### 3.1 数据编码

本节内容按2课时设计。

**第一课时**

**【**教学重点**】**模拟数据的数字化方法 。

**【**教学难点**】**模拟数据的数字化方法。

**【教学过程】**

**一、引入**

1.学生预习，阅读第49页“任务一 认识智能公交系统中的数据”之“活动1 办理市民卡”，填写第50页的表3.1.1（不受活动1限制）。

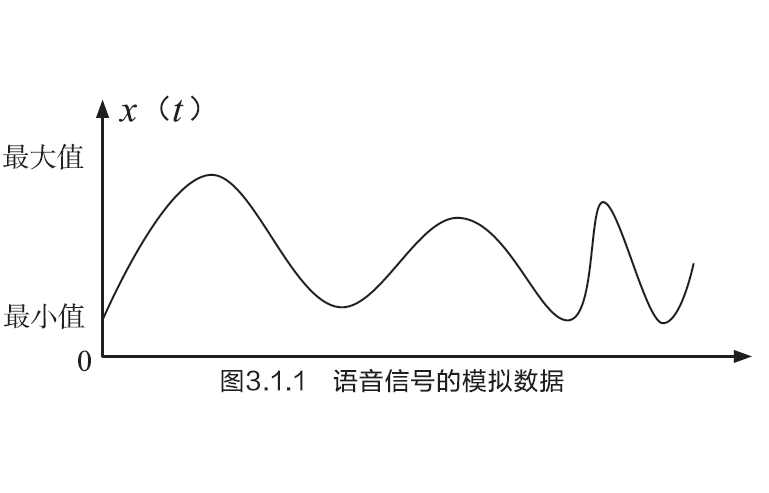
2.教师检查，并评讲填写情况，引出数据的表现形式。

表3.1.1 计算机处理的数据

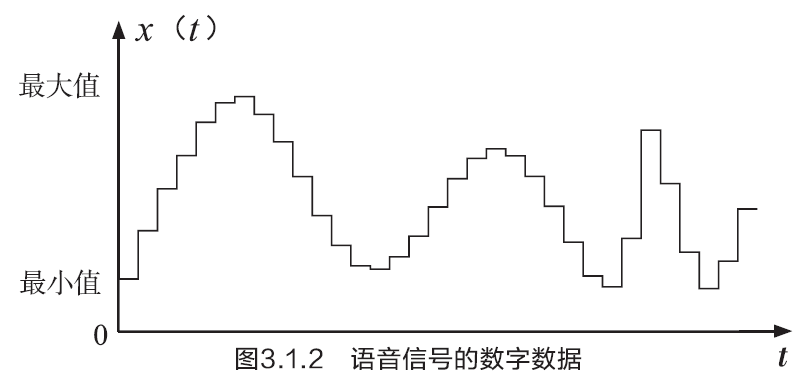
|  |  |  |  |
| --- | --- | --- | --- |
| 表现形式 | 实例 | 采集该类数据的设备 | 用什么软件处理 |
| 文本 | 姓名,家庭住址,性别（也可以用数字表示）、手机号 | 键盘  手写输入  语音输入 | 文字处理软件，电子表格，数据库  OCR识别  语音识别 |
| 数字 | 身高、体重、肺活量、血压,身份证号 | 键盘，各种数字化采集仪  手写输入 | 电子表格，数据库 |
| 图像 | 照片 | 手机，相机，摄像头，扫描仪 | PS,美图秀秀  用软件从视频中抓帧 |
| 声音 | 声音，语音，音乐 | 手机，录音机（录音笔），话筒（声卡） | 录音机，CoolEdit  从有声视频中分离音轨 |
| 视频 | 监控视频 | 手机，摄像机，摄像头 | QQ影音、绘声绘影  Camtasia |

注：考虑到处理各种类型的数据，不仅需要硬件支持，也离不开各种相关软件的支持，所以这里在教材原表3.1.1的基础上增加了一列。本表内容不是标准答案，只是抛砖引玉，启发师生进一步思考。

