汉字信息处理

彭寿全　黄可　编著

电子科技大学出版社
·1994·

[川]新登字 016 号

内 容 提 要

本书从建立汉字信息处理系统的角度出发，介绍了汉字属性、汉字编码、汉字输入方法、词处理、字形存储等基本理论和技术，介绍了汉字信息处理的若干国家标准、推荐的内码方案和 ISO 10646 等；从建立中西文兼容系统出发介绍了汉字 I/O 设备、汉字系统的配置以及计算机系统的汉字化技术和方法。全书共八章，主要内容有汉字信息处理及系统的基本概念、特点、内容和发展现状；汉字属性和汉字编码；词汇处理；汉字输入；汉字字形存储；汉字 I/O 设备；中西文兼容信息处理系统的实现；民族文字信息处理。

本书重点在阐述汉字信息处理的基本理论、技术和方法，并结合当前国内外发展动态，介绍了大、中、小和微型计算机系统及网络系统的汉字化技术和方法，力图反映当前发展技术和水平。

本书适合作为高等院校计算机专业本科教材，增加或删去部分内容可分别作为研究生或大专教材，也可作为相近专业的教学参考书；对从事计算机和汉字信息处理研究、开发和应用的工程技术人员以及汉字系统的使用人员也有参考价值。

汉字信息处理

彭寿全　黄可　编著

*

电子科技大学出版社出版
（中国成都建设北路二段四号）　邮编 610054
西南冶金地质印刷厂印刷
四川省新华书店经销

*

开本 787×1092　1/16　印张 23.375　字数 568 千字
版次 1994 年 8 月第一版　印次 1994 年 8 月第一次印刷
印数 1—6000 册
ISBN 7-81016-920-3/TP・75
定价：15.40 元

前　言

近二十余年来，经过计算机科学和语言学等方面专家的艰苦努力和密切配合，汉字信息处理取得了丰硕的成果，不仅在我国计算机及其应用方面产生了显著的社会和经济效益，而且自身也迅速发展成为一门新的综合性学科，同时也是我国信息处理科学的基础和先导。因此，汉字信息处理的理论和技术已成为我国攻读计算机和计算机信息处理的高等院校学生，以及从事计算机和汉字信息处理研究、开发和应用的工程技术人员应该具备的知识。目前，在国内外不少高等院校中已为研究生或本、专科生开设了“汉字信息处理”课程。本书是在从事汉字信息处理的教学和科研基础上，参阅了大量文献资料特别是近几年的新成果编写而成的。其重点在阐述汉字信息处理的基本理论、基本技术和基本方法，并结合当前国内外的发展动态介绍大、中、小和微型计算机系统，以及网络系统的汉字化方法，力图反映当前的发展技术和水平。

本书编写过程中注意从建立汉字信息处理系统的角度出发来选择教材内容；从汉字输入、存储、加工和输出这一信息流来组织教材；从建立中西文兼容系统来讨论汉字化技术和方法；遵循我国的国家标准，尽力使用推荐的名词术语；既重基础又重实现技术和方法，力图反映新技术。

全书分为四部分。第一部分即第一章，是本书的概述篇，介绍了汉字信息处理及系统的基本概念、特征、内容和发展现状，可作为学习本书的入门；第一章为不需要深入了解汉字信息处理的读者建立一个完整的系统印象。第二部分是基础篇，由第二～第六章构成，讲述了建立汉字信息处理系统所需的基本理论和基本部件，诸如汉字属性、汉字编码、词汇及词汇处理、汉字输入方法、汉字 I/O 设备、汉字字形存储等，其中包括汉字内码推荐方案、ISO 10646 等方面的新内容。第二部分将为建立汉字信息处理系统打下坚实的基础，此外，第四章还介绍了若干汉字编码输入方法，为使用者提供了有益的参考。第三部分即第七章，是本书的系统篇，介绍了中西文兼容信息处理系统的实现技术。其主要内容有系统配置、汉字化技术和方法；系统的分级；大、中、小和微型计算机系统及网络系统的汉字化技术和实例。第四部分即第八章，是民族语言文字信息处理篇，介绍了我国民族语言文字处理的特点和发展现状，以及几种民族语言文字信息处理的基本知识、技术和方法。

