不需要解码所需的信息都在这个块中。如果块解码所需的信息可以被许多块共享，那么可以仅在第一块中传送所需的信息，并在后面的音频块解码时重复使用这个信息。由于各个音频块并不包含全部所需信息，所以在音频帧中的块大小各不相同，但帧内 6 个块的总长度必须固定。某些块可以分配较多的比特，其他块就要相应减少比特，在第 6 块后面余下的任何信息可以作为辅助数据(AUX)。

杜比 AC-3 解码器的原理如图 3-33 所示。

在解码时，首先利用帧同步信息使解码器与编码数码流同步，接着利用循环冗余校验码(CRC)对数据帧中的误码进行纠错处理，使其成为完整、正确的数据，然后进行数据帧的解格式化。在编码中的格式化就是按设定的标准将各种数据捆成一包一包的。一个包就是一

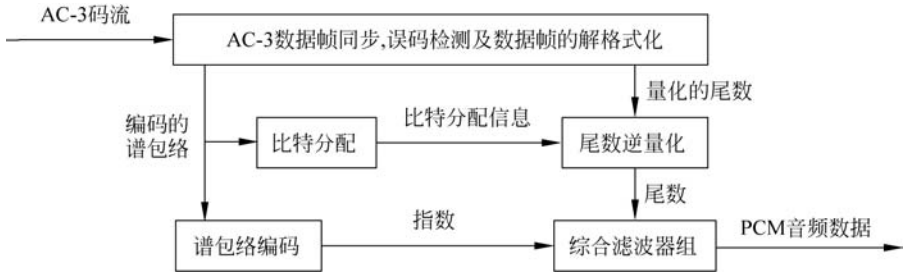

图 3-33 杜比 AC-3 解码器的原理

个数据帧,以包头的同步信息为标志。解码中的解格式化,就是以同步信息为准,将包打开,以便分门别类地处理各种数据,然后运行比特分配例行程序,从编码的谱包络中获得在编码中采用的比特分配信息。

利用此信息便可对量化的尾数进行逆量化处理,还原成原来的尾数,再对谱包进行解码,便获得编码前的各个指数。这些用二进制表示的各种指数和尾数代表了各样本块的256个频域变换系数,最后利用综合滤波器进行离散余弦反变换,将这些变换系数还原成时间域中的 PCM 数字音频信号。

其他一些音频编码标准对比如表 3-2 所示。

表 3-2 各种音频压缩标准比较

标准	采用的主要编码技术	比特率	主要应用
G.711	A 律或 μ 律压扩的 PCM 编码	64kb/s	固定电话语音编码
G.721/3/6	自适应差分脉冲编码调制(ADPCM)	32,24,16kb/s	IP 电话
G.722	子带自适应差分脉冲编码(SB-ADC)	48kb/s	高质量语音信号
G.723.1	代数码激励线性预测(ACELP)、多脉冲最大似然量化机制(MP-MLP)	5.3kb/s	公用电话网、移动网和互联网的语音通信
G.728	低延时码激励线性预测(LD-CELP)、矢量量化	16kb/s	光盘存储、计算机磁盘存储、视频娱乐、视频监控
G.729	共轭结构代数码激励线性预测编码(CS-ACELP)	32kb/s	IP 电话、会议电视、数字音视频监控
MPEC-1 Audio1/2/3	掩蔽模式通用子带集成编码、多路复用(MUSICAM)、自适应频率感知熵编码(ASPEC)	384,256,64kb/s	DAB、ISDN 宽带网络传输
MPEC-2 Audio(BC)	MPEC-1 所有技术、线性 PCM、杜比 AC-3 编码、5.1/7.1 声道	8～640kb/s	数字电视、DVD
MPEC-2 AAC(NBC)	改进离散余弦变换(MDCT)		数字电视、DVD
MPEC-4 Audio	参数编码、码激励线性预测、矢量量化	2～64kb/s	通信、中/短波数字声音广播
MPEC-7 Audio	描述音频内容		建立音频档案、检索
杜比 AC-3	改进离散余弦变换(MDCT)、自适应变换编码(ATC)		DVD、DTV、DBS

