第5章 chapter 5

数字音频技术

声音是多媒体技术研究中的一个重要内容。声音的种类繁多，如人的话音、乐器的声响、动物的叫声、机器产生的声音以及自然界的雷声、风声、雨声、闪电声等。在用计算机处理这些声音时，既要考虑它们的共性，又要利用它们各自的特性。

自从1969年Bell实验室开始数字语音的研究以来，计算机产生音乐以及语音识别、语音合成技术得到了越来越广泛的研究和应用。多媒体数字音频处理技术在音频数字化、语音处理、合成及识别等各个方面都有着很好的发展。

5.1 声音与声音信号数字化

作为一种模拟信号，声音在时间和振幅上都是连续的，即它的振幅能以任意精度并在任何一个时刻进行测量。与之不同的是，数字信号只能在确定的时刻才有意义，其数值也只能取有限的量。

5.1.1 声音与听觉器官

声音的强弱表现在声波压力的大小上，音调的高低表现在声音的频率上。当声音用电信号表示时，在时间和幅度上都是连续的模拟信号。对声音信号的分析表明，声音信号由许多频率不同的信号组成，这类信号称为复合信号，而单一频率的信号称为分量信号。声音信号的一个重要参数就是带宽，用来描述组成复合信号的频率范围，如高保真声音的频率范围为10～20 000Hz，带宽约为20kHz，而视频信号的带宽是6MHz。

声音信号的两个基本参数是频率和幅度。信号的频率是指信号每秒钟变化的次数，用Hz表示。频率小于20Hz的信号称为亚音信号，或称为次音信号；频率范围为20～20kHz的信号称为音频(Audio)信号。虽然人的发音器官发出的声音频率大约是80～3400Hz，但人说话的信号频率通常为300～3000Hz，在这种频率范围的信号称为话音信号；高于20kHz的信号称为超音频信号，或称超声波信号。一般来说，人的听觉器官能感知的声音频率大约在20～20 000Hz之间，在这种频率范围里感知的声音幅度大约为0～120dB。多媒体技术中处理的主要是音频信号，包括音乐、语音和音效(风雨声、鸟叫声和机器声)等。

5.1.2 模拟信号与数字信号

大多数电信号(模拟信号)过去一直是用模拟元部件(如晶体管、变压器、电阻和电容等)进行处理的。但是,开发一个具有相当精度且几乎不受环境变化影响的模拟信号处理元部件相当困难,成本也很高。

话音信号是典型的连续信号,不仅在时间上,而且在幅度上也是连续的。时间上"连续"是指在一个指定的时间范围内声音信号的幅值有无穷多个,在幅度上"连续"是指幅度的数值有无穷多个。我们把在时间和幅度上都是连续的信号称为模拟信号。

如果把模拟信号转变成数字信号,用数字来表示模拟量和对数字信号做计算,那么开发模拟运算部件的问题就转变成了开发数字运算部件的问题,这就出现了数字信号处理器(Digital Signal Processor,DSP)。DSP 与通用微处理器相比,除了结构不同外,它们的基本差别是,DSP 有能力响应和处理采样模拟信号得到的数据流,如做乘法和累加求和运算等。

在数字环境进行信号处理的主要优点是:首先,数字信号计算是一种精确的运算方法,它不受时间和环境变化的影响;其次,表示部件功能的数学运算不是物理上实现的功能部件,而仅仅是用数学运算来模拟,相对容易实现;此外,可以对数字运算部件进行编程,如欲改变算法或改变某些功能,还可对数字部件进行再编程。

5.1.3 声音信号数字化

计算机要处理或合成声音,就必须把模拟的(连续的)声音波形转换成数字(离散化),这个过程称为声音采样(见图 5-1),它是把连续的声波信号通过一种称为模数(A/D)转换器的部件转换成数字信号,供计算机处理,如果需要的话,这种转换后的数字信号又可以通过数模转换(D/A)器,经过放大输出,变成人耳能够听到的声音。