书中的第一、二、三、五、七章和第四章的第一～第五节由彭寿全编写；第六、第八章和第四章的第六、第七节由黄可编写；全书由彭寿全统编。

本书编写过程中，张轴材、李硕、董汉忠、陈代于、杨国纬等专家教

授为本书提供了丰富的技术资料，提出了宝贵的建议；万国根、袁文君、吴晓蓉、张东航、白虹、邓晓帆等同志为本书提供了宝贵的研究成果和有关资料；王远秀、彭立微、彭宇星为书稿抄写、校对、绘图做了大量工作，在此一并致以真诚的谢意。

在本书编写出版过程中，受到电子科技大学出版社、教材科和计算机系的大力支持，编写过程中还参考了不少专家的论文和专著，在此深表感谢。

本书可作为计算机专业本科教材，也兼顾研究生和大专生使用，增加部分内容可作为研究生教材，删去部分章节可供大专生使用。此外，本书也可作为其他相近专业的教学参考书，对从事计算机和汉字信息处理研究、开发和应用的工程技术人员，以及汉字系统的使用人员也有参考价值。

由于编者水平有限，加上汉字信息处理技术发展很快，书中的不妥之处在所难免，敬请读者批评指正。

编　者

1993 年 12 月

目　录

第一章 绪 论

1.1 汉字与汉字信息处理概念

1.1.1 信息与信息处理

信息是自然存在和人类活动所产生的各种状态和消息的总称，它涉及多种范畴。例如，一切自然存在及其现象所包含的各种信息；人类社会活动，如政治、经济、军事、文化、商业等活动所产生的各种信息；科学技术和生产活动，如揭示自然和物质结构的奥秘，从事地质研究、探矿、采矿、生态研究……产生的各种信息。故信息涉及到人们生存的地球及宇宙环境和从事科研、生产、生活等活动的一切方面，它既表示自然存在又是维护人类生产活动、经济活动和社会活动的第三种资源，成为人类社会存在和发展的三大要素（信息、物质和能源），对人类活动有重要意义。

信息的表示形式是多种多样的，它可以有数据、文字、声音、图形和图像等多种形式，这称为信息的多元化表示，这些形式都可以作为信息的载体。人类对信息的处理主要表现在信息传输、存储和加工上。在科技不发达的时代，信息的作用及其利用价值被限制在较低的程度上。如信息传输问题，在电信技术发明前，人们只能用人工通信或其他的简单方式来传递信息；而现在有有线电和无线电通信、光纤通信和卫星通信等先进的信息传递手段，使信息传递的速度大大提高。在信息的存储与加工方面，以前人们用纸来记录信息，靠人或简单的设备对信息进行加工，这很有局限性。现在，由于微电子技术和计算机技术的飞跃发展，人们广泛使用计算机系统来存储信息，并可对不同的信息按不同规律和意义用计算机对它进行加工来达到特定的目的。因此，对信息的加工处理已离不开计算机技术。可以说，信息处理这一术语已和计算机技术紧密地联系在一起。计算机所处理的信息（包括数据、文字、声音、图形和图像等）是以量化的形式出现，即用二进制的 0 和 1 的各种组合来代表信息。因此人们又常常把用于信息处理的电子计算机称为信息处理机。

本书所指的信息主要是文字信息。文字（script）是人类记录和传送语言的书写符号系统，故文字信息是文字和代表这种文字的语音所包含的信息，对它们的处理就称为文字信息处理。目前，计算机文字信息处理技术已应用于各个领域。例如：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，企业管理和办公自动化，文字处理，文字翻译，医疗诊断，公用事业咨询服务，数据通信，工业自动控制，数据库系统和计算机辅助设计等等。文字信息处理技术已逐步渗透到人类思维、生产和生活等活动的一切方面，它同国防、科研、生产、社会活动、家庭生活等的联系十分密切。毫无疑问，这项技术的发展和应用是人类进入信息化社会、国家走向现代化的一个重要标志。