**二、模拟信号与数字信号**



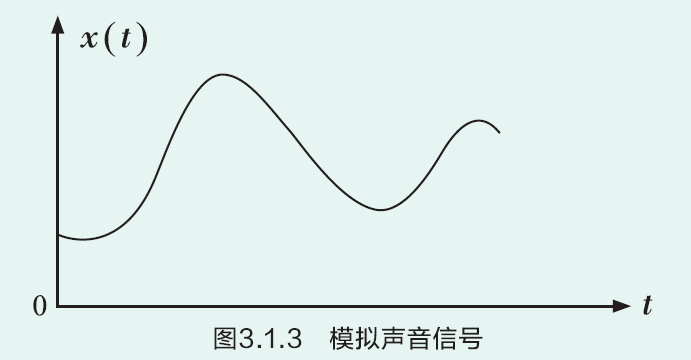
**师：**上图是语音信号（振幅或声压）随时间变化的图像。可见，它是随时间连续变化的，波形光滑，这种信号称为**模拟信号**，模拟信号的值称为模拟数据。用传感器直接获得的信号一般为模拟信号，类似的还有温度、压强、电压等。



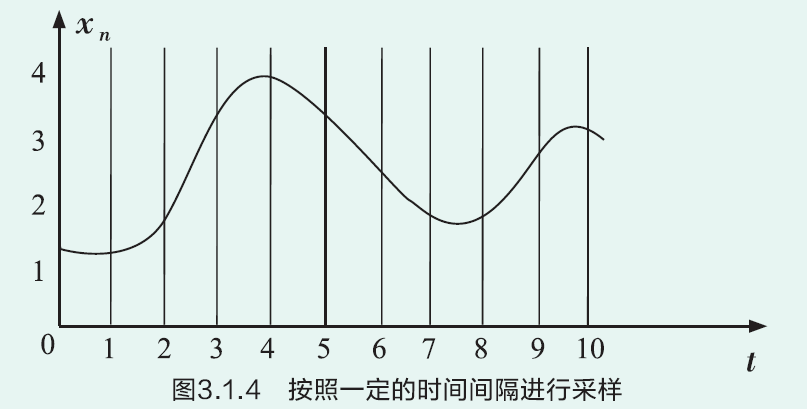
上图是语音信号数字化后的图像。可见，时间及幅值不是连续变化的，而是离散变化的，波形如台阶，这种信号称为**数字信号**。数字信号可以由模拟信号转换得来，数字信号的值即数字数据，直接用计算机所能理解的二进制来表示，这样方便计算机对其进行处理。

综上所述，按照取值特征可以将**信号分为模拟信号和数字信号**。

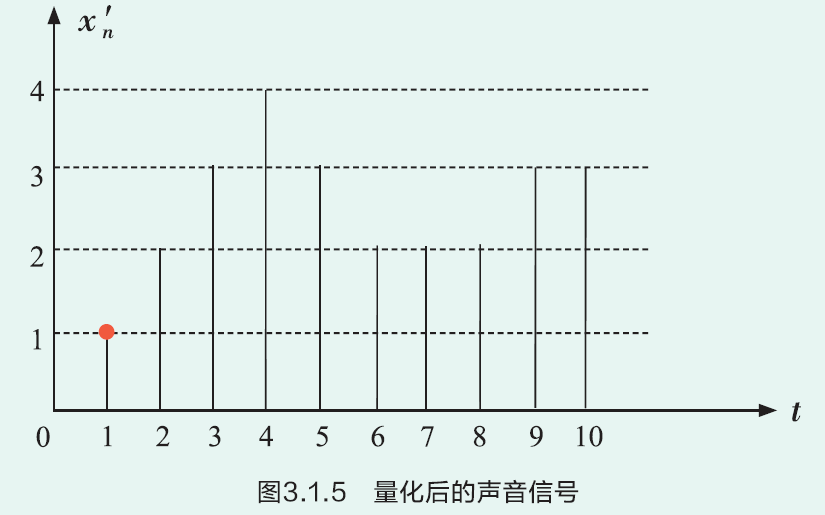
如何将图3.1.3中的模拟声音信号用计算机“忠实地”记录下来？换种说法：如何把某条曲线“精确地”存储到计算机中？



