第4章　汉字编码技术

- “汉字编码”，是指汉字的外码，而非汉字的内码。
- 汉字的外码面向用户，汉字的内码面向系统。
- 用户通过外码来获得汉字。
- 由于当今汉字外码主要是用于汉字的输入，所以通常把它称为输入码。

主要内容

- 汉字编码的发展
- 汉字编码中的几个概念
- 汉字编码理论
- 数码键盘方案
- 编码方案的标准和规范

4.1 汉字编码的发展

- 汉字编码的初步概念
  - 从广义角度看

    早期字典/词典的编码方法，用于在字典和词典中，对被查字/词的快速定位，常见的方法有，部首、笔画、拼音和四角号码等方法。
  - 从狭义角度看

    实现把汉字输入计算机而进行的编码，称为汉字键盘编码，也就是用计算机键盘上的按键为汉字编码，从而达到把汉字输入计算机的目的，目前已经有上千种这样的编码。

汉字编码的发展过程

- **20世纪70年代**

  初期推出大键盘方案，中期发展拼音、简拼、双拼编码，后期出现五笔字形、自然码等；

- **20世纪80年代中期**

  可谓是“万码奔腾”时代，出现了一千多种编码，但普遍水平不高，个别编码获得进步和发展；

- **20世纪90年代**

  汉字编码的萧条期，自然输入方式的研发有较大的进展，对编码研究有一定影响；

- 进入**21**世纪后

  将是又一个发展期，对汉字编码输入有了新的认识，数字编码和拼音码方案将占主导。

4.2 汉字编码中的几个概念

❖ 字符集/字汇和词汇

- 字符集/字汇是指计算机内可以处理的汉字集合，如：**GB2312、ISO10646**等编码字符集；
- 词汇特指被某个编码方案所编码的词组之集合，这里的“词组”是泛指词、 词组、短语。

❖ 码元

- 组成编码的字符称为码元，如拼音码的码元是“**a**”—“**z**”中的任意一个字母；纵横码的码元是“**0**”—“**9**”中的任意一个数字 ；
- 码元增加，重码率会降低，编码记忆量会增加。

码长

- 构成编码的码元的个数称为码长。例如，编码“123”的码长为3；
- 等长编码，如区位码、电报码等；
- 不等长编码，如拼音码、纵横码等。

单码、重码和重码率

- 如果一个编码只对应于一个汉字或词组，这种现象称为单码；
- 一个编码可能对应于多个汉字和词组，那么这种现象称为重码；
- 具有重码的字/词组个数，与被编码的字/词组总数之比，称为重码率。

编码空间

- 所有可能的输入码集合，称为编码空间；
- 编码空间的大小与码元集大小及码长有关：
  - 如果某个编码方案的码元共有K个，编码为等长码，其码长为j，则编码空间$C=K^j$；
  - 区位码和电报码均有10个码元，码长为4，其编码空间$C=10^4=10000$个。

- **编码效率：** 被编码的字/词组的个数，与编码空间的大小之比，区位码被编码的汉字数为**6763**个，则其编码效率为**6763/10000＝67.63％**。

4.3 汉字编码理论

- 汉字的熵
- 汉字键盘编码的依据
- 汉字编码分类
- 键盘编码和键盘

4.3.1 汉字的熵

- 对信息定义的回顾
  - 信息论奠基者香农Claude Shannon认为，信息就是能够用来消除不确定性的东西，是一个事件发生概率的对数的负值；
  - 从定性的意义上说，人们在得知某个消息后，他在事前认为消息中的事件发生的不确定性越大，则认为该消息给他带来的信息量越大。

不确定性　　概率　　对数

一、汉字熵和信息量的概念

- 熵（Entropy）在信息论里被称为信息量或信息熵，从控制论的角度来看，也就是不确定性.
  - 一个事物的状态所具有的可能性越多，它的不确定性就越大，其熵也越大；
  - 不确定性越大的事物，如果最后确定了，那么从中得到的信息就越多，也就是说，其信息量就越大。

