第5章 语音与音频压缩编码

CHAPTER 5

语音编码与音频编码作为目前常用的语音与音频信号数字传输/存储技术，广泛地应用于信号处理、移动通信、IP通信、广播电视以及多媒体互联网等多个领域。近年来，随着计算机网络技术的发展，数字信息服务和多媒体娱乐方式层出不穷，人们已经不满足于单一的语音通信需求，更希望享受兼容语音与音频的通信服务所带来的愉悦。同时，功能日益强大的移动通信设备也为这一需求的实现提供了所需硬件支撑，电信网、有线电视网和计算机网之间相互渗透、互相兼容，并逐步整合成为全球统一的信息通信网络已经成为一种趋势和必然。

然而，受到处理对象、编码模型及应用背景等的限制，传统语音编码和音频编码一直以来都是语音频信号处理领域的两大独立研究分支，由此导致不同网络间存在的码流格式不兼容问题，已经成为制约多媒体技术进一步发展的瓶颈。另外，实际生活中人们所接触的声音信息极为复杂，除了单纯的语音和音频外，还包含了自然声、混合音频等诸多信息，基于语音与音频单独编码的传统服务系统，因无法高效地处理复杂的混合信息，给不同网络、服务系统之间的融合带来障碍。可见，若能够对语音和音频信号采用统一的编码模型进行处理，既能够保证系统高质量地编解码语音和音频信号，又能够为不同服务系统之间码流的兼容提供可能。

语音与音频通用编码，从广义上讲，是在现有语音编码和音频编码技术的基础之上，利用统一的编解码模型，实现对语音、音频以及语音和音频混合信号的无差别编码。从而在同等码率约束条件下，对语音、音频及其混合信号均能够取得高质量的合成音质，以弥补传统单一类型的语音或音频编码器仅适合处理单一信号，对于其他类信号或混合信号无法获得优良编码性能的不足。现如今，对于语音和音频信号的通用编码正逐渐引起国内外学者和研究机构的关注，适合移动音频和网络在线娱乐音频的语音与音频通用编码的标准化也正成为MPEG和ITU-T等国际组织的重点工作之一。

语音与音频通用编码算法是对现有语音编码和音频编码技术的拓展和完善，因此，现有的语音频编码技术、声学感知理论、率-失真理论及其相关的最新研究成果均可以借鉴，从而为通用编码算法的研究提供了坚实的理论基础和技术支撑。同时，对通用编码算法的研究，紧密契合了包括移动音频、手机电视、音视频会议、流媒体音乐、音视频娱乐点播在内的4G或5G多媒体应用对语音和音频编码的需求，具有重要的现实意义。

5.1　语音与音频编码技术概况

5.1.1　语音与音频压缩的必要性

语音与音频压缩技术指的是对原始数字音频信号流运用适当的数字信号处理技术，在不损失有用信息量或所引人损失可忽略的条件下，降低其码率，也称为压缩编码。它必须具有相应的逆变换，称为解压缩或解码。在多媒体音频信号处理中，一般需要对数字化后的声音信号进行压缩编码，使其成为具有一定字长的二进制数字列，并以这种形式在多媒体网络内传输和存储。在播放这些声音时，需要经解码器将二进制编码恢复成原来的声音信号播放。由于数字音频文件的信息量是非常大的，例如，未经压缩的 1min 立体 CD 音乐所需的存储量为(44.1×1000×16)×2×60/8＝10584000B≈10.1MB(存储量的计算公式：存储量＝(采样频率×采样精度×声道数×时间)/8)，数据量大得惊人。例如，一套双声道数字音频若取样频率为 44.1kHz，每样值按 16bit 量化，则其码率为：2×44.1kHz×16bit＝1.411Mb/s，而 1GB 的容量只能存储约 1 分钟的彩色电视信号数据。在通信网络上，大多数远程通信网络的速率都在几兆每秒以下，这样大的数据量不仅超出了计算机的存储和处理能力，更是当前通信信道的传输速率所不及的。为了音频的普及，数字音频压缩技术显得尤为重要，尤其是无损压缩，更符合人们对于音乐的要求。

数字音频压缩编码在保证信号在听觉方面不产生失真的前提下，对音频数据信号进行尽可能大的压缩。数字音频压缩编码采取去除声音信号中冗余成分的方法来实现。所谓冗余成分指的是音频中不能被人耳感知到的信号，它们对确定声音的音色、音调等信息没有任何的帮助。

冗余信号包含人耳听觉范围外的音频信号以及被掩蔽掉的音频信号等。

音频信号是多媒体信息的重要组成部分，它可以分成电话质量的语音信号、调频广播质量的音频信号和高保真立体声信号。音频编解码技术是随着音频信号数字化而产生的，目前主要应用在数字音频通信和数字音频存储两个领域。

由于简单地由连续音频信号抽样量化得到的数字音频信号，在传输和存储时要占用较多的信道资源和存储空间，因此，如何在尽量减少失真的情况下，高效率地对模拟语音信号进行数字表达，即压缩编码，就成为音频编码技术的主要内容。数字语音压缩编码技术由于具有加密容易、保密性强；易于纠错编码、抗干扰能力强、便于传输；便于对语音信号进行数字化处理；有利于提高话路容量等优点，使其广泛应用于多媒体语音通信系统。

声音信号能进行压缩编码的基本依据主要有 3 点：

(1) 声音信号中存在着很大的冗余度，通过识别和去除这些冗余度，便能达到压缩的目的。

(2) 音频信息的最终接收者是人，人的视觉和听觉器官都具有某种不敏感性。舍去人的感官所不敏感的信息对声音质量的影响很小，在有些情况下，甚至可以忽略不计。例如，人耳听觉中有一个重要的特点，即听觉的“掩蔽”。它是指一个强音能抑制一个同时存在的弱音的听觉现象。利用该性质，可以抑制与信号同时存在的量化噪音。

(3) 对声音波形采样后，相邻采样值之间存在着很强的相关性。

语音编码与音频编码的主要目的都是在保证一定主观听觉质量的前提下，最大程度地

去除输入信号的统计冗余和感知冗余来实现数据量的压缩，以满足不同传输和存储条件下的需求。

图 5-1 给出了语音处理的基本过程。国际电信联盟已经制定了多个语音编码标准，表 5-1 详细列出了各种标准的各种参数和性能指标；图 5-2 给出了语音与音频编码标准的速率与质量关系。

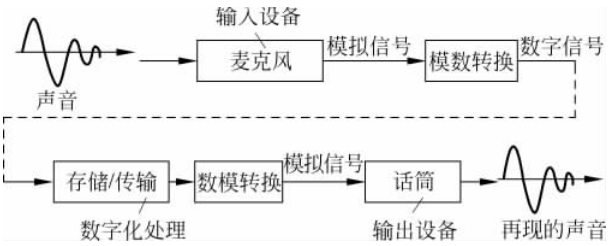

图 5-1　音频处理基本过程

表 5-1　语音编码国际标准及参数

标　准	算　法	码率/(kb·s^{-1})	算法时延/ms	复杂度/MIPS	语音质量/MOS
G.711	PCM	64	0.125	1	4.3
G.723.1	ACELP	5.3	37.5	25	3.5
	MP-MLQ	6.3	37.5	16	3.8
G.726	ADPCM	32	0.125	10	4.0
G.728	LD-CELP	16	0.625	50	4.0
G.729	CS-ACELP	8	15	30	4.0

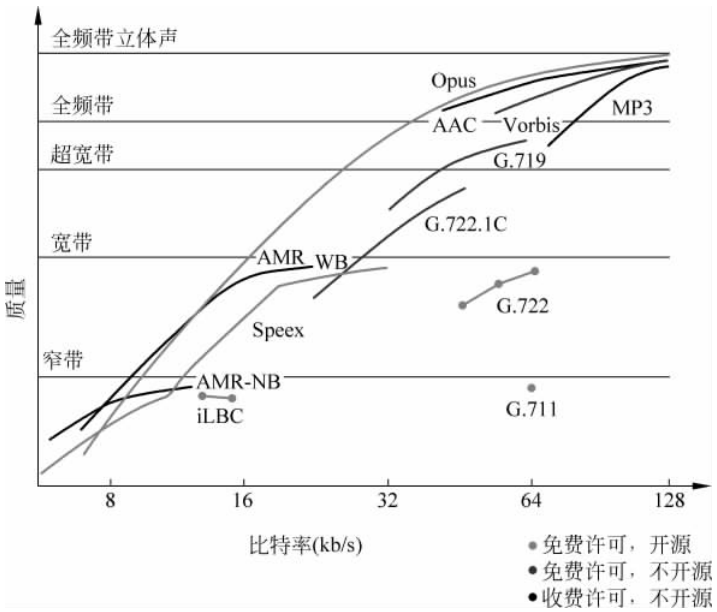

图 5-2　语音与音频编码标准的速率与质量关系

5.1.2　语音与音频压缩的区别

虽然语音编码和音频编码同属信源编码，但由于输入信号特征和应用背景的不同，二者在核心算法上往往存在着巨大的差异。通常，语音编码的输入为由人的发声器官所发出的频段在80Hz～3400Hz之间的语音信号，信号来源单一，频谱结构相对简单。因此，语音编码技术往往基于人类语音的产生模型，通过去除信号远样点间和近样点间的相关性，在较低码率下实现了语音信号的高质量编码，最具典型的例子就是移动通信中普遍使用的码激励线性预测(Code-Excited Linear Prediction，CELP)语音编码。而音频编码的处理对象为人耳可以听到的、频率在20Hz～20kHz之间的音频信号，信号的来源包括了人耳能感觉到的所有声音，声源较多、信号复杂，无法用统一的声源模型来处理。另外，在应用背景方面，语音编码主要应用于数字通信、移动无线电和蜂窝电话等系统，因此算法延迟、数据速率和语音质量是算法设计考虑的重点。音频编码则主要应用于数字广播、网络流媒体和影视音像等娱乐场合，对算法延迟和数据速率的要求较之语音编码相对宽松，在算法设计上更多地侧重于音频信号的合成音质和感知舒适度。

目前，对于语音编码标准的制定，主要由ITU-T来实现；而音频编码的标准则主要由国际电工委员会第一联合技术组(ISO/IEC JTC1)的运动图像专家组MPEG来完成。