最容易想到的办法，就是依次记下曲线中每个点的坐标。但是，曲线中的点有无穷多个，我们却只能记录下有限的个点。怎么办呢？



每间隔一段时间取一个点间隔多少时间取一个点为好呢？1秒？0.1，0.01秒？通常，我们把这种从连续的时间中每间隔一个时间段抽取一个时刻点的操作称为“采样”。这个时间间隔的倒数称为采样频率，单位是赫兹，即秒-1。显然采样频率越高（时间间隔越短），声音还原的效果自然就越真实，但采样点也就越多，需要保存的数据也越多，所占存储空间也越大，计算机的工作量也越大。因此，需要在声音还原效果与存储数据量之间寻找一个平衡点。常用的CD音质的采样频率是44.1kHz，也就是把1秒的时间分成44100等份，每份取其对应的一个幅值。



还有一个问题:每个幅值用什么数据记录下来？精确到什么程度？这个问题同样关联到声音还原效果与存储数据量！CD音质的量化位数为16位，即用16个二进制位记录一个幅值，因此可记录216=65536种不同的数值，即0～65535。我们把每个声音幅值换算为一个对应的整数，这种操作称为量化。

量化得到的整数，当然要转换为二进制数，并把它们用一定的格式存储起来，有些还要按照一定的算法进行压缩处理后存储为文件，这个过程称为编码。

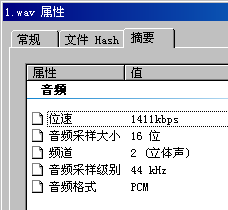
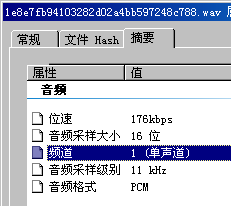
**生：**填表3.1.2，完成简单的进制转换。

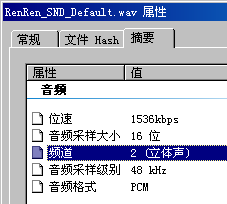
表3.1.2 数字化的声音信号

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 时刻 | 1 | 2 | 3 | 4 | 5 |
| 二进制数值 | 001 | 010 | 011 | 100 | 011 |
| 时刻 | 6 | 7 | 8 | 9 | 10 |
| 二进制数值 | 010 | 010 | 010 | 011 | 011 |

**【拓展】介绍“声道”与“立体声”** 声道是指声音在录制或播放时在不同空间位置采集或回放的相互独立的音频信号，所以声道数也就是声音录制时的音源数量或回放时相应的扬声器数量。双声道就是有两个声音通道。声卡所支持的声道数是衡量声卡档次的重要指标之一，从单声道到环绕立体声。我们知道，人耳之所以能准确辨别发音点的方位，是因为声源与双耳之间存在着角度差。这个角度差使声源传到左右耳的时候产生微小的时间差，而人耳对这微小的时间差非常敏感，从而使人能够产生准确的方向感和距离感。这种原理被称为哈斯效应。正是通过对这种声像定位原理的逆向运用，人们发明了最早的也是最简单的双声道立体声系统，即在录制声音时，在不同的位置用两只话筒进行录音，而在重放时则使用两路独立的放大器和两个扬声器，从而使听者可以较准确地判断出录音中不同音源的准确位置。双声道的典型应用场景是，在录制歌曲时将歌手的声音与乐器的伴奏音分别置于两个声道，用于卡拉OK时关闭歌手的声音；或者把影片的两种语言的配音分置于两个声道，播放时选择其中一种。人类对声音质量的追求是无止境的，后来相继出现了4，4.1，5.1，6.1，乃至7.1声道，立体声的立体感自然就越来越强了。

**【操作】现场录音实现声音的数字化** 我们把话筒连接到电脑的声卡，打开“录音机”软件，就可以录音了，也就是将语音和自然界中的背景声音的模拟信号，转化为计算机中的数字化声音，采样、量化、编码，这些过程是由软件在硬件支持下自动进行的。播放、查看保存的数字声音文件（扩展名均为wav，但具体的音频数据参数不同），对比各种格式wav文件的声音效果及文件大小。

**三、小结**

1.从模拟数据到数字数据，要经过采样、量化、编码三个过程。

2.采样是时间的离散化数字化，量化是幅值的离散化数字化，编码是数据的格式化文件化。

**四、练习**

1.一段时长为1分钟的双声道立体声的无压缩音频（采样频率为44.1kHz，量化位数为16位），占用的存储空间是多少MB（精确到0.1）？

参考解答：

60\*44.1\*1000\*16\*2= 84672000（位，比特）= 10584000（B，字节）≈10.1MB

2.一张CD光盘的容量大约为700MB，能容纳多长时间的无压缩双声道立体声CD音频（采样频率为44.1kHz，量化位数为16位）？

参考解答：

700/10.1≈69.3(分钟)

