第3章 音频处理技术

声音是人类进行交流和认识自然的主要媒体形式，通常的声音主要是指语音、自然声和音乐。如何将声音数字化转换成数字音频，更加方便地进行传输、存储和处理成为多媒体研究的一个重要领域。数字音频信号的处理主要表现在数据采样和编辑加工两个方面。其中，数据采样的作用是把自然声转换成计算机能够处理的数字音频信号；对数字音频信号的编辑加工则主要表现在剪辑、合成、静音、增加混响及调整频率等方面。

3.1 基本概念

声音是指通过一定介质（如空气和水等）传播的一种连续波，其本质是机械振动或气流扰动引起周围弹性媒质发生波动的现象，它是一个随着时间连续变化的模拟信号，在物理学中称为声波。声波具有普通波所具有的特性，即反射（reflection）、折射（refraction）和衍射（diffraction），它有如下几个重要指标，如图3-1所示。

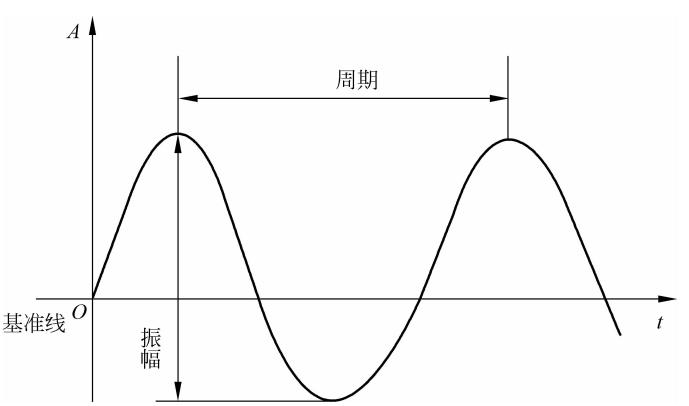

图3-1 声音关键指标

（1）基准线——提供模拟信号的基准点。

（2）振幅（amplitude）——波的高低幅度，表示声音的强弱。

（3）周期（period）——两个相邻波之间的时间长度。

（4）频率（frequency）——每秒钟振动的次数，以Hz为单位。

通常人耳的听力频率范围是20～20 000Hz，如果物体振动频率低于20Hz或高于20 000Hz人耳就听不到了，高于20 000Hz频率的称为超声波，而低于20Hz频率的称为次声波。

3.1.1 声音的基本特点

1. 声音传播方向

声音依靠介质的振动进行传播。声源实际上是一个振动源，它使周围的介质（空气、液体、固体）产生振动，并以波的形式进行传播，人耳如果感觉到这种传播过来的振动，再反映到大脑，就意味着听到了声音。

声音以振动波的形式从声源向四周传播，人类在辨别声源位置时，首先依靠声音到达左、右两耳的微小时间差和强度差异进行辨别，然后经过大脑综合分析而判断出声音来自何方。从声源直接到达人类听觉器官的声音叫做“直达声”，直达声的方向辨别最容易。现实生活中，森林、建筑、各种地貌和景物存在于人们周围，声音从声源发出后，须经过多次反射才能被人们听到，这就是“反射声”。就理论而言，反射声会影响方向的准确辨别。但实际中，反射声不会使人丧失方向感，起关键作用的是大脑的综合分析能力。经过大脑的分析，不仅可以辨别声音的来源，还能丰富声音的层次，感觉声音的厚度和空间效果。

2. 声音的三要素

声音的三要素为音调、音色和音强。就听觉特性而言，这三者决定了声音的质量。

（1）音调——代表声音的高低。音调与频率有关，频率越高，音调越高，反之亦然。当人们提高唱盘的转速时，声音频率提高，音调也提高。当使用音频处理软件对声音进行处理时，频率的改变可造成音调的改变。

（2）音强——代表声音的强度，也称“响度”，“音量”是指音强。音强与声波的振幅成正比，振幅越大，强度越大。CD 音乐盘、MP3 音乐以及其他形式的声音强度是一定的，可以通过播放设备的音量控制改变聆听的响度，也可使用音频处理软件改变声源的音强。定量描述声音强弱的方式有多种，声压和声压级就是其中的两种形式。声压是指在声场中某处由声波引起的压强的变化值，用 P 表示，单位是“帕斯卡”（Pa）。声压越大，声音也就越大。由于人耳对声音强弱的感觉并不与声压的大小成线性关系，而是大体上与声压有效值的对数成正比。因此，为了适应人类听觉的这一特性，通常对声压的有效值取对数，用其对数值来表示声音的强弱即声压级，用 SPL 表示，单位“分贝”（dB），表达式如下：

$$\mathrm{SPL} = 20\lg\frac{P_{\mathrm{rms}}}{P_{\mathrm{ref}}}$$

式中 20 为参考常量，P_{rms}是计量点的声压有效值，P_{ref}是人为定义的零声压级参考声压值，国际协议规定 $P_{\mathrm{ref}}=2\times10^{-5}$（帕），这是大多数具有正常听力的年轻人，刚刚能察觉到的 1kHz 单一频率信号（称为简谐音）存在时的声压值。