图 5-1 声音信号数字化的过程

连续时间的离散化通过采样来实现,就是每隔相等的一小段时间采样一次,这种采样称为均匀采样;连续幅度的离散化通过量化来实现,就是把信号的强度划分成一小段一小段,如果幅度的划分是等间隔的,就称为线性量化,否则就称为非线性量化。图 5-2 表示了声音数字化的概念。

我们把时间和幅度都用离散的数字表示的信号称为数字信号。声音数字化需要回答两个问题:①每秒钟采集多少个声音样本,也就是采样频率是多少;②每个声音样本的位数(bit per sample,bps)应该是多少,也就是量化精度。

采样的速度决定了录制声音的准确性,而采样值的精度则决定了录制声音的精确性。实践证明,采样速度越快,采样值越准确,声音特征复原得就会越好。

常用的几种音频信号数字化的采样率标准是 44.2kHz(CD 音质)、22.05kHz(FM 音质)、11.025kHz(AM 音质)等。为了追求音响品质的完美,减少噪声的干扰,达到理想的

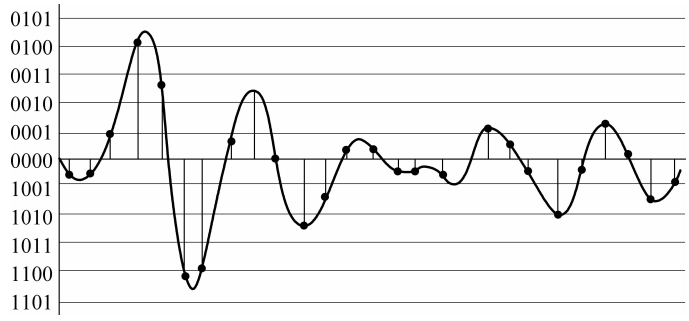

图 5-2　声音的采样和量化

传播声音的环境，国际上制定了一系列判断音质的标准。图 5-3 给出了几种数字声音质量等级的国际标准所对应的频率范围。

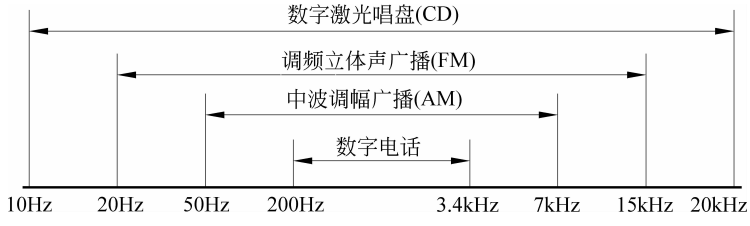

图 5-3　数字声音质量等级对应的频率范围

样本大小是用每个声音样本的位数(b/s)表示的，它反映度量声音波形幅度的精度。例如，每个声音样本用 16 位(2B)表示，测得的声音样本值是在 0～65 536 的范围里，它的精度就是输入信号的 1/65 536。样本位数的大小影响到声音的质量，位数越多，声音的质量越高，而需要的存储空间也越多；位数越少，声音的质量越低，需要的存储空间越少。

采样精度的另一种表示方法是信号噪声比。

原始的音频数据一般需进行编辑加工才能使用。通过编辑可以实现各种声音混合以及消除或降低声音中的畸变等。一般的音频编辑软件都具有设置音量、渐强渐弱处理及多通道混合等常用功能。音频处理主要集中在音频压缩上，最新的语音压缩算法可将原始声音数据压缩 6～8 倍以上。

5.1.4　声音质量与数据率

数字化音频的质量取决于采样频率和量化位数这两个重要参数，反映音频数字化质量的另一个因素是通道(或声道)个数。记录声音时，如果每次生成一个声波数据，称为单声道；每次生成两个声波数据，称为立体声(双声道)，立体声更能反映人的听觉感受。音频数字化的采样频率和量化级越高，结果越接近原始声音，除此之外，数字化音频的质量还受其他一些因素(如扬声器的质量等)的影响。

根据声音的频带，通常把声音的质量分成 5 个等级，由低到高分别是电话、调幅广播(AM)、调频广播(FM)、光盘(CD)和数字录音带(Digital Audio Tape，DAT)的声音。在

