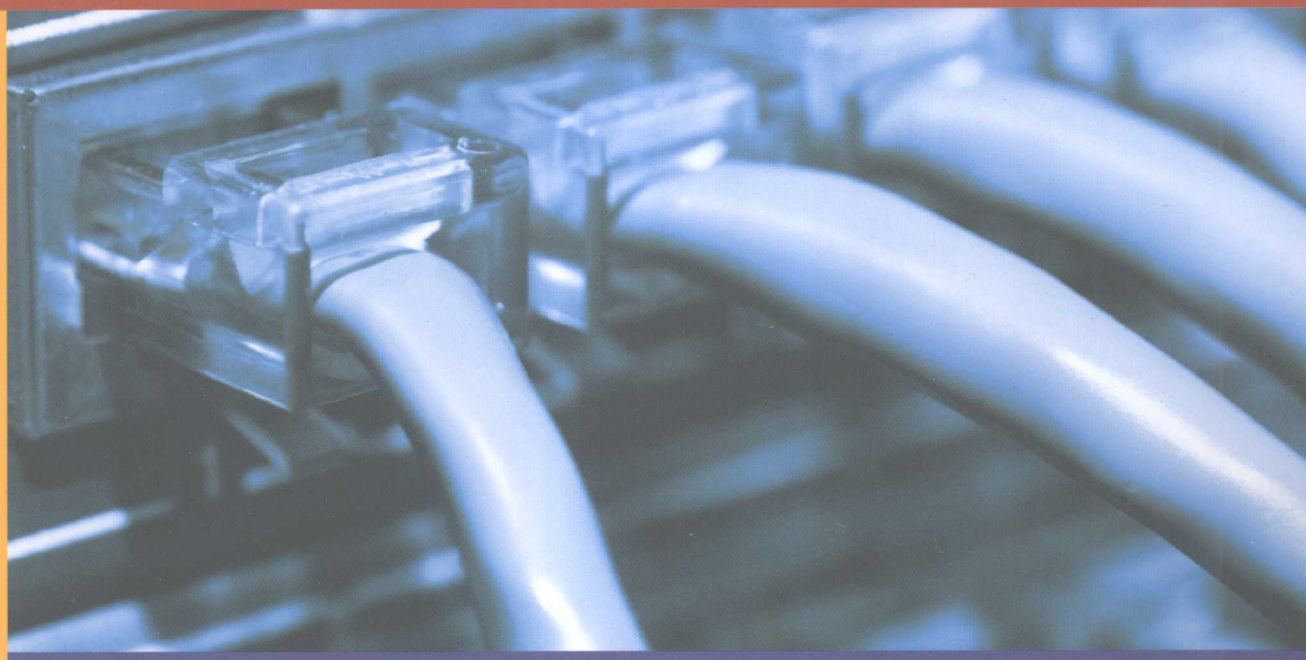
普通高等教育“十一五”国家级规划教材

高等院校信息与通信工程系列教材

信息论基础



曹雪虹 编著

清华大学出版社

普通高等教育“十一五”国家级规划教材

高等院校信息与通信工程系列教材

信息论基础

曹雪虹 编著

清华大学出版社
北京

内 容 简 介

本书重点介绍由香农理论发展而来的信息论的基本理论以及编码理论。全书共分 8 章，在介绍有关信息度量的基础上，重点讨论了信源熵、信道容量和率失真函数，以及无失真信源编码、限失真信源编码、信道编码和密码学中的信息理论，并简单介绍了网络信息论。

本书注重概念，采用通俗的文字，联系目前实际通信系统，用较多的例题和图示阐述了基本概念、基本理论及实现原理，尽量减少繁杂的公式定理证明。在各章的最后还附有小结和大量习题，书后附有习题答案，便于读者学习，加深对概念的理解。

