汉字信息处理系统

李治柱　陈永乐

上海交通大学 微机研究所
科技交流室

1985.11

内 容 简 介

本书全面系统地介绍了汉字信息处理的概况，系统的组成，系统的原理及其设计方法。内容丰富，理论联系实际。可供从事汉字信息研究的同志阅读，或作为高等院校有关专业教学参考。

目　录

第一章 导 论

近几年来，汉字信息处理技术发展十分迅速，出现了各种各样的汉字信息处理系统，这有力地推动了计算机在我国的普及，以计算机为标志的“信息革命”在我国已经到来。“计算机中文化与中文计算机化”已成为各行各业讨论的中心问题。

§1.1 文字信息处理的意义

早期的计算机是以数值计算为主的，本世纪六十年代以来，计算机的应用已突破数值计算的框框，广泛应用于非数值性的数据处理，同样广泛用于各种信息的加工处理。电子计算机所处理的信息是以量变的形态出现的，相应于二进制的“0”和“1”，称为数字信息。而文字信息是文字或代表这种文字的语音所包含的信息。目前计算机文字信息处理技术已广泛应用于社会生活的各个领域，信息处理所包含的内容也愈来愈丰富了。

如：计算机情报资料检索，书刊、报纸、杂志的自动编辑和排版，事务处理、企业管理，办公室自动化，文字处理，文字翻译，医疗诊断，技术咨询，数据通讯等等，可以联想到人们的思维、生产和生活的一切方面。可见文字处理技术和人们的生活是密切相关的，这项技术的发展和应用，将是现代化的一个重要标志，特别是在今天这样一个信息爆炸的时代。

计算机文字信息处理体现为数据处理，用数据来表示的文字或符号，称为代码。例如用ASCII码表示所有26个英文大小写字母和所有符号，所以对文字信息的加工就是对代码数据的加工，大致分为以下三步：

- 输入信息。经输入设备把文字信息转换成代码，並送入计算机。
- 加工处理信息。根据不同的应用，由应用程序对输入的信息加工处理，得出结果。
- 输出信息。以代码形式表示的结果信息，通过输出设备还原成文字。

§1.2 汉字信息处理及国内外概况

西方国家对计算机系统的设备，技术开发得较早，应用文字信息处理技术也比较早，所以目前这方面已相当成熟。

在我国，汉字信息的研究开始于五十年代，是从汉字的查字法及汉语的拼音化开始的。计算机引入我国以后，用计算机处理汉字信息才引起人们的重视。近几年来，汉字处理在技术上和应用上得到全面迅速的发展。从世界范围看，汉字信息处理大致经历了以下几个发展时期：

五十年代中期研制了面向报社、邮电等专业部门的汉字收发装置，开始了汉字信息的早期研究。

六十年代研究并制成了各种大型的汉字照排系统，这种系统一般耗资多、体积大、专用性强。

七十年代中期，以中小型计算机为基础的小型汉字事务处理系统问世。特别在七十年代后期，微处理机和软盘技术的发展使汉字信息处理系统迅速向小型化和普及方向发展，掀起

了汉字信息处理研究和应用的热潮。汉字程序语言、汉字数据库也开始出现。

进入八十年代以来，一方面以微处理机为基础的小型汉字事务处理系统和其它小型、分散的汉字系统有迅速的发展；另一方面，大型汉字联机系统和汉字终端得到广泛的重视和应用。在这同时，自然语言的研究开始兴起，其结果势将从根本上缩小汉字信息处理和西文信息处理之间的差距。

国内汉字信息处理的研究起步较迟，但最近几年发展较快，我国在七十年代中期开始研制汉字照排系统。典型的例子有北大汉字信息处理小组等研制的照排系统、上海汉字信息处理照排系统等。与此同时，不少单位在原有的中小型计算机上进行了汉字信息处理的试验或某些简单应用。七十年代后期我国引进日本、美国的一些汉字系统，同时引进了微处理机技术和一些微机系统。在这些基础上，不少单位提出了在现有的通用微型计算机系统上扩充汉字处理功能的各种方法，并进行了试验。

几年来，我们在汉字研究、汉字编码、汉字信息压缩、汉字输入方式等方面进行了广泛的研究，取得了很大成果，特别是在八一年五月发布的国家标准〈信息交换用汉字编码字符集基本集〉，为进一步开展汉字信息处理的研究和应用打下了良好的基础。

当前，汉字信息处理技术发展的特点和趋向是：①以微处理机为基础的汉字终端和汉字事务处理系统是当前发展的重点，对这些设备的要求是小型化、低价格、高适应性。②有计划地发展在汉字数据库和汉字终端支持下的大型汉字联机系统，开拓汉字信息处理应用的新领域。③发展汉字程序语言和汉字信息处理应用软件。④使系统结构和处理方式由集中向分散方式发展。⑤开展自然语言的研究，改变汉字信息处理的方式，使信息处理向更高级的形式发展。