5.1.3　音频压缩方法

音频信号的压缩方法如图5-3所示。按照压缩原理的不同，声音的压缩编码可分为3类，即波形编码、参数编码和混合型编码。

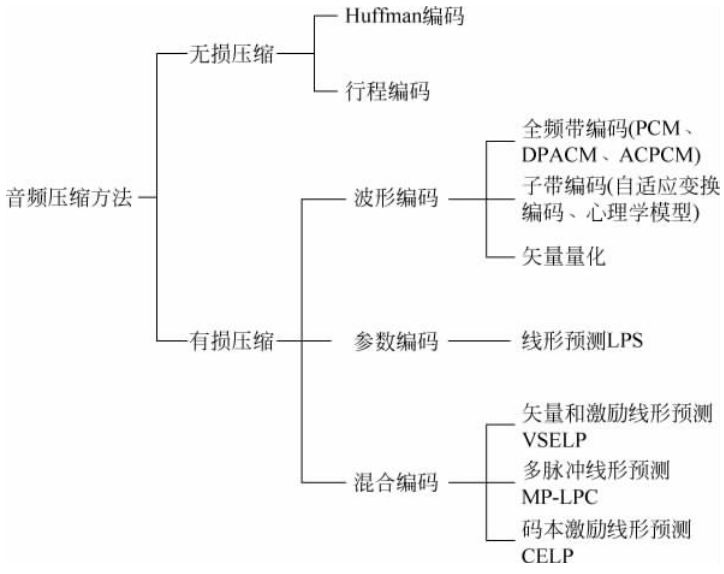

图5-3　音频信号压缩方法

1. 波形编码

这种方法主要利用音频采样值的幅度分布规律和相邻采样值间的相关性进行压缩，目标是力图使重构的声音信号的各个样本尽可能地接近于原始声音的采样值。这种编码保留

了信号原始采样值的细节变化，即保留了信号的各种过渡特征，因而复原的声音质量较高。波形编码技术有脉冲编码调制(Pulse Code Modulation，PCM)、自适应增量调制和自适应差分脉冲编码调制等。

2. 参数编码

参数编码又称为声源编码，是把音频信号表示成某种模型的输出，利用特征提取的方法抽取必要的模型参数和激励信号的信息，并对这些信息编码，最后在输出端合成原始信号。参数编码是一种对语音参数进行分析合成的方法。语音的基本参数是基音周期、共振峰、语音谱、声强等，如果能得到这些语音基本参数，就可以不对语音的波形进行编码，而只要记录和传输这些参数就能实现声音数据的压缩。这些语音基本参数可以通过分析人的发音器官的结构及语音生成的原理，建立语音生成的物理或数学模型通过实验获得。得到语音参数后，就可以对其进行线性预测编码(Linear Predictive Coding，LPC)。

线性预测编码及其各种改进型都属于参数编码。这种编码方式的编码速率可达到 2～4.8kb/s，甚至更低。所付出的代价是计算量大以及语音质量的下降：语音的清晰度尚可，但自然度不好，且对背景噪声相当敏感。但它的保密性能非常好，因此这种编码在军事上获得了广泛应用。

随着一些复杂的算法得以硬件实现，突破了波形编码与参数编码的界线，提出了混合编码。

3. 混合型编码

混合型编码将波形编码和参数编码两者结合起来，很好地解决了两者的缺点，混合型编码是一种在保留参数编码技术的基础上，引用波形编码准则去优化激励源信号的方案。混合型编码充分利用了线性预测技术和综合分析技术，其典型算法有码本激励线性预测、多脉冲线性预测、矢量和激励线性预测等，尽量保留了两者的优点。混合编码可将编码速率压缩到 4～8kb/s，在 4～8kb/s 范围内能达到良好的语音质量。

得到广泛研究的混和编码算法是基于线性预测技术的分析合成编码方法(Linear Prediction Analysis-by-Synthesis，LPAS)。最早实用的 LPAS 方案的是由 Atal 和 Remede 提出的多脉冲线性预测编码，另外较典型的方案还有规则脉冲激励线性预测编码。但最重要的一种 LPAS 算法是由 Atal 和 Schroeder 提出的码激励线性预测编码(Code Excited Linear Prediction，CELP)，也称随机编码、矢量激励编码或随机激励线性预测编码。现在一般把以 LPAS 为基础的采用 VQ(Vector Quantization，向量量化)技术对激励信号进行量化编码的算法统称为 CELP，它不再单指一项特定的编码技术，而是一类重要的编码技术。它在 4～16kb/s 编码速率中可以得到比其他算法更高的重建语音质量，而且以 CELP 为基础的多种算法已成为国际标准，其中包括 G.728 建议的 LD-CELP 和 G.729 建议的 CS-ACELP 算法。

波形编码器试图保留被编码信号的波形，能以中等比特率(32kb/s)提供高品质语音，但无法应用在低比特率场合。声码器试图产生在听觉上与被编码信号相似的信号，能以低比特率提供可以理解的语音，但是所形成的语音听起来不自然。混合编码器结合了两者的优点。

(1) RELP：在线性预测的基础上，对残差进行编码。机制为：只传输小部分残差，在接受端重构全部残差(把基带的残差进行复制)。

(2) MPC(Multi-Pulse Coding,多脉冲激励编码)：对残差去除相关性,用于弥补声码器将声音简单分为 voiced 和 unvoiced,而没有中间状态的缺陷。

(3) CELP(Codebook Excited Linear Prediction)：用声道预测其和基音预测器的级联,更好地逼近原始信号。

(4) MBE(Multiband Excitation)：多带激励,目的是避免 CELP 的大量运算,获得比声码器更高的质量。

5.2　语音与音频编码技术

在进行信源编码时,既希望最大限度地降低码率,又希望尽可能不要对音源造成损伤,两者是矛盾的,随着比特率的进一步压缩,势必要影响信源的失真度。一般来讲,根据压缩后的音频能否完全重构出原始声音,可以将音频压缩技术分为无损压缩及有损压缩两大类。无损压缩时根据统计学观点分析数据流,仅从数据量减少数据率,有损压缩是从声音怎样被听到的基础上来减少数据率,利用人的听觉不能检测某些信号损失,从而可以大量减少比特率。而按照音频压缩编码方式的不同,又可将其划分为时域编码(包括预测编码、增量编码)、频域编码(包括变换编码、子带编码)、统计编码(熵编码、哈夫曼编码)以及多种技术相互融合的混合编码等。对于各种不同的压缩编码方法,其算法的复杂程度(包括时间复杂度和空间复杂度)、重建音频信号的质量、算法效率(即压缩比),编解码延时等都有很大的不同,因此其应用场合也各不相同。下面介绍几种主要的波形编码方式。

5.2.1　时域编码

时域编码是指直接针对音频 PCM 码流的样值进行处理,通过静音检测、非线性量化、差分等手段对码流进行压缩。此类压缩技术的共同特点是算法复杂度低,声音质量一般,压缩比小(CD 音质下将大于 400kb/s),编解码延时最短(相对于其他技术)。此类压缩技术一般多用于语音压缩,低码率应用(源信号带宽小)的场合。

1. 脉冲编码调制(PCM)

下面介绍波形编码方案中常用的 PCM 编码。

香农(Claude E. Shannon)于 1948 年发表的“通信的数学理论”奠定了现代通信的基础。同年贝尔实验室的工程人员开发了 PCM 技术,虽然在当时是革命性的,但今天脉冲编码调制被视为一种非常单纯的无损耗编码格式,音频在固定间隔内进行采集并量化为频带值,其他采用这种编码方法的应用包括电话和 CD。脉冲编码调制是一种对模拟信号数字化的取样技术,将模拟语音信号变换为数字信号的编码方式,特别是对于音频信号。PCM 对信号每秒钟取样 8000 次；每次取样为 8 个位,总共 64kb。

PCM 主要有三种方式：标准 PCM、DPCM(Differential pulse code modulation,差分脉冲编码调制)和自适应 DPCM。

PCM 主要经过 3 个过程：抽样、量化和编码。抽样过程将连续时间模拟信号变为离散时间、连续幅度的抽样信号,量化过程将抽样信号变为离散时间、离散幅度的数字信号,编码过程将量化后的信号编码成为一个二进制码组输出。如图 5-4 所示。图中 $S(k)$的是发送端编码器的输入信号,$S_r(k)$是接收端译码器输出的信号。

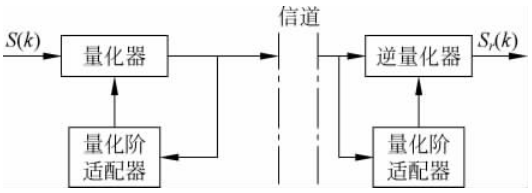

图 5-4　PCM 示意图

2. 差分脉冲编码调制(DPCM)

DPCM 只对样本之间的差异进行编码。前一个或多个样本用来预测当前样本值。用来做预测的样本越多,预测值越精确。真实值和预测值之间的差值叫残差,是编码的对象。

差分脉冲编码调制的思想是,根据过去的样本去估算下一个样本信号的幅度大小,这个值称为预测值,然后对实际信号值与预测值之差进行量化编码,从而就减少了表示每个样本信号的位数。它与 PCM 不同的是,PCM 是直接对采样信号进行量化编码,而 DPCM 是对实际信号值与预测值之差进行量化编码,存储或者传送的是差值而不是幅度绝对值,这就降低了传送或存储的数据量。此外,它还能适应大范围变化的输入信号。

差分脉冲编码调制的概念示于图 5-5。图中差分信号是离散输入信号和预测器输出的估算值之差。(注意:是对的预测值,而不是过去样本的实际值。)DPCM 系统实际上就是对这个差值进行量化编码,用来补偿过去编码中产生的量化误差。DPCM 系统是一个负反馈系统,采用这种结构可以避免量化误差的积累。重构信号是由逆量化器产生的量化差分信号与对过去样本信号的估算值求和得到。它们的和,即作为预测器确定下一个信号估算值的输入信号。由于在发送端和接收端都使用相同的逆量化器和预测器,所以接收端的重构信号可从传送信号获得。