**第二课时**

**【**教学重点**】**认识二进制，了解文本的编码 。

**【**教学难点**】**二进制及与其他进制之间的转换。

**【教学过程】**

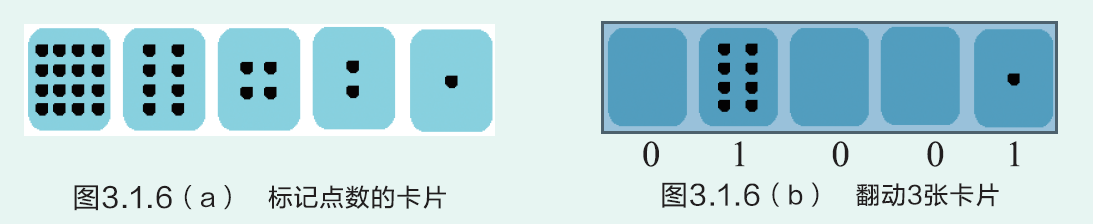
**一、引入**

学生预习，阅读第52页“任务二 查看存储在计算机中的数据”之“活动1 卡片组合游戏”，动手做一做，领会二进制计数的基本原理。

预习思考题：

思考题1 用这5张卡片，最大能表示的十进制数是多少？若用二进制数表示，是多少？如何分别表示十进制数的0，1，2，3，…？（参考答案：最大能表示的十进制数是31，用二进制表示为11111。其他略）

思考题2 现有255个相同的苹果，事先分装成若干袋，每袋装若干个苹果，当用户来买苹果时，无论他要多少个（不超过255），都可以不拆开袋子，直接选取几个袋子给他。那么事先应该怎样装袋？如果他要100个苹果，则应给他哪几袋？（参考答案：每袋分别装1，2，4，8，16，32，64，128个，共8袋。100=64+32+4）



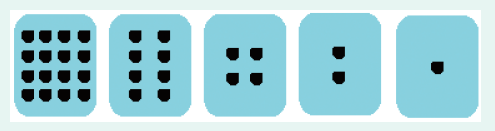
**二、二进制与数制转换**

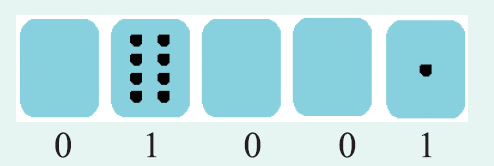
**（一）体会、归纳数制的通用规律**

**师**：把图3.1.6(a)和图3.1.6(b)关联起来看：

**24 23 22 21 20**

**16 8 4 2 1**





二进制数01001所表示的数的含义就是：

1\*1+0\*2+0\*4+1\*8+0\*16=1\*20+0\*21+0\*22+1\*23+0\*24=9

其中20，21，22，23，24，…，称为二进制整数从右到左各位的**权值**，2是二进制数的**基数**。二进制数**有2种不同的数字**：0和1，基本计数规则是**逢二进一**，即1+1=10。

一般地，任意进制数都有这样类似的规律：

如2019(10)=9\*100+1\*101+0\*102+2\*103=2019，十进制数的**基数**是10，十进制整数从右到左各位的**权值**依次是100，101，102，103，…，十进制数**有10种不同的数字**：0，1，2，…，9，基本计数规则是**逢十进一**，即9+1=10.

同理，对于十六进制数来讲，其**基数**是16，要**有16种不同的数字**，除了0，1，2，…，9，还需补充6个符号，一般用A～F(a～f)，分别相当于十进制数的10～15。基本计数规则是**逢十六进一**，即F+1=10。 十六进制整数从右到左各位的**权值**依次是160，161，162，163，…。例如，十六进制整数2F3D(16)=D(13)\*160+3\*161+F(15)\*162+2\*163=12093。