（3）音色——具有特色的声音，它与声波的形状有关，是由混入基音的泛音决定的。通常的声音分为纯音和复音两种类型。所谓纯音，是指振幅和周期均为常数的声音，一般只会出现在专用的电子设备中；复音则是具有不同频率和振幅的混合音，大自然中的声音大部分是复音。复音中最低的频率称为基频，即“基音”，它是声音的基调。其他频率复音称为“谐音”，也叫泛音，复音中的基频和谐音决定了复音的音质和音色。各种声源都有自己独特的音色，如各种乐器、不同的人和各种生物等，即使在同一音高和同一声音强度的情况下，人们也能根据音色辨别声源种类。

3. 声音的频谱与质量

声音的频谱有线性频谱和连续频谱之分。线性频谱是具有周期性的单一频率声波；连

续频谱是具有非周期性的、具有一定频带的所有频率分量的声波。纯粹的单一频率的声波只能在专门的设备中创造出来，声音效果单调而乏味。自然界中的声音几乎全部属于非周期性声波，这种声波具有广泛的频率分量，听起来声音饱满、音色多样且具有生气。

声音的质量简称“音质”，音质的好坏与音色和频率范围有关。悦耳的音色、宽广的频率范围（带宽），能够获得非常好的音质。

4. 声音的连续时基性

声音在时间轴上是连续信号，具有连续性和过程性，属于连续时基性媒体形式。构成声音的数据前后之间具有强烈的相关性。除此之外，声音还具有实时性，这对处理声音的硬件和软件提出了很高的要求。

3.1.2 数字音频文件

多媒体计算机中的声音可分为三类：波形声音（wave）、话音（voice）和音乐（music）。波形声音实际上包含所有的声音形式，它可以把任何声音进行采样量化，并且恰当地恢复出来，相应的文件格式是 WAV 文件或 VOC 文件。人的说话声虽是一种特殊的媒体，但也是一种波形，所以与波形声音的文件格式相同。音乐是符号化的声音，乐谱可转变为符号媒体形式，对应的文件格式有 MID，MP3，CMF 文件等。

数字音频文件是数字化音频的软载体，目前常用的文件格式有：

1. VOC

VOC 文件是 Creative 公司开发的波形音频文件格式，也是声霸卡（sound blaster）所使用的音频文件格式，被 DOS 平台所支持，但随着 Windows 平台的普及，已逐渐被淘汰，取而代之的是 WAV 格式。VOC 文件由文件头块（header block）和音频数据块（data block）组成。文件头包含一个标识、版本号和一个指向数据块起始的指针。数据块分成各种类型的子块，如声音数据、静音、标记、ASCII 码文件、重复、重复的结束及终止标记等。

2. WAV

WAV 文件是 Microsoft 公司开发的音频文件格式，它来源于对声音模拟波形的采样。用不同的采样频率对声音的模拟波形进行采样可以得到一系列离散的采样点，以不同的采样位数（8b 或 16b）把这些采样点的值转换成二进制数，然后存入磁盘，这就产生了声音的 WAV 文件，即波形文件。波形声音是最基本的一种声音格式，该格式记录声音的波形，因此只要采样频率高、采样字节长、机器速度快，利用该格式记录的声音文件能够和原声基本一致，质量非常高，但其文件尺寸太大，多用于存放简短的声音片段。

3. AIF

AIF/AIFF 是音频交换文件格式（Audio Interchange File Format）的英文缩写，是 Apple 公司开发的一种声音文件格式，被 Macintosh 平台及其应用程序所支持，SGI 及其他专业音频软件包同样支持 AIFF 格式。AIFF 支持 ACE2，ACE8，MAC3，MAC6 压缩，支持 16 位 44.1kHz 立体声。

4. MIDI

MIDI 是乐器数字接口（Musical Instrument Digital Interface）的缩写。它是由世界上主要的电子乐器制造厂商共同提出来的一个通信标准，规定了计算机音乐程序、电子合成器和其他电子设备之间交换信息与控制信号的方法。MIDI 文件中包含音符、定时和多达 16