本书可作为理工科高等院校电子、信息、通信工程及相关专业的本科学生的教材，亦可供信息、通信、电子工程等有关专业的科技人员作为参考书。

图书在版编目(CIP)数据

信息论基础/曹雪虹编著. —北京：清华大学出版社，2009.6
(高等院校信息与通信工程系列教材)
ISBN 978-7-302-19745-4

Ⅰ. 信…　Ⅱ. 曹…　Ⅲ. 信息论—高等学校—教材　Ⅳ. G202

中国版本图书馆 CIP 数据核字(2009)第 039083 号

责任编辑：陈国新
责任校对：梁　毅
责任印制：杨　艳
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　　总　　机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185×260　**印　张**：11.75　**字　数**：289 千字
版　　次：2009 年 6 月第 1 版　　**印　　次**：2009 年 6 月第1 次印刷
印　　数：1～3000
定　　价：23.00 元

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题，请与清华大学出版社出版部联系调换。联系电话：(010)62770177 转 3103　　产品编号：023502-01

出版说明

信息与通信工程学是信息科学与技术的重要组成部分。改革开放以来，我国在发展通信系统与信息系统方面取得了长足的进步，形成了巨大的产业与市场，如我国的电话网络规模已占世界首位，同时该领域的一些分支学科出现了为国际认可的技术创新，得到了迅猛的发展。为满足国家对高层次人才的迫切需求，当前国内大量高等学校设有信息与通信工程学科的院系或专业，培养大量的本科生与研究生。为适应学科知识不断更新的发展态势，他们迫切需要内容新颖又符合教改要求的教材和教学参考书。此外，大量的科研人员与工程技术人员也迫切需要学习、了解、掌握信息与通信工程学科领域的基础理论与较为系统的前沿专业知识。为了满足这些读者对高质量图书的渴求，清华大学出版社组织国内信息与通信工程国家级重点学科的教学与科研骨干以及本领域的一些知名学者、学术带头人编写了这套高等院校信息与通信工程系列教材。

该套教材以本科电子信息工程、通信工程专业的专业必修课程教材为主，同时包含一些反映学科发展前沿的本科选修课程教材和研究生教学用书。为了保证教材的出版质量，清华大学出版社不仅约请国内一流专家参与了丛书的选题规划，而且每本书在出版前都组织全国重点高校的骨干教师对作者的编写大纲和书稿进行了认真审核。

祝愿《高等院校信息与通信工程系列教材》为我国培养与造就信息与通信工程领域的高素质科技人才，推动信息科学的发展与进步做出贡献。

北京邮电大学
陈俊亮

前　言

当前信息产业发展得很快，需要大量信息、通信、电子工程类专业的人才，而信息理论是这些专业的基础，必须掌握，它可以指导理论研究和工程应用。

由于信息论介绍的是信息论基础和编码理论，内容本身理论性很强，现有的一些教材除了介绍理论和公式外，都用了大量篇幅来证明这些理论和公式，这些用做研究生教材是比较适合的。而作为电子、信息、通信工程的本科生及相关专业的工程技术人员，由于他们理论基础的不足以及实际应用的需要，不可能花很多精力去研读那些在他们看来是非常难懂而且枯燥乏味的证明，迫切需要一本介绍有关信息理论的基本知识，且与实际应用紧密联系的书籍，本书就是出于这样的目的而编写的。

本书共分 8 章，第 1 章是绪论。第 2 章介绍信息论的一些基本概念，包括自信息量、条件自信息量、互信息量、条件互信息量、平均互信息量、相对熵、单符号熵、随机序列的熵、熵的性质以及连续信源熵、最大熵定理等，对信源的信息给出定量描述，并解释了冗余度的由来及作用。这一章是后续章节的基础。

第 3 章介绍信道的分类及其表示参数，讨论各种信道能够达到的最大传输速率，即信道的容量及其计算方法，并讨论了信源与信道的匹配问题。

第 4 章介绍失真函数和信息率失真函数的定义及性质，给出了在一定失真限度内信源必须输出的最小传输速率。

第 5 章介绍信源编码，首先给出了无失真信源编码定理和限失真信源编码定理，其中无失真信源编码定理包括定长编码定理和变长编码定理，并详细阐述了最佳无失真编码中的香农码、费诺码和哈夫曼码的编码方法及其性能比较。