这 5 个等级中，使用的采样频率、样本精度、通道数和数据率见表 5-1。

表 5-1 声音质量和数据率

质量	采样频率/kHz	样本精度/b/s	单声道/立体声	数据率(未压缩)/kb/s	频率范围/Hz
电话 *	8	8	单声道	8	200～3400
AM	11.025	8	单声道	11.0	20～15 000
FM	22.050	16	立体声	88.2	50～7000
CD	44.1	16	立体声	176.4	20～20 000
DAT	48	16	立体声	192.0	20～20 000

说明：* 电话使用 μ 律编码，动态范围为 13 位，而不是 8 位。

5.2 音乐合成和 MIDI

多媒体音频数据的一个重要来源是 MIDI(乐器数字接口)。从 20 世纪 80 年代初期开始，MIDI 逐步为音乐界广泛接受和使用。MIDI 是乐器和计算机使用的标准语言，是一套指令(即命令)的约定，它指示乐器(即 MIDI 设备)要做什么，怎么做，如演奏音符、加大音量、生成音响效果等。MIDI 不是声音信号，它传送的是发给 MIDI 设备或其他装置让其产生声音或执行某个动作的指令。作为数字音乐的一个国际标准，MIDI 标准规定了电子乐器与计算机之间传送数据的通信协议等规范。MIDI 标准使不同厂家生产的电子合成乐器可以互相发送和接收音乐数据。随着 MIDI 标准的施行，计算机成为电子合成乐器间的控制环节，出现了大量可以记录、存储、编辑和播放乐谱(音符表或音符序列)的计算机软件。

MIDI 音频的处理过程如图 5-4 所示，其主要优点如下：

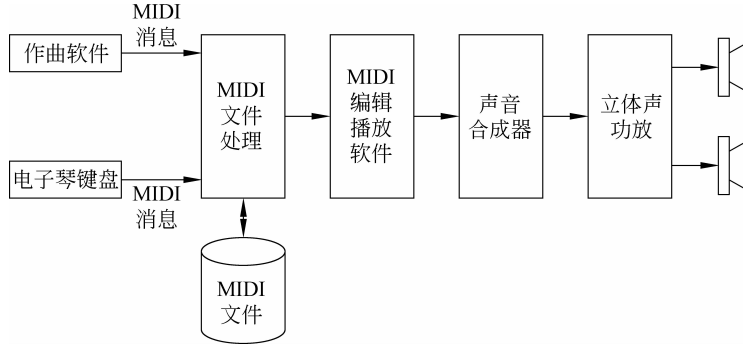

图 5-4 MIDI 音频的处理过程

(1) 生成的文件比较小。由于 MIDI 文件存储的是命令，而不是声音本身，因此它比较节省空间。例如，同样半小时的立体声音乐，MIDI 文件只有 200KB 左右，而波形文件(WAV)则要差不多 300MB。

(2) 容易编辑。因为编辑命令比编辑声音波形要容易得多。

(3) 可以作为背景音乐。MIDI 音乐可以和其他的媒体,如数字电视、图形、动画和话音等一起播放,这样可以加强演示效果。

产生 MIDI 乐音的方法很多,主要有两种:一种是频率调制(Frequency Modulation,FM)合成法;另一种是乐音样本合成法,也称为波形表(wave table)合成法。

5.3 数码音乐 MP3

MP3 的全称是 MPEG-1 Layer3 音频文件。MPEG-1 是活动影音压缩标准,其中的声音部分称为 MPEG-1 音频层,它根据压缩质量和编码复杂度划分为 3 层,即 Layer1、Layer2 和 Layer3,分别对应 MP1、MP2 和 MP3 这 3 种声音文件,并根据不同的用途,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 的压缩率则高达 10∶1～12∶1。也就是说,一分钟 CD 音质的音乐,未经压缩需要 10MB 的存储空间,而经过 MP3 压缩编码后只有 1MB 左右。不过 MP3 对音频信号采用的是有损压缩方式,为了降低失真度,MP3 采取了“感官编码技术”,即编码时先对音频文件进行频谱分析,然后用过滤器滤掉噪音电平,再通过量化的方式将剩下的每一位打散排列,最后形成具有较高压缩比的 MP3 文件,使压缩后的文件在回放时能达到比较接近原音源的声音效果。虽然它是一种有损压缩方式,但它以极小的声音失真换取了较高的压缩比,使得 MP3 能够在互联网上广泛传播。