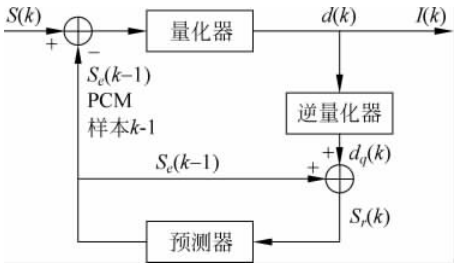

图 5-5　DPCM 示意图

图 5-5 中,差分信号 $d(k)$是离散输入信号 $s(k)$和预测器输出的估算值 $S_e(k-1)$之差。注意,$S_e(k-1)$是对 $S(k)$的预测值,而不是过去样本的实际值。DPCM 系统实际上就是对这个差值 $d(k)$进行量化编码,用来补偿过去编码中产生的量化误差。DPCM 系统是一个负反馈系统,采用这种结构可以避免量化误差的积累。重构信号 $S_r(k)$是由逆量化器产生的量化差分信号 $d_q(k)$,与对过去样本信号的估算值 $S_e(k-1)$求和得到。它们的和 $S_r(k)$,即作为预测器确定下一个信号估算值的输入信号。由于在发送端和接收端都使用相同的逆量化器和预测器,所以接收端的重构信号 $S_r(k)$可从传送信号 $I(k)$获得。

3. 自适应差分脉冲编码(ADPCM)

自适应差分脉冲编码(Adaptive Differential Pulse Code Modulation,ADPCM)。即在

DPCM 的基础上，根据信号的变化，适当调整量化器和预测器，使预测值更接近真实信号，残差更小，压缩效率更高。

它的核心思想是：利用自适应的思想改变量化阶的大小，即使用小的量化阶(step-size)去编码小的差值，使用大的量化阶去编码大的差值；使用过去的样本值估算下一个输入样本的预测值，使实际样本值和预测值之间的差值总是最小。

它的编码简化框图如图 5-6 所示。接收端的译码器使用与发送端相同的算法，利用传送来的信号来确定量化器和逆量化器中的量化阶大小，并且用它来预测下一个接收信号的预测值。

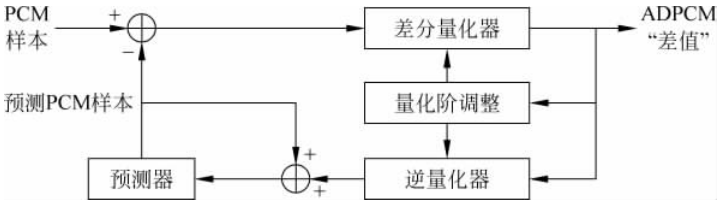

图 5-6　ADPCM 示意图

5.2.2　频带编码

1. 子带编码

子带编码(Sub-band Coding，SBC)是一种以信号频谱为依据的编码方法，是将原始信号由时间域转变为频率域，然后将其分割为若干个子频带，并对其分别进行数字编码的技术。它是利用带通滤波器(BPF)组把原始信号分割为若干(例如 m 个)子频带(简称子带)。将各子带通过等效于单边带调幅的调制特性，将各子带搬移到零频率附近，分别经过 BPF(共 m 个)之后，再以规定的速率(奈奎斯特速率)对各子带输出信号进行取样，并对取样数值进行通常的数字编码，其设置 m 路数字编码器。将各路数字编码信号送到多路复用器，最后输出子带编码数据流。

子带编码理论的基本思想是将信号分解为若干子频带内的分量之和，然后对各子带分量根据其不同的分布特性采取不同的压缩策略以降低码率。SBC 编解码原理图见图 5-7。

在采用子带编码时，利用了听觉的掩蔽效应进行处理。它对一些子带信号予以删除或大量减少比特数目，可明显压缩传输数据总量。例如，不存在信号频率分量的子带，被噪声掩蔽的信号频率的子带，被邻近强信号掩蔽的信号频率分量子带等，都可进行删除处理。另外，全系统的传输信息量与信号的频带范围、动态范围等均有关系，而动态范围则决定于量化比特数，若对信号引进公道的比特数，可使不同子带内按需要分配不同的比特数，也可压缩其信息量。

子带编码技术具有突出的优点。首先，声音频谱各频率分量的幅度值各不相同，若对不同子带分配以合适的比例系数，可以更公道地分别控制各子带的量化电平数目和相应的重建误差，使码率更精确地与各子带的信号源特性相匹配。通常，在低频基音四周，采用较大的比特数目来表示取样值，而在高频段则可分配以较小的编码比特。其次，通过公道分配不同子带的比特数，可控制总的重建误差频谱外形，通过与声学心理模型相结合，可将噪声频谱按人耳主观噪声感知特性来形成。于是，利用人耳听觉掩蔽效应可节省大量比特数。

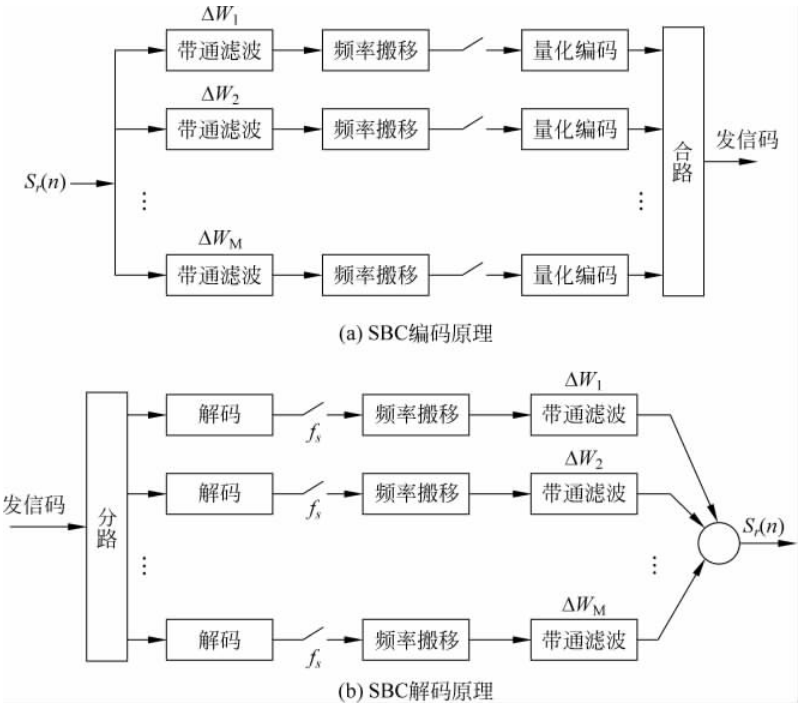

图 5-7　SBC 编码与解码原理

子带编码技术和后面介绍的变换编码技术都是利用人耳的听觉感知特性，使用心理声学模型(Psychoacoustic Model)，通过对信号频谱的分析来决定子带样值或频域样值的量化阶数和其他参数选择的，因此又可称为感知型(Perceptual)音频编码。这两种编码方式相对于时域编码技术而言要复杂得多，同时编码效率、声音质量也大幅提高，编码延时相应增加。一般来讲，子带编码的复杂度要略低于变换编码，编码延时也相对较短。

由于在子带编码技术中主要应用了心理声学中的声音掩蔽模型，因而在对信号进行压缩时引入了大量的量化噪声。然而，根据人耳的听觉掩蔽曲线，在解码后，这些噪声被有用的声音信号掩蔽掉了，人耳无法察觉；同时由于子带分析的运用，各频带内的噪声将被限制在频带内，不会对其他频带的信号产生影响。因而在编码时各子带的量化阶数不同，采用了动态比特分配技术，这也正是此类技术压缩效率高的主要原因。在一定的码率条件下，此类技术可以达到“完全透明”的声音质量(EBU 音质标准)。

2. 变换编码

变换编码是当前音频编码标准普遍采用的压缩技术。变换域编码属于频域编码，把信号从时域变换到频域，再对其频谱系数进行量化编码。变换编码充分利用人耳在频域上的听觉特性(掩蔽效应和临界频带)来实现对音频信号的压缩，是一种高效的编码技术。在标准化组织 ISO/IEC 制定的音频编码标准 MPEG1-3 和 MPEG-AAC 都使用了改进余旋变换 MDCT，可有效消除 DCT(离散余弦变换)的块边界噪声。

变换编码技术与子带编码技术的不同之处在于该技术对一段音频数据进行“线性”的变换，对所获得的变换域参数进行量化、传输，而不是把信号分解为几个子频段。通常使用的变换有 DFT(离散傅氏变换)、DCT、MDCT(改进的离散余弦变换)等。根据信号的短时功

率谱对变换域参数进行合理的动态比特分配可以使音频质量获得显著改善，而相应付出的代价则是计算复杂度的提高。

5.3　目前主流音频压缩编码标准及应用

音频压缩技术的应用面很广，电信、计算机、消费电子产品中大量使用了音频压缩技术，因而人们见到的压缩方法也非常多，甚至不同厂商的压缩标准也不同。本文集中介绍用于广播电视的主流音频压缩编码标准和它们的应用。

5.3.1　MPEG-1

在音频压缩标准化方面取得巨大成功的是 MPEG-1 音频标准(ISO/IEC11172-3)，它是世界上第一个高保真音频数据压缩标准。

MPEG-1 压缩编码原理方框图如图 5-8 所示。它采用的压缩技术方案是子带压缩，子带分割的实现是通过时频映射，采用多相正交分解滤波器组将数字化的宽带音频信号分成 32 个子带；同时，信号通过 FFT(快速傅里叶变换)运算，对信号进行频谱分析；子带信号与频谱同步计算，得出对各子带的掩蔽特性，由于掩蔽特性的存在，减少了对量化比特率的要求，不同子带分配不同的量化比特数，但对于各子带而言，是线性量化。另加上循环冗余校验(Cyclic Redundancy Check，CRC)校验码，得到标准的 MPEG 码流。在解码端，只要解帧，子带样值解码，最后进行频-时映射还原，最后输出标准 PCM 码流。在 MPEG-1 压缩中，按复杂程度规定了三种模式，即层Ⅰ、层Ⅱ、层Ⅲ。层Ⅰ的编码简单，用于数字盒式录音磁带；层Ⅱ的算法复杂度中等，用于数字音频广播(DAB)和 VCD 等；层Ⅲ的编码复杂，用于互联网上的高质量声音的传输，如 MP3 音乐压缩 10 倍。