个通道的乐器定义，每个音符包括键、通道号、持续时间、音量和力度等信息，所以 MIDI 文件记录的不是乐曲本身，而是一些描述乐曲演奏过程的指令。因此，MIDI 音频与波形音频完全不同，它不对声波进行采样、量化与编码，而是将电子乐器键盘的演奏信息（包括键名、力度和时间长短等）记录下来，这些消息称为 MIDI 消息，是乐谱的一种数字式描述。对应于一段音乐的 MIDI 文件不记录任何声音消息，而只是包含一系列产生音乐的 MIDI 消息（描述乐曲演奏过程的指令），播放时只需从中读出 MIDI 消息，生成所需的乐器声音波形，经放大处理即可输出。与波形声音相比，由于 MIDI 数据不是声音而是指令，所以它的文件长度非常小。半小时的立体声音乐，如果用波形文件无压缩录制，约需 300MB 的存储空间，而 MIDI 数据大约只需要 200KB，两者相差 1000 多倍。MIDI 的另一个优点表现在配音方面，由于数据量小，可在多媒体应用中与其他波形声音配合使用，形成伴音的效果，而两个波形文件是不能同时播放的。与波形声音相比，MIDI 声音在编辑修改方面也是十分方便灵活的。例如，可以任意修改曲子的速度和音调，也可改换乐器等。MIDI 的缺陷主要是无法模拟自然界中其他非乐曲类声音，文件的录制比较复杂，需掌握一定的 MIDI 创作及改编作品的专业知识，同时还必须借助于专门的工具如键盘合成器等。

根据 MIDI 的特点，在以下三种情况下比较适合用 MIDI 谱曲：

（1）长时间播放高质量的音乐；

（2）从 CD-ROM 或 DVD-ROM 等装载其他数据的同时，以音乐作为背景音响效果；

（3）用音乐作为背景音响效果，同时播放波形音频或进行文字/语言转换音乐输出。

MIDI 是目前最成熟的音乐格式，实际上已经成为一种行业标准，其在科学性、兼容性和复杂程度等各方面远远超过了其他标准（除交响乐 CD 和 Unplug CD 外），它的 General MIDI 是最常见的通用标准。作为音乐工业的数据通信标准，MIDI 能指挥各种音乐设备的运转，而且具有统一的标准格式，能够模仿原始乐器的各种演奏技巧，甚至可达到人类无法演奏出的效果。MIDI 文件的扩展名为.MID。

MIDI 设备是处理 MIDI 信息所需要的硬件设备，其基本组成包括：

（1）MIDI 端口：一台 MIDI 设备可以有 1～3 个 MIDI 端口，分别称为 MIDI In，MIDI Out，MIDI Thru。

① MIDI In：接收来自其他 MIDI 设备的 MIDI 信息。

② MIDI Out：发送本设备生成的 MIDI 信息到其他设备。

③ MIDI Thru：将从 MIDI In 端口传来的信息转发到相连的另一台 MIDI 设备上。

（2）MIDI 键盘：主要用于 MIDI 乐曲演奏，MIDI 键盘本身并不发音，当作曲人员按下键盘上的按键时，就会发出按键信息，产生的也只是 MIDI 音乐消息，经过音序器录制后才生成 MIDI 文件。这些数据可以进一步加工，也可以和其他的 MIDI 数据合并，经过编辑后的 MIDI 文件就可以送合成器播放。

（3）音序器（sequencer）：用于记录、编辑和播放 MIDI 的声音文件，音序器既有硬件形式也有软件音序器，目前大多为软件音序器。音序器可捕捉 MIDI 消息，将其存入 MIDI 文件。音序器还可以编辑 MIDI 文件。

（4）合成器：MIDI 合成器与 WAV 合成器之间没有任何关系，它们是声卡上两个独立的声音合成器单元。MIDI 文件的播放是通过 MIDI 合成器，合成器解释 MIDI 文件中的指令符号，生成所需要的声音波形，经放大后由扬声器输出，声音的效果比较丰富。MIDI 文

件也可以不经合成器直接送原 MIDI 设备播放。

目前被广泛采用的 MIDI 合成方式有调频合成(FM) 和波形表合成(wave table) 两种。

(1) 调频合成方式：其原理是根据傅里叶级数而来的，即任何一种波动信号都可被分解为若干个频率不同的正弦波，合成器利用硬件产生的若干个正弦波合成某种乐器的声音。

(2) 波形表合成：其原理是 ROM 中已存储着各种实际乐器的声音采样，合成时以查表方式调用这些样本将其还原回放。它可分为硬波表合成与软波表合成。

硬波表合成方式：该合成方式的数字声音样本被保存在 ROM 内存或 RAM(可动态更换) 内。而软波表为数字化样本保存于系统主存中，合成运算靠 CPU 完成，最终的音频合成靠声卡上的 WAV 合成器来完成。