MP3 这种压缩比非常高的数字音频文件不仅能在网上传播,而且还能容易地下载到便携式数字音频设备(MP3 随身听)中。MP3 随身听基于 DSP(数字信号处理器),无须计算机支持便可以实现 MP3 文件的存储、解码和播放。事先可以将创建好的 MP3 文件从计算机或互联网上下载到 MP3 随身听内置的存储器中,当从中选择播放一首 MP3 歌曲时,文件数据将被传送给 DSP,通过它来对文件进行解压缩。所需的解压缩软件被置入 DSP 处理器内部,或者存放在存储体中。DSP 将处理完的数据传给数模转换器,它将二进制的数码信息转换成模拟信号,然后再输出到耳机或扬声器中。

5.4 语音信号与处理

语音是人类沟通的主要方式,可以被人或机器来处理,后者就称为数字语音处理。

语音理解意味着要有效地适应说话人及其说话习惯,包括不同方言和情绪化的发音。语音信号有两个重要的特点可以应用在语音处理中:

(1) 浊语音信号(相对于清语音)在某一个确定的时间间隔上有一个几乎是周期性的结构,因此这种信号保持大约 30ms 的准稳态。

(2) 一些声音的频谱具有特征最大值,通常包括多达 5 个频率。这些在说话时生成的频率最大值被称作共振峰。根据定义,共振峰是一段语音质量的特征成分。

5.4.1 语音输出

语音输出涉及机器如何生成语音的问题，在这方面的主要挑战是，如何使得语音输出系统能够实时地生成语音信号，例如，自动地把文字转化为语音。某些应用（如语音报时）采用有限的词汇表来处理这一任务，但大多数采用的是广泛的词汇表。

机器输出的语音必须是可以听懂的，而且应该听起来很自然。其中可懂性是强制而自然的事情，可以增加用户的接受度。

下面是与语音输出相关的几个重要术语。

（1）语音基本频率。是语音信号中最低周期信号部分。它体现在嗓音中。

（2）音素。是最小的语音单位之一，用于区分语言或方言中的两个发音。它是最小的有意义的语言学单位，但并不携带内容。

（3）音位变体。确定了作为语音环境的函数的音素变化。

（4）词素。是有意义的语音学单位，在自由或受限的形式中都包含的最小且有意义的部分。

（5）嗓音。由声带的振动产生。嗓音强烈地依赖于说话者。

（6）非嗓音。由声带张开产生，这些声音相对独立于说话者。

5.4.2 语言合成

音频技术的一个重要方面是语音合成，即将普通正文合成为语音，如图 5-5 所示。

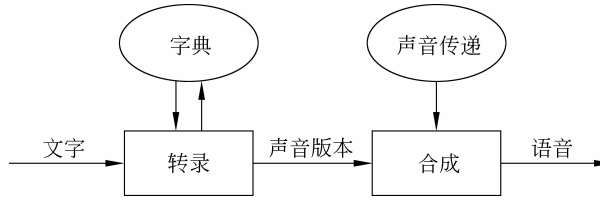

图 5-5 使用时间域声音连接的语音合成系统

第一步涉及转录，或将文本翻译成相应的音标。大部分方法使用一个包含大量单词或仅仅是音节或音调组的词典。这样的词典创建非常复杂，可以是单独实现的或是几个人使用的普通词典，其质量可通过相互作用的用户干预而不断提高。这意味着由用户识别出转换公式的缺陷，人工地改进发音，他们的发现逐渐成为词典的一个集成部分。