世界上，各国使用的文字大多数是不同的。我国是个多民族国家，存在多种文字。中文应该是包括汉字、蒙古文、藏文、朝鲜文、彝文等多种文字的集合。要在我国推广电子

计算机应用必须使各民族的文字都能在电子计算机上实现输入、加工和输出。这些民族的文字中，只有汉字是一种象形文字，其他的都是拼音文字范畴。拼音文字的字符最多不超过数十种。而现代，我国使用的汉字仅常用部分就有六七千字，而且字形复杂，这就使计算机实现汉字信息处理存在较多的困难。如果解决了汉字信息处理的技术问题，那么对于我国其他民族的文字信息处理也比较容易解决。本书讲述“汉字信息处理”，它包括对汉语书面形式和语音形式两种信息的处理。此外，对少数民族文字信息处理也作一些介绍。

1.1.2 汉字的基本特征

汉语（Chinese）是汉族的语言，属汉藏语系。汉语是中国境内主要的通用语言，也是国际通用语言之一，故国际上把中文（Chinese）特指汉语。汉字（Chinese character，Hanzi）是记录汉语的书写符号系统。汉字也被其他一些国家或民族用作书写符号。通常把这一书写符号系统中可以独立使用的书写符号称为汉字字符。如果把一个简体字（如专）和它对应的繁体字（專）看成是两个不同的汉字，那么一个汉字字符就是某个汉字的代表，故汉字是汉字字符的集合。而在日常生活中，又常把汉字作为汉字字符的简称。本书一般把“汉字”作为集合看待，考虑到人们的习惯用法，有时“汉字”也作为单个汉字字符看待。

汉字最基本的特征是每个汉字字符有其字形、字音和字义三方面的信息，形、音、义三者是构成汉字的三要素（见图 1-1-1）。所以，汉字是把形、音、义融为一体构成的一种优美文字。汉字的字形是汉字特有的形体，汉字的字音是该汉字的语音的记录，汉字的字义是该汉字所表达的语义。正因为汉字有形、音、义结合的特点，在研究汉字信息处理时要特别注意汉字的这一基本特征。

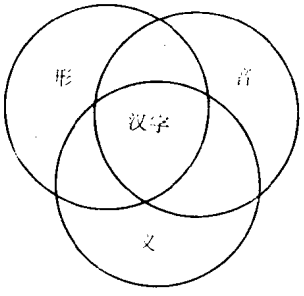
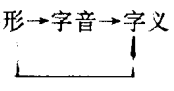


图 1-1-1 汉字的三要素

每个汉字有其特定的形体，它由点、横、竖、撇、弯和拐等简单的笔画构成，其形体酷似方块，所以汉字是方块字。汉字字符的形体结构复杂，其字形可分成多个层次，层次越高表示一个汉字所用的符号就越少。通常，把汉字分为单字、字根和笔画三个层次，当用字根组字时，“江”可以用“氵”和“工”两个字根拼成；而用笔画组字时，则要用六画才能构成。所以，人们又把汉字归为拼形文字。这种拼形文字是以平面结构方式非线性地构成汉字字符，可以认为汉字是二维空间的图形。与此相反，拼音文字则是按照发音的先后顺序用少量的字母（如英文为 26 个）从左到右排成一维的序来组成“字”（即单词）。因此，拼音文字是用字形记录字音，用字音表达字义。汉字集合中拥有不同形体的汉字字符数达六万多个，这些汉字字符除可以用形表音，以音达义外，还可以用字形直接表达字义（例如“钟”字，其偏旁“钅”指示出“钟”字是表示一种金属物），这是与拼音文字不同的一个突出的特征，可以图示如下：