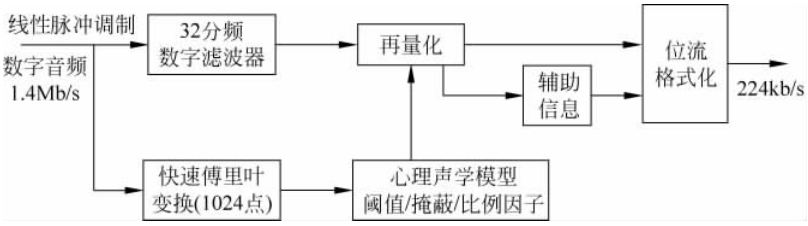

图 5-8　MPEG-1 编码器方框

1. 层Ⅰ

数字的多相正交滤波器组把信号分成 32 个子带信号，因为层Ⅰ是均匀地划分，所以每个子带频宽为 625Hz。这种划分与关键频宽段的概念不一样，在低端只有一个子带 625Hz，这样对低频的量化比较简单，容易引起低频端的量化误差。心理声学模型：使用 512 个点的 FFT 变换得到信号的短时频谱功率信息，输出的电平和时频映射的子带样值同步计算，得到每个子带的掩蔽阈值。最后将该子带的最大信号/掩蔽阈值率输入给量化器。VCD 中使用 MPEG-1 层Ⅰ的音频压缩方案。

2. 层Ⅱ(即 MUSICAM，又称 MP2)

时频映射和层Ⅰ类似，不同之处在于每个子带不是均匀频带宽，低频取的带宽窄，即意

味着对低频有较高频率分辨率，在高频端时则相对有较低的分辨率。这样的分配，更符合人耳的灵敏度特性，可以改善对低频端压缩编码的失真。但这样做，需要较复杂一些的滤波器组。其心理声学模型和层Ⅰ类似，但是使用的 FFT 精度高一些，是 1024 点的 FFT（快速傅里叶变换）运算方式，提高了频率的分辨率，得到原信号的更准确瞬间频谱特性。MUSICAM 广泛应用在数字演播室、数字音频广播（Digital Audio Broadcasting，DAB）、数字视频广播（Digital Video Broadcasting，DVB）等数字节目的制作、交换、存储、传输中。

3. 层Ⅲ（又称 MP3）

层Ⅲ比层Ⅱ更为复杂，它使用了多相正交滤波器组之外，还使用了 DCT 变换滤波器组，提高频率的分辨率，还应用了预测心理声学模型，使用更为复杂的量化和编码，允许不同的帧码流。MP3 广泛应用于数字无线电广播的发射和接收，数字声音信号的制作与处理，声音信号的存储，Internet 传输，消费电子产品（MP3 播放机）等方面。

5.3.2 MPEG-2

1. MPEG-2 概述

MPEG-2 定义了两种音频压缩算法。一种称为 MPEG-2 后向兼容多声道音频编码（Backward Compatible Multichannel Audio Coding，MPEG-2BC），它与 MPEG-1 音频编码算法是兼容的，考虑前、后兼容以及多声道环绕声等特点，在压缩算法承袭了 MPEG-1 的绝大部分技术，并为在低码率条件下进一步提高声音质量，还采用了多种动态传输声道切换、动态串音等新技术。事实上，正是由于 MPEG-2BC 与 MPEG-1 的兼容性，使其不得不以牺牲数码率的代价来换取较好的声音质量，一般情况下，MPEG-2BC 需 640kb/s 以上的码率才能基本达到 EBU（European Broadcasting Union，欧洲广播联盟）“无法区分”声音质量的要求。由于 MPEG-2BC 标准化的进程过快，其算法自身仍存在一些缺陷。这一切都成为 MPEG-2BC 在世界范围内得到广泛应用的障碍。另一种称为 MPEG-2 先进音频编码（Advanced Audio Coding）标准，简称 MPEG-2 AAC，它放弃了原有的兼容性要求，显著地提高编码的效率，因为它与 MPEG-1 音频编码算法是不兼容的，又被称为 MPEG-2 NBC（Non Back Compatible）编码。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。AAC 标准可支持 48 个主声道、16 个低频增强声道、16 个配音声道和 16 个数据流。MPEG-2 AAC 在压缩比为 11∶1，即每个声道的数码率为（44.1×16）/11＝64kb/s，5 个声道的总数码率为 320kb/s 的情况下，很难区分还原后的声音与原始声音之间的差别。与 MUSICAM 相比，MPEG-2 AAC 的压缩比可提高 1 倍，而且音质更好；与 MP3 相比，在质量相同的条件下数码率是它的 70%。

MPEG-2 AAC 压缩编码原理：为了实现低比特率的数据流、提高编码效率，采用去除声音信号中的冗余度及无关分量的做法是基本原则。但因采用的措施不同，降低比特率的程度也随之不同。AAC 采用音频采样信号和采样样本统计特性之间的关系除去冗余；利用人耳听觉系统在频域和时域中的掩蔽效应除去不可闻的无关分量以及利用心理声学模型对声音信号进行量化和无噪声编码。AAC 编码原理方框图如图 5-9 所示，时域里的 PCM 信号先通过滤波器组（进行加窗 MDCT 变换）分解成亚采样频谱分量，变成频域信号，同时时域信号经过心理声学模型获得各子带的掩蔽阈值、M/S 以及强度立体声编码需要的控制信息，还有滤波器组中应使用长短窗选择信息。时域噪声整形（TNS）模块将噪声整形为与

能量谱包络形状类似，控制噪声的分布。强度立体声编码和预测以及 M/S 立体声编码都能有效降低编码所需比特数，随后的量化模块用两个嵌套循环进行了比特分配，并控制量化噪声小于掩蔽阈值，之后就是改进了的哈夫曼编码。这样，与前面各模块得到的边带信息一起，就能构成 AAC 码流了。把上述过程逆过来就是解码。

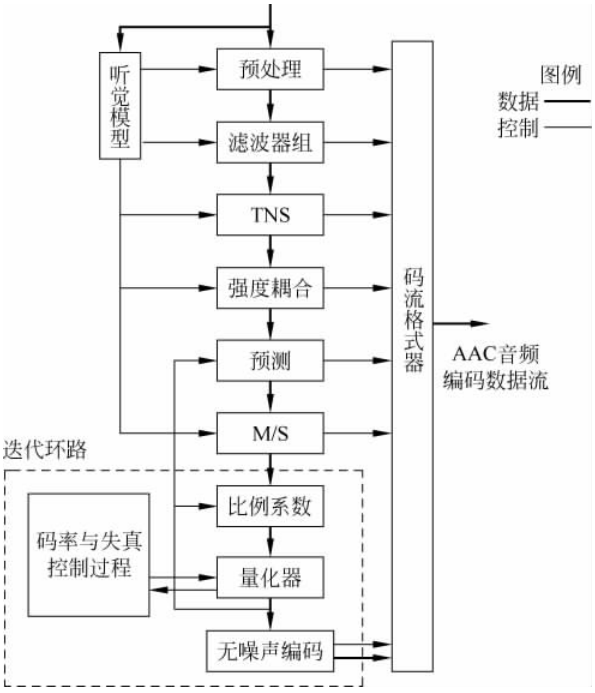

图 5-9　MPEG-2 ACC 编码器方框图

为了能够适应于不同的应用场合，在 AAC 标准中定义了三种不同复杂度的子集(Profile)框架。分别为：

(1) 主子集(Main Profile)或主框架，在这种框架具有最高的复杂度，可以用于存储量和计算能力都很充足的场合。在这种框架中，利用了除增益控制以外的所有编码工具来提高压缩效率。

(2) 低复杂性子集(Low Complexity Profile)或 LC 框架，这种框架用于要求在有限的存储空间和计算能力的条件下进行压缩的场合。在这种框架中，没有预测和增益控制这两种工具，TNS 的阶数比较低。

(3) 可伸缩采样子集(Scalable Sample Rate Profile)或 SSR 框架，在这种框架中使用了增益控制工具，但是预测和耦合工具是不被允许的，具有较低的带宽和 TNS 阶数。对于最低的一个 PQF 子带不使用增益控制工具。当带宽降低时，SSR 框架的复杂度也可降低，特别适应于网络带宽变化的场合。

2. MPEG2 AAC 系统描述

1) 系统框图

编码框图其整体 AAC 编解码系统，如图 5-9 所示，其编码流程概述如下：当一音频信号送至编码端时，会分别送至听觉心理模型以求得编码所需之相关参数及增益控制(Gain

Control)模块中,使信号做某个程度的衰减,以降低其峰值大小,如此可减少 Pre-echo 的发生。之后,再以 MDCT 将时域信号转换至频率域,而送入至 TNS(Temporal Noise Shaping,暂时噪音成形)模块中,来判断是否需要启动 TNS,此模块系利用开回路预测(Open-loop Prediction)来修饰其量化噪声,如此可将其量化噪声的分布,修饰到原始信号能量所能包括的范围内,进一步减少 Pre-echo 的发生,若 TNS 被启动,则传出其预测差值;反之,则传出原始频谱值。AAC 为了提升其压缩效率,则使用了 Joint Stereo Coding 与预测(Prediction)模块来进一步消除信号间的冗余成分。在 Joint Stereo Coding 中又可分为 Intensity Stereo Coding 与 M/S Stereo Coding。在 Intensity Stereo Coding 模块中,是利用信号在高频时,人耳只对能量较敏感,对于其相位不敏感之特性,将其左右声道之频谱系数合并,以节省使用之位;在 M/S Stereo Coding 模块中,利用左右声道之和与差,做进一步地压缩,若其差值能量很小,如此便可以用较少之位编码此一声道,将剩余之位应用于另一声道上的编码,如此来提升其压缩率。而预测模块的主要架构是使用 Backward Adaptive Predictors,利用前两个音频帧来预测现在的音频帧,若决定启动此模块,则传出其预测差值,如此一来可以减少其数据量,达数据压缩之目的。经过上述处理频谱信号上的压缩 tools 程序后,则将其数据予以量化与编码,为了达到量化编码的最佳化,AAC 使用了双巢状式循环(Two Nested Loop)的量化编码结构,以得最佳的压缩质量,最后则将其位串送至解码端,而完成整个编码程序。

表 5-2 三种不同 profile 所需使用的 tools

Took Name	Main	LC	SSR
Noiseless	coding	Used	Used
Quantizer	Used	Used	Used
M/S	Used	Used	Used
Prediction	Used	Not	Use
Intersity/Coupling	Used	Not	Use
TNS	Used	Limited	Limited
Filter	Bank	Used	Used
Gain	Control	Not	Use