- **熵的度量：**
  - 最简单的事物只有2（即2^1）种可能性，非此即彼，我们以这种事物的信息量为单位，叫1比特（bit）；
  - 如果一个事物有4（即2^2）种可能性，则其信息量就为2比特；
  - 如果事物有N种可能性，当$N=2^n$ 时，那么其信息量H就是n比特，也就是，信息量H等于可能性数目N的以2为底的对数，即：

$$H=\log_2 N = \log_2 2^n = n$$

- 信息量也可以用以10为底的对数来计算，即:

$$H=\log_2 N=\lg N/\lg 2$$

  - 若N为3种可能性，则H= lg3/lg2=1.585

- 描述事物的数字符号的信息量
  - 二进制数：一位二进制数有2种状态，其信息量为1比特，n位二进制数有2^n种状态，其信息量则为n比特；
  - 十进制数：一位十进制数有10种状态，其信息量为$\log_2 10=\lg 10/\lg 2=3.32$比特，n位十进制数有$10^n$种状态，其信息量则为：

$$\log_2 10^n = n\log_2 10=3.32n\text{比特}$$

- 结论：
  - 数字符号的信息量H，为一位数字所具有的状态个数n的以2为底的对数，即：$H= \log_2 n$；
  - 由上可得出，一位n进制数的信息量为$\log_2 n$比特，k位n进制数的信息量为$k \log_2 n$比特。

- 我们也可以通过状态出现的概率来表示信息量H，在一位数字的n个状态中，每个状态出现的概率p均为1/n（即：n=1/p)，故有：

$$H = \log_2 n = \log_2(1/p) = \log_2 1 - \log_2 p$$

$$= 0 - \log_2 p = -\log_2 p$$

- 信息就是能够用来消除不确定性的东西，是一个事件发生概率的对数的负值。

描述事物的文字符号的信息量

- 设有n个文字符号，当其中每个符号出现的概率都为p时，那么文字符号的信息量H也为$\log_2 n$，或者为$-\log_2 p$；
- 其实，在日常使用中，每个文字符号出现的概率是不同的，设符号W_i的出现概率为p_i，其中i=1,2,…,n，则W_i的信息量$H_i = -\log_2 p_i$；
- 文字符号的平均信息量H，就为各个文字符号信息量H_i的加权平均，即：

$$H=\sum p_i H_i = -\sum p_i \log_2 p_i \quad (i=1,2,\ldots,n)$$

各种语言中字母的信息量

法文	3.98比特
意大利文	4.00比特
西班牙文	4.01比特
英文	4.03比特
德文	4.10比特
罗马尼亚文	4.12比特
俄文	4.35比特
中文	9.65比特

二、汉字熵的概率分布

- 假设给定一个汉字字符集HZ，其中汉字数为n，则该字符集内汉字的平均熵为：

$$H = -\sum p_i \log_2 p_i \quad (i=1,2,\ldots,n)$$

其中，p_i为单个汉字在汉语文本中出现的概率，则$-\log_2 p_i$就是第i个汉字在文本中出现时的信息量，$-\sum \log_2 p_i$ 就是文本中所有汉字在不考虑前后相关性时所给出的全部信息量，H是这些信息量的加权平均，即该集合中的每个汉字的平均信息量。

- p_i是从随机样本中得出的一个随机变量，

  令$h_i = -\log_2 p_i$， 则h_i也是一个随机变量，使得H也是随机量，把w个字看作是w个独立同分布的随机变量，则有：

  $$X = p_1h_1+p_2h_2+\ldots+p_wh_w = \sum p_ih_i \quad (i=1,2,\ldots,w)$$

  即为w个独立同分布的随机变量之和

- 在语言中某个冷僻字的出现比常用字给出的信息量大，但是每个字的信息量不会无限大，即其熵的分布方差是有限的，因此，不管h_i原来的分布如何，它的样本平均值----熵X，在大样本情况下服从中心极限定理，即服从正态分布。