数制的这种规律，实现了任何一个整数只有一种表示法，而且任何一种表示法只表示一个整数，即整数与表示法之间是一一对应关系。

**生：**【练习1】填写下表，熟悉二进制、十六进制的写法。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 十进制 | 二进制 | 十六进制 | 十进制 | 二进制 | 十六进制 |
| 0 |  |  | 16 |  |  |
| 1 |  |  | 17 |  |  |
| 2 |  |  | 18 |  |  |
| 3 |  |  | 19 |  |  |
| 4 |  |  | 20 |  |  |
| 5 |  |  | 21 |  |  |
| 6 |  |  | 22 |  |  |
| 7 |  |  | 23 |  |  |
| 8 |  |  | 24 |  |  |
| 9 |  |  | 25 |  |  |
| 10 |  |  | 26 |  |  |
| 11 |  |  | 27 |  |  |
| 12 |  |  | 28 |  |  |
| 13 |  |  | 29 |  |  |
| 14 |  |  | 30 |  |  |
| 15 |  |  | 31 |  |  |

显然，把R(R为不小于2的整数)进制整数转换为十进制整数，使用上面的“按权展开求和法”。

【练习2】请将8进制数2071转换为十进制。（参考答案：1081）

在Python中，可使用函数int(num,R)将R进制的整数num转换为十进制，这里num要使用字符型(常数两端加上引号)。例如：

>>> int('11011010',2) #将二进制整数11011010转换为十进制

218 #返回结果为十进制整数

>>> int('11011010',8) #将八进制整数11011010转换为十进制

2363912 #返回结果为十进制整数

>>> int('FF',16) #将十六进制整数FF转换为十进制

255 #返回结果为十进制整数

**（二）比特、字节、文件大小的单位**

计算机中数据的最小单位是二进制位，称为比特(bit,简写为b)。

计算机存储数据的基本单位是字节（Byte，简写为B），1B=8b。

因为计算机内数据都是以二进制形式存储的，所以文件大小的单位也是字节，但字节单位太小，常用的有KB,MB,GB等。

1KB=210B=1024B

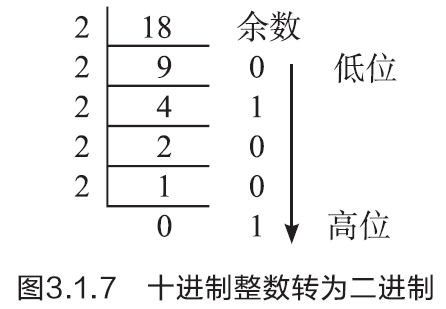
1MB=210KB=220B=1,048,576B

1GB=210MB=220KB=230B=1,073,741,824B

递进倍数是210=1024，因此民间把10月24日戏称为“程序员节”。

**（三）十进制整数转换为其他进制数**

把十进制非负整数转换为二进制数，可使用短除法，即“除二取余”法。例如，把十进制整数18转换为二进制数的过程如图3.1.7所示，结果为18（10）=10010（2）。



【练习3】仿照图3.1.7，将十进制整数175转换为16进制。（参考答案：AF）

一般地，十进制非负整数转换成R进制数（R≥2）均可采用“除R取余法”，即将图3.1.7中的除数2改为R。当R>10时，可能会出现余数为二位数（十进制）的情况，此时要转换为对应的R进制的单个数字。（如10→A,11→B,…,15→F）