§1.3 汉字信息处理研究的主要内容

汉字信息处理是一门新兴的综合性科学，必须圆满的解决汉字信息的输入、存贮、处理和输出，其基本内容包括汉字输入输出，中文文件生成及管理，汉字数据库，汉字信息的传递及通讯，汉字程序语言的开发以及各种汉字应用软件的研究。

西文是拼音文字，字母少，汉字是表意文字，字多且繁，所以汉字信息处理比西文信息处理要复杂。但汉字信息处理的目的和基本原理与西文是相同的。在代码一级的处理上，如文本编辑，文件管理等方面几乎是一样的，所以建立汉字信息处理系统的关键是解决好汉字的输入、输出以及汉字库。汉字信息处理系统的软件，可以在西文软件的基础上加以改造或扩充，得到中西文兼容的系统软件。以下分四个问题加以叙述：

- 汉字的输入
- 汉字的存贮
- 汉字的输出
- 汉字信息处理系统

§1.3.1 汉字的输入

汉字输入通常分为自然输入、编码输入和汉字键盘输入三大类，自然输入是指对汉语的汉字和汉语的语音识别，属于模式识别范畴，也是汉字输入的最高形式和最终形式，目前还处于试验阶段。

编码输入，由于汉字字数多，编码方法是一个重要的问题，虽然可以用多种方法实现汉

字编码，但要得到一种功能上全优，并且适用面很广的汉字编码方法却非容易之事。

汉字编码输入按照所采用的具体方法，可以分成许多种，例如按照字形特征编码的，称为字形编码法；按照汉字发音特征编码的，称为声韵编码法；另外，有用形音结合作为编码依据的，还有用汉字的其它特征作为编码依据的，等等，方法很多，统称为汉字编码输入法。

汉字键盘输入方法的优点是直观，操作者容易学习、掌握，没有重码问题。但它所用的键盘是通称的整字键盘，体积较大，造价贵，并且输入速率较低。整字键盘有两种主要型式，一种是早期使用的主辅键式(或称移位键式)汉字键盘，盘面上布置有约400个键，每个键上收容9～12个汉字，用双手操作，由于这种键盘体积大，造价高，不易推广使用。目前流行的汉字键盘是一种笔触式汉字键盘，盘面上可收容3000～4000个键位，每位一字，体积可以做得较小，造价也低于前一种整字键盘，但国标码规定的汉字为6763个，故整字汉字键盘除了盘内字外，还要解决盘外字的输入问题。目前一般采用直接送入汉字代码或者汉字字根组合编码的方法实现盘外字的输入。

汉字编码输入方法很多，如字形编码输入法、声韵编码输入法、形音结合法、字形双拼法以及联想式编码法等。以下对几种编码输入方法作简要介绍。

(一)笔画(Stroke)编码法。按组成汉字的基本笔画(五种或八种)编码，输入组成一个汉字的笔画时，可按照笔顺的先后。这种方法的优点是输入一个汉字就像书写这个字的过程，容易学习操作。整个汉字的输入码是组成该字基本笔划代码的组合，输入码较长，而且是不等长码。

(二)汉字声韵编码法。是以汉字发音为基础的编码输入方法，它也包括很多种类。例如，可以按照较早流行的汉字四角号码发音规则来编码。国内近几年来发展的声韵双拼编码法，以汉字普通话发音的声母、韵母为基础，分别以相应的字母代表各个声母、韵母作为音符。这类编码输入方法都可以用字母数字键盘作为输入工具。

(三)形音混合编码法。混合编码可以利用汉字形、音两属性的各自特点，在区分字种上有可能节省信息或得到其它效果。有的混合编码方案中，除了形、音特征外，再添加其它信息。例如字义。

(四)联想式汉字编码(Conceptualize Chinese Characterencoding)输入法。这是一种联机操作的汉字输入方法。把在一起使用频度高的汉字代码连接起来，输入操作时，键入为首的一个汉字的编码，就在荧光屏上显现一串汉字，再用光笔或键盘按所需次序挑选出来，就完成了在系统中输入这些汉字的目的。

(五)字形双拼码。目前国内外的各种编码方案，各有长处，但是不论哪种方案，在构造汉字信息处理系统时，都存在汉字字数多，字库容量大等问题，而且都需要经过专门训练才能使用，故一时难以推广。为此我们交大微机所提出一个字形双拼方案，并以此为基础研制了Mic-58c汉字智能终端和Mic-PC/XT中西文微型机系统。

字形双拼方案的特点首先是拼，但最多只能双拼。在字形双拼法中，一个汉字不是整字就是双拼字，二者必居其一，其他情况是没有的，这就使方案基本上保留了拼字法的优点，即大量减少了基本字的数目。同时，由于在拼法上进行了节制，因而克服了一般拼字方案的缺点，做到拼法简单，拼而不乱，学起来非常容易，用起来直观方便，任何人(包括外国人)均可不经专门训练就能够在双拼中文输入键盘上，见字触键，将汉字输入计算机，并立即在显示屏幕上显示出该字。根据统计，常驻一汉字系统的汉字字数在7000字左右，即可充分满