三、汉字熵的意义

- 汉字的平均信息量（熵）就是存储或表示汉字字符所需要的平均二进制位的个数，约为9.65个二进制位。
- 在汉字集内，每个汉字的平均熵都不同，故可以根据每个汉字的平均熵，通过对汉字采用不等长编码来减小汉字编码的平均码长：
  - 对常用汉字，由于其信息熵较小，故可以采用较短的编码；
  - 对非常用汉字，由于其信息熵较大，故可以采用较长的编码。

- 通过增加编码码元集内的码元数（即增大了码元的信息量），也可以减小汉字编码的平均码长，设码元数为n，平均码长为Lavg，经实验得出：

  当n=12时， $L_{avg} \in$（3.00，4.00）

  当n=26时， $L_{avg} \in$（2.05，3.05）

  当n=39时， $L_{avg} \in$（1.87，2.87）

  当n=47时， $L_{avg} \in$（1.73，2.73）

4.3.2 汉字键盘编码的依据

- 汉字键盘编码的重要依据有：
  - 汉字信息熵；
  - 人的心理因素；
  - 汉字本身的特征，主要是字形和字音。

一、心理依据

- 从心理学角度来看
  - 人们使用根据汉字字音信息的汉字编码时，大脑中无需进行任何间接思考，就可以直接输入汉字，所以这是一种直接的输入方式；
  - 人们使用根据汉字字形信息的汉字编码时，大脑就需要把语言转换为字形，然后才能输入汉字，所以这是一种间接的输入方式。
- 汉字心理学和模糊心理学的研究表明，人认字时，对字的上半部要优于下半部，字的外围优于中间。
- 从排列心理学角度来看，希望汉字的编码能尽量唯一，并能尽量表达汉字的本身特征。

二、汉语拼音

- 汉语拼音的语言形式有三个要素：声母、韵母和声调，三者构成一个音节。其中声母有21个，韵母有35个，声调有五种：阴平、阳平、上声、去声和轻声。声韵结合起来共有417个基本音节，如果再考虑声调，则总共有1330个左右的音节。
- 所有的计算机用汉字的发音都在这些音节范围内，这就是汉字同音字/词多的根本所在，例如在GBK 13000.1的汉字集（20902个汉字）中，拼音“yi”共有460多个对应的汉字。

- 同样也存在大量的同音词，这就造成重码多和输入不方便，这是以音为编码要素所存在的主要问题；
- 由于汉语拼音在我国已经成为小学生识字前的必学内容，所以拼音输入法就成为大家不要专门学习就会使用的汉字输入法，故其拥有最大量的用户；
- 近年来，社会上出现了多种具有人工智能技术的拼音输入法，它们能够有效地克服传统拼音输入法的短处，使其成为一种易用、高效的汉字输入法。

三、笔画/笔顺

- 笔画/笔顺编码是选取汉字的基本笔画（如五种或八种），把笔画定义到汉字的数字键和字母键上，然后依汉字的笔画和笔顺来给汉字编码，笔画编码常见的有两种：
  - “札”字法：把笔画分为“横、竖、撇、捺（点）、折”
  - “永”字法：把笔画分为“横、竖、钩、挑、撇、捺、折”
- 笔画输入近年来受到了特别的重视，主要是手机迅速普及所致；
- 笔画输入的优势在于简单，无需学习和记忆；
- 笔画输入的困难在于单字输入重码多、码长较长、笔顺要求严，词组输入效率低，句子输入更困难。

四、汉字部件

- 部件—由笔画组成的具有组配汉字功能的构字单位
  - 汉字的字形分成三级：笔画、部件、整字；
  - 部件由笔画构成，整字由部件构成；
  - 有的笔画既是部件又是字，比如：“一”、“乙”；
  - 有的部件也是字，比如：“马”、“牛”、“土”；
  - 除了几百个独体字外，其余的汉字都可以拆分成若干个部件。