为了允许其系统可对音频质量与内存/处理功率要求之间做一舍取,因此 AAC 系统提供了三种 profiles:Main profile、Low Complexity(LC) profile、Scaleable Sampling Rate (SSR)profile。且每一种 profile 所使用的 tools 皆不同,表 5-2 表示其三种不同 profile 所需使用的 tools。

2) MPEG2 AAC 码流格式与数据结构层次

MPEG2 AAC 规定了 2 种码流格式:ADIF 和 ADTS,前者用于属于文件格式用于存储;后者属于流格式,用于传输。

MPEG2 AAC 规定 1024 个 sample 数据为一个 frame,一个 frame 的 sample 从时域通过 MDCT 映射到频域时,由于引入 50%交叠,所以变成 2048 个谱线。如果是长块变换,则一个 frame 只有一个 window group,每个 window group 有一个 window,每个 window 有 2048 个谱线。如果是短块变换,则可能有若个 window group,每个 window group 可能有若干个 window,但是所有 window group 的 window 个数加一起一定为 8 个,此时每个

windows有256个谱线。需要注意的是：分window group的意义在于同一个window group的谱线数据使用一个scalefactor。而每个windows又可以分为n个section($1 \leqslant n \leqslant$ max_sfb,“一个frame内最多的scalefactor band的个数”),每个section有若干个谱线数据(Spectral Data),但需要注意,section的边界必须和scalefactor band的边界重合,所以也可以说每个section有若干个scalefactor band。提出section的意义在于统一个section的谱线数据(Spectral Data)使用同一个huffman table编码。

MPEG2 AAC提出的window group和section的个数都是不确定的,所以编码端要在比特流中加入相关的side info用来指示window group和section分割方法。在isc_info()中的scale_factor_grouping和section_data()中的sect_len_incr就是起到这样的作用。

3) 码流解析

码流可以分为side info的解析和压缩数据的解析,side info解析出的状态信息和控制信息都使用定长码。解码只要按照格式解析出来即可。由于解码简单和篇幅限制,本书就不再提及,请查阅13818-7标准语法部分。其次是对压缩数据的解析,压缩数据属于无损编码,主要是变长码。

3. MPEG-2 AAC的主要技术

1) 滤波器组与块交换(Filter Bank and Block Switching)

滤波器组(Filter Bank)是MPEG-2 AAC中一个基本的组件,扮演着将音频信号从时间域转换至频谱域之表示,其在解码端则反向处理。对Filter bank而言,它必须具备对音频编码有着完美的重建的特性,然而,有时其音频还原似乎不是如此完美,其主要因素在于,处理时间域转换至频谱域时的音频信号,是以逐帧(frame by frame)的方式送至Filter bank,也就是将目前的音频信号切割成多个音频帧来处理,因而会造成音频帧间的边缘信号,给予不同精确度的编码,并造成信号的不连续性,都将造成日后还原时,所发生的质量影响。这种效应,称之为块效应(Blocking Effect),为了解决此问题,其块间的信号在送入Filter bank之前,一个overlapping windowing的方式将被采用,以减少其信号不连续性。

2) MDCT and IMDCT

在AAC或其他音频信号的编解码器上,最普遍解决上述问题的filter bank设计,即为在编码端上的MDCT(Modified Discrete Cosine Transform)及解码端上的IMDCT(Inverse Modified Discrete Cosine Transform)。MDCT/IMDCT使用了一种技术,称为TDAC(Time Domain Aliasing Cancellation),它使用了一种名为window-overlape-add的处理方式来消除时间域上的交迭(aliasing),如图5-10所示为AAC Filter bank的框图表示,对一个输入音频信号的目前音频帧,是取前一个音频帧的后面50%与目前输入的音频帧音频值前50%作为此次处理的音频。

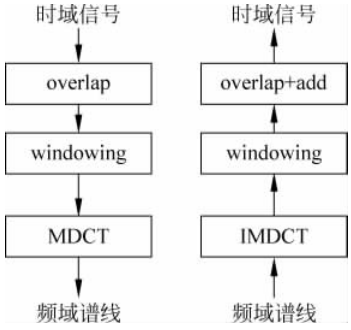

图5-10　AAC Filter bank的框图表示

3) 窗块切换(Windowing and Block Switching)

对一个MDCT filter bank的频率响应的分辨率改善,进来的音频信号在经MDCT转换前,需经过一个window function相乘后才送至MDCT。AAC支持两种window shapes,

即 sin window 及 KBD(Kaiser-Bessel Derived) window，KBD window 可以比 sine window 更准确地重建出原始的时间域的信号。在 MPEG-2 AAC 系统中，可以允许其 KBD 及 sin window 的切换，来达到最好的用来接受输入信号的状态，而得到更好的音质重建结果。

另外，MPEG-AAC 编码器中，为了在声音特性、编码效率与声音压缩质量上取得适合的块长度，总共提供 $N=256$(短块)与 $N=2048$(长块)两种块长度作为选择。其块的选择，是根据听觉心理模型(Psychoacoustic Model)的 PE 值来决定。

通常，长块的使用可以被选择来减少其信号的冗余部分，并得到较高的频率分辨率，来改善编码质量，但是也可能对于某些瞬时信号产生问题。一般地，当音频信号在时间域上有变化较大的瞬时信号(Transient Signal)时，则以连续的 8 个短块来处理，可以提升在音频压缩时的精确度，并减少 pre-echo 的发生；相对地，当音乐数据属于稳态的信号(Stationary signal)，则使用长块来处理。而在长短块转换中，还存在着两种缓冲块，长块切换到短块必须经过起始块(Start Block)才切换到短块，从短块切换到长块也必须经过停止块(Stop Block)才切换到长块。图 5-11 显示了其块切换方式。

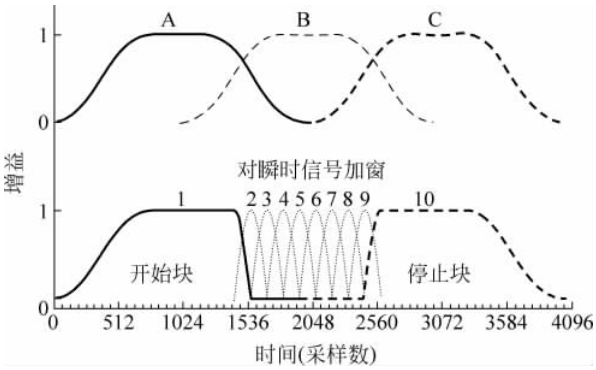

图 5-11　短块与长块块切换方式

4) TNS

由于 MPEG-2 AAC 的块大小比 MPEG-1 layer3 的还要大，因此，一般在处理单一长块信号时，假如在时间上有一个急剧变化的信号变化时，经由在时间域与频率域上的信号转换，再经量化后，转回其时间域时，有可能会增加造成 pre-echo 的现象发生。而 pre-echo 的发生，从时间域上的遮蔽效应可发现，若一较高的能量是在转换长块的前半部时，其经由量化所产生的噪声，可能被 post-masking 遮蔽，但是若较高的能量是在长块的后半部时，则散布到前半部的噪声将无法被 pre-masking 遮蔽，这就是由于对长块而言，其在时间域上的分辨率较低，因此噪声分布范围超过 pre-masking 的遮蔽范围，而造成量化的噪声将被人耳所听到，此现象，就是称为 pre-echo。

如图 5-12 所示为 pre-echo 现象发生所造成时间域上信号的失真。减少 Pre-echo 现象有许多种方式，如经由动态地切换块大小可解决此一问题，另外，在 MPEG-2 AAC 中加入了 TNS，也是用来减少 pre-echo 的现象。而 TNS 概念是使每个单一块再经过 TNS 编解码后，将量化噪声的分布能被原信号所遮蔽。

在编码端，首先将经过 MDCT 模块的频域信号送入，利用 Levinson- Durbin recursion

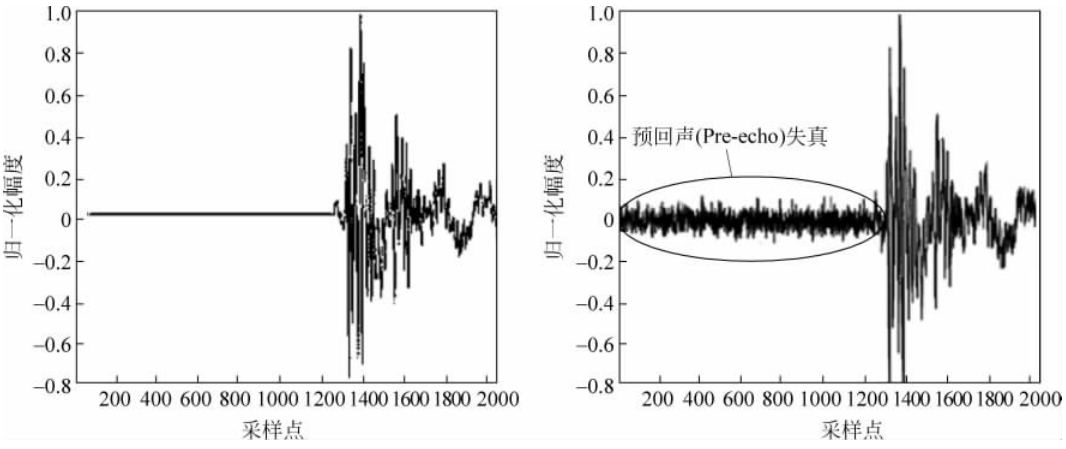

图 5-12　pre-echo 现象

方法取得此音频块的反射系数(Reflection Coefficients)与预测增益(Prediction Gain),当求得的预测增益大于 MPEG2-AAC 标准中所设定的常数值,则使用 TNS 模块。首先,为了减少反射系数传送所需的比特率,将反射系数进行量化,然后再经过 Truncate coefficients 来调整 TNS 系数的阶数,开始会根据不同 profile 所定义的系数阶数,将后面系数小于 0.1 的系数值舍去,来调整系数阶数,使得产生的 LPC 预测系数阶数少于 profile 定义的阶数,最后将反射系数经过计算求出 TNS 的预测系数,送入 TNS filter(MA)中。如果没有启动 TNS,则将原始的频域谱线送出。在编码端只需要传送量化后所需要的反射系数阶数以及整数的索引值,而不必传出所有反射系数的信息。给 Index 及 TNS order 的信息将使用在解码端,用来还原量化后的反射系数。当解码端所接收到的反射系数阶数大于 0,代表有使用 TNS 编码,在解码端就会启动 TNS 解码模块,求得编码时的预测系数送入 TNS filter(AR)中,解码出频率域上信号的数据。