第 6 章介绍信道编码，分析错误概率与译码规则和编码方法的关系，介绍了两种典型的译码规则，从平均差错率导出了信道编码定理，并简单介绍了纠错编码的基本知识，以及联合信源信道编码定理。

第 7 章介绍密码体制的基础知识及其与熵的关系，讨论系统保密的理论要求和实际要求。

第 8 章介绍网络信息论的基本思想和部分结论，首先讨论网络信道的几种类型及其信道容量域，然后分析网络中相关信源的信源编码问题。

本书注重基本概念，用较通俗的文字解释其物理意义，辅以一定的例题和图示说明，不再用繁杂的公式来证明这些早已非常成熟的公理，联系当前实际通信技术来讲述，研读本书后使读者概念清晰，可有目标地将基本概念应用于实际工作中。

本书在编写的过程中，得到了张宗橙教授、徐澄圻教授的大力帮助，在此表示衷心的感谢。

限于编者的水平，书中不妥或谬误之处难免，恳请广大读者提出宝贵意见，以便进一步完善。

编　者
2008 年 12 月
caoxh@njupt.edu.cn
caoxh@njit.edu.cn

目　录

第 1 章　绪　论

科学技术的发展使人类跨入了高度发展的信息化时代。在政治、军事、经济等各个领域，信息的重要性不言而喻，有关信息理论的研究正越来越受到重视。

人们在自然和社会活动中，获取信息并对其进行传输、交换、处理、检测、识别、存储、显示等操作，研究这方面的科学就是信息科学，信息论(information theory)是信息科学的主要理论基础之一，它主要研究可能性和存在性的问题，为具体实现提供理论依据。与之对应的是信息技术(information technology)，主要研究如何实现和怎样实现的问题。

通过本章的学习，可以了解下列内容：信息论的形成和发展，信息论研究的内容及信息的基本概念。本章还结合通信系统模型介绍了模型中各部分的作用、编码的种类和研究内容。

1.1　信息论的形成和发展

信息论理论基础的建立，一般来说开始于香农(C. E. Shannon)在研究通信系统时所发表的论文。随着研究的深入与发展，信息论具有了更为宽广的内容。

信息在早些时期的定义是由奈奎斯特(Nyquist，H.)和哈特利(Hartley，L. V. R.)在20世纪20年代提出来的。1924年奈奎斯特解释了信号带宽和信息速率之间的关系；1928年哈特利最早研究了通信系统传输信息的能力，给出了信息度量方法；1936年阿姆斯特朗(Armstrong)提出了增大带宽可以使抗干扰能力加强的观点。这些工作都给香农很大的影响，他在1941～1944年对通信和密码进行了深入研究，用概率论的方法研究通信系统，揭示了通信系统传递的对象就是信息，并对信息给以科学的定量描述，提出了信息熵的概念，指出通信系统的中心问题是在噪声下如何有效而可靠地传输信息，以及实现这一目标的主要方法是编码等。这一成果于1948年以“通信的数学理论”(a mathematical theory of communication)为题公开发表。这是一篇关于现代信息论的开创性的权威论文，为信息论的创立作出了独特的贡献，香农也因此成为信息论的奠基人。

20世纪50年代信息论在学术界引起了巨大的反响。1951年美国IRE成立了信息论组，并于1955年正式出版了信息论汇刊。60年代信道编码技术有了较大进展，使它成为了信息论的又一重要分支，它把代数方法引入到纠错码的研究，使分组码技术发展到了高峰，找到了大量可纠正多个错误的码，而且提出了可实现的译码方法。70年代卷积码和概率译码有了重大突破，提出了序列译码和Viterbi译码方法，并被美国卫星通信系统采用，使香农理论成为真正具有实用意义的科学理论。1982年Ungerboeck. G提出了将信道编码和调制结合在一起的网格编码调制，无须增大带宽和功率，以设备的复杂度增加换取编码增益，这种方法受到了广泛关注，在目前的通信系统中占据统治地位。