软波表合成方式：该合成方式表实际上是针对合成 MIDI 音乐而开发的一套软件，其主要作用是控制 CPU 来完成波表 MIDI 合成器的部分功能。

波表与 FM 的最大区别就在于，FM 通过对简单正弦波的线性控制来模拟音乐乐器、鼓和特殊效果，而波表采用真实的声音样本进行回放，因此采用波表合成的 MIDI 音乐听上去更加接近自然、更具真实感，而 FM 合成的 MIDI 音乐则多带有人工合成的色彩。

5. MP3

MP3 的全称是 MPEG-1 Layer3 音频文件，是目前最流行的声音文件格式。MPEG 即动态视频压缩标准，其中的声音部分称 MPEG-1 音频层，它根据压缩质量和编码复杂度划分为三层，即 Layer1，Layer2，Layer3，分别对应 MP1，MP2，MP3 三种声音文件，并且根据不同的用途，使用不同层次的编码。MPEG 音频编码的层次越高，对应的编码器越复杂，压缩率也越高。MP1 和 MP2 的压缩率分别为 4∶1 和 6∶1～8∶1，而 MP3 的压缩率高达 10∶1～12∶1 或更高。举例来说，一个未经压缩的 50MB 的 WAV 文件，压缩成 MP3 文件时可能只有 5MB。不过，MP3 采用的是有损压缩方式，与 CD 相比音质差强人意。

由于 MP3 是压缩后产生的文件，因此需要一套 MP3 播放软件进行还原。目前 Windows 自带的媒体播放器和 Winamp 等很多软件都支持这种声音文件格式。为了降低失真度，MP3 采取“感官编码技术”，以极小的声音失真换取了较高的压缩比，这使得 MP3 既能在 Internet 上自由传播，又能把它轻而易举地下载到便携式数字音频设备(如 MP3 随身听) 中，这种便携式数字音频设备是基于数字信号处理器(DSP) 的，无须计算机支持便可实现 MP3 文件的存储、解码和播放。MP3 文件的扩展名为.MP3。

6. MP4 音乐

在 MP3 日益成为一种主流的音乐格式之后，现在又出现了 MP4。MP4 并不能望文生义地理解为 MPEG-4 或者 MPEG-1 Layer4 格式。从技术层面讲，MP4 使用的是 MPEG-2 AAC 技术，简称 A2B 的技术。它的特点是音质更加完美而压缩比更大(15∶1～20∶1)。MPEG-2 AAC 是在采样频率为 8～96kHz 时，可提供 1～48 个声道可选范围的高质量音频编码。AAC 就是先进音频编码(Advanced Audio Coding) 的缩写，它适用于从比特率为 8kb/s 单声道电话语音音质到 160kb/s 多声道超高质量音频信号范围内的编码，并且允许对多媒体进行编码/解码。它增加了诸如对立体声的完美再现、比特流效果音扫描、多媒体控制和降噪等 MP3 没有的特性，使得在音频压缩后仍能完美地再现 CD 的音质。

MP4 真正的含义由来是因为版权问题，对唱片公司来说，MP3 的缺陷就是忽视了著作者和出版者应享有的版权待遇。于是，GMO(Global Music One) 公司针对 MP3 提出了基

于 AT&T 公司授权的 AAC 改良技术——A2B 的音频压缩方法和应用，并将其命名为 MP4，其用意大概是想表明 MP4 是继 MP3 之后的一种升级换代技术，这正好契合了人们的习惯思维。

A2B 技术主要由三个部分组成：第一，AT&T 的音频压缩技术专利，它可以将 AAC 压缩比提高到 20∶1 而不损失音质；第二，安全数据库，它可以为 A2B 音乐文件创建一特定的密钥，并将此密钥置于其数据库中，只有 A2B 的播放器才能播放含有这种密钥的音乐；第三，协议认证，这个认证包含了复制许可、允许复制副本数量、歌曲总时间、歌曲可以播放时间以及经营销售许可等信息。

7. Real Audio 文件——RA/RM/RAM

Real Audio 文件是由 Real Networks 公司开发的主要适用于网络实时数字音频流技术的文件格式，如今已成为网上在线收听的标准。它将音频文件大大压缩，所以在高保真方面远不如 MP3，不过由于体积小，适合实时收听。与 MP3 相同，它也是为解决网络传输带宽资源而设计的，因此主要追求压缩比和容错性，其次才是音质。