在Python中，可使用函数bin(num),oct(num),hex(num),将十进制整数num分别转换为二、八、十六进制，这里num为整数。例如：

>>> bin(100) #将十进制整数100转换为二进制

'0b1100100' #返回结果为字符串，0b表示二进制

>>> oct(100) #将十进制整数100转换为八进制

'0o144' #返回结果为字符串，0o表示八进制

>>> hex(100) #将十进制整数100转换为十六进制

'0x64' #返回结果为字符串，0x表示十六进制

**（四）算法演示：十进制整数与其他进制整数之间的互相转换（详见配套资源“进制转换.fld”）**

1. 任意进制整数转换为十进制。

2．十进制整数转换为任意进制。

**三、文本数据的编码**

【学生活动】查看数据的编码

在Python中，通过ord()函数查看字符的编码。

>>> ord('1') #显示字符“1”的编码值

49

>>> ord('A') #显示字符“A”的编码值

65

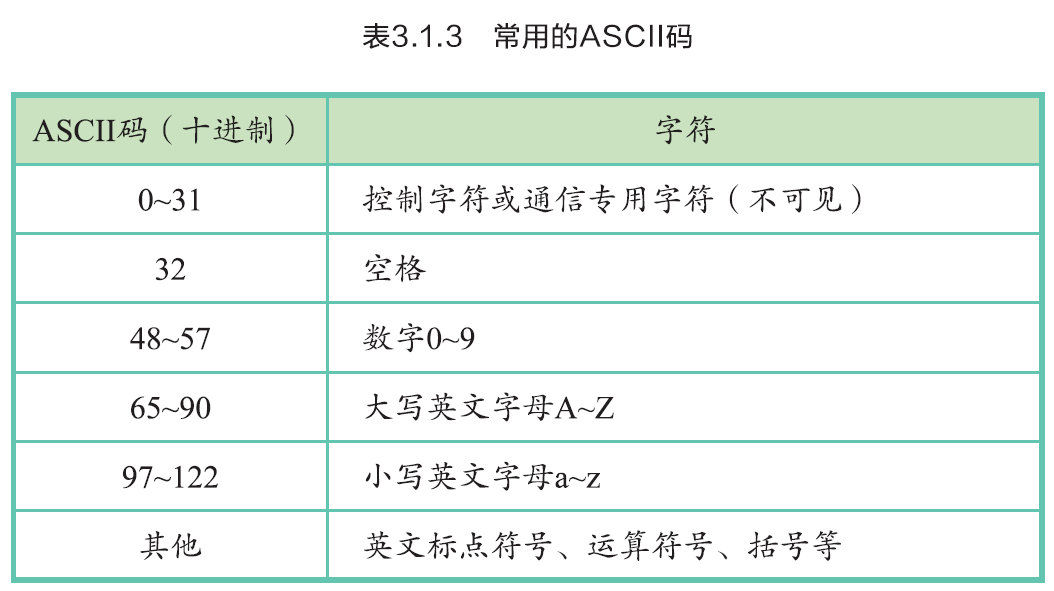
>>> ord('男') #显示字符“男”的编码值

30007

>>> chr(30007) #显示编码值为30007的字符

'男'

**1.ASCII码**



**2.Unicode码**

汉字编码的国家标准如下：

（1）GB 2312—1980字符集共收录6763个汉字。

（2）GBK字符集，兼容GB 2312—1980标准，收入21003个汉字。

（3）GB 18030—2000字符集，包含GBK字符集和CJK统一汉字扩充A的汉字，共计27533个汉字。

（4）GB 18030—2005字符集，在GB 18030—2000的基础上，增加了CJK统一汉字扩充B的汉字及其他汉字， 共计70244个汉字。

Unicode字符集，是全球可以共享的编码字符集，涵盖了世界上主要文字的字符，其中包括简繁体汉字，共计74686个汉字。如汉字“男” 的Unicode编码为30007（十进制），用十六进制表示为7537，用二进制表示为0111010100110111。

在Python中，函数ord()返回的是字符的Unicode码值，函数chr()返回Unicode码对应的字符。

【练习4】在Python中输出自己的姓名，但编程或表达式中不得直接出现这些汉字。

参考解答：事先用ord()函数查看姓名中各汉字的Unicode编码，再用chr()函数获取这个汉字。

>>> chr(24352)+chr(39134)

'张飞' #返回“张飞”

**3.汉字的其他编码及相互关系（拓展）**

①**区位码**

在GB2312-80《信息交换用汉字编码字符集》中，所有的国标汉字与符号组成一个94×94的矩阵。在此方阵中,每一行称为一个“区”,每一列称为一个“位”,因此,这个方阵实际上组成了一个有94个区(区号分别为1到94)、每个区内有94个位(位号分别为1到94)的汉字字符集。一个汉字所在的区号和位号简单地组合在一起就构成了该汉字的“区位码”。在汉字的区位码中，高两位为区号，低两位为位号。 在区位码中，01～09区为682个特殊字符，16～87区为汉字区，包含6763个汉字 。其中16～-55区为一级汉字(3755个最常用的汉字，按拼音字母的次序排列)，56～87区为二级汉字(3008个汉字，按部首次序排列)。

如“热”字的区位码为4040（第40区第40位）。

汉字区位码可在网上查询。

②**国标码**

国标码是由区位码稍作转换得到,其转换方法为:先将十进制区码和位码转换为十六进制的区码和位码，再将这个代码的第一个字节和第二个字节分别加上20H，就得到国标码。如：“热”字的国标码为4848H，它是经过下面的转换得到的：4040D->2828H->+2020H->4848H。