信源编码的研究落后于信道编码。香农在1948年的论文中提出了无失真信源编码定理，也给出了简单的编码方法——香农码。1952年费诺(Fano)和哈夫曼(Huffman)分别提出了各自的编码方法，并证明其方法都是最佳码。1959年香农的文章“保真度准则下的离散信源编码定理”(coding theorems for a discrete source with a fidelity criterion)系统地提出了信息率失真理论和限失真信源编码定理，它们是数据压缩的数学基础，为各种信源编码的研究奠定了基础。1971年伯格尔(T. Berger)给出了更一般性的率失真编码定理。随着传输内容和传输信道的发展，人们针对各种信源的特性，产生了大量实用高效的信源编码方法。

香农又于1961年发表了论文“双路通信信道”，开拓了网络信息论的研究，从点与点间的单用户通信推广到多用户系统的研究。1972年盖弗(Cover)发表了有关广播信道的研究，以后陆续有关于多接入信道和广播信道模型和信道容量的研究。近三十多年来，这一领域研究活跃，发表了大量的论文，使多用户信息论的理论日趋完整。

香农在1949年发表了论文“保密通信的信息理论”，首先用信息论的观点对信息保密问题作了全面的论述。但由于通信保密研究当时主要用于政府和军方，成果很少对外公布，因此公开发表的论文也很少。直到1976年迪弗和海尔曼发表了论文“密码学的新方向”，提出了公钥密码体制之后，保密通信问题才得到公开、广泛的研究。尤其是现在，信息安全已成为一个关系到信息产业发展的重大问题。因此，密码学以及信息安全已经成为各国科学家研究的重点和热点。

可见，信息论主要研究的是通信的一般理论，在信息可以度量的基础上，研究有效地和可靠地传递信息的科学，它涉及信息度量、信息特性、信息传输速率、信道容量、干扰对信息传输的影响等方面的知识。

1.2 信息理论研究的内容

信息理论是信息科学的基础，强调用数学语言来描述信息科学中的共性问题及解决方案。目前，这些共性问题分别集中在狭义信息论、一般信息论和广义信息论中。

狭义信息论主要总结了香农的研究成果，因此又称为香农信息论。它在信息可以度量的基础上，研究如何有效、可靠地传递信息。有效、可靠地传递信息必然贯穿于通信系统从信源到信宿的各个部分，狭义信息论研究的是收、发端联合优化的问题，而重点在各种编码。它是通信中客观存在的问题的理论提升。

一般信息论研究从广义的通信引出的基础理论问题，它除了香农信息论外，还包括其他人的研究成果，其中最主要的是维纳(Wiener)的微弱信号检测理论。微弱信号检测又称为最佳接收，是为了确保信息传输的可靠性，研究如何从噪声和干扰中接收信道传输的信号的理论。它主要研究两个方面的问题：从噪声中判决有用信号是否出现和从噪声中测量有用信号的参数。该理论应用近代数理统计的方法来研究最佳接收的问题，系统和定量地综合出存在噪声和干扰时的最佳接收机结构，并推导出这种系统的极限性能。除此之外，一般信息论的研究还包括噪声理论、信号滤波与预测、统计检测与估计理论、调制理论、信号处理与信号设计理论等。可见它总结了香农和维纳以及其他学者的研究成果，

是广义通信中客观存在的问题的理论提升。

无论是狭义信息论还是一般信息论，讨论的都是客观问题。然而，当讨论信息的作用、价值等问题时，必然涉及主观因素。广义信息论研究包括所有与信息有关的领域，如心理学、遗传学、神经生理学、语言学、社会学等。因此，有人对信息论的研究内容进行了重新界定，提出从应用性、实效性、意义性或者从语法、语义、语用方面来研究信息，分别与事件出现的概率、含义及作用有关，其中意义性、语义、语用主要研究信息的意义和对信息的理解，即信息所涉及的主观因素。

广义信息论从人们对信息特征的理解出发，从客观和主观两个方面全面地研究信息的度量、获取、传输、存储、加工处理、利用以及功用等，理论上可以说是最全面的信息理论，但由于主观因素过于复杂，很多问题本身及其解释尚无定论，或者受到人类知识水平的限制，目前还得不到合理的解释，因此广义信息论还处于正在发展的阶段。