汉字中绝大多数是形声字，字形中的形旁（或称义符）用来表意，故汉字又属以形表意的表意文字。汉字字形虽然复杂，但在以字造词和以字词造句时，汉字本身没有字形变化，其

字与字、词与词之间没有空格，这与英语的书写格式也是截然不同的。汉语句子书写上的这一特点为汉语的自然语言理解带来了困难。

每个汉字有一定的字音。形声字的声旁（如“钟”字的声旁“中”）指示出该字的大致读音，以达到以字形记录字音的目的。汉字的读音绝大多数是单音节的，而英文单词绝大多数是多音节。在现代汉语中，有417个基本音节，带上阳平、阴平、上声、去声和轻声共有1330个左右的音节。要用这些音节来标注六万多个汉字的字音必然出现大量的同音字。在汉字集合中有10%左右的汉字是多音字，即同一字形的一个汉字有不同的读音。汉字属表意文字，其“以形达意”比“以音表意”的成分大得多，这就造成不同地区的人们往往按自己的方言来读音，因此，汉字的方言多。

每个汉字有一定的字义。一个汉字一般有2～5个意义，多的达6～9个；有少数汉字，一字多达十几种意义。每个汉字所表达的意义是与其字形和字音结合在一起的。同一个字形，由于读音不同意思也就不同，属“以音表意”。例如“行”字，读“xíng”，表示行走的意思；读“hāng”，表示行业或行列等意思。同一读音的字，由于字形不同意义也就不相同，这是“以形达意”。例如“坚”和“艰”都读“jiān”，可是“坚”表示结实、硬的意思，而“艰”字却表示困难的意思。

1.1.3 汉字信息处理

1. 基本概念

汉字信息处理(Chinese character information processing)是指用计算机对汉字表示的信息进行操作和加工，如汉字输入、输出、识别等。汉字字符处理（Chinese character processing)是指用计算机对汉字字符进行操作和加工，例如对汉字字符的编码、输入、显示、打印、存储、传输、识别等等。因此，在汉字信息处理中，汉字字符处理是基础，它是前者最基本的处理层次。在此基础上进行的面向汉字信息的处理更为重要，因为它已不仅是单个汉字或词语，而是用汉字及其集合来表征自然存在和社会活动中产生的信息，并利用计算机对这些信息进行分类、编目、存储、检索、分析、综合决策、打印输出、辅助设计以及提供咨询服务等等。当前，我国的计算机广泛用于事务处理、管理信息系统、办公自动化、情报检索、轻印刷、激光照排、科学研究、工程设计和国防建设等各个领域。汉字信息处理已成为这一系列应用的基础，不实现汉字信息的计算机处理，我国的计算机应用就难于推广。现代社会中的一切领域若不能及时掌握信息，就无法有效地进行工作。目前不少发达国家已进入信息化社会，社会现代化是以社会的信息化为标志，没有信息化，就没有现代化。因此，汉字信息处理的研究和开发对我国尽快实现社会信息化、现代化有极其重要的意义。

2. 汉字信息处理研究的主要内容

汉字信息处理是一门跨学科、多学科交叉的边缘学科，它涉及语言学（包括词汇学、词义学、语用学、音韵学、识别与理解、方言学、风格学、翻译等)、心理学、声学、计算机科学、数学、电子学、逻辑学和人工智能等学科。它是一个多学科密切结合的综合性学科，而多学科交叉研究又产生了计算机语言学和语言工程学等新的边缘学科。汉字信息处理又是一门高技术学科，它建立在计算机、大规模集成器件和设备的最新技术上。汉字是中华民族的通用语言文字，也是联合国正式使用的语言之一，它受到国内外的极大关注。因此，

从多方面看，对汉字信息处理的研究，其广泛性大大超过了其他学科。

汉字信息处理是在多层次上开展研究的，其基础研究、高层次研究、应用研究和产品开发彼此密切配合，相互促进，使汉字信息处理技术的水平不断提高。通常可以把汉字信息处理所涉及的范围和内容概括为如下几个方面：