足各种用户的要求。通过对大量汉字进行分析、归纳和统计得知，按字形双拼方案来处理汉字时，基本字数为1787字，便可组成通用的7200多汉字，我们就是以此为依据，进行字形双拼码中文微型机系统的设计和研究的。该字形双拼方案能处理的汉字字数超过7200字，不仅包含了全部已公布的国家标准信息处理交换码，而且很多生辟专业字均可用此法合成。

由于用不同的输入方法得出的汉字键盘码差别很大，为了便于不同的系统相互交换汉字代码，需要确定一个统一的码制，或称为标准汉字交换码。我国已于1981年颁布了作为国家标准GB2312《信息交换用汉字编码字符集(基本集)》，目前国外研究的大多数汉字系统都采用国标码，基本集中包括汉字，几种外文字母，数字和符号，总数为7145个，所有国标交换码对每个字或符号用二字节表示，实际是用14个二进制位。

§1.3.2 汉字存贮问题

所谓汉字存贮是指两层意思，一个是指汉字代码信息的存贮，一个是指汉字字形信息的存贮。前者与汉字数据库设计有关，后者与汉字字形库设计有关。

汉字数据库中的汉字代码信息存贮，可以采用等长码，以节省数据库的存贮空间。换句话说，有存贮数据压缩问题。

计算机信息处理用的汉字字模，按用途可以分成精密型字模和通用型字模两大类。精密型字模用于精密汉字编辑排版系统，这种汉字字模，对字形、字体、字号等都有严格的要求，必须适合印刷出版行业所定的规格，其精密度要求分辨率至少在64×64点阵以下。通用型汉字字模适用于一般的汉字信息处理系统，应用面广。目前，我国已经制订了通用型汉字点阵字模的国家标准。在字模存贮设备和输出印字设备能达到的技术条件下，要求较高的文字质量。通用型的汉字字模需用的点阵(Dot matrix)结构有以下几种。

简易型：15×16点；

普通型：24×24点；

提高型：32×32点；

一个汉字系统必须设定一个存贮汉字的字形信息(Chinese Ideogrphic information)的存贮体系，称为汉字字模库(Chinese Font Bank)或汉字字形发生器(Chinese Character generator)，其存贮容量是比较大的，需采用多种信息压缩技术。

对于汉字字模存贮介质的要求主要有两点：一是单位存储量的成本要低；二是读取信息的速度要快。在普遍采用大规模集成电路存储器(如EPROM,ROM或RAM)以前，主要采用磁芯存贮器和磁盘存贮器存贮汉字字模。磁芯存贮器虽然可以提供较快的读取速度，但是包括译码和驱动线路在内，单位存贮量的成本高，并且体积大，消耗电流大，不能推广应用。用磁盘存贮器(包括硬盘和软盘)存贮汉字字模，虽然单位存贮量的成本较低，但从磁盘中取出字模的平均等待时间较长，字模输出速度低。近几年来，由于大规模集成电路存贮器的价格不断降低，用它作为汉字字模的存贮是一种较为理想的介质。起初人们把常用字(例如国标码中的一级汉字，或其中的大部分)存入大规模集成电路存贮器，非常用字(如国标码中的二级字)存入磁盘存贮器。这种存贮方法，称为两级存贮体制。因为使用频度高的字模是从高速的集成电路存贮器中取出的，平均来说仍可得到较高的字模输出速度，又使整个汉字字库的成本不致过高。目前这种方法已用得不多。一般的汉字信息处理系统所用的汉字库都是固化在EPROM内作成硬字库；或者存在软盘上，用时从软盘上一次调入内存。这样字形输出速度问题就基本上解决了。

还有压缩信息(Information Compression)的存贮方式。汉字压缩信息的原理和技术有多种，例如采用只收存汉字字根的汉字字模库如嵌套式汉字字模发生器；利用向量组字法产生汉字字模；用霍夫曼树型压缩法存贮汉字信息等。特别是用向量组字法和字根字模库相结合的汉字字模库技术，可以得到较大的压缩比，目前已得到较广泛的应用。利用这种汉字字模库技术，要组成和输出某个汉字时，用地址链的方法把组成该汉字的有关字根从存贮器中取出，这些字根是用坐标、斜率、长度等原始信息记录的，经过信息恢复；比例尺寸选择；字根间相对位置的确定后，就得出相应的字模。这样的汉字字模库，可以较多地节省存贮成本，特别是在要求收容的汉字字数较多时，可以不把汉字分级，不需用磁盘存贮器作辅助存贮，只需用数量不大的ROM或EPROM组件(一般在30～60K 字节)就可以产生7000～20,000个汉字。这种汉字字模库的缺点是产生的汉字字模质量不如整字存贮的汉字质量好。此外，由于需用软件方法组成汉字，所以输出速度较低。