信息在传输、存储和处理的过程中，不可避免地要受到噪声或其他无用信号的干扰，信息理论就是为能可靠地、有效地从数据中提取信息，提供必要的根据和方法。这就必须研究噪声和干扰的性质以及它们与信息本质上的差别，噪声与干扰往往具有按某种统计规律的随机特性，信息则具有一定的概率特性，如度量信息量的熵值就是具有概率性质的。因此，信息论、概率论、随机过程和数理统计学，都是信息论应用的基础和工具。

本书讲述的信息理论的基本内容是与通信科学密切相关的狭义信息论，涉及信息理论中很多基本问题，例如：

① 什么是信息？如何度量信息？

② 在信息传输中，基本的极限条件是什么？

③ 对于信息的压缩和恢复的极限条件是什么？

④ 从环境中抽取信息的极限条件是什么？

⑤ 设计什么样的设备才能达到这些极限？

⑥ 实际上接近极限的设备是否存在？

信息论主要应用在通信领域，在含噪信道中传输信息的最优方法到今天还不十分清楚，特别是当数据的信息量大过信道容量的情况更是毫无所知，这是经常遇到的情况。因为从信源提取的信息常常是连续的，即信号的信息含量为无限大。在一般信道中传输这样的信号是不可能不产生误差的。引入信道容量和信息量的概念以后，这类问题就可以得到满意的解释，并可给出一个通信系统的最佳效果。因而就为设计这样的系统提供了理论依据。

在通信理论中经常会遇到信息、消息和信号这三个既有联系又有区别的名词，下面将它们的定义比较一下。

信息：信息是指各个事物运动的状态及状态变化的方式。人们从对周围世界的观察得到的数据中获得信息。信息是抽象的意识或知识，它是看不见、摸不到的。例如，人脑的思维活动产生的一种想法，当它仍存储在脑子中的时候它就是一种信息。

消息：消息是指包含有信息的语言、文字和图像等，例如，我们每天从广播节目、报纸和电视节目中获得各种新闻及其他消息。在通信中，消息是指担负着传送信息任务的单个符号或符号序列。这些符号包括字母、文字、数字和语言等。单个符号消息的情况，例

如用 x_1 表示晴天，x_2 表示阴天，x_3 表示雨天；符号序列消息的情况，例如“今天是晴天”，这一消息由 5 个汉字构成。可见消息是具体的，它载荷信息，但它不是物理性的。

信号：信号是消息的物理体现，为了在信道上传输消息，就必须把消息加载（调制）到具有某种物理特征的信号上去。信号是信息的载荷子或载体，是物理性的，如电信号、光信号等。

在通信系统中传送的本质内容是信息，发送端需将信息表示成具体的消息，再将消息加载至信号上，才能在实际的通信系统中传输。信号在接收端（信息论里称为信宿）经过处理变成文字、语音或图像等形式的消息，人们再从中得到有用的信息。在接收端将含有噪声的信号经过各种处理和变换，从而取得有用信息的过程就是信息提取，提取有用信息的过程或方法主要有检测和估计两类。载有信息的可观测、可传输、可存储及可处理的信号，均称为数据。