8. CD Audio 音乐

CD Audio 音乐是 CD 唱片采用的格式，又叫“红皮书”格式，是目前音质最好的音频格式，其扩展名为.CDA。在大多数播放软件的“打开文件类型”中，都可以看到 *.CDA 格式，这就是 CD Audio 了。CD 音轨可以说是近似无损的，因此它的声音基本上是忠于原声的。CD 光盘可以在 CD 唱机中播放，也能用计算机中的各种播放软件播放。一个 CD 音频文件即一个 *.CDA 文件，这只是一个索引信息，并不真正的包含声音信息，所以不论 CD 音乐的长短，在电脑上看到的“*.CDA 文件”都是 44 字节长。注意，不能直接复制 CD 格式的 *.CDA 文件到硬盘上播放，需要使用像 EAC 这样的抓音轨软件把 CD 格式的文件转换成 WAV 格式的文件。如果光盘驱动器质量过关而且 EAC 的参数设置得当的话，这个转换过程可以说基本上是无损抓音频。CD Audio 音乐的缺点是无法编辑、文件太大。

9. AAC 文件

AAC(Advanced Audio Coding)“高级音频编码”，出现于 1997 年，基于 MPEG-2 的音频编码技术。由 Fraunhofer IIS、杜比实验室、AT&T 和 Sony(索尼) 等公司共同开发，目的是取代 MP3 格式。2000 年，MPEG-4 标准出现后，AAC 重新集成了其特性，加入了 SBR 技术和 PS 技术，为了区别于传统的 MPEG-2，AAC 又称为 MPEG-4 AAC。

10. 其他音频文件格式

除了上述常见的音频文件格式以外，还有以下几种格式：

RMI 文件：Microsoft 公司的 MIDI 文件格式，它可以包括图片、标记和文本。

SND 文件：另一种计算机的波形声音文件格式，Apple 计算机上音频文件的存储格式。

AU 文件：Sun 和 NeXT 公司的声音文件存储格式，主要用于 UNIX 工作站上。

3.1.3 音质与数据量

本书中所讲的数字音频主要指 WAV 格式的波形音频文件，它是其他格式音频文件转换的基础。数字音频的声音质量好坏，取决于采样频率的高低、表示声音的基本数据位数和声道形式。音频文件的数据量由下式算出：

$$v = fbs/8$$

式中 v 代表数据量；f 是采样频率；b 是数据位数；s 是声道数。

例如 CD 质量的参数为：$f=44.1\text{kHz}$，$b=16\text{b}$，$s=2$，则每秒钟的数据量为：

$$v=(44\,100\text{Hz}\times 16\text{b}\times 2)\div 8=176\,400\text{B}(\text{约合 }172\text{KB})$$

如果以 CD 激光盘音质(44 100Hz 的采样频率，16 位，立体声，172KB/s)记录一首 5min(300s) 的乐曲，则数据量为：

$$172\text{KB/s}\times 300\text{s}=51\,600\text{KB}(\text{合 }50.39\text{MB})$$

由计算结果看出，音频文件的数据量问题不容忽视。为节省存储空间，通常在保证基本音质的前提下，适当降低采样频率。在一般场合，人的语音采用 11.025kHz 的采样频率、8b、单声道已足够；如果是乐曲，22.05kHz 的采样频率、8 位、立体声就已满足要求。

3.2 数字音频

将时间上连续的模拟音频(自然声或其他种类的声音）转换成时间上不连续的数字音频的过程，称为音频的数字化。只有将模拟音频转换为标准数字音频信号，计算机才能进行处理。因此，无论现在的多媒体计算机功能如何强大，其内部也只能处理离散的数字信号。音频的数字化过程包括采样、量化和编码三大步骤。音频的数字化过程所用到的主要硬件设备便是模拟/数字转换器 ADC(Analog to Digital Converter）。目前记录声音主要有两种技术即模拟录音技术与音频数字化技术。

3.2.1 模拟录音

模拟磁性录音技术在数字化音频技术以前已使用多年，这一技术被广泛地用于采集和播放各种各样的声音，如音乐、配音及特殊的声音效果，至今在某些领域还被广泛应用。模拟磁性录音过程就是声→电→磁的转换过程。以录音机为例，其工作过程如图 3-2 所示。