随着集成电路ROM器件的集成度的提高和价格的不断下降，目前已普遍采用ROM器件作为汉字整字字模库的存贮介质，並制定了字形和点阵规格标准，成批制作了国标一、二级汉字的标准字库。至於字根式汉字字模库，还可以进一步改善字形质量，提高软件功能和汉字组成速度，这样，这种汉字字库在某些应用场合，仍有它的使用价值。

§1.3.3 汉字的输出

汉字输出技术主要包括汉字显示(Chinese Character display)和汉字打印(print)。由于汉字字模的点阵较西文字符的点阵密，因此对于显示器和印刷机的技术要求比显示和打印字符的同类设备自然要高些。

汉字显示目前主要采用荧光屏显示技术。对于汉字显示器的技术指标，主要是一帧画面能显示的汉字数，这和荧光屏的分辨率(Resolution)显示器的视频通带等指标有关。目前，一般都采用光栅扫描的显示方式。通常，一幅画面的字数在500～1,000范围，字数愈多，要求荧光屏的分辨率愈高，是有一定限制的。例如，显示12行汉字，每行40字，若显示的汉字点阵为15×16点，则要求荧光屏的分辨率为320×640点；汉字点阵为24×24点，即使每屏的显示字数降为360字(30字×12行)，则要求分辨率为480×800点以上。要求再增多显示字数，或使显示器兼有图形显示功能时，需要再提高显示器的分辨率，同时也会提高显示器的成本。此外，视频范围，与一帧显示的汉字数，汉字点阵多少，以及显示器帧频等成比例增加，其上限值有一定限制。汉字显示器的刷新存贮(retlosh storage)方式，分为缓存汉字在字形发生器中地址的刷新方式和对画面信息逐点缓存的刷新方式两种，后一种方式可以兼容对图形的显示，对于要求分辨率较高的荧光屏，和一定视频范围内显示，目前技术上都可以解决。

对于汉字印刷机，主要技术指标是打印分辨率和印字速度、用纸要求等。目前的汉字印刷机，对印字分辨率的要求在4～10线/毫米的范围。而根据不同的印字原理和机制，各种类型的汉字印刷机，其印字速度可以在每秒数十字到数千字的很宽范围内变化。按照印字速度可以把它们分成低、中、高三档。

低速档：印字速度为每秒数十到上百字，属于这档的印刷机如针式汉字打印机；热敏式汉字打印机；喷墨式汉字打印机等，其中针式汉字打印机是目前应用得多的一种。这档汉字打印机分辨率为4～5线/毫米。

中速档：印字速度为每秒数百到一千字，如简易型激光扫描汉字印刷机，发光二极管静电转印式汉字印刷机等，分辨率在5～8线/毫米。

高速档：印字速度为每秒一千字以上，最高可达每秒上万字。如高精度激光扫描汉字打印机，光纤管(OFT)转印汉字印刷机等。这档印刷机分辨率在8～10线/毫米。

上述各种类型的汉字打印机中，虽然针式打印机的技术较低，但是由于它结构简单，成本低，使用维护方便，维持费用低，所以目前绝大多数汉字系统，特别是微型、小型机汉字系统，都配用这种打印设备。目前流行的汉字针式打印机，根据汉字字模点阵规格不同，设计成16针头式，24针头式，交叉排成两纵列，打印速度在40～60字/秒。另有一种称为梳齿状的汉字针式打印机有多个针头间隔一定距离排成横列，每个针头可以在一段距离内几个汉字的区间中移动，这些针头同时在横向打印，打出横向的一排点阵，纵向点阵由走纸动作形成。这种针头打印机的输出速度可达100字/秒以上。这种打印方式的另一个优点是点阵打印的输出方式和显示器的电视光栅扫描方式是一致的，在输出控制技术上较为方便。

某些微型机汉字系统，也采用普通7针或9针的字符打印机打印汉字，用较少的针头靠往返打印两次或三次形成一行汉字。

其它类型的汉字印刷机中，像激光扫描汉字印刷机是一种较新发展的机种，在印字速度和分辨率方面都可以达到较高指标。它使用普通纸，纸幅可以在较大的范围内变化。在降低造价和提高工作可靠性的情况下，对于小型或中型、大型机汉字系统的配置，可望得到较广的应用。

§1.3.4 汉字信息处理系统

汉字信息处理系统是指利用计算机等技术按不同的应用目的建立起来的处理汉字信息的实用系统。

一个汉字信息处理系统由计算机，汉字库、汉字输入输出设备等组成，其中计算机具有支持汉字处理的系统软件、实用程序及应用软件。从信息处理的角度看，汉字信息处理和西文信息处理无本质区别。为了加速汉字系统的研制开发，目前建立汉字信息处理系统的理想办法是充分考虑汉字双字节特点，对西文计算机系统做适当改造，做到中西文兼容，共享国际信息处理的成果，充分利用已有西文计算机的全部资源，这是开发汉字信息处理系统的捷经。