加入 TNS 模块后,也有一些 side information 的项目需加入至位串(bit stream)里,以提供解码端使用,如表 5-3 所示。由于 TNS 预测级数对于 LONG window 而言,最多为 20,对于 SHORT window 而言,最多为 7,因此,TNS 在编码中对 side information,最多增加的位数目为:

- LONG window: $1+1\times(2+1+6+5+1+1+4)=97$bits;
- SHORT window: $1+8\times(1+1+4+3+1+1+7\times4)=313$bits(Joint Stereo Coding)。

MPEG AAC 的系统为了提升其编码效率及压缩质量,Joint Stereo Coding 利用了左右声道的特性,对立体音编码引进了两种技术,即 M/S Stereo 与 Intensity Stereo。

表 5-3　side information

side information	位宽	注释
TNS Present or not	1	
Number of filters	2/1	长/短窗
TNS coefficients resolution	1	
TNS filter length in band	6/4	长/短窗

续表

side information	位宽	注释
TNS filter order	5/3	长/短窗
TNS filter direction	1	
Coefficient compress or not	1	
Bit per coefficient	4	

5) M/S Stereo

在 MPEG-2 AAC 系统中，M/S(Mid/Side) Stereo coding 被提供在多声道信号中，每个声道对(Channel Pair)的组合(也就是每个通道对)是对称地排列在人耳听觉的左右两边，其方式简单，且对位串不会引起较显著的负担。一般地，当其在左右声道数据相似度大时，常被用到，并需记载每一频带的四种能量临界组合，分别为左、右、左右声道音频合并($L+R$)及相减($L-R$)的两种新的能量。然后再利用听觉心理学模型与滤波器来处理。一般地，若所转换的 Sid 声道的能量较小时，M/S Stereo coding 可以节省此通道的位数，而将多余的位应用于另一个所转换的声道，即 Mid 声道，进而可提高此编码效率。对 M/S Stereo coding，可以选择性地切换其在时间域上块与块间是否使用的时机，其切换的旗标(ms_used)将被设定与否而传送至解码端上。

6) Intensity Stereo

对低频信号而言，人类听觉系统对信号的能量与相位皆较敏感，相对于在高频信号，人耳只对其能量较为敏感，而对其相位较不敏感。Intensity Stereo coding 就是利用人耳的这种特性，被使用在高频区域里，声道对之间的不相关性条件下，这个方式在过去对立体声或多声道编码中已广泛使用，又可称为 dynamic crosstalk 或 channel coupling。其编码是利用这一因素来完成，也就是在高频声音组件的接收感觉，主要是依赖在它们的能量分析上，即时间封包(Time Envelopes)。因此，它对某些型式的信号就有可能仅需传送单一频谱值来达到，其他音频的声道在不影响其质量的情况下，可以虚拟地由此一频谱值被表示出来。而原始编码声道的能量，即 time envelopes，对每一个 scalefactor band，经由一个调整(Scaling)大小的运算因子，近似地被表示而储存，使得在解码端，对每一个声道的信号，可借由此一因子来重建。

7) 量化编码

在完成之前的频谱处理的工具后，实际位率减少是在量化处理中来达到，这个模块主要的目的是量化频谱上的数据，使得量化噪声能够满足声音心理模式的要求。迭代循环(Iteration Loop)模块被用来决定量化的 step size，并保证其允许的失真不会超过，并在满足迭代循环后，非线性的量化函数被执行。另外，对于每一个音频帧被量化的有效位数，也需在某个临界之下，一般其值与取样率及所要求的位率有关。在每个音频帧开始计算时，先将一些所需的变量初始化，如果此音频帧里所有的频域数据皆为 0，则可以跳过此音频帧不作处理，如果有频域数据，则将进入 outer iteration loop，开始进行频域数据的量化与位计算，最后将未使用的位数，保留到下一个音频帧时继续使用。

8) 无损解码

无损解码 ics_info()的参数如表 5-4 所示。

表 5-4　解码 ics_info()

	位宽	作　　用
ics reserved bit	1	一定为 0
window sequence	2	窗口类型 00：长窗　01：起始窗 10：短窗　11：结束窗
window shape	1	决定使用正弦窗还是 KBD 窗 0：正弦窗 1：KBD 窗
max sfb	4/6	短窗下 4 位，其他时 6 位。表示每个窗组内的 scalefactor band 的个数
scale factor grouping	7	在短窗时有效。指明 window group 的分割方式。7 个 bit 表示 8 个窗中的 1-7 窗的分组情况。即 bit(8－n)表示 window(n)的分组属性，当 bit(8－n)＝'1'表示 window(n)和 window(n－1)是同一个组，若 bit(8－n)＝'0'表示 window(n)和 window(n－1)是不是同一个组
predictor data present	1	指示码流中是否出现预测数据
predictor reset	1	指示预测器是否全部复位
predictor reset group number	5	指示预测器组是否复位
prediction used	1	指示每个 scalefactor band 是否是由预测器

5.3.3　MPGE-4 HE-AAC

MPEG-2 AAC 通过进一步改善和增补，增加了知觉噪声代替(Perceptual Noise Substitution，PNS)等技术，使之发展成为 MPEG-4 音频标准，MPEG-4 高效先进音频编码(High Efficiency Advanced Audio Coding，HE-AAC)是由 AAC 主体，加上频带复制(Spectral Band Replication，SBR)技术组合而成的编码算法。以往的声音压缩受限于感知编码，使得高频段的声音容易产生失真，尤其是在使用低码率压缩时，因数据量大幅减少，声音质量让使用者无法接受。SBR 是数字音频中一种提高效率的压缩工具，它可以大大提升使用低码率压缩时的声音质量。

HE-AAC 承袭了 AAC 的所有优点，利用 SBR 这一种独特的带宽扩展技术来改善音频中高频段的失真现象，而所谓的高效率在于编码器仅需对低频部分编码，高频部分利用低频信号，配合一组数据量极少的参数来重建，SBR 能够使编码器仅以一半的比特率传送同质量的音频信号。

如图 5-13 所示，在编码部分，对原始的音频输入信号进行分析，其高频的谱包络及与低频相关的特性将被编译，形成 SBR 数据，与核心的编码数据流进行多路复用。在解码部分，SBR 数据首先被分离解码，解码过程分成两个阶段，第一阶段，核心的编码数据流产生低频带信号；第二阶段，以 SBR 解码作为后处理进行操作，利用解出的 SBR 数据指导频带复制过程，从而得到全频带输出信号。

MPEG-4 HE-AAC v2 是 HE-AAC 的完全扩展集，它是由 MPEG-4 音频标准技术中的 AAC、SBR、PS 三种技术结合而成的最高效的音频编码方法。为进一步提高低速率立体声

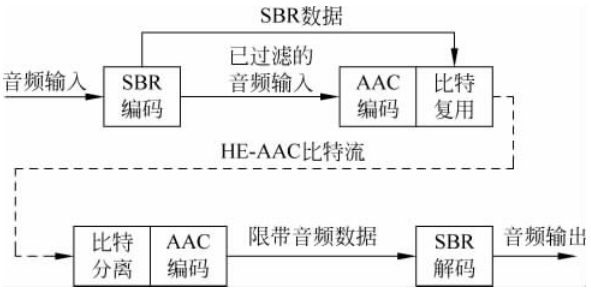

图 5-13　HE-AAC 编解码方框图

编码的性能，增加基于 SBR 框架的参数立体声(Parametric Stereo，PS)编码技术。参数立体声编码的基本思想就是传输一种描述立体声图像的数据，而这种数据是以混合单声道的边信息的形式传送的，这种立体声边信息非常简洁，虽然只占整个比特流很小的一部分，却可以使整个比特流的单声道信号获得最好的品质。PS 技术能够使低比特率的立体声信号，在编解码的效率上增加一倍。

SBR 和 PS 都是兼有前向兼容和后向兼容特性的技术，可以提高任何一种音频的编解码效率。正因为采用这几种新的高压缩比技术，HE-AAC v2 能够在速率为 128kb/s 时传输流媒体和可下载的 5.1 多声道音频信号，在速率为 32kb/s 时传输准 CD 音质，在速率为 24kb/s 时传输优质立体声音质，甚至在速率低于 16kb/s 的单声道方式下还能对混合音频内容传输较好的音质。

MPEG-4 HE-AAC 在许多国际标准化组织中都已经广泛采用。在第三代移动通信合作伙伴计划(3rd Generation Partnership Project，3GPP)中，MPEG-4 HE-AAC v2 被指定为高效音频编解码标准，新一代的数字广播 DAB＋中也采用 HE-AAC 作为信源编码技术。此外，它还在互联网流媒体联盟(Internet Streaming Media Alliance，ISMA)、3GPP2、HDTV、DVB、DVD 论坛以及其他许多标准化组织、论坛中都采纳为其规范之一。2009 年杜比实验室最新推出的 Dolby Pulse 技术也是建立在 MPEG-4 HE-AAC 标准开放音频技术之上，并与之兼容，Dolby Pulse 将 HE-AAC 先进的编码效率与杜比音频技术的性能、特点、一致性、兼容性很好地集于一身。

5.3.4　MPEG 通用语音与音频编码算法

为进一步促进语音频通用编码技术的发展，运动图像专家组于 2007 年首次提出了构建语音频通用编码器的倡议。对此，Fraunhofer IIS 和 VoiceAge 公司提出了基于类型判别的混合编码方式，该编码器将输入信号分为语音、音频两类来分别处理，采用 HE-AAC 与 AMR-WB＋相结合的方式来实现对音频和语音信号处理的无缝切换。最终，这种 AMR-WB＋与 HE-AAC 相结合的参考模型被 MPEG 组织所采纳，成为了 MPEG 通用语音频编码(Unified Speech and Audio Coding，USAC)算法，其基本原理如图 5-14 所示。

如图 5-14 所示，编码器对于输入的语音信号采用 AMR-WB＋编码，其编码原理如前所述；音频信号采用 HE-AAC 编码方法，该编码方法在低频段以 AAC 编码方法为内核，AAC 编码器是一种变换域编码方法，主要包括三个模块：