信息的基本概念在于它的不确定性，任何已确定的事物都不含有信息。其特征有：

① 接收端在接收到信息之前，对它的内容是不知道的，所以信息是新知识、新内容；

② 信息是能使认识主体对某一事物的未知性或不确定性减少的有用知识；

③ 信息可以产生，也可以消失，同时信息可以被携带、存储及处理；

④ 信息是可以度量的，信息量有多少的差别。

各类通信系统，例如电报、电话、广播、电视、雷达、遥测等传输的是各种各样的消息。消息的形式可以不同，但它们都是能被传输的，能被人们的感觉器官（如眼、耳、触觉等）所感知的，而且消息表述的是客观物质和主观思维的运动状态或存在状态。在各种通信系统中，其传输的形式是消息。但消息传输过程的一个最基本、最普通却又不十分引人注意的特点是收信者在收到消息以前是不知道消息的具体内容的。在收到消息以前，收信者无法判断发送者将会发来描述何种事物运动状态的具体消息，也更无法判断是描述这种状态还是那种状态。再者，即使收到消息，由于干扰的存在，也不能判断所得到的消息是否正确和可靠。总之，收信者存在着“不知”、“不确定”或“疑问”。通过消息的传输，收信者知道了消息的具体内容，原先的“不知”、“不确定”和“疑问”消除或部分消除了。因此，对收信者来说，消息的传输过程是一个从不知到知的过程，或是从知之甚少到知之甚多的过程，或是从不确定到部分确定或全部确定的过程。如果不具备这样一个特点，那就根本不需要通信系统了。试想，如果收信者在收到电报或电话之前就已经知道报文或电话的内容，那还需要电报、电话系统吗？

1.3 通信系统的模型

图 1-1 是目前较常用的、也是较完整的通信系统模型。下面介绍模型中各个部分的作用及需要研究的核心问题。

（1）信源

信源（source）是向通信系统提供消息 u 的人和机器。信源本身是十分复杂的，在信息论中我们仅研究信源的输出。信源输出的是以符号形式出现的具体消息，它载荷信息。信源输出的消息可以有多种形式，但可归纳成两类：离散消息，例如由字母、文字、数字等

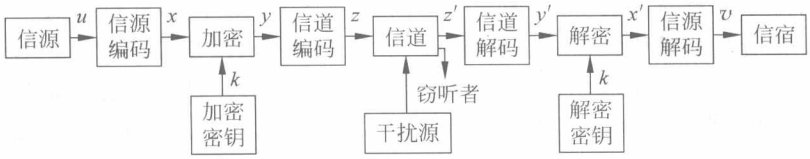


图 1-1　通信系统的物理模型

符号组成的符号序列，或者单个符号；连续消息，例如语音、图像和在时间上连续变化的电参数等。因为通信系统的接收端（信宿）在收到消息之前并不知道信源所发出消息的内容，所以一般地说信源发出的是随机性的消息。但因信源发出的消息都携带着信息，可见消息的变化是具有一定规律性的，因此严格地说信源发出消息并不是完全随机性的。信源的核心问题是它包含的信息到底有多少，怎样将信息定量地表示出来，即如何确定信息量。

（2）信宿

信宿（receiver）是消息传递的对象，即接收消息的人或机器。根据实际需要，信宿接收的消息 v 的形式可以与信源发出的消息 u 相同，也可以不相同，当两者形式不相同时，v 是 u 的一个映射。信宿需要研究的问题是能收到或提取多少信息。

（3）信道

信道（channel）是传递消息的通道，也是传送物理信号的设施。信道可以是一对导线、一条同轴电缆、传输电磁波的空间、一条光导纤维等传输信号的媒质。信道的问题主要是它能够传送多少信息，即信道容量的大小。

（4）干扰源

干扰源（noise）是整个通信系统中各个干扰的集中反映，用以表示消息在信道中传输时遭受干扰的情况。对于任何通信系统而言，干扰的性质和大小是影响系统性能的重要因素。

（5）密钥源

密钥源（key source）是产生密钥 k 的源。信道编码器输出信号 x 经过 k 的加密运算后，就把明文 x 变换为密文 y。若窃听者未掌握发送端采用的密钥 k，则很难从窃听到的信号 z' 解出明文。而接收端的信宿则因知道事先已约定好的密钥 k，因此能从收到的信号 z' 解出明文。对于二进制的代码而言，加密相当于 $y=z\oplus p$ 运算（其中序列 p 通常是受密钥控制的伪随机序列），而解密则相当于 $x'=y'\oplus p$ 运算。这里 x'、y'、z' 之所以不同于发送端的 x、y、z，是考虑到信号 z 在信道中传输时所受到的干扰影响。但在正常通信条件下，总会有 $x'\approx x$，$y'\approx y$，$z'\approx z$ 的结果。