为了充分开发和利用中文信息资源，利用计算机进行现代化管理，办公事务管理，生产管理等等，迅速推广计算机在我国的应用，必须建立各种各样的汉字信息处理系统，通用的、专用的、分布式的等等。

目前我国已经出现了许多形式的汉字系统。汉字系统就其处理方式来看可分集中处理和分散处理两种类型。

在集中处理方式下，一切处理工作，包括汉字的输入输出，各种汉字例行程序以及应用程序的执行等均由主机完成。王安公司的VS中文系统即可看作一个集中处理的例子。该系统可带多个终端，这些终端以分时方式享用主机，汉字输入输出、文件编辑等一概由主机来处理，汉字库也在主机上，离开主机终端不能进行任何工作。国内有直接用DJS—6，DJS—130等中小型机进行汉字信息处理试验的。这些试验中通常让主机来承担包括汉字输入输出在内的各项任务，也属于集中处理方式。集中处理的缺点是由于汉字输入输出的开销很大，严重限制了应用程序的开发，大大降低了系统的效率和性能，此外，系统的灵活性差，造价也高。

分散处理方式是将汉字信息处理的任务分解，并交给几个处理机去分别完成。由于汉字输入输出和文本编辑占用大量运行时间和存贮空间，而且是个较为独立的环节，所以分散处

理方式中最常见的就是用专门的处理机来处理汉字的输入输出和文本编辑。微处理机的出现为分散处理提供了有利的条件。特别是在微处理机的基础上建立起来的汉字智能终端，确实是解决汉字信息处理中的输入输出和建立硬字库等基本问题的有效手段。近几年来国内外已研制了多种以微处理机为基础的汉字智能终端，如：MiC—58C等。

MiC—58C是上海交大自行设计研究的汉字智能终端。它采用从终端级实现联机的先进技术，采用RS-232C标准接口，挂接主机方便，在不改主机硬件和软件的情况下实现中英文兼容。汉字库为双拼压缩硬字库。显示器为12英寸绿色高分辨率显示器。输入键盘为字形双拼码笔触式中键盘。

汉字信息处理系统按功能也可分为专用和通用两种。专用汉字系统包括汉字照排系统及一些用于单项业务的系统。通用汉字系统又有独立汉字系统和联机汉字系统。前者一般是以微型计算机为核心构成的系统，类似于西文的字处理系统。后者是指大中小型机及汉字终端构成的汉字联机系统。

下面我们简单介绍一下汉字微型机系统和汉字联机系统的组成。

1.汉字微型机系统：

这种汉字系统和西文微型机系统一样，它是一种通用的能独立处理汉字信息的系统。作为汉字信息处理的专用设备主要有汉字输入键盘，汉字字库及汉字输出用的CRT和打印机等。系统中所有其他设备都是通用类型，如八位或十六位微机，软盘驱动器，硬盘驱动器，磁带机等，系统的硬件连接基本上与通常的微型机系统相同。当然，如果采用汉字编码输入，汉字键盘也可以用国际标准字符键盘代替，如果不要求印字速度或字形美观，汉字打印也可以采用通用的西文字符打印机，用它分几次打印出汉字。但是作为一个完善的汉字信息处理系统，还是需要配上上述几种汉字专用设备。总之，作为一个处理汉字信息的通用系统，其硬件系统应该有：

(1)汉字信息输入设备。考虑到目前汉字输入方案的实际情况和用户的方便，在系统配置上既要有为非专业人员使用整字输入键盘，又要有为专业人员使用的小键盘，而这种小键盘需与通用的ASCII字符键盘兼容，并实现多种编码方案的输入。另外，整字键盘还要考虑盘外字的输入方法及用户自定义键区，词组区等。

(2)汉字字库。系统最好能配置有二种点阵形式的汉字库，15×16点阵用于显示输出，24×24点阵用于打印输出。字库容量至少4000汉字，一般应包含国家标准交换码字符集的6763个汉字。在系统运行时，至少有4000字在EPROM或RAM内，以保证字形输出速度。字模应采用国家标准字模。

(3)汉字输出设备。显示器每屏至少应能显示汉字480个以上，以满足一般文件或报表的需要。显示器可采用12吋或14吋以上的单色或彩色显象管，无闪烁感，采用光栅扫描体制。在同一屏上应能同时显示西文、汉字及图象。打印机打印速度应达到35字/秒以上。在同一页上印出大、中、小三种字形。