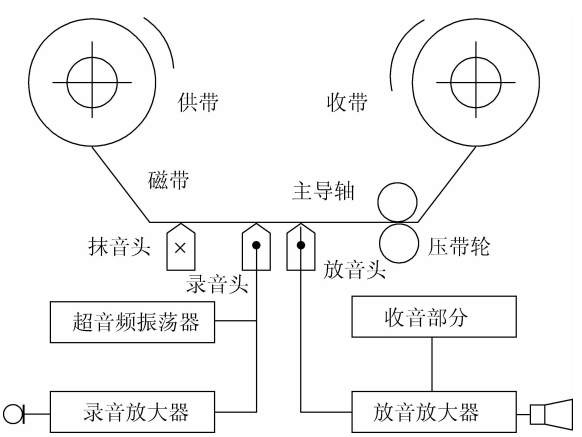

图 3-2 录音机工作过程

这种模拟录音方式是直接记录音频信号的波形，重放时用唱针扫描唱片槽纹或者用放音磁头拾取信号。模拟磁性录音性能受电磁性能的影响较大。磁带的频率特性微小的变化就会对音质产生影响。目前模拟录音动态范围可达 80dB。若进一步提高录音和放音的音质，需借助于音频数字化技术。

3.2.2 音频数字化

音频数字化与音频磁记录对于声源产生模拟电信号的捕获方式相同，所不同的是在对这种捕获后的电信号的处理方式。音频数字化处理中，并不是利用磁头以及磁头线圈进行相关的处理，而是利用硬件按照固定的时间间隔截取该音频电信号的振幅值，振幅值采用若干位二进制数表示，从而将模拟声音信号变成数字音频信号，这样就将连续变化的振动波的模拟声音信号转化为阶跃变化的离散的数字音频信号，如图 3-3 所示。

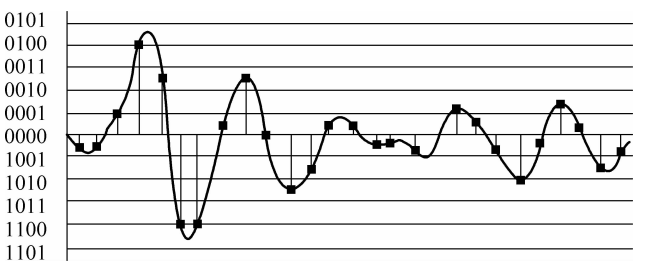

图 3-3　采样过程

截取模拟声音信号振幅值的过程称为“采样”，得到的振幅值称为“采样值”。采样值用二进制数的形式表示，该表示形式称为“量化编码”。具体的实现过程如下：

1. 采样

根据傅里叶定理，只要在连续的信号量上等间隔地取足够多的“点”，就能逼真地模拟出原来的连续量，这个取点的过程称为“采样”。每秒钟所抽取的模拟音频幅度的样本次数称为采样频率，单位为 Hz(赫)，通常使用 kHz(千赫)，即 1kHz＝1000Hz。采样频率的高低决定了声音失真程度的大小，采样精度越高(“取点”越多)，数字声音越逼真，音质就越好。当然，采集的样本数量越多，数字化声音的数据量也越大。如果为了减少数据量而过分降低采样频率，音频信号增加了失真，音质就会变得很差。采样过程如图 3-4 所示。

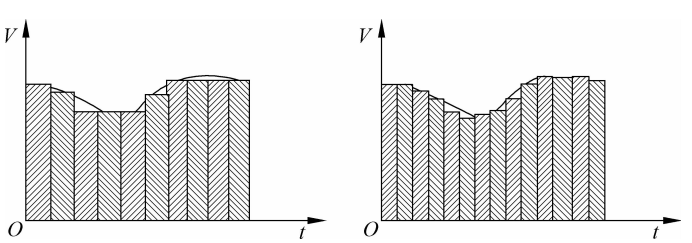

图 3-4　采样过程

采样频率的高低是根据奈奎斯特理论(Nyquist theory) 和音频信号本身的最高频率决定的。奈奎斯特理论指出，采样频率不应低于音频信号最高频率的两倍，这样就能把以数字表达的音频还原成原来的音频，这叫做无损数字化(lossless digitization) 。采样定律用公式表示为：

$$f_s \geqslant 2f \quad \text{或者} \quad T_s \leqslant T/2$$

其中 f 为被采样信号的最高频率。可以这样来理解奈奎斯特理论，例如音频信号可以看成由许多个正弦波组成的，则振幅为 A、频率为 f 的正弦波至少需要两个采样样本表示，因此

音频数据的采样频率 f 采样与声音还原频率 f 还原的关系就可以表示如下：

$$f_{采样} = 2 \cdot f_{还原}$$

目前经常用到的采样频率有 11.025kHz，22.05kHz，44.1kHz，48kHz 等。例如，人耳的可听频率范围在 20Hz～20kHz，根据奈奎斯特采样定理，为保证声音不失真，采样频率至少应保证不低于 40kHz。此外，由于每个人听力范围不同，20Hz～20kHz 只是一个参考范围，因此还要留有一定余地，所以 CD 音频通常采用 44.1kHz 的采样频率，这样的采样频率可以保证即使是采样 22.05kHz 的超声波也不会产生失真。