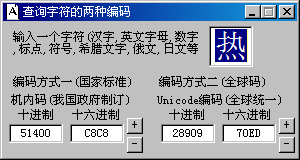
③**机内码**

国标码是汉字信息交换的标准编码，但因其前后字节的最高位为0，与ASCII码发生冲突。如“热”字，国标码为4848H，而英文字符“H”的ASCII也为48H，现假如内存中有两个字节为48H和48H，这到底是一个汉字，还是两个英文字符“H”？于是就出现了二义性，显然，国标码是不可能在计算机内部直接采用的。于是，汉字的机内码采用变形国标码，其变换方法为：将国标码的每个字节都加上128（即80H），即将两个字节的最高位由0改1，其余7位不变，如上面我们知道，“热”字的国标码为4848H，加上8080H，因此，“热”字的机内码就是C8C8H。



利用Word的“插入-字符”工具查看汉字“热”的编码

“热”字的Unicode码为70EDH。Unicode码与机内码（基于我国国标编码）是两套独立的编码体系，相互之间没有固定的转换机制。那么，怎样同时获得某个汉字的机内码与Unicode码呢？可使用下面的小软件（**详见配套资源“查询字符的两种编码.exe”**）：



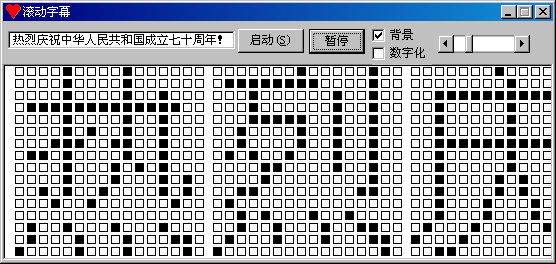
查询汉字的两种编码

④**输入码**

在使用键盘录入汉字时，要用到输入码。“热”字的全拼输入法编码为re，五笔字型输入法编码为rvyo。

⑤**字形码**

显示、打印汉字时要用到字库字形。下面的软件，调用了宋体16点阵字库，来显示汉字（**详见配套资源“滚动字幕.exe”**）：



查看汉字的点阵字形

【练习5】你能写出图中汉字“热”的字形编码吗？用十六进制表示（直接将二进制转换为十六进制，每4位二进制数字对应1位十六进制数字）。汉字“热”的字形编码占多少字节？

（参考答案：0840 0840 0848 7FFC 0848 0A48 1CC8 6848 08A8 088A 290A 1204 0000 4888 4446 8442，占32字节）

这些“编码”之间既相互联系又有区别，在不同环节下使用，为计算机处理汉字形成了完整的数据链，而这些“编码”就是这条链路上的关键“节点”。

**四、小结**

1.二进制的优点。

（1）使计算自动化成为可能。二进制只有0和1两个状态，具有两个稳定状态的电子器件很多，如开关的接通和断开、晶体管的导通和截止、磁元件的正负剩磁、电位电平的高与低等，都可表示0、1两个数码。

（2）使自动化计算装置简便可靠。二进制数的运算法则少，运算简单，使计算机运算器的硬件结构大大简化（十进制的乘法九九口诀表55条公式，而二进制乘法只有4条规则）。

（3）使自动化逻辑运算简单易行。由于二进制0和1正好和逻辑代数的假（False）和真（True）相对应，有逻辑代数作为理论基础，用二进制能很方便地进行逻辑运算。

基于这三条，计算机系统采用二进制。

2.二进制的缺点。

（1）用二进制表示一个数时，位数太多，太繁琐。故有16进制。

（2）难于记忆，可读性差。因此，与人类交互时，计算机系统已经尽量隐藏了二进制的内核，尽可能人性化了（数值则已自动转为十进制）。

**五、练习**

【练习1】进制转换

10110100(2)=( )(10) 3D(16)=( )(10)

255(10)=( )(16) 100(10)=( )(2)

E2(16)=( )(2) 11100101(2)=( )(16)

参考答案：180，61，FF，1100100，11100010，E5

【练习2】我们最熟悉的是十进制，在十进制中，有21\*12=252，那么这个结论还在哪些进制中成立？为什么？

参考解答：在R(R≥6)进制中都成立。证明如下：

21(R)\*12(R)=(2\*R+1)\*(1\*R+2)= (2R+1)(R+2)=2R2+5R+2=252(R)

由于式子中出现的最大数字为5，所以R≥6。