第二步将音素记录转换成声学的语音信号，其中连接可以发生在时域或频域。通常第一步用软件来解决，第二步则涉及信号处理器或专门的处理器。

除了副发音和韵律产生的问题外，语音识别还必须注意发音模糊问题。解决这个问题的唯一方式就是提供有关上下文的附加信息。

5.4.3 语音输入与识别

在语音输入处理的各种应用中，需要正确回答以下 3 个问题：

（1）谁？语音输入依赖说话者的某种特性，这意味着语音输入能识别出说话者。计算机可用于识别说话者的声音指纹。

（2）什么？语音输入的关键是检测语音内容本身。通常输入的语音序列产生一块文本。典型的应用有语言翻译系统。

（3）怎么样？这个问题有关如何研究语音采样。其典型应用如测谎仪。

音频技术中难度最大、也最具应用前景的当属语音识别，其潜在的商业应用前景使之一直是音频技术研究关注的热点。语音识别和语音合成相结合，实现了媒体转换。

语音识别一般是通过各种比较来完成的。利用现有技术，可以实现一个包含有大约25 000个词汇的依赖于讲话者的识别系统。语音识别中影响识别质量的问题主要是方言、情绪化的发音以及环境噪声等。要改善语音识别和语音生成的质量，需要弥合人类大脑与高性能计算机之间的相当大的性能差异，这仍需要一定的时间。

语音识别的原理如图5-6所示，是将个人发音的特殊特征和由以前抽取的语音元素组成的句子做比较。这意味着这些特征通常被量化，用于被研究的语音序列。这一结果与现有的参考做比较，以将它定位于现有的语言单元之一。识别出的言词作为参数化的语言单元序列被存储，传输或处理。

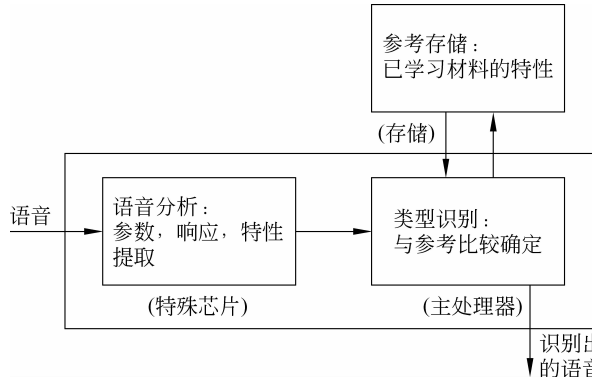

图5-6　语音识别原理

具体操作通常使用专门的元件或信号处理器抽取特征信息。比较和决定一般由系统的主处理器处理，但具有参考特征的词典通常位于计算机的二级存储单元。大多数具体的实现方法在定义特征信息时会有所不同。语音识别的组成部分如图5-7所示。

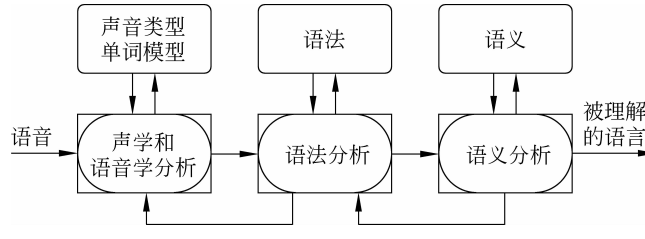

图5-7　语音识别组成部分

语音输入中的一个特殊问题是房间的声学特性，即环境噪声，此外，必须定义字边界，但这并不容易做到，因为大多数人说话并不强调一个字的开始和结束，同一个字也可

以被说得有快有慢。

依赖于特定人的识别系统比独立于说话者的系统能识别更多的字，但这是以提前“训练”系统为代价的。为训练系统使之适应说话者，通常要求他读特定的语音序列。目前的语音识别系统有大约半个小时的训练时间。大多数依赖说话者的系统能识别出 25 000 个字或者更多，而独立于说话者的系统则命中率接近 1000 个字。注意，现实的系统评估还应包括环境因素。