3.4 数字音频文件的常见格式

1. WAV 波形音频文件

WAV 波形音频文件是微软公司和 IBM 公司共同开发的 PC 标准声音格式，文件扩展名为.wav，是一种通用的音频数据文件。

WAV 格式通常用来保存一些没有压缩的音频，也就是经过 PCM 编码后的音频，因此也称为波形文件，按照声音的波形进行存储，要占用较大的存储空间。CD 唱片包含的就是 WAVE 格式的波形数据，只是扩展名没写成.wav，而是.cda。WAV 文件也可以存放压缩音频，但其本身的文件结构使之更加适合存放原始音频数据并用来做进一步的处理。

WAV 文件组成如下。文件头：标明是 WAVE 文件、文件结构和数据的总字节数；数字化参数：如采样频率、声道数、编码算法等；实际的波形数据。

WAV 文件特点为易于生成和编辑，但是在保证一定音质的前提下压缩比不够，不适合在网络上播放。

2. MP3 文件

MP3 是人们比较熟知的一种数字音频格式，网络中大多数歌曲、音乐文件都采用这种格式，是一种流式音乐文件格式。它是 MPEG 制定的 MPEG-1 Audio Layer 3 压缩标准，是第一个使用有损音频压缩编码的。MP3 采用的是高压缩比(10∶1 或 12∶1)，但是能够保持良好的音质，利用人耳的特性，削减音乐中人耳听不到的成分，同时尝试尽可能地维持原来的声音，几乎达到了 CD 音质标准。

3. WMA 文件

WMA 是 Windows Media Audio 的缩写，是微软定义的一种流式声音格式，扩展名为.wma，相对于 MP3 的主要优点是在较低的采样频率下音质要好些。

4. RA 文件

RA 是 Real Audio 的缩写，是由 Real Networks 公司推出的一种文件格式，最大的特点是可以实时传输音频信息，尤其是在网速较慢的情况下，仍然可以较流畅地传送数据。因此，Real Audio 主要适用于网络上的在线播放。现在的 Real Audio 文件格式主要有 RA(Real Audio)、RM(Real Media，Real Audio G2)、RMX(Real Audio Secured)等 3 种，这些文件的共同性在于根据网络带宽的不同而改变声音的质量，在保证大多数人听到流畅声音的前提下，令带宽较宽的听众获得较好的音质。

5. MID 文件

MID 是通过数字化乐器接口 MIDI 输入的声音文件的扩展名，这种文件只是像记乐谱一样记录下演奏的符号，所以它占用的存储空间是所有音频格式中最小的。

MID 文件结构如下。文件头：描述文件的类型和音轨数等；音轨：记录 MIDI 数据，主要是命令序列，每个命令包括命令号、通道号、音色号和音速等。

MID 文件特点主要有：WAV 文件记录声音数据，MID 文件记录一系列乐谱指令；数据量小，占用存储空间极小，适合在网络上传输；编辑修改灵活方便，可通过音序器自由修改 MIDI 文件的曲调、音色、速度等，甚至可以改换不同的乐器；MIDI 声音仅适于重现打击乐或一些电子乐器的声音。

6. AIFF 文件

音频交换文件格式(Audio Interchange File Format,缩写为 AIF 或 AIFF),是苹果公司开发的一种标准声音文件格式,被 Macintosh 平台及其应用程序所支持。它属于 Quick Time 技术中的一部分,而且是一种优秀的文件格式,投入使用后便很快得到微软公司的青睐,Netscape Navigator 浏览器中的 Live Audio、SGI 及其他专业音频软件包都支持它。

AIF/AIFF 支持 16 位,44.1kHz 立体声,现在几乎所有的音频编辑软件和播放软件都支持这种格式。

7. DVD Audio

DVD Audio 是新一代的数字音频格式,与 DVD Video 尺寸以及容量相同,为音乐格式的 DVD 光碟,采样频率为 48kHz,96kHz,192kHz 和 44.1kHz,88.2kHz,176.4kHz(可选择),量化位数可以为 16,20,24b,它们之间可自由地进行组合。低采样率的 96kHz 虽然是两声道重播专用,但它最多可收录到 6 声道。而以两声道 192kHz/24b 或 6 声道 96kHz/24b 收录声音,可容纳 74min 以上的录音,动态范围达 144dB,整体效果出类拔萃。