(4)内存容量一般要求512KB以上，以满足汉字及图象处理对内存的开销，同时又能留给用户一定容量的存储空间。

(5)联机接口是采用异步/同步的串行接口，标准的转换电平及通用的标准通讯接口。传输波特率为54～9600bPS范围内。

目前，主要是通过改造西文微型机系统来实现上述对硬件的要求，同时又不破坏原系统

的功能。

构成汉字系统更大量的工作是软件的研制。通用汉字系统的设计目标是：做到中西文兼容，能对中西文进行混合处理，不降低原系统的效率。汉字信息处理软件应该包括：汉字的输入、输出，屏幕编辑，文件管理，打印控制及联机程序等。其中屏幕编辑应包括格式编辑和非格式编辑二种。前者用来编辑所输入的报表或记录，后者用来编辑一般文件的输入，编辑命令能对中西文混合进行编辑；文件管理主要有：显示汉字文件目录，按名存取文件，删去一文件，给一文件改名，改变文件的保护特性，两个文件合并成一个文件，文件传送等；打印控制主要控制打印机启动，走纸，字体变形，打印格式选择或报表语言等；联机程序完成汉字系统与大中型机的通讯联接。

另外，还需要一些汉字实用程序支持。如：汉字库的建立与维护，以便用户增加外字字形到字库中去，选择常用字组成常用字库，修改字库中的汉字字形等；还有各种代码转换表，建立汉字属性库等。

这些支持汉字处理的软件，可以建立在用户级上，也可以把主要的汉字处理模块建立在操作系统一级上。

总之，一个汉字系统应有能处理中西文的操作系统，有汉化的高级语言以及中文数据库管理系统，有各种应用软件。

2.汉字联机系统：

它是由大、中、小型机为主机加若干个汉字终端所组成，其系统框图如图1-1所示。汉字终端有简易终端及智能终端二种，简易终端由CPU，汉字库，汉字显示器，汉字输入键盘及联机通讯接口等几个部份组成，它只有在联机时才能进行汉字的输入输出，汉字的文本编辑等功能。汉字智能终端除了具有上述设备外，一般还带有磁盘驱动器及一定容量内存，还配有汉字打印机。它能脱机进行汉字输入输出及中西文文本的编辑。一般讲，它具有一个汉字信息处理管理程序，而不配备汉化的高级语言，中文数据库管理系统及各种中文数据处理的

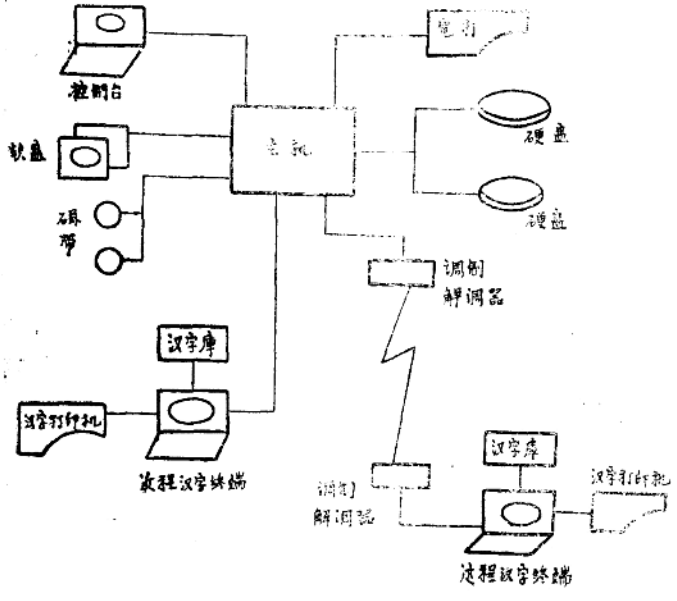


图1—1　汉字联机系统框图

应用软件，也没有处理汉字的操作系统，这些软件资源，汉字终端在联机时可以共享。

为了实现联机运行，汉字终端上有通讯接口，能与主机进行近程或远程通讯。主机系统是一台大、中、小型机，它不配汉字字库，主机与终端之间的汉字通讯是采用汉字交换码，它要求主机的操作系统既能处理一字节的西文字符，又能处理二字节的汉字代码，要求主机有中文数据库管理系统及丰富的应用软件，实际上就是要在主机的软件系统中增加两字节长的汉字数据类型，并将它与一字节的西文字符类型区别开来。

与简易终端相联的主机系统还应包括汉字智能终端在处理汉字信息方面所具有的功能。

有些汉字联机系统是用汉字前端处理机与终端相连，前端机是一个具有一定功能的小型机系统，利用它来代替主机对远程终端进行控制，这样可使运行在分时系统中的主机从慢速的输入输出中解放出来，提高系统运行效率。

汉字联机系统的重要部分是联机通讯。通讯方式多采用半双工通讯方式，也可以采用双工方式，传输率目前一般为75～9600bpS之间，联机的接口采用国际通用的标准接口RS232C，也可在终端上设置专用接口，后者可用硬件实现数据传输中的某些基本功能，如传送码出错校验等。主机和终端可采用中断方式或查询方式进行通讯，通讯时必须按照主机相应规定的通讯规程来应答，当处于查询方式时，汉字终端就可像主机其他西文终端一样，调用主机的命令，在主机操作系统支持下工作。