- 绝大多数根据汉字字形的编码方案，基本上都是基于部件的编码，如“五笔”、“表形码”等；
- 基于部件的汉字编码需要解决的问题是：汉字如何拆分？这就必然会带出部件的规范问题；
- 国家“语言文字工作委员会”在1998年5月发布了GF3001－“信息处理用GB13000字符汉字部件规范”，共定义了560个独立使用的部件。

4.3.3 汉字编码的分类

- 流水码
- 音码
- 形码
- 音形码/形音码

4.3.4 海曼公式与汉字编码时间

- 海曼公式（Hyman）的一般形式为：

$$T = a + bH(k)$$

  - T为平均选择反应时间；
  - k是选择信号的个数；
  - H(k)为每一个信号的平均信息量；
  - a是简单反应时间，即从刺激到器官反应的时间；
  - b是选择反应系数，即外界条件的影响。

- 汉字编码时间的公式：

$$T=a+b\log_2 k+c$$

  - a为大脑发出指令冲动到肌肉动作所需时间，即击键时间；
  - $b\log_2 k$理解为“选择时间与信息量成正比”的适用条件下，选择等概率键位所需时间的一种可采用的表达方式， 其中b是选择反应系数，即外界条件的影响；
  - c代表一个码元的平均“编码时间”，反应了思维时间和检索时间的长短。

4.3.5 键盘分区图

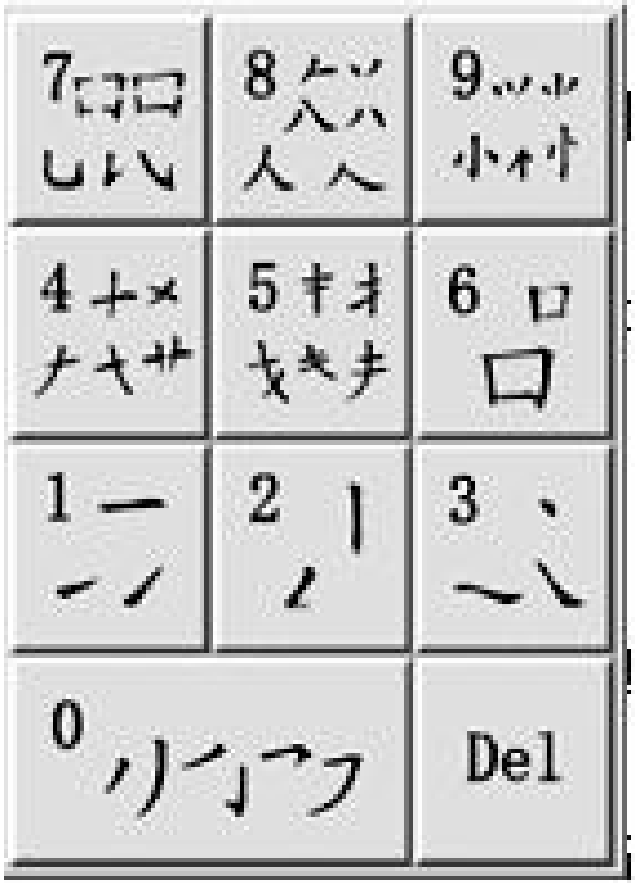
4.3.6 大键盘编码

- 如果一种编码的码元集合为“a”—“z”这26个字母或它的子集，那么我们称这种码元的键盘映射方式为大键盘编码。如全拼和智能ABC等拼音编码、五笔字形、郑码等等；
- 各种音码一般都采用大键盘编码；
- 形码也有不少采用大键盘的，如五笔字形就是一个最典型的例子；
- 采用大键盘的编码的码长一般为3—4，其平均码长一般不会超过4。

4.3.7 小键盘编码

- 码元采用键盘右边的数字区内的“0”—“9”这10个数字的编码，称为小键盘编码。比如：区位、纵横、字原、五笔数码等均是小键盘编码；
- 小键盘编码中以形码居多，它们一般都采用笔画编码；
- 音码在计算机的小键盘上应用较少，但是在数码产品的键盘上用得较多，如手机、GPS等，实现时，把26个字母依次分配到10个数字按键上。