5.5 声音文件的存储格式

在因特网和各种计算机上使用的声音文件格式很多，但比较流行的主要是 WAV、AU(audio)、AIFF(Audio interchangeable File Format)和 SND(sound)文件格式。WAV 格式用于 PC，AU 用于 UNIX 工作站，AIFF 和 SND 用于苹果机和 SGI 工作站。

为便于读者辨认文件的属性，表 5-2 列出了部分声音文件的后缀。

表 5-2 常见的声音文件扩展名

文件的扩展名	说 明
AU	Sun 和 NeXT 公司的声音文件存储格式(8 位 μ 律编码或者 16 位线性编码)
AIFF	Apple 计算机上的声音文件存储格式
CMF(Creative Music Format)	声霸(SB)卡带的 MIDI 文件存储格式
MCT	MIDI 文件存储格式
MFF(MIDI Files Format)	MIDI 文件存储格式 1/2
MID(MIDI)	Windows 的 MIDI 文件存储格式
MP2	MPEG Layer 1、Layer 2
MP3	MPEG Layer 3
MOD(Module)	MIDI 文件存储格式
RM(Real Media)	Real Networks 公司的流放式声音文件格式
RA(Real Audio)	Real Networks 公司的流放式声音文件格式
ROL	Adlib 声卡文件存储格式
SND(sound)	Apple 计算机上的声音文件存储格式
SEQ	MIDI 文件存储格式
SNG	MIDI 文件存储格式
VOC(Creative Voice)	声霸卡的声音文件格式存储
WAV(Waveform) *	Windows 采用的波形文件存储格式
WRK	Cakewalk Pro 软件采用的 MIDI 文件存储格式

5.6 声　　卡

在多媒体计算机中，所有的音乐与音效都需要经过声卡来处理。声卡使用大规模集成电路技术，将音频技术范围的各类电路制成芯片而组成，以便直接插入计算机的扩展槽里，使用方便。声卡的主要工作就是把数字信号转换成模拟信号，然后送到喇叭上发出声音；另一方面，声卡也可以对计算机上的各种音频进行“混音”，例如串联电子合成乐器，或是从麦克风输入声音后与 CD 音乐一起由喇叭放出来等。

声卡通过反复地检测和记录声音信号的幅度来实现称为“采样”（实际上每秒钟要做几万次这样的操作）的录音过程，将声音信号转化为大量的幅度随时间变化的数字，并存储在磁盘上。

播放声音的过程与录音正好相反，计算机将一串数字传给声卡，声卡将它们转换成模拟信号，根据数字量的大小改变模拟信号的幅度，经放大后由音箱播出。

5.6.1 主要技术指标

评价声卡的主要技术指标如下：

（1）采样频率。为记录信号的精确细节，声卡必须以极快的速率进行采样。声卡的采样频率通常有 3 个标准：11.025kHz、22.05kHz 和 44.1kHz，目前一般的声卡都能达到 44.1kHz 的采样频率。

（2）采样位数（量化位数）。另一个影响声音质量的重要因素是每个采样点幅度的准确性。采样位数越多，声音的幅度就会越精确，但占用的存储空间也就越多。目前通常使用的有 8 位（低档）、16 位（中档）和 32 位（高档）3 种量化精度的声卡。

（3）声道数。分单声道和双声道，双声道可以播放立体声信号。一般的声卡都是双声道的。

（4）MIDI（数字化乐器接口）和游戏杆接口。该接口能够利用计算机控制和演奏电子乐器，或利用诸如 Windows 提供的实用程序记录电子乐器演奏的音乐，然后进行回放。游戏杆接口用来与游戏操纵杆相接，在声卡上一般与 MIDI 接口共享。

（5）合成器。音色是区别不同乐器的重要特征之一，声卡上的合成器能将各种不同频率的声音混合起来，形成某种特定乐器的音色。例如，安装在计算机上的控制软件对于同一组电子琴乐曲，可以同时选择小号演奏效果和钢琴演奏效果。