在早期的联机系统中，不改变主机的操作系统等系统软件，只是在主机中用汇编语言编制汉字服务程序。用它实现联机控制功能，即接受汉字终端送来的信息或向汉字终端发送信息，解释各种控制命令，对汉字进行输入输出处理等。主机中的汉字是有一个特定标志符开始的字符串，只有当输出汉字时才能对这串字符作汉字输出处理。

第二章　汉字输入技术及设备

§2.1　概述

计算机的汉字输入技术是实现中文信息计算机处理的关键问题之一。目前汉字输入计算机的基本方法可分为三大类：汉字编码输入，汉字键盘输入和自然语言输入。

1.汉字编码输入：

汉字编码输入一般是指小键盘编码输入，即借助于编码技术，在小键盘上，用26个字母键或10个数字键完成汉字的输入。如何用这三十个左右键符的排列组合来构成汉字的输入码呢？通常是根据汉字的各种属性，分别进行编码后组合而成汉字输入代码。汉字的属性包括：音(声、韵、调)、形(边旁、部首、笔形、部位、结构)、频(使用频率)和搭配关系(组词能力)等。汉字输入编码基本类型如下：

① 形码——是根据汉字的字形结构特征信息进行编码。典型的形码有：

笔形编码法，三角编码法，仓颉编码法等。

② 音码——是利用汉字的字音信息进行编码的方法。典型的音码方案如：

拼音电报方案，声韵双拼编码法等。

③ 音形结合码——利用音、形、义和频度信息结合编码法。如：

见字识码法，SYX编码法等。

2.汉字键盘输入：

采用字根、字元从键盘拼形输入或整字从大键盘直接输入。根据键盘的键位多少可以分为大、中、小键盘。

3.自然语言输入：

采用语音和文字识别手段的高级输入方式，目前还在实验阶段。形码和拼形码发展结果将是文字识别；而音码发展结果将是语音识别；词汇码发展结果将是自然语言的处理。

§2.2 小键盘编码输入

所谓小键盘就是指标准的西文计算机键盘，其中包括26个英文字母键，10个数字键，若干个常用标点符号键和一些特殊的功能键。目前，此类小键盘的键位数随着计算机应用的发展而有日趋增多的趋势。

新增加的键位一般都是越来越复杂的功能键，其基本标准键位是不容轻易改变的。如果采用此类标准小键盘作为中文计算机的输入设备，则按键指法与西文统一，操作人员熟悉了某种汉字编码输入方法后，可以实现盲打，得到较高的输入速度，适合于专业人员使用。其缺点是，按键次数多，输入码长，信息的冗余度大。下面简述几种较常用的编码输入法。

§2.2.1 三角编码法

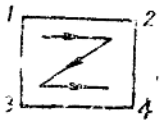
三角编码法根据取角和取形二原则，两相配合处理。取角原则注重顺序和地位；取形原则在於选取基本符号，以确定号码(每个字的号必须为六位)。王安公司的中文信息处理系统采用三角码输入方案。此方法的缺点，是要死记基本符号，需经过一段时间的训练，才能熟练运用。

一、取角原则

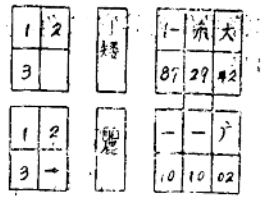
取角原则的基本顺序是从左到右，由上到下。如下图所示：

1.若一字的四角均属基本符号，则只取前三角，最后一角不取。

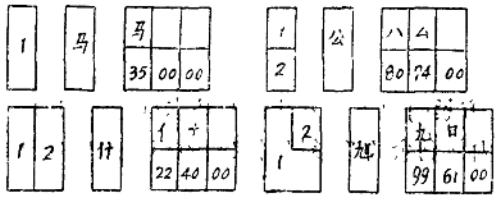
如：



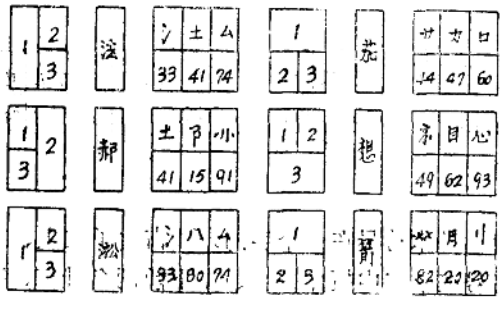
1．左上角
2．右上角
3．左下角
4．右下角



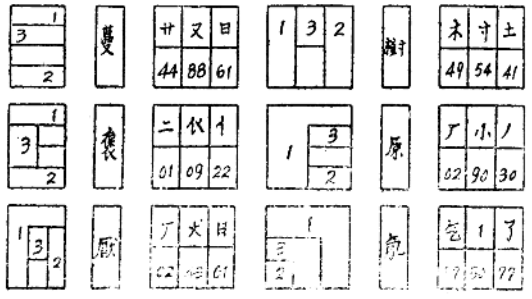
2.若一个字由一个或二个基本符号组成，最后四位或二位号码，应用0000或00补足。如：