- 时频变换，将时域音频信号映射到变换域；

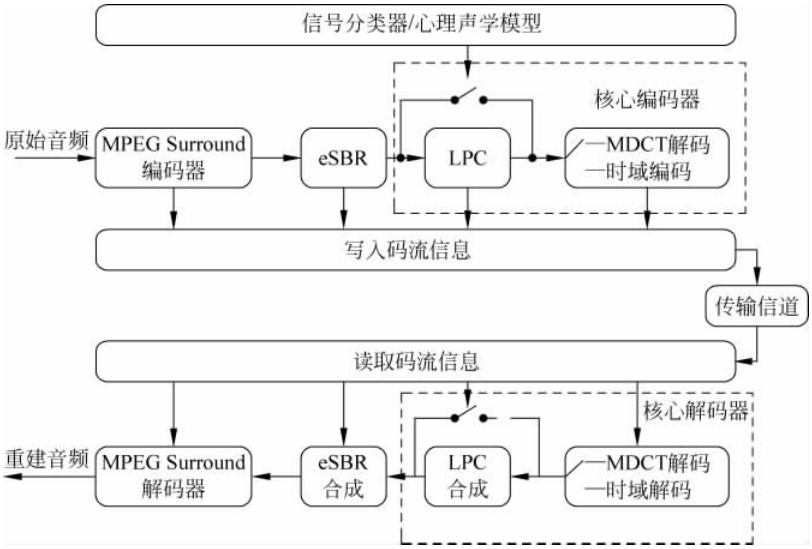

图 5-14　MPEG USAC 编/解码原理框图

- 参数量化，该模块所产生的量化误差由心理声学模型计算得到的输入信号的听觉感知特性来控制；
- 熵编码，对于量化后的频域参数和边信息通过熵编码的方式写入码流。

因此，AAC 编码器是一种由源信号控制的可变码率的编码器，结合信号自身的统计信息及听觉感知特性实现对音频信号的高质量恢复。而对于高频段音频信号，则利用频带复制技术(Spectral Band Replication，SBR)，通过复制低频段频谱信息，同时结合信号噪声和谐波性参数来恢复高频信息，从而实现码率压缩。

该编码方法充分结合了 HE-AAC 和 AMR-WB＋编码器对音频和语音信号处理的优势，对于输入的信号首先进入 SBR 模块，进行高频参数的提取，对于信号的低频成分则采用 LPC 和 MDCT 编码方式进行处理。利用信号分类模块以及心理声学模型对输入信号进行判别，在编码时，时域波形编码与 MDCT 域编码方法同时进行，当输入信号为语音信号时，则启动 LPC 处理模块，若输入信号为音频信号，则仅采用 MDCT 域编码方式进行处理。由于 HE-AAC 和 AMR-WB＋编码器处理帧长不同，LPC 编码与 MDCT 编码处理方式也存在很大区别，因此利用该模型实现对音频和语音信号的通用编码所需解决的核心问题是如何实现不同处理方法之间的快速切换和平滑过渡。当出现两种处理模块的切换，需要解决如下问题：

(1) MDCT 域(HE-AAC)，时域和 LPC 滤波器(AMR-WB＋)之间的过渡；

(2) 消除模式切换所产生的人工产物；

(3) 在帧长不变的情况下，实现两种模式的平滑过渡。

为解决以上问题，该编码方法通过增加特殊的修正离散余弦变换(Modified Discrete Cosine Transform，MDCT)窗来消除因各帧编码模式不同所带来的块效应。然而该算法依赖于分类判别的准确率，对于语音与音频混合信号处理效果并不理想。

5.3.5* 语音频编码的未来发展方向

通过对语音和音频编码技术发展历程的回顾，可以发现，语音频编码技术在传输信号带宽、编码速率和应用场景方面都取得了长足的发展。在传输信号带宽方面，语音频编码标准的发展表现出从窄带(8kHz 采样)到宽带(16kHz 采样)，再到超宽带(32kHz 采样)，最终到全频带(48kHz 采样)的发展趋势；在编码速率方面，由起初的定速率编码(G.711、G.721)，到多速率编码(G.722.1、G.722、MPEG-1/2)，最终发展到具有可变速率、可分层的嵌入式编码(G.729.1、G.718)；在应用场景方面，从 IP 电话通信到移动互联网通信，再到如今的 3G/4G/5G 移动通信网络。

上述语音、音频编解码技术发展状况和趋势表明，随着 3G/4G/5G 移动通信的发展，以手机电视、移动音乐、流媒体音乐，以及移动音视频会议等为代表的诸多移动多媒体应用将快速发展，同时大容量存储器和宽带网络的发展，使得人们对传输带宽和传输速率的要求逐渐放宽，而随之提高了对音频质量的要求。在这种趋势下，当前对语音、音频编码的研究主要集中在空间音频编码和适合未来移动通信的语音与音频通用编码等方面。

空间音频编码(Spatial Audio Coding，SAC)是一种基于空间听觉线索的压缩编码技术，它通过高效提取和重建空间听觉信息，实现低码率高质量多声道音频编码。空间音频编码的目标是利用空间听觉冗余，以尽可能低的码率传送高质量的多声道音频信号，通过与空间视频编码技术相结合，在保证重建音频质量的前提下，完成对现实场景的完美重现。空间音频编码的基础理论是由 C. Faller 和 F. Baumgarte 提出的双耳线索编码理论，通过提取声道间的差异信息以及相关度信息实现多声道压缩编码。随后，Coding Technology、Fraunhofer IIS、Philips 以及 Agere 等 4 家国际研究机构对空间音频编码展开研究，并于 2005 年共同提出了 MPEG 空间音频编码架构，进而通过融合其他技术，最终形成 SAC 的参考模型。2006 年 7 月，MPEG 经过对 SAC 模型的不断校正和改进，颁布了世界上第一个空间音频编码标准 MPEG Surround。与传统多声道编码相比，空间音频编码在相同的音质下可有 1/2～2/3 的码率下降，在满足人们较高音质需求的同时，也减轻了目前高质量多声道音频信号对传输和存储上的压力，从而使其在广播、Internet 流媒体等领域有着巨大的应用前景。

而语音与音频通用编码作为当前语音、音频编码的另一个研究热点，其研究目标是利用统一的编码框架，实现对包括语音、音乐、语音和音乐的混合(混合音频)等复杂音频信号的高效编码，从而弥补单一类型的语音、音频编码方式仅适合处理一种类型信号的不足。

在语音与音频编码领域，目前的压缩算法大致可以分为两类，一类为基于线性预测的参数编码；另一类为基于变换的编码。基于线性预测技术的编码通常基于语音信号发声的源-系统模型，通过分析/合成的方式去除语音信号远样点和近样点之间的冗余，实现对语音信号的高效压缩。然而，由于该模型不符合音乐信号的产生机理，因此利用其处理音乐信号时，会产生明显的编码噪声。基于变换的编码通常采用基于心理声学模型的波形编码方法，适合对音乐信号进行编码，但它所需码率较高。利用其对语音信号进行编码时，在取得相同音质的条件下，压缩效率远低于基于线性预测技术的编码方法。然而，在移动多媒体应用场

---

* 编辑注：章节号标有“*”号，表示本章节为选读内容。全书同。

景中涉及的音频内容较为复杂，并非是单纯的语音或音乐信号，通常包括自然音、语音、音乐以及语音和音乐的混合音频，因此要求编码算法必须能够无差别地对上述较为复杂的音频信号实现高效编码。

基于此，人们提出了多种解决方案，如将线性预测与变换编码相结合的变换预测编码(TPC)算法。该算法以线性预测编码技术为核心，通过开环或闭环的方式，将预测残差在频域量化，通过在时域分辨率和频域分辨率之间取得折中，实现对语音和音频信号的通用编码，其典型代表为第三代合作伙伴计划 3GPP 于 2005 年制定的扩展的自适应多码率宽带(Extended Adaptive Multi-Rate-Wideband，AMR-WB＋)语音/音频编码标准。另外，Fraunhofer IIS 和 VoiceAge 公司也于 2009 年联合提出了一种基于信号分类的混合编码方法，该方法通过在多个不同的编码器中使用开环信号分类法选择最佳的编码器编码，实现对语音和音频的通用编码。

5.4* 常用的音频信号处理软件

1. Adobe Audition

它是一个专业音频编辑和混合环境，原名为 Cool Edit Pro。被 Adobe 公司收购后，改名为 Adobe Audition。Audition 专为在照相室、广播设备和后期制作设备方面工作的音频和视频专业人员设计，可提供先进的音频混合、编辑、控制和效果处理功能。最多混合 128 个声道，可编辑单个音频文件，创建回路并可使用 45 种以上的数字信号处理效果。Adobe Audition 功能强大，控制灵活，使用它可以录制、混合、编辑和控制数字音频文件，也可轻松创建音乐、制作广播短片、修复录制缺陷。通过与 Adobe 视频应用程序的智能集成，还可将音频和视频内容结合在一起。

Adobe Audition(见图 5-15)是一个非常出色的数字音乐编辑器和 MP3 制作软件。不少人把它形容为音频“绘画”程序。你可以用声音来“绘”制：音调、歌曲的一部分、声音、弦乐、颤音、噪音或是调整静音。而且它还提供有多种特效使你的作品增色：放大、降低噪音、压缩、扩展、回声、失真、延迟等。你可以同时处理多个文件，轻松地在几个文件中进行剪切、粘贴、合并、重叠声音操作。使用它可以生成的声音有噪音、低音、静音、电话信号等。该软件还包含有 CD 播放器。其他功能包括支持可选的插件、崩溃恢复、支持多文件、自动静音检测和删除、自动节拍查找、录制等。另外，它还可以在 AIF、AU、MP3、Raw PCM、SAM、VOC、VOX、WAV 等文件格式之间进行转换，并且能够保存为 RealAudio 格式。

2. GoldWave

GoldWave 是一个集声音编辑、播放、录制和转换的音频工具，体积小巧，功能却不弱。可打开的音频文件格式种类较多，包括 WAV、OGG、VOC、IFF、AIF、AFC、AU、SND、MP3、MAT、DWD、SMP、VOX、SDS、AVI、MOV 等，也可从 CD、VCD、DVD 或其他视频文件中提取声音。内含丰富的音频处理特效，从一般特效(如多普勒、回声、混响、降噪)到高级的公式计算(理论上可以利用公式产生任何你想要的声音)，效果很多。后续版本在处理速度上有了很大提高，而且能够支持以动态压缩保存 MP3 文件。目前的版本为 GoldWave V6。