合成器的主要参数是合成的复音数和用语音合成的操作数目。一般有 20 种复音就可以满足大多数用户的需要了，复音成分越多越适合于专业音乐工作者。

（6）内部声音混合调节器。主要功能是将来自不同输入源的声音信号进行混合和音量调节。该混合器可以编程和控制。

（7）CD-ROM 接口。若用户需要使用 CD-ROM 来播放 CD、VCD 节目，应将声卡上的 CD-ROM 接口和 CD-ROM 上的声卡接口用专用的三芯电缆连接起来。连接时应注意各接口的规格，因为不同的 CD-ROM 声卡接口标准可能略有不同。

5.6.2 功能和分类

声卡的心脏是音效芯片，它有多个音频接口，通过音频线和光驱或其他的音频输入相连。一般声卡上都有 CD-ROM 接口，在声卡的挡板上，可以看见一排输出/入端子及游戏杆连接口。

声卡的主要功能包括：录音、编辑和回放数字音频文件；控制各声源的音量并加以混合；在记录和回放数字音频文件时进行压缩和解压缩；采用语音合成技术让计算机朗读文本；具有初步的语音识别功能；具有 MIDI 接口、输出功率放大等。

声卡主要根据其数据采样量位数来确定其分类，通常分为 8 位、16 位和 32 位等。位数越大，其量化精度越高，音质就越好。声卡通常带有自己的 CPU，具有较高的智能性和灵活性。声卡的关键技术包括数字音频、音乐合成、MIDI 与音效。数字音频部分具有的基本功能有 44.1kHz 的采样率，8 位以上的分辨率，录音和播放声音信号，同时具有压缩采样信号的能力。最常用的压缩方法是自适应脉冲编码调制。数字音频的实现有不同的方法和芯片，大多数采用的是 CODEC 芯片，它具有硬件压缩功能，部分采用的是 DSP＋ADC 方法，利用软件方法压缩数字音频信号。声卡上的音乐合成器也有许多不同的类型，目前主要采用两种合成技术：FM 与波形表。波形表合成使用了 DSP 技术，它要求大容量的 ROM，以获得高质量的演奏效果；通用 MIDI 要求支持 128 种乐器；不少声卡采用音效芯片，从硬件上实现回声、混响、和声等，使声卡发出的声音更生动。

声卡的种类很多，其功能不尽相同，但在相应软件支持下，应具备以下大部分或全部功能：

(1) 录制、编辑和回放数字声音文件。声卡可将来自话筒、收录音机以及激光唱盘等的声源采样，保存成数字文件，并由相应的软件对声音文件的数据进行编辑、混合或回放。

(2) 控制、混合各声源的音量。通常随声卡提供的软件有一个 Mixer 程序，它显示有多个滑键的控制板，用来控制和混合各声源的音量，用鼠标可调节话筒、激光唱盘和其他音源的输入音量，以及调节 MIDI、WAV 文件回放和主输出电路音量，除话筒之外均为双通道立体声调节。

(3) 在记录和回放数字文件时压缩和解压缩。这样可以节省存储空间。以立体声为例，其数字声音文件每分钟可占多达 10MB 的磁盘空间，因此，声音文件的压缩与解压缩是多媒体领域研究的一个重要课题。一般声卡的压缩算法固化在卡上，也有的以软件形式提供给用户。

(4) 采用语音合成技术让电脑朗读文本。在相应软件的支持下，采用语音合成技术，可让大部分声卡朗读英文或中文文本，用来帮助用户检查文章中的句法和语法错误，这是一般的拼写检查功能所无法做到的。常用的语音合成技术有两种：一种是基于字典技术，根据单词查到发音代码并送到合成器上去；另一种是基于规则将文本转换成语音。

声卡一般只能合成英文语音，国内在汉语语音识别、汉语语音合成方面做了多年的研究，已经取得了较好的成果。一般能够利用声卡通过软件把汉字国标代码置换成较自然的汉语语音，并具有语音信箱的功能，大大扩展了语音合成技术的应用范围。

(5) 具有 MIDI 接口。一台计算机可以控制多台带 MIDI 接口的电子乐器，利用计算