2. 量化

如图 3-5 所示，把整个声波振幅划分成有限个小等份，每一个小等份赋予一个相同的值，则每个取样点可用振幅的等份数量来描述精度，这些等份值在计算机中用若干位二进制数来表示，这一过程称为量化。从图中可以看出取样后的离散值用二进制表示要损失一些精度，量化级别越多，则损失越少，音质就越好，声音就越清晰。量化级别是用量化位数表示每个采样点能够表示的数据范围，常用的有 8 位、12 位、16 位、24 位以及 32 位甚至是 64 位等。要注意的是，8 位(1 个字节) 不是说把纵坐标分成 8 份，而是分成 $2^8=256$ 份，同理 16 位是把纵坐标分成 $2^{16}=65\,536$ 份。通常 16 位的量化级别足以表示从人耳刚能听到的最细微的声音到无法忍受的巨大的噪音这样的声音范围了。无论量化精度有多高，量化过程必然会产生一定的噪音，这个称为量化噪音，但只要选择适当的量化精度，量化噪音就可以控制在人耳感觉不出来的范围内。

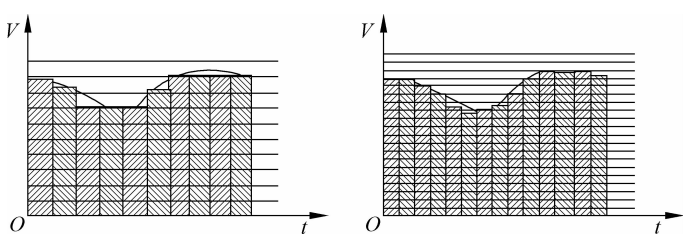

图 3-5　量化过程

采样精度的另一种表示方法是信号噪声比，简称为信噪比(Signal-To-Noise，SNR) 。狭义来讲是指放大器的输出信号的电压与同时输出的噪声电压的比，通常用分贝数表示。一般来说，信噪比越大，说明混在信号里的噪声越小，声音回放的音质越好，否则相反。信噪比一般不应该低于 70dB，高保真音箱的信噪比应达到 110dB 以上。并用下式计算：

$$\mathrm{SNR} = 10\lg[(V_{\mathrm{signal}})^2/(V_{\mathrm{noise}})^2] = 20\lg(V_{\mathrm{signal}}/V_{\mathrm{noise}})$$

其中 V_{signal} 表示信号电压，V_{noise} 表示噪音电压，SNR 的单位为分贝(dB)。

例如，假设 $V_{\mathrm{noise}}=1$，采样精度为 1 位表示，则 $V_{\mathrm{signal}}=2^1$，SNR=20lg2≈6dB。

再如，假设 $V_{\mathrm{noise}}=1$，采样精度为 16 位表示，则 $V_{\mathrm{signal}}=2^{16}$，SNR=20×16lg2≈96dB。

3. 编码

采样与量化后的二进制音频数据需要按一定的规则进行组织，以便于计算机进行处理，这就是编码。最简单的编码方案是直接使用二进制的补码表示，也称脉冲编码调制 PCM (Pulse Code Modulation)，它属于非压缩编码。在多媒体计算机中用这种编码方法存储的未压缩的音频数据文件大小可用下面的公式来计算：

文件存储量(B)＝时间(s)×采样频率(Hz)×采样精度(b)×声道数/8

4. 声道数

声道数是声音通道的个数，指一次采样的声音波形个数。单声道一次采样一个声音波形，双声道(立体声)一次采样两个声音波形，双声道比单声道多一倍的数据量。

3.3 声音的输出与识别

随着计算机科学技术的发展，人们已不再满足于仅仅通过键盘和显示器与计算机交互信息，而是迫切需要一种更加自然的接受多数人的声音同计算机进行交互。让计算机能听懂人说的话，或者用语音控制各种自动化系统，即用人类最直接、最方便的交换信息的形式——语言，来与计算机进行通信，这是人类一直以来的梦想，针对此，诞生了一门新的学科——计算机语音学(computer phonetics)。人们对于计算机语音学的研究主要包括以下几个方面：语音编码(speech coding)、语音合成(speech synthesis)、语音识别(speech recognition)、语种识别(language identification)、说话人识别(speaker recognition)或说话人确认(speaker verification)等。