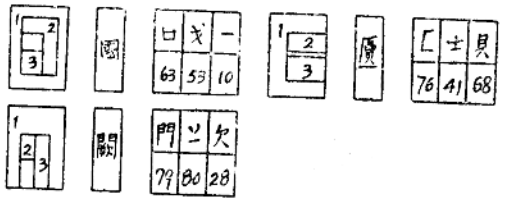
3.若一字中相邻二角属于同一个基本符号，这二角应并作成一个角，即该字只有三个角，因此恰合三角编码原则，其编码取角有下列各种形状：



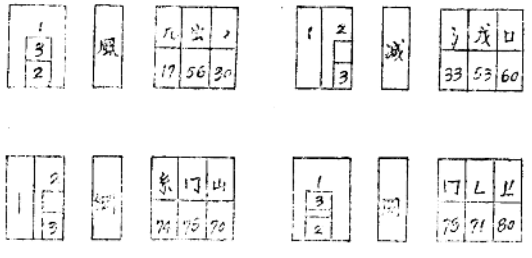
4.若二个基本符号完全占去一个字的四个角(此四角可称外角)，最后二位号码应取字中剩余的最左上方的基本符号编码。这是第三角(此角称为内一角)。



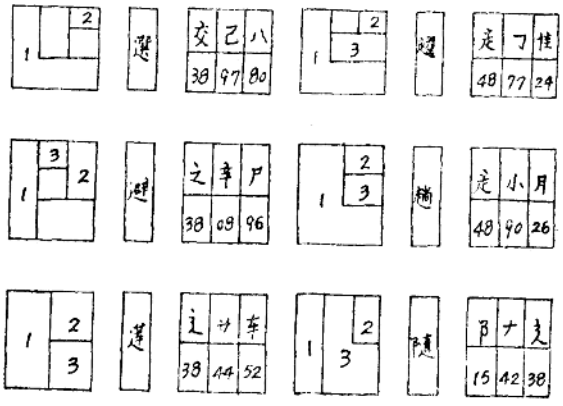
5.若一个基本符号完全占去了一字的四个角，如口、上、门等，最后四位号码应取剩余内部的基本符号作为其它二角，这二个内角的编号顺序仍须按基本取角原则和下述取形原则继续依次编号。



6.若遇左、上、右三方紧闭而下一方开口的基本符号，如凵、几、戊等所构成的字，最后的几位号码，应就该字内部剩余的笔形根据取角的相反顺序(即由下而上、从右到左的顺序)依次编号。



7.若遇有辶、廴、走，基本符号在一个字的外围占去了左上、左下及右下三个角的位置时，则辶、廴、走基本符号在该字中概作左上、左下前两角计算，剩下的四位号码仍应按照基本取角原则继续依次编号(若辶、廴、走，基本符号在字的内部时，则辶、廴、走仍以占三角论)。



二、取形原则：

1.凡已使用过的基本符号应视为全部从该字中移去，不再留用。

2.在同一角有两个或两个以上的基本符号可以取用时，应采取最多笔画的基本符号予以编号。

若笔画相同时，则取代码最大的基本符号。

	正			误		
在	𠂇	丨	土	𠂇	亻	土
	42	20	41	42	22	41
雀	少	亻	王	少	住	
	92	20	14	92	24	00
求	寸	丶	氷	十	丶	水
	54	31	32	54	31	33
親	辛	目	八	辛	目	小
	08	62	80	08	52	90

	正			误		
産	文	厂	生	文	厂	生
	04	02	25	04	02	14
流	氵	冫	川	氵	冫	儿
	33	01	99	33	01	17
爾	尒	冂	×	尒	八	乂
	19	78	43	12	80	43

3.为采用最多笔画的基本符号时，可以把竖笔(其它笔画则绝对不可)分段至有其它笔画阻挡时为止(即应就该角位置，就近采用一个最完整、最多笔画的基本符号)。

4.凡遇有钩、拐等笔形之笔画(即基本符号表中70～79之主符号)及(口)等笔形均不可分拆。

	正			误		
事	申	尹		申	彐	一
	51	96	00	51	73	10
毛	壬	乚		丿	毛	
	14	71	00	30	50	00
東	亻	小	一	木	彐	一
	30	90	10	29	73	10
段		殳	+	丿	殳	+
	76	88	40	30	88	40

	正			误		
勃	十	力	寸	土	力	寸
	40	47	54	41	47	54
庚	广	丈	彐	广	大	1
	02	42	73	02	58	20
技	木	十	又	木	土	又
	49	40	88	49	41	43
至	一	土	厶	一	土	一
	10	41	74	10	40	10
由	丨	凵		丰	𠃌	
	20	64	00	50	70	00