(1) 基础理论工作

·汉字属性研究（包括汉字字量、字形分解、汉字字体、使用频度、汉字字音、汉字索引、排序等）；

·汉字信息处理的国家标准；

·词和词组研究（包括词的种类、意义、数量、排序和使用频度等）；

·汉字信息输入的理论及评测标准；

·汉字字形识别和汉字语音识别研究；

·汉语的自然语言处理（包括用计算机进行的汉语语法分析和语义分析、汉语人机接口、自动作文摘、文稿自动编辑、机器翻译等等）；

·计算语言学和语言工程学研究。

(2) 汉字信息处理的基本环境及其实现

·中西文兼容的实现技术；

·汉字信息的输入输出技术和设备；

·汉字信息的存储与汉字库；

·汉字信息的机内处理；

·中西文兼容的系统软件；

·建立汉字信息处理系统；

·建立汉字信息处理开发平台；

·汉字本地通信和远程通信技术等等。

(3) 汉字信息处理的应用开发

·各类电子印刷排版系统的研制与开发；

·情报检索和档案管理；

·CMIS 系统的建立与开发；

·办公自动化系统；

·专家系统；

·翻译系统；

·模式识别技术（如汉字语音识别系统和字形识别系统）；

·汉字信息通信系统；

·其他应用项目（如订票系统、公用咨询服务系统、电话查号系统、广告宣传系统等）。

3. 汉字信息处理过程

汉字信息处理的流程可以概括成图 1-1-2 所示的三大部分。

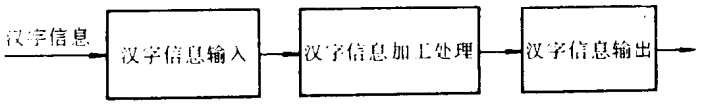


图 1-1-2 汉字信息处理流程

(1) 汉字信息输入

通常借助于键盘、模式识别设备、汉字终端、光笔等输入设备把用汉字字形或语音所代表的汉字信息变换成计算机内所使用的代码(称为内部码或内码),然后送入计算机进行加工处理。

(2) 汉字信息加工处理

对汉字信息加工是借助计算机执行相关的程序来完成的。汉字信息处理包括多种不同的处理要求,一是对文稿中单个汉字的增、删、改操作;二是对若干字或整段的增、删、改操作;三是对字符串的处理,如分类、合并、比较、排序、检索以及对齐等。上述三点,实质上是对汉字字符的加工处理,是基础性工作。对更高层次上的汉字信息加工处理,包括文稿中句子的语法和语义分析,句子和全文的理解,自动作文摘,机器翻译等等。

(3) 汉字信息的输出

汉字信息的输出形式主要有字形和字音两大类。在独立型汉字信息处理系统中,汉字信息输出设备把计算机处理后的结果信息以汉字字形形式或字音形式输出。当汉字信息处理系统要进入公用网交换汉字信息时,该汉字系统输出的汉字信息一般用某种汉字代码(如汉字交换码)来表示。

在汉字信息处理的全过程中,使用了多种表征汉字字符的汉字代码。这些不同类型的汉字代码采用了不同的编码规则。例如当用标准键盘输入汉字字符时,一般采用编码输入。不同的编码输入方案其编码规则和使用的码元数均不相同,故同一个汉字,当编码输入方案不同时,用来表征该汉字字符的西文字符及其数量均不一定相同。不管编码输入方案如何,用来实现汉字字符输入的汉字代码均称为输入码,而在机器内部用来表示汉字字符的代码称为内部码。通常,内部码和输入码的编码规则是不相同的。