3.3.1 语音输出

实现计算机语音输出通常有两种方法：一是录音/重放，二是文字→语音的转换。若采用第一种方法，首先要把模拟语音信号转换成数字序列，编码后，暂存于存储设备中(录音)。需要时，再经解码，重建声音信号(重放)。录音/重放可获得高音质声音，并能保留特定人或乐器的音色，但所需的存储容量随发音时间线性增长。第二种方法是基于声音合成技术的一种声音产生技术，它可用于语音合成和音乐合成。文字→语言转换是语音合成技术的延伸，它能把计算机内的文本转换成连续自然的语声流。若采用这种方法输出语音，应预先建立语音参数数据库和发音规则库等。需要输出语音时系统按需求先合成出语音单元，再按话音学规则或语言学规则，连接成自然的话流。文字→语言转换的参数库不随发音时间增长而加大，但规则库却随语音质量的要求而增大。基于语音合成技术的方法众多，根据不同分类标准有不同的合成方法，从研究技术上来分，有发音参数合成、声道模型参数合成和波形编辑合成；从合成策略上来分，有频谱逼近和波形逼近方法。

3.3.2 语音识别

语音识别技术，也被称为自动语音识别(Automatic Speech Recognition，ASR)，其目的是要将人类语音中的词汇内容转换为计算机可读取的输入，如按键、二进制编码或者字符序列等。这与说话人识别及说话人确认不同，因为后者并不关心语音中所包含的词汇内容，它只对发出语音的说话人进行识别或确认。一般来说，语音识别技术包括语音拨号、语音导航、室内设备控制、语音文档检索、简单的听写数据录入等。在实际应用，它往往与机器翻译、语音合成技术等自然语言处理技术相结合，以解决具体需求，如语音到语音的翻译等。语音识别技术所涉及的领域广泛，不同领域上的研究成果都对语音识别的发展做出了贡献，主要包括信号处理、模式识别、概率论和信息论、发声机理和听觉机理、人工智能等。语音识别技术与语音合成技术结合可以使人们甩掉键盘，通过语音命令进行操作，现在已成为一个

很具有竞争性的新型高技术产业。

1. 语音识别系统的分类

语音识别系统可以根据对输入语音的限制加以分类。如果从说话者与识别系统的相关性考虑，可以将识别系统分为3类：

(1) 特定人语音识别系统：仅考虑对于专人的话音进行识别。

(2) 非特定人语音系统：识别的语音与人无关，通常要用大量不同人的语音数据库对识别系统进行学习。

(3) 多人的识别系统：通常能识别一组人的语音，或者称为特定组语音识别系统，该系统仅要求对要识别的那组人的语音进行训练。

如果从说话的方式考虑，也可以将识别系统分为3类：

(1) 孤立词语音识别系统：要求输入每个词后要停顿。

(2) 连接词语音识别系统：要求对每个词都清楚发音，一些连音现象开始出现。

(3) 连续语音识别系统：连续语音输入是指进行自然流利的连续语音输入时，大量连音和变音就会出现。

如果从识别系统的词汇量大小考虑，也可以将识别系统分为3类：

(1) 小词汇量语音识别系统：通常包括几十个词的语音识别系统。

(2) 中等词汇量语音识别系统：通常包括几百个词到上千个词的识别系统。

(3) 大词汇量语音识别系统：通常包括几千到几万个词的语音识别系统。

随着计算机与数字信号处理器运算能力以及识别系统精度的提高，识别系统根据词汇量大小进行分类也会不断发生变化。

2. 语音识别的几种基本方法

一般来说，语音识别的方法有三种：基于声道模型和语音知识的方法、模板匹配的方法以及利用人工神经网络的方法。

1) 基于声道模型和语音知识的方法

该方法起步较早，在语音识别技术提出的开始，就有了这方面的研究，但由于其模型及语音知识过于复杂，目前还没有达到实用的阶段。通常认为常用语言中有有限个不同的语音基元，而且可以通过其语音信号的频域或时域特性来区分。该方法分为两步实现：

第一步，分段和标号。把语音信号按时间分成离散的段，每段对应一个或几个语音基元的声学特性。然后根据相应声学特性对每个分段给出相近的语音标号。

第二步，建立词序列。根据第一步所得语音标号序列得到一个语音基元网格，从词典得到有效的词序列，也可结合句子的文法和语义同时进行。

2) 模板匹配的方法

模板匹配的方法发展比较成熟，目前已达到了实用阶段。在模板匹配方法中，要经过4个步骤：特征提取、模板训练、模板分类和判决。常用的技术有三种：动态时间规整(DTW)、隐马尔可夫(HMM)理论和矢量量化(VQ)技术。

(1) 动态时间规整

语音信号的端点检测是进行语音识别中的一个基本步骤，所谓端点检测就是正确地标注出语音信号中的各种段落(如音素、音节、词素)的始点和终点的位置，从语音信号中排除无声段。在早期，进行端点检测的主要依据是能量、振幅和过零率。但效果往往不明显。20