纵横码的键位图

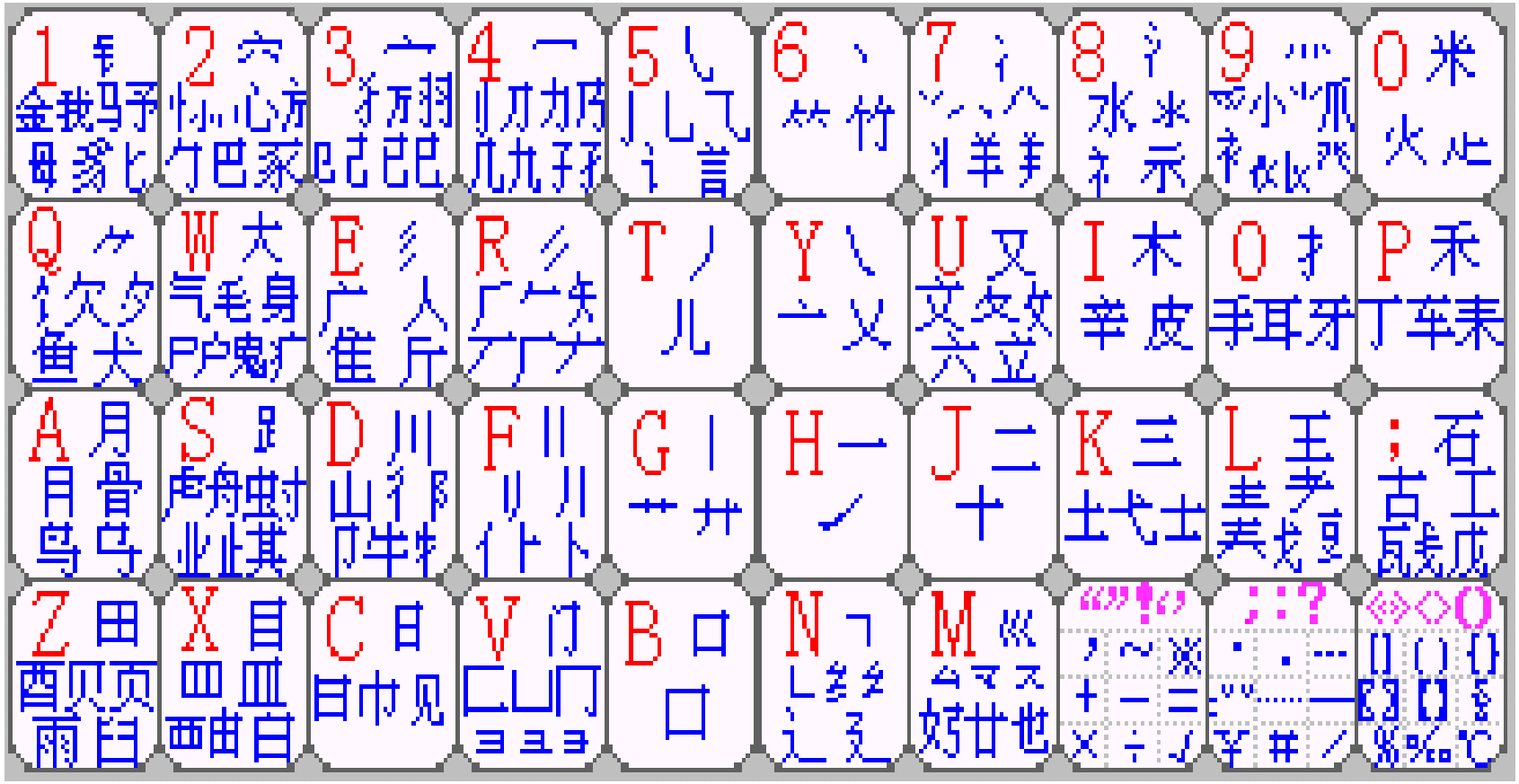


字母数字映射图

4.3.8 大大键盘编码

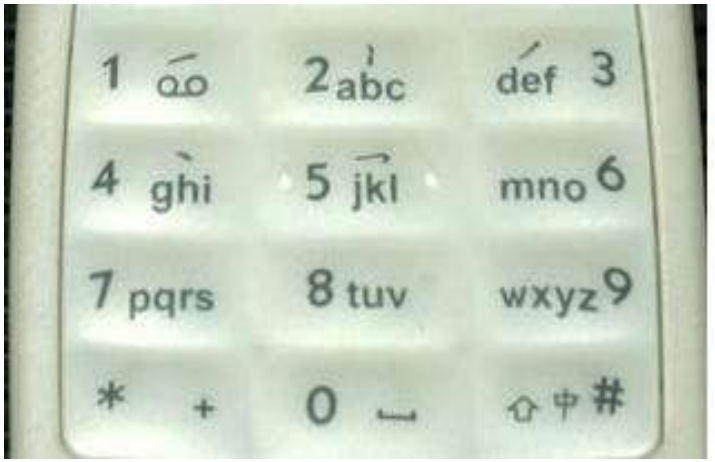
- 大大键盘编码是指编码的码元除了大键盘上的26个字母之外，还包括10个数字和其它符号，这类编码方案大多是早期的编码方案，如字元编码、钱码、陆码、绿色拼形等；
- 提出这类编码方案的目的是为了增加编码空间，从而降低重码率，实现汉字快速输入；
- 目前，这类编码方案已经很少见了，因为这类编码方案不易学和较难记忆。

绿色拼形编码的键位分布图



4.3.9 小小键盘编码

- 小小键盘编码是指只用**5**个数字来进行编码，也就是说，其只有**5**个码元。
- 小小键盘编码方案主要是应用在手机等数码设备上，几乎所有的手机笔形编码都采用小小键盘编码。



Nokia 笔画输入法小小键盘图

4.4 数码键盘方案

- 汉字数码是指用“0”—“9”这10个数字对汉字的单字和词组进行编码，使得只需用小键盘就可以完成汉字的输入，并可移植到手机以及各类PDA产品上使用。

4.4.1 纵横码

- 在纵横汉字编码方案中，把笔形分为10类，分别用“0”到“9”这10个数字表示。笔形与数字代码的关系可通过下列口诀记忆：