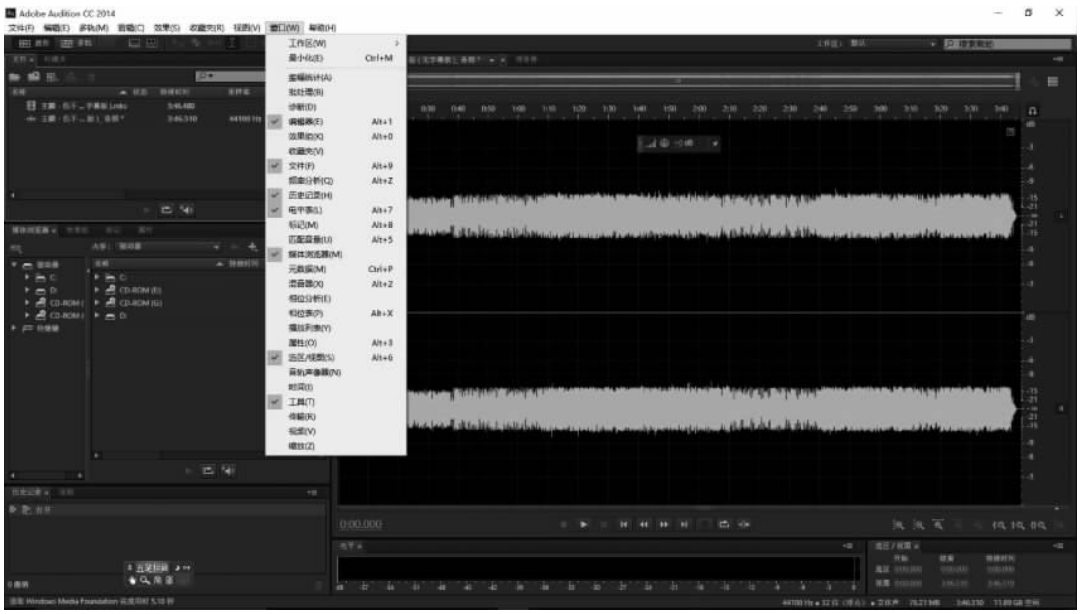

图 5-15　Adobe Audition 界面

3. NGWave Audio Editor

NGWave Audio Editor 是一个功能强大的音频编辑工具，采用下一代的音频处理技术，使用它可以在一个可视化的真实环境中精确快速进行声音的录制、编辑、处理、保存等操作，并可以在所有的操作结束后采用创新的音频数据保存格式，将其完整而高品质地保存下来。

4. All Editor

All Editor 是一个专业的多功能的声音编辑和录音工具，All Editor 是一款超级强大的录音工具，All Editor 还是一个专业的音频编辑软件，它提供了多达 20 余种音频效果供你修饰你的音乐，例如淡入淡出、静音的插入与消除、哇音、混响、高低通滤波、颤音、震音、回声、倒转、反向、失真、合唱、延迟、音量标准化处理等。

5. Total Recorder Editor

Total Recorder Editor 是 High Criteria 公司出品的一款优秀的录音软件，其功能强大，支持的音源极为丰富。不仅支持硬件音源，例如：麦克风、电话、CD-ROM 和 Walkman 等，还支持软件音源，例如：Winamp、RealPlayer、Media player 等，而且它还支持网络音源，例如：在线音乐、网络电台和 Flash 等。除此之外，还可以巧妙地利用 Total Recorder 完成一些“不可能完成”的任务。总之，“全能录音员”这一称号对 Total Recorder 来说一点都不过分。音频玩家最关心的还是录音质量，Total Recorder 的工作原理是利用一个虚拟的“声卡”去截取其他程序输出的声音，然后再传输到物理声卡上，整个过程完全是数码录音。因此，从理论上来说，不会出现任何的失真。

6. AD Stream Recorder

AD Stream Recorder 是一款流媒体录音工具，它可以对实况流媒体进行录音或者可视化分析。与同系列产品 AD Sound Recorder 可谓相辅相成之作。MP3 编码时使用的是

LAME 3.93 的 DLL 版本。它能录制 internet 主流媒体、Windows 媒体播放器播放的电影和音乐。录音和监视过程中可实时显示信号，有助于录制高质量的音频。

7. Audio Recorder Pro

Audio Recorder Pro 是一款实用、快速和容易使用的录音工具。它可录制音乐、语音和任何其他声音，并保存成 MP3 或 WAV 格式，支持从麦克风、Internet、外部输入设备（如 CD、LP、音乐磁带、电话等）或者声卡进行录制。允许预设置录音质量以帮助快速设定和管理录音参数；允许定时录制，内置增强的录音引擎，允许在录音前预设定录音设备。

8. Audacity

Audacity 是一个免费的跨平台（包括 Linux、Windows、Mac OS X）音频编辑器。可以使用它来录音、播放、输入输出 WAB、AIFF、Ogg Vorbis 和 MP3 文件，并支持大部分常用的工具，如剪裁、贴上、混音、升/降音以及变音特效等功能。具有混合音轨和给录音添加效果的功能。它还有一个内置的封装编辑器、一个用户可自定义的声谱模板和实现音频分析功能的频率分析窗口。

其他音频处理软件还有 Sound Forge、Logic Audio、Samplitude、Vegas Audio、Nuendo、Band in a Box、Guitar Pro、T-Racks 等，此处不再一一介绍。

5.5* 常见的音频格式

按照压缩后的数据是否有信息丢失，数据的压缩编码分为无损压缩和有损压缩。其中，简单来说，无损压缩就是对压缩数据进行还原之后得到的数据与原来的数据是完全相同的，也称冗余压缩方法，它利用数据的统计冗余进行压缩，解码后的数据与压缩编码前的数据严格相同，没有失真，是一种可逆运算。这类方法的压缩比例一般不高，仅使用无损压缩方法不可能解决音频数据的存储和传输问题。有损压缩方法也称信息量压缩方法，它利用了人类听觉对声音的某些频率成分不敏感的特性，允许压缩编码过程中损失一定的信息。换句话说，解码数据和原始数据是有差别的。

5.5.1 无损压缩的音频编码文件格式

1. WMA

WMA（Windows Media Audio）格式是微软公司开发的基于互联网流媒体应用的数字音频压缩算法，音质要比 MP3 格式更好，以减少数据流量但保持音质的方法来达到比 MP3 压缩率更高的目的，WMA 的压缩比一般都可以达到 18∶1 左右。WMA 的另一个优点是内容提供商可以通过 DRM（Digital Rights Management，数字版权管理）方案加入防复制保护，这种内置的版权保护技术可以限制播放时间和播放次数，甚至限制播放的机器等等。另外，WMA 还支持音频流技术，适合在网络上在线播放，作为微软抢占网络音乐的开路先锋。

2. APE

APE 是流行的数字音乐无损压缩格式之一，因出现较早，在全世界特别是中国大陆有着广泛的用户群。与 MP3 这类有损压缩格式不可逆转地删除（人耳听力不敏感的）数据以缩减源文件体积不同，APE 这类无损压缩格式，是以更精练的记录方式来缩减体积，还原后

的数据与源文件相同，从而保证了文件的完整性。APE 由软件 Monkey's audio 压制得到，开发者为 Matthew T. Ashland，源代码开放，因其界面上有只“猴子”标志而出名。

3. FLAC

FLAC(Free Lossless Audio Codec)是一种开源的无损压缩音频编码解码技术，FLAC 注重解码的速度。解码只需要整数运算，并且相对于大多数编码方式而言，对计算速度要求很低。在很普通的硬件上就可以轻松实现实时解码。

其他无损压缩音频还有 TTA、TAK、ALAC、WMA Lossless、WavPack 等。

5.5.2 有损压缩的音频编码文件格式

1. MP3

MP3 格式诞生于 20 世纪 80 年代的德国，所谓的 MP3 也就是指的是 MPEG 标准中的音频部分，也就是 MPEG 音频层。MP3 格式压缩音乐的采样频率有很多种，可以用 64Kb/s 或更低的采样频率节省空间，也可以用 320Kb/s 的标准达到极高的音质。MP3 的编码方式是开放的，用户可以在这个标准框架的基础上自行选择不同的声学原理进行压缩处理。

2. VQF

VQF 格式实际指的是 TwinVQ(全称为 transform-domain weighted interleave vector quantization)，是日本 ntt(全称为 nippon telegraph and telephone)集团属下的 ntt human interface laboratories 开发的一种音频压缩技术。VQF 格式技术受到 yamaha 公司的支持，VQF 是其文件的扩展名。VQF 格式和 mp3 的实现方法相似，都是通过采用有失真的算法来将声音进行压缩，不过 VQF 格式与 mp3 的压缩技术相比却有着本质上的不同：VQF 格式的目的是对音乐而不是声音进行压缩，因此，VQF 格式所采用的是一种称为“矢量化编码(vector quantization)”的压缩技术。该技术先将音频数据矢量化，然后对音频波形中相类似的波形部分进行统一与平滑化，并强化突出人耳敏感的部分，最后对处理后的矢量数据标量化再进行压缩而成。VQF 的音频压缩率比标准的 MPEG 音频压缩率高出近一倍，可以达到 18∶1 甚至更高。

但是 VQF 不使用如合适的比特分配、可变长度编码等技术。虽然 VQF 的解码/还原软件体积很小(NTT 公司称其能够对应所有的 CPU)，但是 VQF 的编码/压缩软件需要极其强劲的 CPU。但 VQF 压缩速度慢。MP3 压缩速度相对较快，在压缩速度方面 VQF 比不上 MP3 文件。

其他的有损压缩的音频编码文件格式还有 Real Media、MIDI、Ogg Vorbis、AIFF、AU、VOC 和 VOX 等，此处不再一一介绍。

习题五

5-1 声音信号能进行压缩编码的基本依据有哪些？

5-2 简述语音与音频压缩的区别。

5-3 时域编码与频带编码的特征主要有哪些？

5-4 简述差分脉冲编码调制的思想及其编码原理。

5-5　简述子带编码理论的基本思想及其编码原理。

5-6　简述自适应差分脉冲编码的基本思想及其原理。

5-7　目前主流音频压缩编码标准主要有哪些？试论述它们的区别。

5-8　简述 MPEG-2 AAC 压缩编码的基本原理。

5-9　试总结现有主流无损压缩的音频编码文件格式类型。

5-10　简述矢量化编码的原理。