1.2 汉字信息处理系统

1.2.1 汉字信息处理系统的特点和中西文兼容

当汉字字符和西文字符被数字化后,从信息处理角度来看,汉字与西文(例如英文)没有什么本质区别。原则上,现有的西文信息处理系统都可以用来处理中文(或汉字)信息。但实际上,汉字有字数多、字形复杂、笔画严格等特点,因此给汉字信息带来了固有的特殊性。其最重要的表现是一个汉字字符至少需要用两个字节来表示,而一个西文字符只用一个字节。那么,怎样建立汉字信息处理系统呢?其办法有两类:一类是独立设计一个专门处理汉字信息的计算机系统,从系统设计开始就考虑系统是处理两字节的汉字字符,故

硬、软件与现在的计算机系统都有极大的差异；另一类是在现有西文计算机系统的基础上，考虑汉字字符为双字节特点，对系统硬、软件作适当改造，做到汉字、西文共享国际信息处理的成果。然而，由于历史原因，汉字信息处理系统不宜单独地自成系统，而必须在国际通用的西文字符系统的基础上开发。这是由于不论是系统硬设备和软件，通用的西文计算机系统已有了相当基础。若撇开它独立地开发汉字计算机系统。在技术上并非不能实现。但是，这样做，工作的起点太低，大量已成熟的国际上通用的各种软件资源就不能加以利用，限制了系统功能的发展，而且也不利于和国际上的标准技术相兼容。因此，开发汉字信息处理技术，必须走和国际上通用技术相兼容的道路。同时，这样做也可以站在较高的起点上开发汉字信息处理系统，收到事半功倍的效果。这项技术可称为中西文兼容技术，它的出发点是完全保留并利用原西文计算机系统的一切硬、软件功能。在此基础上，再增加汉字信息处理功能，把汉字信息和西文、数字信息的处理功能兼容于同一系统中，并且不损失原西文系统的功能，使系统能方便地处理汉字、西文混合的信息流。西文泛指西方国家的文字，本书特指英文。

在西文计算机系统的基础上扩充汉字信息处理功能，必须解决汉字信息的输入和输出问题（包括技术和设备）、汉字字形数据的存储问题、汉字字符的数据格式和机内表示问题、中西文软件兼容问题。这些问题受到一定的约束。第一，汉字字符的数据格式必须能为西文计算机系统所接受。这一点比较容易解决，GB2312-80《信息交换用汉字编码字符集　基本集》中规定的汉字交换码是标准字符代码（即 ASCII 或 GB1988 代码）扩充而来。即用两个 ASCII 图形字符代码交叉组合成汉字交换码，允许 94×94＝8836 个，但在 GB2312-80 中，实际只规定了 6763 个汉字字符。这些汉字字符都是七位二进制信息表示的代码，其区别是英文字符用单字节表示，而汉字字符用双字节表示。数据格式相同，故英文计算机系统可以接受汉字字符。第二，必须使系统能准确地区分英文字符（包括数字和符号）代码和汉字字符代码。因为英文字符和汉字字符都是用 ASCII 图形字符（七位二进制信息）表示，若不加其他标识信息，则无法对两者加以区分。因此，汉字信息进入系统后，应对汉字代码添加相应的标识信息。最简单的办法是利用汉字的国家标准交换码的第 8 位来区分两类代码，即每个汉字字符的两个字节中，将其最高位（第 8 位）置 1，作为汉字代码的标识。若系统判断一个字节的最高位为 0，则认为是英文字母或数字代码；否则，认为是汉字字符代码。这种方法实际上是把汉字字符代码纳入字符代码的数据类型，可以看成是字符代码的扩充。此方法简单，加工处理方便，但第 8 位不能用作代码信息位，并有若干局限性。例如，在一些中、大型计算机系统以及网络通信环境中，要利用字节的最高位作为奇偶校验位时，就不能简单地用这种方法来表示机器内部使用的汉字代码。另一类方法是在一串汉字代码的前和后分别加上起始符号和结束符号，这种符号可以利用某个不常用的字符代码或功能码的组合来代表。这类方法的优点是第 8 位能用作代码信息位，使表示的汉字数量增多。但标识符在系统中可能出现二义性，且多占用存储单元，加工处理也较麻烦。上述两类表示汉字内部码的方法各有其优缺点。目前，在主机或终端设备上通常采用第一类标识方式来标识汉字字符代码（内部码）；在联机或网络通信时，再将第一类方式所标识的汉字字符转换成为第二类标识方式表示的汉字字符。