  “一横二竖三点捺，叉四插五方块六，七角八八九是小，撇与左钩都是零。”

- 取码规则是将汉字看成一个方块字，取汉字四个角的笔形为有关编码。类似四角号码取码规则。
- 部分汉字的取码实例：

  人（8） 中（5） 十（4）重（01） 要（14）

  喜（46） 事（50）

4.4.2 五笔数码

- 五笔数码按照笔画进行编码。笔画分为“横”、“竖”、“撇”、“捺”、“折”五种，分别用“1”、“2”、“3”、“4”、“5”作为代码。下表为基本笔画代码表：

代号	基本笔画	名称	笔画走向	笔画变形
1	一	横	左→右	
2	丨	竖	上→下	亅
3	丿	撇	右上→左下	
4		捺	左上→右下	丶
5	乙	折	带转折	𠂆𠂇乚𠃌冂

6键6码键盘图

9键9码键盘图

4.4.3 统一码

- 该方案取5种基本笔画：“横（一）”（含“提”）、“竖（丨）”（含“竖勾”）、“撇（丿）”（包括“啄”）、“点（丶）”（含“捺”）和“折（乙）”（包括左折和右折），并且将这五种笔画赋予顺序值“1”—“5”。
- 数字统一码将汉字结构归纳概括为四种基本结构。它们是：上下结构，左右结构，包围结构，嵌套结构。同时规定一个汉字可以取一至六码。
- 字445576　各359251　右689