一般地说，通信系统的性能指标主要包括有效性、可靠性、安全性和经济性。通信系统优化就是使这些指标达到最佳。除了经济性外，这些指标正是信息论的研究对象，可以通过各种编码处理来使通信系统的性能最优化。根据信息论的各种编码定理和上述通信系统的指标，编码问题可分解为三类：信源编码、信道编码和加密编码。

（1）信源编码

信源编码器（source encoder）的作用有两个：一是把信源发出的消息变换成由二进制码元（或多进制码元）组成的代码组，这种代码组就是基带信号；另一个作用是通过信

源编码可以压缩信源的冗余度(即多余度),以提高通信系统传输消息的效率。信源编码可分为无失真信源编码和限失真信源编码。前者适用于离散信源或数字信号;后者主要用于连续信源或模拟信号,如语音、图像等信号的数字处理。从提高通信系统的有效性意义上说,信源编码器的主要指标是它的编码效率,即理论上所需的码率与实际达到的码率之比。一般来说,效率越高,编译码器的代价也将越大。信源译码器(source decoder)的作用是把信道译码器输出的代码组变换成信宿所需要的消息形式,它的作用相当于信源编码器的逆过程。

(2) 信道编码

信道编码器(channel encoder)的作用是在信源编码器输出的代码组上有目的地增加一些监督码元,使之具有检错或纠错的能力。信道译码器(channel decoder)具有检错或纠错的功能,它能将落在其检错或纠错范围内的错传码元检错或纠错,以提高传输消息的可靠性。信道编码包括调制解调和纠错检错编译码。信道中的干扰常使通信质量下降,对于模拟信号,表现在接收到信号的信噪比下降;对于数字信号,就是误码率增大。信道编码的主要方法是增大码率或频带,即增大所需的信道容量。这与信源编码恰好相反。

(3) 加密编码

密码学(cryptology)是研究如何隐蔽消息中的信息内容,以便在传输过程中不被窃听,提高通信系统的安全性。将明文变换成密文,通常不需要增大信道容量,例如在二进制码信息流上叠加一密钥流。但也有些密码要求占用较大的信道容量。

在实际问题中,上述三类编码应统一考虑,以提高通信系统的性能。这些编码的目标往往是相互矛盾的。提高有效性必须去掉信源符号中的冗余部分,此时信道误码会使接收端不能恢复原来的信息,这就需要相应提高传送的可靠性,不然会使通信质量下降;反之,为了可靠而采用信道编码,往往需扩大码率,也就降低了有效性。安全性也有类似情况。编成密码,有时需扩展码位,这样就降低了有效性;有时也会因不同步而使授权用户无法获得信息,必须重发而降低有效性,或丢失信息而降低可靠性。从理论方面来说,若能把三种码合并成一种码来编译,即同时考虑有效性、可靠性和安全性,可使编译码器更理想化,在经济上可能也更优越。这种三码合一的设想是当前众所关心的课题;但从理论上和技术上的复杂性,要取得有用的结果,还是相当困难的。值得注意的是,信息论分析的问题是存在性问题,即符合条件的编码是存在的,但并没有给出如何去寻找的方法。

本书讨论了编码问题,着重介绍信源和信道的编码定理。限于课时,主要从概念上解释了这些定理的结论,并没有从严格意义上加以证明。而对于密码学仅介绍了保密通信中的一些基本知识。这里首先举几个例子来说明编码的应用,如电报常用的莫尔斯(Morse)码就是按信息论的基本编码原则设计出来的;又如在一些商品上面有一张由粗细条纹组成的标签,从这张标签可以得知该商品的生产厂家、生产日期和价格等信息。这些标签是利用条形码设计出来的,非常方便有用,应用越来越普遍;再如,计算机的运算速度很高,要保证它几乎不出差错,相当于要求有 100 年的时间内不得有一秒的误差,这就需要利用纠错码来自动及时地纠正所发生的错误;每出版一本书,都给定一个国际标准书号(ISBN),这大大方便了图书的销售、编目和收藏工作。可以说,人们在日常生活和生产实践中,正在越来越多地使用编码技术。