概括起来，汉字信息处理系统的特点和要求主要有以下几个方面：

（1）具备英文计算机系统原有的全部功能；

(2) 该计算机系统能输入和输出汉字信息；

(3) 要解决信息量很大的汉字字形在系统内的存储；

(4) 在系统技术上，要解决汉字和英文信息的兼容问题，要求系统能处理汉字和英文混合的信息流；

(5) 汉字信息处理系统的技术必须走和国际标准相兼容的道路，以便汉字信息处理能共享原英文系统的硬、软件资源。

1.2.2 汉字信息处理系统的组成框图

要在我国推广计算机应用和进行汉字信息处理的工作，就必须建立各类汉字信息处理系统，而通用型汉字信息处理系统（general-purpose Chinese character information system）就是用得最多最广的一类。这类系统是指适用于各种数据处理和汉字信息处理的计算机系统。其特点是通用性强，汉字输入输出手段多，操作方便。本书主要介绍如何在英文计算机系统的基础上实现通用型汉字信息处理系统。这既涉及到汉字输入、输出和存储等设备及其相关技术，又涉及如何实现汉字和西文兼容的各类软件。

汉字信息处理系统是以计算机为中心，再配上汉字输入设备、汉字字模库和汉字输出设备等硬件，以构成处理汉字信息的硬件系统。同时，还必须配置汉字信息处理软件，才可能构成一个基本的汉字信息处理系统，其组成框图示于图 1-2-1。从图可见，汉字信息（如汉字输入码、字形、字音或交换码）通过相应的输入设备（如键盘、汉字识别设备、语音识别设备或通信接口）变换成汉字内部码送入计算机，经计算机加工处理后送输出设备进行输出处理，最后由汉字显示器、汉字打印机、语音合成设备或通信接口输出汉字信息。当前，通用型汉字信息处理系统主要还是用键盘实现汉字信息输入，用汉字显示器和汉字打印机完成汉字信息输出。倘若系统需要与其他汉字系统通信时，则借助通信接口来连接。

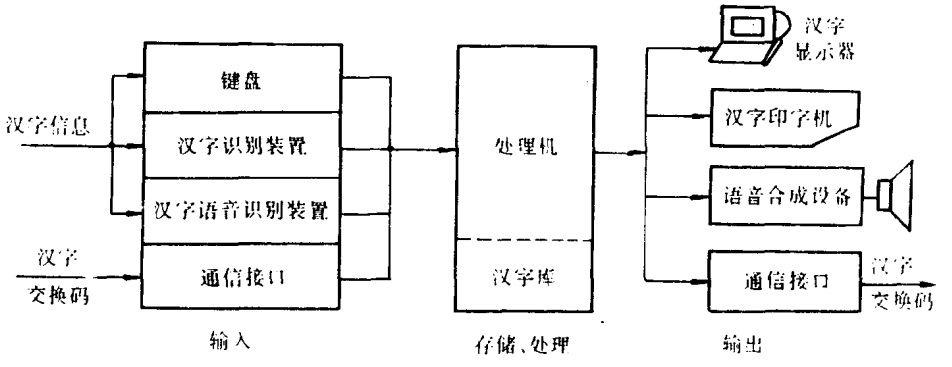


图 1-2-1 汉字信息处理系统组成框图

1.2.3 汉字输入

汉字输入（Chinese character input）是指利用汉字的形、音或相关信息通过各种方式把汉字输入到计算机中去的过程。由于汉字数量多，字形复杂，形体相近字多，同音字多，使得汉字输入技术成为汉字信息处理的一个主要难题。虽然目前已在汉字编码输入方面得到了解决，并已进入实用阶段，但一个为人们共同接受的实用的汉字输入方法还在研究和开