4.5 汉字编码国家标准

国家标准	说明
GB13000.1	《信息技术多八位编码字符集（UCS）》
GB18030	《信息技术 信息交换用汉字编码字符集 基本集的扩充》
GB/T18031	《信息技术 数字键盘汉字输入通用要求》
GB15834	《标点符号用法》
GB/T19246	《信息技术 通用键盘汉字输入通用要求》，

4.5.1 GB/T 18031

- GB/T18031《数字键盘汉字输入通用要求》
  - 被编码的汉字应包括GB2312或GB13000.1或GB18030中定义的全部汉字；
  - 汉字编码的编码元素（码元）只能是“0”—“9”；
  - 定义了五种基本笔形和汉语拼音符号的键位。

4.5.2 GB/T19246

- GB/T19246《通用键盘汉字输入通用要求》
  - 被编码的汉字应包括GB18030中定义的全部汉字和现代汉语标点符号；
  - 汉字编码的编码元素（码元）只能是“a”—“z”这26个字母，“0”—“9”可以用作编码的辅助信息，如汉语声调、重码选择等；
  - 定义了GB15834中的23个标点符号的键位。

4.5.3 国家语委的规范

规范	说明
GF3001	《信息处理 GB13000.1字符汉字部件规范》
GF3002	《GB13000.1字符集汉字笔顺规范》
GF3003	《信息处理用汉语拼音方案表示规范通用键盘》

- GF3001《GB13000.1字符集汉字部件规范》
  - 对GB13000.1中的20902个汉字进行逐个拆分，经归纳和统计后定义了560个基本部件；
  - 基本部件是末部件，是最小的不可拆分的部件；
  - 基本部件可以组成成字部件使用，但不得组成非字部件；
  - 字拆分成部件时，应遵循“相离、相接可拆，交重不拆”的原则.

- GF3002《GB13000.1字符集汉字笔顺规范》
  - 定义了汉字的五种基本笔形是：横、竖、撇、点、折，分别用1、2、3、4、5来表示它们；
  - 给GB13000.1中的20902个汉字分别定义了规范笔顺。
- GF3003《汉语拼音方案表示规范通用键盘》
  - 分别用1、2、3、4、5来表示汉语拼音中的五个声调；
  - 用字母“v”来表示汉语拼音中的字母“yu”（“u”上面加两点）。

4.5.4 汉字键盘编码性能指标

- 易学性
  - “学会使用汉字编码输入系统的时间应尽量短，并应符合使用汉语作为母语的使用者的思维习惯”。GB/T18031对数字编码更进一步提出要求：“做到上手能用”。
- 汉字输入平均码长
  - 定义：在输入给定的测试样本时，测得的输入每个汉字的平均击键次数；
  - 平均码长＝输入样本的击键次数/测试样本总字数

GB/T19246《通用键盘汉字输入通用要求》给出的指标

编码类型	平均码长（键/字）
汉语拼音，笔画为主的简易编码	＜3.2
形码（部件码）、音形码（形音码）、双拼	＜2.2

GB/T18031（数字键盘）给出的指标

编码类型	平均码长（键/字）
逐字字段输入字	＜6
字、词混合输入	＜4

- 重码字词键选率
  - 定义：在输入给定测试样本过程中，通过重码选择键确认输入的汉字字数与测试样本总字数的百分比；
  - 重码字词键选率＝(重码选择确认的字数/测试样本总字数)×100％

GB/T19246《通用键盘汉字输入通用要求》给出的指标

编码类型	重码字、词键选率（%）
汉语拼音，笔画为主的简易编码	＜6
形码（部件码）、音形码（形音码）、双拼	＜1.5

GB/T18031（数字键盘）给出的指标

输入方式	重码字、词键选率（%）
逐字字段笔画、部件码输入	＜10
字、词混合笔画、部件码输入	＜8
逐字字段拼音输入（10键位）	＜13
字、词混合拼音输入（10键位）	＜12
逐字字段拼音输入（8键位）	＜14
字、词混合拼音输入（8键位）	＜14

作业

- P73. 1-10