8. MD

MD(MiniDisc)由日本索尼公司开发。MD 之所以能在一张盘中存储 60～80min 采用 44.1kHz 采样的立体声音乐,就是因为使用了自适应声学转换编码算法(Adaptive Transform Acoustic Coding,ATRAC)压缩音源。这是一套基于心理声学原理的音响解码系统,它可以把 CD 唱片的音频压缩到原来数据量的大约 1/5,而声音质量没有明显的损失。ATRAC 利用人耳听觉的心理声学特性(频谱掩蔽特性和时间掩蔽特性)以及人耳对信号幅度、频率、时间的有限分辨能力,将人耳感觉不到的成分不编码、不传送,这样就可以相应减少某些数据量的存储,从而既保证音质,又达到缩小体积的目的。

9. AAC

AAC 是高级音频编码(Advanced Audio Coding)的缩写。AAC 是由 Fraunhofer IIS-A、杜比和 AT&T 共同开发的一种音频格式,它是 MPEG-2 规范的一部分。AAC 所采用的运算法则与 MP3 不同,通过结合其他的功能来提高编码效率。AAC 的音频算法在压缩能力上远远超过了以前的一些压缩算法(如 MP3 等)。它还同时支持多达 48 个音轨、15 个低频音轨,具有更多种采样率和比特率、多种语言的兼容能力和更高的解码效率。总之,AAC 可以在比 MP3 文件缩小 30%的前提下提供更好的音质。

10. OGG 格式

OGG 的全称是 OGG Vobis,它是一种新的音频压缩格式,类似于 MP3 等现有音乐格式。但有所不同的是,它是完全免费、开放和没有专利限制的。OGG Vobis 有一个很出众的特点,就是支持多声道。随着它的流行,以后用随身听来听 DTS 编码的多声道作品将不再是梦想。OGG Vobis 在压缩技术上比 MP3 好,使它很有可能成为一个流行的趋势,这也正是一些 MP3 播放器支持 OGG 格式的原因。另外,如果采用相同速率录制音频,MP3 和 OGG 不分上下,OGG 采用更先进的算法,还可能会好一些。

11. APE 格式

APE 是 Monkey's Audio 提供的一种无损压缩格式。Monkey's Audio 提供了 Winamp 的插件支持,这就意味着压缩后的文件不再是单纯的压缩格式,而是和 MP3 一样可以播放的音频文件格式。APE 的压缩比大约为 2∶1,但能够做到真正无损,因此获得了不少发烧

友的青睐。令人满意的压缩比以及飞快的压缩速度，成为不少发烧友私下交流音乐的唯一选择。

12. SACD 格式

SACD 格式是由 Sony 公司正式发布的。它的采样率为 CD 格式的 64 倍，即 2.8224MHz。SACD 重放频率带宽达 100kHz，为 CD 格式的 5 倍，24b 的量化位数也远远超过 CD，声音的细节表现更丰富、清晰。

13. VQF 格式

VQF 是由 YAMAHA 和 NTT 共同开发的一种音频压缩技术，它的压缩比能够达到 18∶1。因此，相同情况下压缩后的 VQF 文件体积比 MP3 小 30%～50%，更便于网上传播。VQF 格式的音质极佳，接近 CD 音质(16b，44.1kHz 立体声)。但 VQF 未公开技术标准，至今未能流行起来。

14. WMA 格式

WMA 格式是以减少数据流量但保持音质的方法来达到更高的压缩比，其压缩比一般可以达到 18∶1。此外，WMA 还可以通过数字版权管理(Digital Rights Management，DRM)方案加入防止复制，或者加入播放时间和播放次数的限制，甚至是播放机器的限制，有力地防止盗版。

15. MP4 格式

MP4 在文件中采用了保护版权的编码技术，只有特定的用户才可以播放，有效地保证了音乐版权的合法性。另外，MP4 的压缩比达到了 15∶1，体积较 MP3 更小，音质却没有下降。不过因为只有特定的用户才能播放这种文件，因此其流传度与 MP3 相比差距甚远。

16. CDA 格式

大家都很熟悉 CD 这种音乐格式了，其扩展名为.cda，采样频率为 44.1kHz，16b 量化。CD 存储采用了音轨的形式，又叫“红皮书”格式，记录的是波形流，是一种近似无损的格式。

3.5 本章小结

本章主要介绍了数字音频信号压缩的可行性，按照压缩品质、编码方式对数字音频压缩方法进行分类，再介绍了波形编码、参数编码、混合编码等编码原理，又对 MPEG 数字音频压缩标准和杜比 AC-3 音频压缩算法进行了详细说明，最后简单介绍了数字音频文件的常见格式，如 WAV、MP3、WMA 和 RA 等。